Learning Expectation of Label Distribution for Facial Age and Attractiveness Estimation

Bin-Bin Gao\textsuperscript{a}, Xin-Xin Liu\textsuperscript{b}, Hong-Yu Zhou\textsuperscript{a}, Jianxin Wu\textsuperscript{b,c}, Xin Geng\textsuperscript{c}

\textsuperscript{a}YouTu Lab, Tencent, Shenzhen 518075, China
\textsuperscript{b}National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China.
\textsuperscript{c}MOE Key Laboratory of Computer Network and Information Integration, Southeast University, Nanjing 211189, China

Abstract

Facial attributes (e.g., age and attractiveness) estimation performance has been greatly improved by using convolutional neural networks. However, existing methods have an inconsistency between the training objectives and the evaluation metric, so they may be suboptimal. In addition, these methods always adopt image classification or face recognition models with a large amount of parameters, which carry expensive computation cost and storage overhead. In this paper, we firstly analyze the essential relationship between two state-of-the-art methods (Ranking-CNN and DLDL) and show that the Ranking method is in fact learning label distribution implicitly. This result thus firstly unifies two existing popular state-of-the-art methods into the DLDL framework. Second, in order to alleviate the inconsistency and reduce resource consumption, we design a lightweight network architecture and propose a unified framework which can jointly learn facial attribute distribution and regress attribute value. The effectiveness of our approach has been demonstrated on both facial age and attractiveness estimation tasks. Our method achieves new state-of-the-art results using the single model with 36× fewer parameters and 3× faster inference speed on facial age/attractiveness estimation. Moreover, our method can achieve comparable results as the state-of-the-art even though the number of parameters is further reduced to 0.9M (3.8MB disk storage).

Keywords: label distribution, deep learning, convolutional neural network, age estimation, attractiveness estimation.

1. Introduction

The human face contains a lot of important information related to individual characteristics, such as identity, expression, age, attractiveness and gender. Such information has been widely applied in real-world applications such as video surveillance, customer profiling, human-computer interaction and person identification. Among these tasks, developing automatic age and attractiveness estimation methods has become an attractive yet challenging topic in recent years.

Why is it a challenging task to find age/attractiveness from facial images? First, compared with image classification \cite{11} or face recognition \cite{22,33,44}, the existing facial attribute datasets are always limited because it is very hard to gather a completely and sufficiently labeled dataset. In Table \ref{tab:1} we list detailed information of various facial attribute datasets. For example, there are only 2476 training images in the ChaLearn15 apparent age estimation challenge \cite{55}. Second, the number of images is very imbalanced in different label groups. What is more, distributions of different datasets are also very different. As Fig. \ref{fig:1} depicts, there are two peaks on ChaLearn15 (the early one at around 2 years old and the latter one at 26 years old) and Morph (the early one at around 20 years old and the latter one at 40 years old), while ChaLearn15 has only one peak. Similar phenomenon also appears in facial attractiveness datasets. These imbalances bring a serious challenge for developing an unbiased estimation system. Third, compared to other facial attributes, such as gender or expression, age/attractiveness estimation is a very fine-grained recognition task, e.g., we humans very hardly sense the change of one person’s facial characteristics when he/she grew from 25 to 26 years old.

Can we use a unified framework to effectively estimate age and attractiveness from facial images? First, these two tasks are both based on facial images thus it is possible to use a unified base model. Second, the labels are ordinal in either the age estimation or the facial attractiveness one. Generally, they are both integrated into ordinal regression or ordinal classification topics whose main character is that the categories are related in a natural or implied order. Common examples of such tasks are movie and facial attractiveness ratings (e.g., from 1 star to 5 stars) or facial age (e.g., from baby to child to adults) or im-

![Figure 1: Illustration of facial attributes’ distributions. (a) illustrates distributions of three facial age datasets, and (b) illustrates distributions of two facial attractiveness datasets (Best viewed in color).](image-url)
age aesthetics (e.g., from “unacceptable” to “professional” to “exceptional”). Third, there is uncertainty information between any two adjacent labels for facial age and attractiveness tasks and their evaluation metric are also the same commonly using MAE. Therefore, we try to integrate facial age and attractiveness estimation into a unified framework in this paper.

The common evaluation metric of age/attractiveness estimation is the Mean Absolute Error (MAE) between the predicted value and ground-truth. Thus, it is very natural to treat facial attributes estimation as a metric regression problem [6] which minimizes the MAE. However, such methods usually cannot achieve satisfactory performance because some outliers may cause a large error term, which leads to an unstable training procedure. Later, Rothe et al. [7] trained deep convolutional neural network (CNN) for age estimation as multi-class classification, which maximizes the probability of ground-truth class without considering other classes. This method easily falls into over-fitting because of the imbalance problem among classes and limited training images [8].

Recently, ranking CNN [9–10] and deep label distribution learning (DLDL) [18] techniques achieved state-of-the-art performance on facial age estimation. Both methods use the correlation information among adjacent ages at different levels. The ranking method transforms single-value estimation to a series of binary classification problems at the training stage. Then, the output of the rankers are aggregated directly from these binary outputs at predication stage. DLDL firstly converts a real-value to a discrete label distribution. Then, the aim of the training is to fit the entire distribution. At inference stage, like [7], an expected value over the predicted distribution is taken as the final output. We can easily find that there is an inconsistency between the training objectives and the evaluation metric in all these methods. Thus, they may be suboptimal. We expect to improve their performance if this inconsistency is removed.

In addition, we observe that almost all state-of-the-art facial attributes estimation methods [14–15] are initialized by a pre-trained model which is trained on large-scale image classification (e.g., ImageNet [1]) or face recognition (e.g., VGGFace [2]) datasets, and fine-tuned on the target dataset. These pre-trained models adopt some popular and powerful architectures (e.g., VGGNet [16]). Unfortunately, these models often have huge computational cost and storage overhead. Taking VGG-16 for example, it has 138.34 million parameters, taking up more than 500MB storage space. Therefore, it is hard to be deployed on resource-constrained devices, e.g., mobile phones. Recently, some researchers devoted to compressing these pre-trained models so that both reducing the number of parameters and keeping accuracy are possible [17]. Unlike these compression methods, we directly design a thin and deep network architecture and train it from scratch.

In this paper, we integrate label distribution learning [23] and expectation regression into a unified framework to alleviate the inconsistency between training and evaluation stages with a simple and lightweight CNN architecture. The proposed approach effectively and efficiently improves the performance of the previous DLDL on both prediction error and inference speed for facial attributes estimation, so we call it DLDL-v2. Our contributions are summarized as follows.

- We provide, to the best of our knowledge, the first analysis and show that the ranking method is in fact learning label distribution implicitly. This result thus unifies existing state-of-the-art facial attributes estimation methods into the DLDL framework;
- We propose an end-to-end learning framework which jointly learns label distribution with the correlation information among neighboring labels and regresses single label ground-truth in both feature learning and classifier learning;
- We create new state-of-the-art results on facial age and attractiveness estimation tasks using single and small model without external age/attractiveness labeled data or multi-model ensemble;
- Our proposed framework is partly interpretable. We find the network employ different patterns to estimate age for people at different age stage. Meanwhile, we also quantitatively analyze the sensitivity of our approach to different face regions.

We organize the rest of this paper as follows. Related works on facial attributes (e.g., age and attractiveness) estimation are introduced in Section 2. Then, Section 3 presents the proposed DLDL-v2 approach including the problem definition, the relationship between existing methods, and our joint learning framework and its model architecture. After that, the experiments are reported in Section 4. In Section 5 we discuss how DLDL-v2 makes the final determination for an input facial image and analyze why it can work well. Finally, the conclusion is given in Section 6. Some preliminary results have been published in a conference presentation [24].

### 2. Related Works

In the past two decades, many researchers have worked on facial attributes estimation. Earlier researches are two stage solutions, including feature extraction and model learning. Recently, deep learning methods are proposed, which integrate both stages into an end-to-end framework. In this section, we briefly review these two types of frameworks.

**Two stage methods.** The task of the first stage is how to extract discriminative features from facial images. Active appearance model (AAM) [25] is the earliest method through extracting shape and appearance features of face images. Later, the Bio-inspired feature (BIF) [26], as the most successful age feature, is widely used in age estimation. But, in face attractiveness

| Dataset    | #Images | Label range |
|------------|---------|-------------|
| ChalLearn15 [4] | 2476+1136 | 3-85 |
| ChalLearn16 [18] | 5613+1978 | 0-96 |
| UTKFace [19] | 24108 | 0-116 |
| Morph [20] | 55134 | 16-70 |
| SCUT-FBP [21] | 500 | 1-5 |
| CFD [22] | 597 | 1-7 |

Table 1: Details of facial age and attractiveness datasets.
analysis, geometric features [27] and texture features [28] depended on facial landmark positions are widely used, since the BIF feature may be suboptimal for facial attractiveness prediction. Obviously, the drawback of hand-designed features is that one needs to re-design a feature extraction method when facing a new task, which usually requires domain knowledge and a lot of efforts. The second stage is how to exactly estimate facial attributes using these designed features. Classification and regression models are often used to estimate facial attributes. The former includes k-nearest neighbors (KNN), multilayer perceptron (MLP) and support vector machine (SVM), and the latter contains quadratic regression, support vector regression (SVR) and soft-margin mixture regression [29]. Instead of classification and regression, ranking techniques [30, 31, 32, 33, 34] utilize the ordinal information of age to learn a model for facial age estimation.

In addition, Geng et al. proposed a label distribution learning (LDL) approach to utilize the correlation among adjacent labels, which improved performance on age estimation [35] and beauty sensing [36]. Recently, some improvements of LDL [37,38] have been proposed. Xing et al. [37] used logistic boosting regression instead of the maximum entropy model in LDL. Meanwhile, He et al. [38] generated age label distributions through weighted linear combination of the label of input image and that of its context-neighboring images. These methods only learn a classifier, but not the visual representations.

**Single stage methods.** Deep CNNs have achieved impressive performance on various visual recognition tasks. The greatest success is learning feature representations instead of using hand-crafted features via the single stage learning strategy. Existing facial attribute estimation techniques fall into four categories: metric regression (MR) [6], multi-class classification (DEX) [7], Ranking [9] [10] [11] and LDL [8].

MR treats age estimation as a real-valued regression problem. The training procedure usually minimizes the squared difference between the estimated value and the ground-truth. DEX adopts a general image classification framework which maximizes the probability of the ground-truth class during training. In the inference stage, Rothe et al. [7] empirically showed that the expected value over the softmax-normalized output probabilities can achieve better performance than the class prediction of maximum probabilities. However, both MR and DEX easily lead to an unstable training [8].

Ranking methods transform facial attribute regression as a series of binary classification problems. Niu et al. [9] proposed a multi-out CNN via integrating multiple binary classification problems to a CNN. Then, Chen et al. [10] [11] trained a series of binary classification CNNs to get better performance. Given a testing image, the output of the rankers are aggregated directly from these binary outputs.

DLDL converts a single value to a label distribution and learns it in an end-to-end fashion. Recently, Shen et al. [39] proposed LDDLs via combining DLDL and differentiable decision trees. Hu et al. [40] exploited age difference information to improve the age estimation accuracy. These approaches have achieved state-of-the-art performance on age estimation. In addition, Yang et al. [41] proposed a multi-task deep framework via jointly optimizing image classification and distribution learning for emotion recognition. However, these methods may be suboptimal, because there is an inconsistency between the training objectives and evaluation metric.

In this paper, we focus on how to alleviate or remove this inconsistency in a deep CNN with fewer parameters. Age and attractiveness estimation from still face images are suitable applications of the proposed research.

### 3. Our Approach

In this section, we firstly give the definition of the joint learning problem. Next, we show that ranking is implicitly learning label distribution. Finally, we present our framework and network architecture.

#### 3.1. The Joint Learning Problem

**Notation.** We use boldface lowercase letters like \( \mathbf{p} \) to denote vectors, and the \( i \)-th element of \( \mathbf{p} \) is denoted as \( p_i \). \( \mathbf{I} \) denotes a vector of ones. Boldface uppercase letters like \( \mathbf{W} \) are used to denote matrices, and the element in the \( i \)-th row and \( j \)-th column is denoted as \( W_{ij} \). The circle operator \( \circ \) is used to denote element-wise multiplication.

The input space is \( \mathbf{X} = \mathbb{R}^{hw \times c} \), where \( h, w \) and \( c \) are height, width and the number of channels of an input image, respectively. Label space \( \mathbf{Y} = \mathbb{R} \) is real-valued. A training set with \( N \) instances is denoted as \( \mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N \), where \( \mathbf{x}_n \in \mathbf{X} \) denotes the \( n \)-th input image and \( y_n \in \mathbf{Y} \) its corresponding label. We omit the image index \( n \) for clarity. The joint learning aims to learn a mapping function \( \mathbf{F} : \mathbf{X} \rightarrow \mathbf{Y} \) such that the error between prediction \( \hat{y} \) and ground-truth \( y \) be as small as possible on a given input image \( x \).

However, metric regression often cannot achieve satisfactory performance. We observe that people usually predict another person’s apparent age in a way like “around 25 years old” in real life, which indicates using not only 25 but also neighboring ages (e.g., 24 and 26) to describe the face. Similar case also happens in facial attractiveness assessment. Based on the observation, label distribution learning methods can utilize the information via transforming the single value regression problem to a label distribution learning problem.

To fulfill this goal, instead of outputting a single value \( y \in \mathbf{Y} \) for an input \( x \), we quantize the range of possible \( y \) values into several labels. For example, it is reasonable to assume that \( y \in [0, 100] \) in age estimation. Thus, we can define \( \Delta l = [0 : \Delta l : 100] \) (MATLAB notation) as the ordered label vector, where \( \Delta l \) is a fixed real number. A label distribution \( \mathbf{p} \) is then \((p_1, p_2, \ldots, p_K)\), where \( p_i \) is the probability that \( y = l_i \) (i.e., \( \Pr(y = l_i) \) for \( 1 \leq i \leq K \)) [8]. Since we use equal step size \( \Delta l \) in quantizing \( y \), the probability density function (p.d.f.) of normal distribution is a natural choice to generate the ground-truth \( \mathbf{p} \) from \( y \) and \( \sigma \):

\[
p_k = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(l_k - y)^2}{2\sigma^2} \right),
\]

(1)
where $\sigma$ is a hyper-parameter. The goal of label distribution learning is to maximize the similarity between $p$ and the CNN generated distribution $\hat{p}$ at training stage. In the prediction stage, predicted distribution $\hat{p}$ is reversed to a single value by a special inference function. It is suboptimal because there exists inconsistency between training objective and evaluation metric. We are interested to not only learn the label distribution $p$ but also regress a real value $y$ in one framework in an end-to-end manner.

3.2. Ranking is Learning Label Distribution

The ranking-based [9][10][11] and DLDL-based [8][42][39][44] methods have achieved state-of-the-art performance in facial age/attractiveness estimation problems. In this section, we analyze the essential relationship between them.

We explore their relationship from the perspective of label encoding. In DLDL-based approaches, for a face image $x$ with true label $y$ and hyper-parameter $\sigma$, the target vector $p^d$ (i.e., label distribution) is generated by a normal p.d.f. (Eq. (1)). For example, the target vector of a 50 years old face is shown in Fig. 2a where $l = [0, 1, \ldots, 100]$. In ranking CNN, $K - 1$ binary classifiers are required for $K$ ranks because the $k$-th binary classifier focuses on determining whether the age rank of an image is greater than $l_k$ or not. For a face image $x$ with true label $y \in f(l_{k-1}, l_k]$, the target vector with length $K - 1$ is encoded as $p^{\text{rank}} = \{1, \ldots, 0, \ldots, 0\}$, where the first $k - 1$ values are 1 and the rest being 0. The target ranking vector of a 50 years old face is shown in Fig. 2c as the dark line.

As we all know, for a generic normal distribution with p.d.f. $p$, mean $y$ and deviation $\sigma$, the cumulative distribution function (c.d.f.) is

$$c_k = 0.5 \left[1 + \text{erf} \left( \frac{l_k - y}{\sigma \sqrt{2}} \right) \right],$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$. Fig. 2b shows the c.d.f. corresponding to the p.d.f. in Fig. 2a. From Eq. (2), we know

$$\begin{cases} 1 - c_k > 0.5, & \text{if } l_k < y \\ 1 - c_k \leq 0.5, & \text{if } l_k \geq y \end{cases}$$

As shown in Fig. 2c, the curve of $1 - c$ is very close to that of $p^{\text{rank}}$ when $\sigma$ is a small positive real number. Thus, $p^{\text{rank}} \approx 1 - c_k$.

Therefore, there is a linear relationship between Ranking encoding and label distribution. The label distribution encoding $p^d$ can represent more meaningful age/attractiveness information with different $\sigma$, but ranking encoding $p^{\text{rank}}$ does not. Furthermore, DLDL is more efficient, because only one network has to be trained.

However, as discussed earlier, all these methods may be suboptimal because there exists inconsistency between training objective and evaluation metric.

3.3. Joint Learning Framework

In order to jointly learn label distribution and output the expectation, in this section we propose the DLDL-v2 framework.

3.3.1. The Label Distribution Learning Module

In order to utilize the good properties of label distribution learning, we integrate it into our framework to formulate a label distribution learning module. As shown in Fig. 3, this module includes a fully connected layer, a softmax layer and a loss layer. This module follows the DLDL method in [8].

Specifically, given an input image $x$ and the corresponding label distribution $p$, we assume $f = \mathcal{F}(x; \theta)$ is the activation of the last layer of CNN, where $\theta$ denotes the parameters of the CNN. A fully connected layer transfers $f$ to $x \in \mathbb{R}^K$ by

$$x = W^T f + b.$$ 

Then, we use a softmax function to turn $x$ into a probability distribution, that is,

$$\hat{p}_k = \frac{\exp(x_k)}{\sum_i \exp(x_i)}.$$ 

Given an input image, the goal of the label distribution learning module is to find $\theta$, $W$, and $b$ to generate $\hat{p}$ that is similar to $p$.
We employ the Kullback-Leibler divergence as the measurement of the dissimilarity between ground-truth label distribution and prediction distribution. Thus, we can define a loss function on one training sample as follows \cite{8}:

\[
L_d = \sum_k p_k \ln \frac{p_k}{\hat{p}_k}. \tag{9}
\]

3.3.2. The Expectation Regression Module

Note that the label distribution learning module only learns a label distribution but cannot regress a precise value. In order to reduce the inconsistency between training and evaluation stages, we propose an expectation regression module to further refine the predicted value. As shown in Fig. 3, this module includes an expectation layer and a loss layer.

The expectation layer takes the predicted distribution and label set as input and emits its expectation

\[
\hat{y} = \sum_k \hat{p}_k l_k, \tag{10}
\]

where \(\hat{p}_k\) denotes the prediction probability that the input image belongs to label \(l_k\). Given an input image, the expectation regression module minimizes the error between the expected value \(\hat{y}\) and ground-truth \(y\). We use the \(\ell_1\) loss as the error measurement as follows:

\[
L_{er} = |\hat{y} - y|, \tag{11}
\]

where \(|\cdot|\) denotes absolute value. Note that this module does not introduce any new parameter.

3.3.3. Learning

Given a training data set \(D\), the learning goal of our framework is to find \(\theta\), \(W\) and \(b\) via jointly learning label distribution and expectation regression. Thus, our final loss function is a weighted combination of the label distribution loss \(L_d\) and the expectation regression loss \(L_{er}\):

\[
L = L_d + \lambda L_{er}. \tag{12}
\]

where \(\lambda\) is a weight which balances the importance between two types of losses. Substituting Eq. (9), Eq. (10) and Eq. (11) into Eq. (12), we have

\[
L = - \sum_k p_k \ln \hat{p}_k + \lambda \sum_k |\hat{y}_k - y_k|. \tag{13}
\]

We adopt stochastic gradient descent to optimize parameters of our model. The derivative of \(L\) with respect to \(\hat{p}_k\) is

\[
\frac{\partial L}{\partial \hat{p}_k} = -\frac{p_k}{\hat{p}_k} + \lambda l_k \text{sign}(\hat{y} - y). \tag{14}
\]

For any \(k\) and \(j\), the derivative of softmax (Eq. (8)) is well known, as

\[
\frac{\partial \hat{p}_k}{\partial x_j} = \hat{p}_k(\delta_{k,j} - \hat{p}_j), \tag{15}
\]

where \(\delta_{k,j}\) is 1 if \(k = j\), and 0 otherwise. According to the chain rule, we have

\[
\frac{\partial L}{\partial x_j} = (\hat{p}_j - p_j) + \lambda \text{sign}(\hat{y} - y) \hat{p}_j(l_j - \hat{y}), \tag{16}
\]

i.e., \(\frac{\partial L}{\partial x} = \hat{p} - p + \lambda \text{sign}(\hat{y} - y) \hat{p} \circ (I - \hat{y}I). \tag{17}\)

Applying the chain rule for Eq. (7) again, the derivative of \(L\) with respect to \(W\), \(b\) and \(\theta\) are easily obtained, as

\[
\frac{\partial L}{\partial W} = \frac{\partial L}{\partial f} \frac{\partial f}{\partial b} = \frac{\partial L}{\partial f} \frac{\partial f}{\partial \theta} = \frac{\partial L}{\partial f} \frac{\partial f}{\partial \theta} W^{\dagger} \frac{\partial f}{\partial \theta}. \tag{18}
\]

Once \(W\), \(b\) and \(\theta\) are learned, the prediction value \(\hat{y}\) of any new instance \(x\) is generated by Eq. (10) in a forward network computation.

3.4. Network Architecture

Considering both model size and efficiency, we modify VGG-16 \cite{16} from four aspects as follows. VGG-16 consists of 13 convolution (Conv) layers, five max-pooling (MP) layers and three fully connected (FC) layers, and each Conv layer and FC layer is followed by a ReLU layer.

First, we observe that the three FC layers roughly contain 90% parameters of the whole model. We remove all FC layers and add a hybrid-pooling (HP) layer which is constructed by an MP layer and a global avg-pooling (GAP) layer. We find that the HP strategy is more effective than single GAP. Second, to further reduce model size, we reduce the number of the filters in each Conv layer to make it thinner. Third, batch normalization (BN) \cite{43} has been widely used in the latest architecture such as ResNet \cite{44}. Thus, we add a BN layer after each Conv layer to accelerate network training. Last but not least, we add the label distribution learning module and the expectation regression module after the HP layer, as shown in Fig. 3.

Since we design the network for age/attractiveness estimation and its architecture is thinner than the original VGG-16, we call our model ThinAgeNet or ThinAttNet which employs the compression rate of 0.5 and has 3.7M parameters. \footnote{0.5 compression rate means every Conv layer has only 50% channels as that in VGG-16.} We also train a very small model with the compression rate of 0.25, and we call it TinyAgeNet or TinyAttNet which only has 0.9M parameters.

4. Experiments

In this section, we conduct experiments to validate the effectiveness of the proposed DLDL-v2 approach on facial age and attractiveness datasets, based on the open source framework Torch7. All experiments are conducted on an NVIDIA M40 GPU. In order to re-produce all results in this paper, we will release source code and pre-trained models upon paper acceptance.
4.1. Implementation Details

Pre-preprocessing. We use multi-task cascaded CNN\cite{45} to conduct face detection and facial points detection for all images. Then, based on these facial points, we align faces to the upright pose. Finally, all faces are cropped and resized to 224×224. Before feeding to the network, all resized images are to subtract mean and divide standard deviation for each color channel.

Data Augmentation. There are many non-controlled environmental factors such as face position, illumination, diverse backgrounds, image color (i.e., gray and color) and image quality, especially in the ChaLearn datasets. To handle these issues, we apply data augmentation techniques to every training image, including random horizontal flipping, random scaling, random color/gray changing, random rotation and standard color jittering.

Training Details. We pre-train a deep CNN model with soft-max loss for face recognition on a subset of the MS-Celeb-1M dataset\cite{3}. One issue is that a small part of identities have a large number of images and others have only a few in this dataset. To avoid the imbalance problem among identities, we cut those identities whose number of images is lower than a threshold. In our experiments, we use about 5M images of 54K identities as training data.

After pre-training is finished, we remove the classification layer of the network and add the label distribution learning and expectation regression modules. Then, fine-tuning is conducted layer of the network and add the label distribution learning and threshold. In our experiments, we use about 5M images of 54K cropped and resized training image, including random horizontal flipping, random scaling, random color/gray changing, random rotation and standard color jittering.

4.2. Evaluation Metrics

MAE is used to evaluate the performance of facial age or attractiveness estimation,

\[
\text{MAE} = \frac{1}{N} \sum_{n=1}^{N} |\hat{y}^n - y^n|, \quad (19)
\]

where \(\hat{y}^n\) and \(y^n\) are the estimated and the ground-truth of the \(n\)-th testing image, respectively. In addition, a special measurement (\(\epsilon\)-error) is defined by the ChaLearn competition, as

\[
\epsilon\text{-error} = \frac{1}{N} \sum_{n=1}^{N} \left[ 1 - \exp \left( - \frac{(\hat{y}^n - y^n)^2}{2 \sigma^2} \right) \right], \quad (20)
\]

where \(\sigma^2\) is the standard deviation of the \(n\)-th testing image.

We also follow\cite{21,13} to compute Root Mean Squared Error (RMSE) and Pearson Correlation (PC), which can be computed as:

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (\hat{y}^n - y^n)^2}, \quad (21)
\]

\[
\text{PC} = \frac{\sum_{n=1}^{N} (\hat{y}^n - \overline{y}) (y^n - \overline{y})}{\sqrt{\sum_{n=1}^{N} (\hat{y}^n - \overline{y})^2 \sum_{n=1}^{N} (y^n - \overline{y})^2}}. \quad (22)
\]

where \(\overline{y} = \frac{1}{N} \sum_{n=1}^{N} y^n\), and \(\overline{y} = \frac{1}{N} \sum_{n=1}^{N} \hat{y}^n\) are the mean values of the ground-truth and predicted scores over all testing images. These two evaluation metrics are only utilized to evaluate the performance of facial attractiveness estimation.

4.3. Experiments on Age Estimation

4.3.1. Age Estimation Datasets

Three types of datasets are used in our experiments. The first type contains two small-scale apparent age datasets (ChaLearn15\cite{5} and ChaLearn16\cite{18}) which are collected in the wild. The second type is a large-scale real age dataset (Morph\cite{20}). We follow the experimental setting in\cite{24} for evaluation. The third one is UTKFace\cite{19} which the ground truth of age are estimated through the DEX algorithm and double checked by a human annotator. We utilize the split employed in\cite{46}, with 20% test images and 80% images for training.

4.3.2. Age Estimation Results

We compare our approach with the state-of-the-art in both prediction performance and inference time.

Low Error. Table\cite{2,3,4,5} report the comparisons of the MAE and \(\epsilon\)-error performance of our method and previous state-of-the-art methods on four age estimation datasets. In the ChaLearn15 challenge, the best result came from DEX. DEX method’s success relies on a lot of external age labeled training images (260282 additional photos). Under the same setting (without external data), our method outperforms DEX by a large margin in Table\cite{11}. On ChaLearn16, the \(\epsilon\)-error 0.267 of our approach is closest to the best competition result 0.241\cite{13} on the testing set. Note that our result is only based on a single model without external age labeled data. In\cite{15}, they not only used external age labeled data but also employed multi-model ensemble. On Morph, our method creates a new state-of-the-art 1.969 MAE. To our best knowledge, this is the first time to achieve below two years in MAE on the Morph dataset.
In short, our DLDL-v2 (ThinAgeNet) outperforms the state-of-the-art methods without external age labeled data and multi-model ensemble on ChaLearn15, ChaLearn16, Morph, and UTKFace.

High Efficiency. We measure the speed on one P40 GPU accelerated by cuDNN v7.6. The number of parameters, forward/backward memory size, computational complexity (GFlops) and actual inference time of our approach and some previous methods are reported in Table 6. Since [9] and [10] do not release pre-trained models, we cannot test the running time and report the number of parameters of these models. [8] 48 ] 49 50 ] 51 ] 52 ] 53 ] 55 ] 57 ] all used similar network architecture (i.e., VGG-16 or VGGFace). Since [15] employed 14 models, it’s model size and running time is 14 times of [7] and [8]. Compared to the state-of-the-art, DLDL-v2 (ThinAgeNet) achieves the best performance using single model with 36× fewer parameters and 3× reduction in inference time. Furthermore, we also report DLDL-v2’s TinyAgeNet results which achieve a better result (150× fewer parameters and 6× speed improvement) than the original DLDL [8].

In addition, we compare ThinAgeNet and TinyAgeNet with some recent network architectures such as ResNet-50 [60] 62 ] 47 ] 49 ] 50 ] 52 ] 53 ] 55 ] 57 ] all used similar network architecture (i.e., VGG-16 or VGGFace).

4.3.3. Visual Assessment

Fig. 4 shows some examples on ChaLearn16 testing images using our DLDL-v2 ThinAgeNet. In many cases, our solution is able to predict the age of faces accurately. Failures may come from some special cases such as occlusion, low resolution, heavy makeup and extreme pose.

4.4. Experiments on Attractiveness Estimation

4.4.1. Attractiveness Estimation Datasets

To further demonstrate the effectiveness of the proposed DLDL-v2, we perform extensive experiments on two facial attractiveness datasets: SCUT-FBP [21] and CFD [22].

SCUT-FBP [21] is a widely used facial beauty assessment dataset. It contains 500 Asian female faces with neutral expressions, simple backgrounds, no accessories, and minimal occlusion. Each face is scored by 75 workers with a 5-point scale, in which 1 means strong agreement about the face being the
experiments. We use 80% images for training and the remain-

7 pants with diverse background in a 7-point scale (1

65. Similar to SCUT-FBP, each faces is scored by some partici-

pation in [13] and [21], 80% images are randomly selected as the

corresponding standard deviation are given. We follow the set-

ting in [13] and [21], 80% images are randomly selected as the

training set, and the remain 20% as the testing set.

CDF [22] provides high-resolution and standardized pho-

tographs with meaningful annotations (e.g., attractiveness,

babyfacedness and expression etc.). Unlike SCUT-FBP, this
dataset includes male and female faces of multiple ethnicity
(Asian, Black, Latino, and White) between the ages of 17-

65. Similar to SCUT-FBP, each faces is scored by some partici-

pants with diverse background in a 7-point scale (1 = Not at all,

7 = Extremely). In this study, we employ all 597 faces with
natural expression and the corresponding attractiveness scores
for experiments. We use 80% images for training and the remain
20% for testing.

least attractive and 5 means strong agreement about the face
being the most attractive. For each face, its mean score and the

remaining standard deviation are given. We follow the setting

in [13] and [21]. 80% images are randomly selected as the

training set, and the remain 20% as the testing set.

CDF [22] provides high-resolution and standardized pho-

tographs with meaningful annotations (e.g., attractiveness,

babyfacedness and expression etc.). Unlike SCUT-FBP, this
dataset includes male and female faces of multiple ethnicity
(Asian, Black, Latino, and White) between the ages of 17-

65. Similar to SCUT-FBP, each faces is scored by some partici-

pants with diverse background in a 7-point scale (1 = Not at all,

7 = Extremely). In this study, we employ all 597 faces with
natural expression and the corresponding attractiveness scores
for experiments. We use 80% images for training and the remain
20% for testing.

4.4.2. Attractiveness Estimation Results

In Table 7 and 8, we report the performance on SCUT-FBP
and CFD and compare with the state-of-the-art methods in the

literature.

Comparing with those methods using hand-crafted features,
such as Regression [21] and SLDDL [6], the proposed DLDL-

v2 (ThinAttNet) achieves 0.930 PC and 0.212 MAE on SCUT-

FBP. It outperforms Regression [21] by 0.282 in PC, and im-

proves SLDDL [6] by 0.135 in RMSE. What is more, for

those methods using deep label distribution, such as LDL

(ResNet50) [13] as one of the state-of-the-art methods, our

DLDL-v2 still outperforms it. Furthermore, our method is com-

parable to the fusional solution of deep features and geomet-
ic features in [13]. There are two major reasons. First, our

pre-trained model is trained on a face recognition dataset which

is closer to facial attractiveness than those object classification

datasets (ResNet50 is trained by ImageNet) in [13]. Second,

we jointly learn label distribution and regress the facial attrac-
tiveness score in DLDL-v2, which can effectively erase the

inconsistency between training objective and evaluation met-

ric (MAE).

4.4.3. Visual Assessment

In order to intuitively visualize the prediction performance
of our DLDL-v2 on facial attractiveness task, we show the
top eight and bottom eight test images based on the prediction
scores of DLDL-v2 with ThinAttNet in Fig. 5. On selected
16 testing images, prediction scores of 12 images highly match
with those of human raters. This result qualitatively demon-

strates that our DLDL-v2 is able to generate human-like results.

In addition, some possible facial attractiveness cues may be ob-

served via comparing between the top and bottom faces with

attractiveness score. Generally speaking, faces with higher at-

tractive scores have smoother and lighter skin, oval face with

larger eyes, narrower nose with a pointed tip, and better har-

mony in facial organs than those with lower scores.

Figure 4: Examples of apparent age estimation using DLDL-v2 (ThinAgeNet) on ChaLearn16 testing images. The left ten columns show good age estimations and
the right five columns are poor cases.

Table 6: Comparisons of model parameters and forward times with state-of-the-art methods. M means million (10^6), G means billion (10^9), MB means megabyte and Time denotes the average forward times of n images in milliseconds on one P40 GPU.

Table 7: Comparisons of state-of-the-art methods for facial attractiveness estimation on SCUT-FBP dataset. ↑ indicates methods using ten-fold cross validation,
90% of images for training and 10% for evaluation and ↓ indicates methods using multi-features fusion.

Table 8: Comparisons with state-of-the-art methods for facial attractiveness estimation on CFD dataset.
4.5. Ablation Study and Diagnostic Experiments

DLDL-v2 (ThinAgeNet) is employed for ablation study on facial age datasets in this section. We firstly investigate the efficacy of the proposed data augmentation and the pooling strategy. For fair comparison, we fix $\lambda_1 = 1$ and $\lambda = 1$. Then, to investigate the effectiveness of the proposed joint learning mechanism, we compare it with two stage and single stage methods under the same setting. At last, we also explore the sensitivity of hyper-parameters in our DLDL-v2.

4.5.1. Influence of Data Augmentation

Data augmentation techniques increase the amount of training data using information only in training set, which is an effective way to reduce the over-fitting of deep models. From Table 9, we can observe 0.26−0.27 MAE improvements on parent age datasets and 0.38 MAE improvement on Morph usings data augmentation. This indicates that data augmentation can greatly improve the performance of age estimation.

4.5.2. Effectiveness of Pooling Strategy

GAP is one of the most popular and simple method for aggregating the spatial information of feature maps in state-of-the-art network architecture such as ResNet [44]. It outputs the spatial average of each feature map of the last convolution layer. Max-pooling takes the maximal value of each small region. Thus, the feature of HP is more discriminative than that of GAP. If we directly use global max-pooling instead of HP, then all discriminative features are aggregated by a max-pooling and a GAP layer. HP firstly encourages the network to learn a discriminative feature in a small region via max-pooling, then all discriminative features are aggregated by GAP. Thus, the feature of HP is more discriminative than that of GAP. If we directly use global max-pooling instead of HP, the training of network easily fall into over-fitting. To explore the effect of the pooling strategy, we further use the HP to replace the traditional GAP when combining data augmentation. It can be seen in Table 9 that the proposed HP can consistently reduce the prediction error on all datasets.

Table 9: Comparison of different methods for age estimation.

| Factors | ChaLearn15 | ChaLearn16 | Morph |
|---------|------------|------------|-------|
| Aug Pooling | MAE | $\epsilon$-error | MAE | $\epsilon$-error | MAE |
| $\times$ HP | 3.399 | 0.303 | 3.717 | 0.290 | 2.346 |
| $\sqrt{}$ GAP | 3.210 | 0.282 | 3.539 | 0.274 | 2.039 |
| $\sqrt{}$ HP | 3.135 | 0.272 | 3.452 | 0.267 | 1.969 |

4.5.3. Comparisons with Two Stage Methods

We compare the proposed approach with two stage methods considering two types of features. The first one is the BIF [26], as the most successful hand-crafted age feature, which was widely used in age estimation. The second one is CNN features which are extracted from our pre-trained face recognition model. For BIF, we adopt 6 bands and 8 orientations [26], which produces 4616-dimensional features. The CNN features are extracted from the hybrid pooling layer of the pre-trained model and their dimension is 256. These features are normalized by $f_2$ without using any dimensionality reduction technique.

We choose three classical age estimation algorithms, including SVR [26], OHRank [30] and BFGS-LDL [35]. For SVR and OHRank, the Liblinear software is used to train regression or classification models. For BFGS-LDL, we use the open source LDL package. Instead of age prediction with the maximal probability in [35], we use the expected value over prediction distribution because it has better performance.

The experimental results are shown in Table 10. First, OHRank and BFGS-LDL using BIF and CNN features have similar performances on all datasets. This further validates our previous analysis that ranking is learning label distribution. Second, our proposed approach significantly outperforms all baseline methods. The major reason is that two stage methods cannot learn visual representations. This suggests that it is crucially important to jointly learn visual features and recognition model using an end-to-end manner. At last, OHRank and BFGS-LDL are much better than SVR, which indicates learning [30].
ing label distribution can really help us to improve estimation performance.

4.5.4. Comparisons with Single Stage Methods

We employ six very strong methods under the same setting as baselines:

- **MR:** In MR, the ground-truth label $y$ is projected to $[-1, 1]$ by a linear transform. For MR, we need to make a little modification to DLDL-v2. Specifically, we add an FC layer with single output after HP, and follow a hyperbolic tangent activation function $f(x) = \tanh(x)$ for speedup convergence. The $\ell_2$ and $\ell_1$ loss function is used to train MR.

- **DEX:** In DEX, true label $y$ is quantized to different label group, which is treated as a class. To train DEX, we only need remove the expectation module and modify loss function to cross-entropy loss in DLDL-v2. In inference time, an expected value over prediction probabilities is used for final estimation.

- **Ranking:** In [10, 11], multiple binary classification networks are independently trained, which lead to time-consuming of training and storage overhead of model. We propose a new multiple output CNN and jointly train these binary classifiers. Specifically, we firstly remove the label distribution and expectation module in DLDL-v2. Then, we add an FC layer with $K$ – 1 output units and follow a sigmoid layer. For training Ranking CNN, we employ $K$ – 1 binary cross-entropy loss. In inference stage, the prediction is computed by $\hat{y} = l_i$, where $i = 1 + \sum_{k=1}^{K-1} [\delta(x) > 0.5]$. [$\cdot$] denotes the truth-test operator, which is 1 if the inner condition is true, and 0 otherwise. Our experiments showed that this new setup has lower MAE than that in [9,10,11].

- **ER ($\ell_1$):** We only employ the expectation regression (ER) loss $L_{er}$ to optimize DLDL-v2’s parameters via removing label distribution loss $L_d$ in Eq. (12).

- **DLDL:** We set $\lambda = 0$ in Eq. (12) to learn DLDL-v2. Table 11 reports the results of all single stage methods. We can see that the MAE and $\epsilon$-error of Ranking, ER and DLDL methods are significantly lower than that of MR and DEX on all datasets. This indicates that utilizing label distribution is helpful to reduce age estimation error. Meanwhile, we also find that the prediction error of Ranking is close to that of DLDL, which conforms to the analysis in Section 3.2. Furthermore, the performance of DLDL is better than that of Ranking, which suggests that learning p.d.f. is more effective than learning c.d.f. It is noteworthy that ER ($\ell_1$) and DLDL are two extreme cases of our DLDL-v2. DLDL-v2 consistently outperforms ER ($\ell_1$) and DLDL on all datasets, which indicates the joint learning can ease the difficult of network optimization. In Table 11, we can see that the proposed joint learning achieves the best performance among all methods. It means that erasing the inconsistency between training and evaluation stages can help us make a better prediction.

4.5.5. Sensitivity of Hyper-parameters

We explore the influence of hyper-parameters $\lambda$ and $\Delta l$, where $\lambda$ is a weight which balances the importance between label distribution and expectation regression loss, and $\Delta l$ refers to the number of discrete labels ($K$). In Table 12 we report results on all three age datasets with different value of $\lambda$ and $\Delta l$. We can see that our method is not sensitive to $\lambda$ and $\Delta l$ with $0.01 \leq \lambda \leq 10$ and $0.25 \leq \Delta l \leq 4$. Note that, too many discrete labels lead to little training samples for per class in DEX [7], which may make prediction less precise. However, our method can ease the problem, because the training samples associated with each class label is significantly increased without actually increase the number of the total training examples. Surprising, there is also a good enough performance when the number output neurons (i.e., $K$) is 26. In our experiment, we fixed hyper-parameters $\Delta l = 1$ and $\lambda = 1$ without carefully tuning them. In practice, it is convenient to find optimal hyper-parameters using a hold-out set.

5. Understanding DLDL-v2

We have demonstrated that DLDL-v2 has excellent performance for facial age and attractiveness estimation. A natural problem is how DLDL-v2 makes the final decision for an input facial image. In this section, we try to answer this question. Then, we analyze why it can work well when compared with existing methods.

5.1. How Does DLDL-v2 Estimate Facial Attributes?

In order to understand how DLDL-v2 makes the final decision for an input facial image, we visualize a score map that can intuitively show which regions of face image are related to the network decision. To obtain the score map, we firstly employ a class-discriminative localization technique [66] that can

| Table 11: Comparisons with single stage methods for age estimation (lower is better). |
|-----------------|-------------|-------------|-------------|-------------|-------------|
| **Methods** | ChaLearn15 | ChaLearn16 | Morph |
| MAE | $\epsilon$-error | MAE | $\epsilon$-error | MAE |
| MR ($\ell_2$) | 3.685 | 0.337 | 3.296 | 0.294 | 2.282 |
| MR ($\ell_1$) | 3.655 | 0.334 | 3.272 | 0.301 | 2.347 |
| DEX | 3.588 | 0.306 | 3.163 | 0.332 | 2.311 |
| Ranking | 3.365 | 0.298 | 3.645 | 0.290 | 2.164 |
| ER ($\ell_1$) | 3.287 | 0.291 | 3.641 | 0.282 | 2.214 |
| DLDL | 3.228 | 0.285 | 3.509 | 0.272 | 2.132 |
| DLDL-v2 | 3.135 | 0.272 | 3.452 | 0.267 | 1.960 |

| Table 12: The influences of hyper-parameters for our DLDL-v2. |
|-----------------|-------------|-------------|-------------|-------------|-------------|
| **Hyper-param** | ChaLearn15 | ChaLearn16 | Morph |
| $\lambda$ | $\Delta l$ ($K$) | MAE | $\epsilon$-error | MAE | $\epsilon$-error | MAE |
| 0.01 | 1 (101) | 3.223 | 0.282 | 3.493 | 0.270 | 1.960 |
| 0.10 | 1 (101) | 3.188 | 0.278 | 3.455 | 0.268 | 1.972 |
| 1.00 | 1 (101) | 3.135 | 0.272 | 3.452 | 0.267 | 1.969 |
| 10.00 | 1 (101) | 3.144 | 0.273 | 3.487 | 0.270 | 1.977 |
| 1.00 | 4 (26) | 3.182 | 0.276 | 3.473 | 0.270 | 1.963 |
| 1.00 | 2 (51) | 3.184 | 0.274 | 3.484 | 0.271 | 1.963 |
| 1.00 | 0.50 (201) | 3.184 | 0.278 | 3.484 | 0.269 | 1.992 |
| 1.00 | 0.25 (401) | 3.167 | 0.274 | 3.459 | 0.265 | 2.028 |
generate class activation maps. Then, these activation maps are aggregated by predicted probabilities.

Let us briefly review our framework. The last convolution block produce activation maps $F^i \in \mathbb{R}^{K \times w \times h}$. These activations are then spatially pooled by a hybrid pooling and linearly transformed (i.e., Eq. (7)) to produce probabilities $\hat{p} \in \mathbb{R}^k$ with a label distribution module. To produce class activation maps, we apply linearly transform layer to $F^i$ as follows

$$A^k = \sum_j w_j F^j + b_k .$$

Then, the score map can be derived by

$$S = \sum_k \hat{p}_k A^k .$$

In Eq. (24), the value of $S(j)$ represents the contribution of the network’s decision at position of $i$-th row and $j$-th column. Bigger values mean more contributions and vice versa. For comparing the correspondence between highlighted regions in $S$ and an input image, we scale $S$ to the size of an input image.

In Fig. 5, we visualize the score maps of testing images (ChaLearn16) coming from different age group. We can see that the highlighted regions (i.e., red regions) are significantly different for different age group faces. For infants, the highlighted region locates in the center of two eyes. For adults, the strong areas include two eyes, nose and mouth. For senior people, the highlighted regions consist of the forehead, brow, two eyes and nose. In short, the network uses different patterns to estimate different age.

We also show some examples coming from SCUT-FBP testing images in Fig. 7. We can observe that it is not significant for the highlighted regions between these faces with higher attractiveness score and that of lower score. An explanation is that DLDL-v2 may be able to estimate facial attractiveness through simply comparing the difference of the common facial traits such as eyebrows, eyes, nose, mouth etc. In fact, the SCUT-FBP dataset indeed has the lower complexity (female faces with simple backgrounds, no accessories, and minimal occlusion) than age estimation on ChaLearn16.

**5.2. Sensitivity to Different Face Regions**

To further quantitatively analyze the sensitivity of DLDL-v2 to different face regions. We occlude different portions of the input image by setting it to mean values of all training images.

Specifically, we use two type of occlusions, small square region (size of 32x32) and horizontal stripe (size of 32x224), as in [67]. We occlude the input images (size of 224x224) using this two type of occlusions in a sliding window fashion. In all, we obtain 49+7 occluded inputs for each input image. For each occluded input, we record prediction performance (i.e., MAE) on all testing images. Finally, we compute the relative performance loss between with and without occlusions to measure the sensitivity of occlusion region.

In Fig. 8, we show the quantitative results under different occlusions. First, we observe that larger values usually appear in some specific regions such as forehead, two eyes, nose, mouth, and chin. This indicates the decision of DLDL-v2 heavily depends on these crucial regions. Second, these values are significantly different in the different regions. For example, on ChaLearn16 testing images, the largest and second largest value appear around nose and eyes, which suggests nose and eyes are the most important facial traits for age estimation. Third, although SCUT-FBP and CFD both are used to evaluate facial attractiveness estimation datasets, the distributions of the largest value are greatly different. The largest value of the former appears the region of eyes, and that of the latter appears the region of mouth and chin. In fact, the faces of SCUT-FBP come from Asia female, which is scored by Chinese, while CFD dataset consists multi-race faces and is scored by diverse background annotators. Therefore, this difference may be due to the phenomenon that different races may have an inconsistent understanding of facial attractiveness.

**5.3. Why does DLDL-v2 make good estimation?**

Compared to MR, the training procedure of our DLDL-v2 is more stable because it not only regresses the single value
with expectation module but also learns a label distribution. Compared to DEX, through introducing label distribution learning module to DLDL-v2, the training instances associated with each class label is significantly increased without actually increasing the number of the total training images, which effectively alleviate the risk of over-fitting.

For Ranking and DLDL-based methods, we have proved that they are both learning a label distribution from different levels. Therefore, they both share the advantages of label distribution learning. However, there are three major differences in the network architectural between these methods and our DLDL-v2. First, these methods depend heavily on a pre-trained model such as VGGNet or VGGFace with more parameters while DLDL-v2 has a thinner architecture with fewer parameters. Therefore, DLDL-v2 has higher efficiency in inference time and lower storage overhead. Second, DLDL-v2 effectively avoids the inconsistency between training objective and evaluation metric via introducing the expectation regression module. Third, DLDL-v2 is a fully convolutional network, which removes all but the final fully connected layer. It is very helpful to understand that how DLDL-v2 makes the final decision because of the parameter sharing mechanism between Eq. 7 and 23. In a word, these differences make DLDL-v2 have good performance on accuracy, speed, model size and interpretability.

6. Conclusion

In this paper, we proposed a solution for facial age and attractiveness estimation problems. We firstly analyze that Ranking-based methods are implicitly learning label distribution as DLDL-based methods. This result unifies existing state-of-the-art facial age and attractiveness estimation methods into the DLDL framework. Second, our proposed DLDL-v2 framework can effectively erase the inconsistency between training and evaluation stages via jointly learning label distribution and regressing single value with a thin and deep network architecture. It creates new state-of-the-art results on facial age and attractive estimation tasks with fewer parameters and faster speed, which indicates it is easy to be deployed on resource-constrained devices. In addition, our DLDL-v2 is also a partly interpretable deep framework which employs different patterns to estimate facial attributes.

It is noteworthy that our approach is easily scalable to others label uncertainty tasks, such as skeletal maturity assessment on pediatric hand radiographs [68], head pose estimation [69], popularity of selfie [70], image aesthetic assessment [71] etc. In addition, a further theoretical study between the ranking-CNN and DLDL will also be our future work.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[2] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition, in: Proc. British Machine Vision Conference, 2015, pp. 1–12.
[3] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: A dataset and benchmark for large-scale face recognition, in: Proc. European Conference on Computer Vision, 2016, pp. 87–102.
[4] M. S. Shakeel, K.-M. Lam, Deep-feature encoding-based discriminative model for age-invariant face recognition, Pattern Recognition 93 (2019) 442–457.
[5] S. Escalera, J. Fabian, P. Parada, 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: Proc. IEEE International Conference on Computer Vision Workshop, 2015, pp. 1–9.
[6] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, R. Chellappa, An all-in-one convolutional neural network for face analysis, in: Proc. IEEE Conference on Automatic Face and Gesture Recognition, 2017, pp. 17–24.
[7] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, International Journal of Computer Vision 126 (2) (2016) 144–157.
[8] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Transactions on Image Processing 26 (6) (2017) 2825–2838.
[9] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output CNN for age estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4920–4928.
[10] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, Using Ranking-CNN for age estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5183–5192.
[11] S. Chen, C. Zhang, M. Dong, Deep age estimation: From classification to ranking, IEEE Transactions on Multimedia 20 (8) (2017) 2209–2222.
[12] K. Li, J. Xing, W. Hu, S. I. Maybank, D2c: Deep cumulatively and comparatively learning for human age estimation, Pattern Recognition 66 (2017) 95–105.
[13] Y.-Y. Fan, S. Liu, B. Li, Z. Guo, A. Samal, J. Wan, S. Z. Li, Label distribution-based facial attractiveness computation by deep residual learning, IEEE Transactions on Multimedia 20 (8) (2017) 2196–2208.
[14] P. Rodriguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, J. Gonzalez, Age and gender recognition in the wild with deep attention, Pattern Recognition 72 (2017) 563–571.
[15] G. Antipov, M. Baccouche, S.-A. Berrani, J.-L. Dugelay, Apparent age estimation from face images combining general and children-specialized deep learning models, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2016, pp. 96–104.
[16] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition, in: Proc. International Conference on Learning Representations, 2015.
[17] J.-H. Luo, J. Wu, W. Lin, ThNet: A filter level pruning method for deep neural network compression, in: Proc. IEEE International Conference on Computer Vision, 2017, pp. 5058–5066.
[18] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneu, M. Oliu, M. Ali Bagheri, et al., ChaLearn looking at people and faces of the world: Face analysis workshop and challenge 2016, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2016, pp. 1–8.
[19] Z. Zhang, Y. Song, H. Qh, Age progression/regression by conditional adversarial autoencoder, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5810–5818.
[20] K. Ricanek, T. Tesafaye, Morph: A longitudinal image database of normal adult age-progression, in: Proc. IEEE Conference on Automatic Face and Gesture Recognition, 2006, pp. 341–345.
[21] D. Xie, L. Liang, L. Jin, J. Xu, M. Li, SCUT-FBP: A benchmark dataset for facial beauty perception, in: IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp. 1821–1826.
[22] D. S. Ma, J. Correll, B. Wittenbrink, The Chicago face database: A free stimulus set of faces and norming data, Behavior research methods 47 (4) (2015) 1122–1135.
[23] X. Geng, Label distribution learning, IEEE Transactions on Knowledge and Data Engineering 28 (7) (2016) 1734–1748.
[24] B.-B. Gao, H.-Y. Zhou, J. Wu, X. Geng, Age estimation using expectation of label distribution learning, in: Proc. International Joint Conference on Artificial Intelligence, 2018, pp. 712–718.
[25] T. F. Cootes, G. J. Edwards, C. J. Taylor. Active appearance models, IEEE
