Mean-Variance Loss for Deep Age Estimation from a Face

Hongyu Pan[1,2], Hu Han*,[1], Shiguang Shan[1,2,3], and Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
hongyu.pan@vipl.ict.ac.cn, {hanhu,sgshan,xlchen}@ict.ac.cn

Abstract

Age estimation has wide applications in video surveillance, social networking, and human-computer interaction. Many of the published approaches simply treat age estimation as an exact age regression problem, and thus do not leverage a distribution's robustness in representing labels with ambiguity such as ages. In this paper, we propose a new loss function, called mean-variance loss, for robust age estimation via distribution learning. Specifically, the mean-variance loss consists of a mean loss, which penalizes difference between the mean of the estimated age distribution and the ground-truth age, and a variance loss, which penalizes the variance of the estimated age distribution to ensure a concentrated distribution. The proposed mean-variance loss and softmax loss are jointly embedded into Convolutional Neural Networks (CNNs) for age estimation. Experimental results on the FG-NET, MORPH Album II, CLAP2016, and AADB databases show that the proposed approach outperforms the state-of-the-art age estimation methods by a large margin, and generalizes well to image aesthetics assessment.[1]

1. Introduction

Age estimation from facial images has broad application scenarios, such as video surveillance, social networking, and human-computer interaction. Studies on age estimation from face images can be dated back to 1994 [28], in which different facial regions and skin wrinkles were used to estimate age group. A widely used feature representation by the early age estimation approaches is the biologically inspired features (BIF) [14], which reported much better accuracy than the traditional descriptors, such as LBP [34], Gabor [8], and SIFT [32]. In recent years, CNNs have shown great success on various computer vision tasks, such as face recognition [44, 36], object detection [37, 31], and scene segmentation [19]. Age estimation using CNNs has also obtained increasing attentions [23, 5], particularly with the promotion of the ChaLearn looking at people challenge[1].

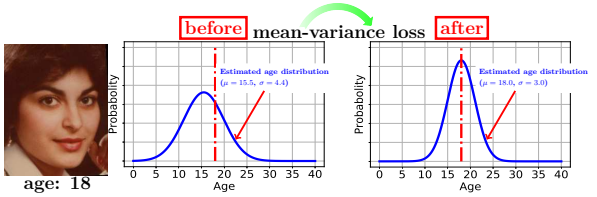

Figure 1. An example of age distribution learning using the proposed mean-variance loss. Our mean-variance loss aims to learn an age distribution which has not only a mean value close to the ground-truth age (red dotted line), but also a concentrated shape.

Existing approaches for age estimation can be grouped into three categories: classification based methods, regression based methods, and ranking based methods. Classification based methods are often used to estimate the age group of the subject in a face image [29, 48], which treat different ages or age groups as independent classes; therefore, the costs of classifying a young subject as middle-aged subject and old subject are the same. Apparently, such a modeling method is not optimum for the age estimation task.

Regression based methods are widely used to estimate the exact age of the subject in a face image [13, 7, 45]. Many of the existing regression based methods use a Euclidean loss (L_2 loss), which penalizes the differences between the estimated ages and the ground-truth ages. However, this type of loss defined based on a single image does not explicitly make use of the ordinal relationship among face images with individual ages.

In recent years, a few ranking based methods were pro-

*H. Han is the corresponding author.

[1]We plan to put the code into public domain: http://www.escience.cn/people/hhan/publication.html

[1]http://chalearnlap.cvc.uab.es

posed for age estimation from a face image [4, 2, 5]. These approaches treat the age value as a rank ordered data, and use multiple binary classifiers to determine the rank of the age in a face image. Different from the L_2 loss commonly used in regression methods, ranking based methods could explicitly make use of the ordinal relationship among face images with individual ages.

Despite a large amount of work on age estimation, their accuracy in unconstrained scenarios is still not sufficiently high in real application scenarios. This is due to the nature of the complicated face aging processing caused by both internal factors such as gene and external factors, such as living environment and lifestyle, as well as the ambiguity issue in the age label space. For age estimation by humans, it is relatively easy to give an age estimate with a particular confidence interval, such as a Gaussian distribution with a particular mean age and a standard deviation (see Fig. 1). Inspired by this observation, in this paper, we propose a mean-variance loss for age estimation which penalizes not only the difference between the mean of an estimated age distribution and the ground-truth age, but also the variance of the estimated age distribution. As a result, the estimated age distribution is expected to have a mean value as close to the ground-truth age as possible, and take a concentrated distribution as sharp as possible. The proposed approach is evaluated on a number of challenging databases (*e.g.*, FG-NET [35], MORPH Album II [25] and CLAP2016 [6]), and it achieves much better age estimation accuracy than the state-of-the-art methods. The main contributions of this work are three-fold: (i) Different from the existing methods which aim to estimate an exact age for a face image [14, 17, 49], we propose a new loss, named as mean-variance loss, aiming at the estimate of an age distribution with its mean as close to the ground-truth age as possible, and its variance as small as possible; (ii) Different from the age distribution learning methods such as [12, 46], the proposed approach does not require that each training image must have a mean age and a variance (neither real nor assumed) labels during model training, but it can still give a distribution estimate for a face image; (iii) The proposed loss can be easily embedded into different CNNs, and the network can be optimized via SGD [30] end-to-end.

2. Related Work

2.1. Age Estimation

Kwon and Lobo [28] did the very early work on age estimation from a face, in which the ages are divided into only three groups (*i.e.*, babies, young adults, and senior adults). After that, age estimation from a face image has attracted increasing attention. Lanitis *et al*. [25] employed an Active Appearance Model (AAM) [35] to combine the shape and texture (*e.g.*, wrinkles) information for age estimation. AAM was also used in [29], and multiple classifiers, such as shortest distance, quadratic functions, and artificial neural networks were used for age estimation. Guo *et al*. [14] used multi-directional and multi-scale Gabor filters followed by feature pooling to extract BIF features for age estimation. While BIF was reported to have promising age estimation results on several public-domain face databases such as FG-NET [35], MORPH II [25], it remains a hand-crafted feature representation, and thus may not be optimum for the age estimation task.

With the great success of deep learning methods in a number of computer vision tasks, such as object detection [37, 31, 19], image classification [27, 41], and face recognition [44, 22], deep learning methods are also being used in age estimation. Similar to [28], Yi *et al*. [49] used CNNs models to extract features from several facial regions, and used a square loss for age estimation. Niu *et al*. [33] utilized the ordinal information of ages to learn a network with multiple binary outputs for age estimation. Ordinal information was also used in [5] to learn multiple binary CNNs, and the outputs were aggregated. Rothe *et al*. [38] used the weights of the softmax classifier to calculate a weighted average age for age estimation, which was found to have better performance than using softmax for age classification. Han *et al*. [16, 43] proposed an effective deep multi-task learning approach for joint estimation of a large number of attributes, which consists of shared feature learning and attribute group specific feature learning. Yang *et al*. [46] and Huo *et al*. [23] proposed to perform age estimation via distribution learning, in which each age was represented as a distribution, and KL divergence was used to measure the similarity between the estimated and ground-truth distributions. However, in real applications, the mean and variance of a distribution are usually not available for a face image except for apparent ages collected via crowdsourcing.

2.2. Distribution Learning

In our approach, we also use the weights of softmax to calculate a weighted average age, but different from [38], we use the method in both the training and the testing phases. Besides, we penalize the variance so that the estimated age distribution could take a sharp shape, which is important to obtain an age estimate as accurate as possible. In addition, the proposed approach does not require each training face image to have a mean age and a variance value.

Distribution learning is proposed to address problems due to label ambiguity [10]. Different from single label learning or multi-label learning [42], which assigns a single label or multiple labels to an object, distribution learning assigns a label distribution to an object. Compared to single label learning and multi-label learning, distribution learning is able to leverage the relative relationship of a sequence of

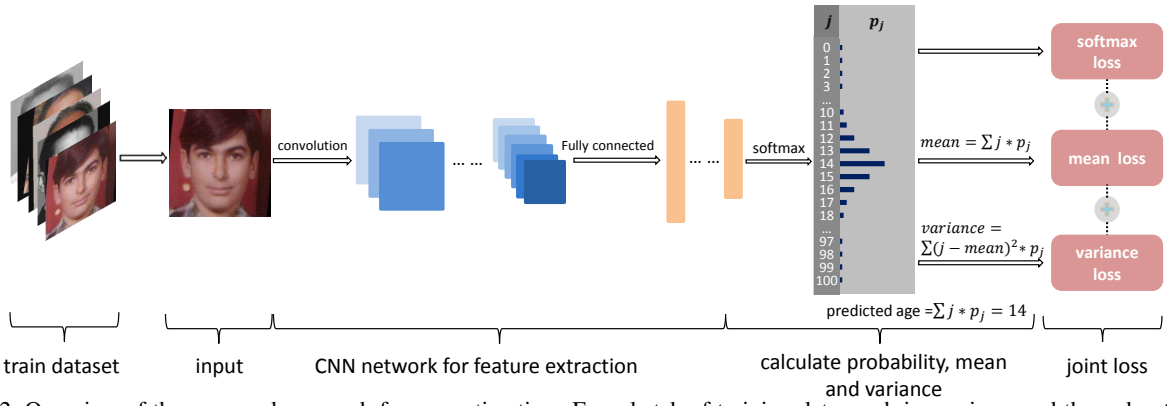

Figure 2. Overview of the proposed approach for age estimation. For a batch of training data, each image is passed through a CNN for feature extraction. A joint loss consisting of softmax and mean-variance loss is then used for backpropagation, in which the mean-variance loss penalizes not only the difference between the mean of an estimated age distribution and the ground-truth age, but also the variance of the estimated age distribution.

values in the label space, leading to more robust estimation.

Distribution learning has been utilized in a few computer vision tasks, such as expression recognize [50], head pose estimation [11], and age estimation [46, 47]. Zhou *et al*. [50] proposed a distribution learning based approach for degree estimation of all the basic emotions. Geng *et al*. have proposed a series of label distribution learning (LDL) methods for age estimation, and shown their effectiveness in resolving a number of issues, such as label ambiguity and data-dependent modeling [10, 9, 20]. Geng and Xia [11] used multivariate label distribution to alleviate the problem of inaccurate pose labels in the training set, and boost pose estimation accuracy without increasing the total amount of training data. Yang *et al*. [46] proposed a label distribution learning approach to model the uncertainty in the age labels collected via crowdsourcing, and then an age distribution instead of a single age value is estimated. We notice that while these approaches benefit from the robustness of representing a single label as a label distribution, there are still limitations: (i) a mean and variance are often assumed to be available for each sample in the training dataset; and (ii) the mean and variance of a distribution are not jointly considered and optimized during the model learning. We propose a distribution learning approach for age estimation which is able to address these issues.

3. Proposed Method

Fig. 2 gives the overview of our approach, in which the proposed mean-variance loss, together with the softmax loss, is embedded into a CNN for end-to-end learning. The details of our approach are given below.

3.1. Mean-Variance Loss

Formally, let x_i denote the feature vector of the i-th sample, $y_i \in \{1,2,...,K\}$ denote the corresponding age label, and $f(x_i) \in \mathbb{R}^{N\times M}$ denote the output of a CNN ahead of the last fully connected (FC) layer. The output of the last FC layer ($z \in \mathbb{R}^{N\times K}$), and a typical softmax ($p \in \mathbb{R}^{N\times K}$) probability can be computed using

$$z = f(x_i)\theta^T, \qquad p_{i,j} = \frac{e^{z_{i,j}}}{\sum_{k=1}^{K} e^{z_{i,k}}}, \tag{1}$$

where $\theta \in \mathbb{R}^{K\times M}$ is the parameter of the last FC layer, and $z_{i,j}$ is one element of z; $j \in \{1,2,...,K\}$ denotes the class labels (here it denotes the age); So p_i denotes the estimated age distribution for sample i over all the K classes, and $p_{i,j}$ denotes the probability that sample i belongs to class j.

Based on Eq. 1, we can compute the mean (m_i) and the variance (v_i) of a distribution (p_i) as follow

$$m_i = \sum_{j=1}^{K} j * p_{i,j}, \tag{2}$$

$$v_i = \sum_{j=1}^{K} p_{i,j} * (j - m_i)^2. \tag{3}$$

As shown in Figure 1, the mean-variance loss aims at penalizing not only the difference between the mean (m_i) of an estimated age distribution and the ground-truth age, but also the variance (v_i) of the estimated age distribution.

Mean Loss. The mean loss component in our mean-variance loss penalizes the difference between the mean of an estimated age distribution and the ground-truth age. Based on Eq. 2, the mean loss can be computed as

$$L_m = \frac{1}{2N}\sum_{i=1}^{N}(m_i - y_i)^2 = \frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=1}^{K} j * p_{i,j} - y_i)^2, \tag{4}$$

where N is the batch size. Different from softmax loss which focuses on classification tasks, our mean loss emphasizes on regression tasks, and we use the L_2 distance to measure the distance between the mean of an estimated age distribution and the ground-truth age. Therefore, it is complementary to the softmax loss.

Variance Loss. The variance loss component in our mean-variance loss penalizes the dispersion of an estimated

age distribution. Based on Eqs. 2 and 3, the variance loss can be computed as

$$L_v = \frac{1}{N}\sum_{i=1}^{N} v_i = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{i,j} * (j - \sum_{k=1}^{K} k * p_{i,k})^2. \quad (5)$$

Such a variance loss requires that an estimated distribution should be concentrated at a small range of the mean. Take a Gaussian distribution as an example, the variance loss makes it as sharp as possible. This is helpful to obtain an accurate age estimation with a small confidence interval but a high confidence.

3.2. Embedding into CNNs

Since face aging is a complicated process which is affected by both internal factors such as gene and external factors, for instance, living environment, lifestyle, etc. [16], the mapping from the face image space into the age label space can be quite nonlinear. Therefore, it is reasonable to use CNNs as our basic feature representation approach to model the complicated face aging process. Specifically, we embed our mean-variance loss into the architecture of CNNs, and use the softmax loss (*i.e.*, L_s) and mean-variance loss jointly as the supervision signal

$$\begin{aligned} L &= L_s + \lambda_1 L_m + \lambda_2 L_v \\ &= \frac{1}{N}\sum_{i=1}^{N} -log p_{i,y_i} + \frac{\lambda_1}{2}(m_i - y_i)^2 + \lambda_2 v_i, \end{aligned} \quad (6)$$

where λ_1 and λ_2 are two hyper-parameters, balancing the influencing of individual sub-losses in the joint loss. In addition, the mean-variance loss and softmax loss have very different scales, normalization of individual losses are necessary to assure stable network training. The reason why we use softmax and the proposed mean-variance jointly is that a randomly initialized network with mean-variance loss alone may have large fluctuation at the early stage of training, and thus a joint use of softmax and mean-variance losses can help the network converge as early as possible. We perform SGD [30] to optimize the weights of the network. In the inference phase, the age of a test image is estimated as

$$y_t = r(\sum_{j=1}^{K} j * p_j), \quad (7)$$

where $p_j, j \in \{1, 2, ..., K\}$ is the output of the softmax layer in the network, and $r(\cdot)$ is a round function.

The reasons why the proposed mean-variance loss benefits the age estimation network training can be summarized into three aspects:

(i) **Shift the estimated distribution to the ground-truth.**

By this effect, we mean that in the iteration progresses, the network could gradually generate a better distribution with its mean getting closer to the ground-truth age. According to Eq. 4, the gradient of mean loss L_m w.r.t. $p_{i,j}$ can be computed as

$$\frac{\partial L_m}{\partial p_{i,j}} = \frac{1}{N}(m_i - y_i) * j. \quad (8)$$

The gradient of $p_{i,j}$ with respect to $z_{i,r}$ can be computed as

$$\frac{\partial p_{i,j}}{\partial z_{i,r}} = \frac{e^{z_{i,r}}}{\sum_{k=1}^{K} e^{z_{i,k}}} - (\frac{e^{z_{i,r}}}{\sum_{k=1}^{K} e^{z_{i,k}}})^2 = p_{i,r} - p_{i,r}^2, \; if \;\; r = j, \quad (9)$$

and

$$\frac{\partial p_{i,j}}{\partial z_{i,r}} = -\frac{e^{z_{i,j}}}{(\sum_{k=1}^{K} e^{z_{i,k}})^2} * e^{z_{i,r}} = -p_{i,r} * p_{i,j}, if \;\; r \neq j. \quad (10)$$

Based on Eqs. 8, 9 and 10, the gradient of L_m w.r.t. $z_{i,j}$ could be written as

$$\begin{aligned} \frac{\partial L_m}{\partial z_{i,j}} &= \frac{(m_i - y_i)}{N}(j * (p_{i,j} - p_{i,j}^2) - \sum_{k=1,k!=j}^{K} k * p_{i,k} * p_{i,j}) \\ &= \frac{(m_i - y_i)}{N} p_{i,j}(j - m_i). \end{aligned} \quad (11)$$

According to the Eq. 11, for an estimated distribution with mean value m_i, if $m_i < y_i$, the network will be updated to increase the probabilities of the classes j ($j > m_i$) via their negative gradients, and decrease the probability of those classes j ($j < m_i$) via their positive gradients. In this way, the mean value of the estimated distribution will be increased, and becomes closer to y_i. Similarly, if $m_i > y_i$, the network will be updated so that the mean value of the estimated distribution will be decreased and gets closer to y_i.

(ii) **Squeeze the estimated distribution from both sides.**

This effect means that in the iteration progresses, the network could generate a sharp distribution. In other words, the closer the class to m_i, the larger its probability. Specifically, the gradient of variance loss L_v w.r.t. $p_{i,j}$ can be computed as

$$\frac{\partial L_v}{\partial p_{i,j}} = \frac{1}{N}((j - m_i)^2 - 2 * j * \sum_{k=1}^{K} p_{i,k}(k - m_i)) = \frac{1}{N}(j - m_i)^2. \quad (12)$$

Finally, based on Eqs. 9, 10 and 12, the gradient of L_v w.r.t. $z_{i,j}$ could be written as

$$\begin{aligned} \frac{\partial L_v}{\partial z_{i,j}} &= \frac{1}{N}((j - m_i)^2 (p_{i,j} - p_{i,j}^2) - \sum_{k=1,k\neq j}^{K} (k - m_i)^2 * p_{i,k} * p_{i,j}) \\ &= \frac{1}{N} p_{i,j}((j - m_i)^2 - \sum_{k=1}^{K} (k - m_i)^2 * p_{i,k}), \end{aligned} \quad (13)$$

where $\alpha = \sum_{k=1}^{K} (k - m_i)^2 * p_{i,k}$ is a nonnegative constant for a certain sample i. So Eq. 13 could be simplified as

$$\frac{\partial L_v}{\partial z_{i,j}} = \frac{1}{N} p_{i,j}((j - m_i)^2 - \alpha), \quad (14)$$

The gradient in Eq. 14 has the following properties:

$$j \in (m_i - \sqrt{\alpha}, m_i + \sqrt{\alpha}), \frac{\partial L_v}{\partial z_{i,j}} < 0, \quad (15)$$

and

$$j \in [1, m_i - \sqrt{\alpha}) \cup (m_i + \sqrt{\alpha}, K], \frac{\partial L_v}{\partial z_{i,j}} > 0. \quad (16)$$

Eq. 15 shows that, the network will be updated to increase the probabilities of the classes j close to m_i ($j \in (m_i - \sqrt{\alpha}, m_i + \sqrt{\alpha})$) via their negative gradients. On the contrary, Eq. 16 shows that the network will be updated to decrease the probabilities of the classes j far away from m_i ($j \in [1, m_i - \sqrt{\alpha}) \cup (m_i + \sqrt{\alpha}, K]$) via their positive gradients.

(iii) **Assign different degrees of contributions to individual classes.**

This effect reflects the differences between the age estimation problem and general classification problems. For general softmax loss, given one positive class, all the negative classes are treated with no difference when updating the network. Differently, in our task, age labels are ordinal and comparable. Therefore, as shown in Eqs. 11 and 13, the gradient of each class is weighted according to its distance ($j - m_i$) from the current age label. In other words, each class is assigned a different degree of contribution when updating the network, which can benefit distinguishing the different classes (ages).

4. Experiments

We provide extensive evaluations of the proposed age estimation approach and comparisons with the state-of-the-art methods on several public-domain face aging databases including MORPH Album II [25], FG-NET [35], and CLAP2016 [6]. In addition, we evaluate the generalization ability of the proposed approach to other tasks, *i.e.*, image aesthetics assessment on AADB [26].

4.1. Datasets

MORPH Album II is one of the largest longitudinal face databases in the public domain, which contains 55,134 face images of 13,617 subjects and the range from 16 to 77 [25]. We use two types of widely used testing protocols in our evaluations. One is the five-fold random split (RS) protocol for all the images [5, 4, 33, 38, 3]; the other is the five-fold subject-exclusive (SE) protocol [15, 18, 16]. The latter testing protocol is more challenging since it assures the images of one subject only appear in one fold.

FG-NET database was a very early database used for age estimation, which contains 1,002 face images from 82 individuals and the ages range from 0 to 69 [35]. We follow a widely used leave-one-person-out (LOPO) protocol [4, 38, 3] in our experiments.

Age Range	MORPH II	FG-NET	CLAP2016
0-19	7,469	710	1,394
20-39	31,682	223	4,362
40-59	15,649	61	1,423
60-69	334	8	366
≥ 80	0	0	46
Total Image	55,134	1,002	7,591

Table 1. Age distributions of the face images in the MORPH II, FG-NET and CLAP2016 databases.

CLAP2016 dataset was released in 2016 at the ChaLearn Looking at people challenge, which contains 4,113, 1,500, and 1,979 face images in the training set, validation set, and testing set, respectively [6]. Different from the MORPH II and FG-NET databases, the ages provided in the CLAP2016 dataset are apparent ages collected via crowdsourcing, so there is a mean age and a variance for each face image.

The age distributions of the MORPH II [25], FG-NET [35], and CLAP2016 [6] dataset are shown in Table 1.

AADB contains 10,000 natural images, each containing a aesthetic quality rating, and attribute assignments provided by five different raters. We follow the same protocol as [26] to perform image aesthetics assessment.

4.2. Evaluation Metrics

We report the mean absolute error (MAE) [23] and cumulative score (CS) [14] on the MORPH II and FG-NET databases. MAE is defined as the mean absolute error between the estimated age ($\hat{y}_i$) and ground-truth age (y_i). CS measures the age estimation accuracy given a tolerance of absolute error. For the CLAP2016 dataset, we use the ε-error [6] defined in the standard testing protocol $\varepsilon = 1 - \frac{1}{N}\sum_{i=1}^{N} e^{-\frac{(y_i-\mu_i)^2}{2\sigma_i^2}}$, where μ and σ are the ground-truth mean age and standard deviation, respectively.

For image aesthetics assessment on AADB, we report the ρ value [26] provided with the standard testing protocol $\rho = 1 - \frac{6\sum d_i}{N^3 - N}$, where $d_i = r_i - \hat{r}_i$, r_i and $\hat{r}_i$ denote the real and estimated ranks calculated by the overall score and the estimated score, respectively.

4.3. Experiment Settings

We align all the face images based on five facial landmarks detected using an open-source SeetaFaceEngine [2], and resize all the face images into $256 \times 256 \times 3$.

Two CNNs, *i.e.*, AlexNet [27] with batch normalization [24] and VGG-16 [41, 36] are used in our age estimation approach. Both models are pre-trained on ImageNet 2012 [39]. Besides, the VGG-16 model [41, 36] is also pre-trained using IMDB-WIKI, which is a large scale

[2] https://github.com/seetaface/SeetaFaceEngine

Method	RS		SE	
	MAE	CS(θ=5)	MAE	CS(θ=5)
Softmax loss	3.324	78.96%	4.043	73.24%
Euclidean loss	3.289	79.52%	3.932	74.73%
Proposed loss + softmax loss	**2.514**	**88.90%**	**3.086**	**83.72%**

Table 2. Comparisons of the age estimation MAEs (in years) and CS ($\theta = 5$) (in %) by different losses on the MORPH II database.

face database with age and gender labels [38].[3] We use an initial learning rate of 0.001 and a batch size of 64 for both AlexNet and VGG-16, and reduce the learning rate by multiplying 0.1 for every 10 epochs (AlexNet), and 15 epochs (VGG-16). The input face images are randomly cropped to 224×224 and 227×227, respectively.

4.4. Age Estimation Results

4.4.1 Comparisons of Different Losses

To validate the effectiveness of our mean-variance loss, we first compare it with two widely used losses in age estimation task, *e.g.*, softmax loss and Euclidean loss by performing age estimation with AlexNet [40] on the MORPH II [25] database using both the RS and SE protocols. The MAE and CS ($\theta = 5$) of the three different losses are shown in Table 2.

We can see that Euclidean loss outperforms softmax loss for age estimation. This is reasonable because softmax loss does not differentiate between classifying a 10-year old subject into 15-year old and into 50-year old. Using a joint loss of softmax and mean-variance leads to the best performance, *i.e.*, 2.5 years MAE, and 3.1 years MAE under the RS and SE protocols, respectively. This shows the benefit of using a distribution learning over the single label learning for age estimation from a face image. Beside the reason explained in Section 3.2, the reason why softmax loss, instead of Euclidean loss, is used jointly with our mean-variance loss is that Euclidean loss and the mean loss component in our loss are essential the same type of loss. Thus, a joint use of softmax loss and our mean-variance loss provides better complementarity.

4.4.2 Influences of the parameters λ_1 and λ_2

Since hyper-parameters λ_1 and λ_2 in Eq. 6 balance the three loss components (softmax, mean, and variance) during network learning, we evaluate their influences in age estimation on MORPH II using the AlexNet model.

[3]This database is large, but the age labels can be quite noisy because they are calculated based on the date of birth of the public figures and the timestamps of the photos crawled from the Internet.

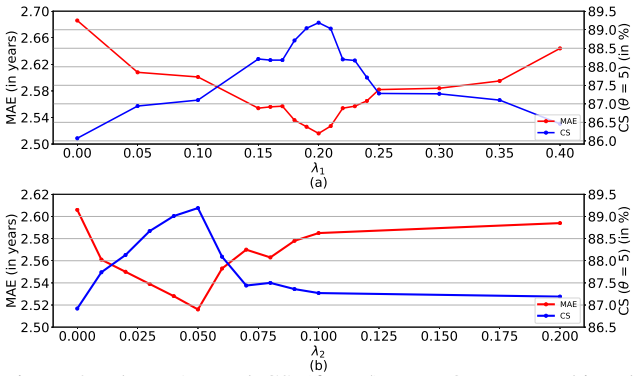

Figure 3. The MAE and CS ($\theta = 5$) on MORPH II achieved by AlexNet using (a) different λ_1 and fixed $\lambda_2 = 0.05$, and (b) different λ_2 and fixed $\lambda_1 = 0.2$.

We first fix λ_2 to 0.05 and change λ_1 from 0 to 0.4 to learn different models. The MAE and CS ($\theta = 5$) of these models are shown in Figure 3 (a). We can see that only using the softmax and variance losses (*i.e.*, $\lambda_1 = 0$) is not a good choice. But λ_1 should be not too big; otherwise, there will be a big performance degradation. As we explained in Section 3.2, the main reason is that the network becomes difficult to converge if the mean-variance loss takes a dominant role. We choose to use $\lambda_1 = 0.2$ in our experiments.

We then fix the λ_1 to 0.2 and change λ_2 from 0 to 0.2 to learn different models. The MAE and CS ($\theta = 5$) of these models are shown in Figure 3 (b). We can see that only using the softmax and mean losses (*i.e.*, $\lambda_2 = 0$) is not a good choice, but compared with the mean loss, the variance loss has a relatively smaller impact on the performance of the network. This is understandable if we look at Eq. 7, in which the final age estimate is calculated as a weighted mean of the entire age distribution. Finally, we choose to use $\lambda_2 = 0.05$ in our following experiments.

4.4.3 Comparisons with the State-of-the-art

We compare the proposed method with a number of the state-of-the-art methods such as Ranking-CNN [5], DEX [38], RED-SVM [3], and DIF [16], for age estimation on FG-NET using AlexNet [40], MORPH II and CLAP2016 using VGG-16 [41, 36], respectively.

Tables 3 and 4 show the MAEs of individual methods on MORPH II and FG-NET. The results suggest that ranking-based methods, such as [5, 4, 33], usually perform better than classification or regression based methods [38, 3]. This is reasonable because ranking-based methods utilize the ordinal relationship and improve the age estimation robustness. Our method performs the best among all the approaches, because our method benefits from not only distribution learning but also the additional constraints introduced to the distribution via mean-variance loss. In addition, we notice that our models pre-trained on ImageNet and

Method	MAE	Protocol
RED-SVM [3]	6.49	RS
OHRank [4]	6.07	RS
OR-CNN [33]	3.27	RS
DEX [38]	3.25	RS
DEX*	2.68	RS
DIF [16]	3.00	SE
Ranking-CNN [5]	2.96	RS
Proposed	**2.41/2.80**	RS/SE
Proposed*	**2.16/2.79**	RS/SE

Table 3. Comparisons of the age estimation MAEs by the proposed approach and the state-of-the-art methods on the MORPH II. * The IMDB-WIKI dataset was used to pre-train the model.

Method	MAE	Protocol
RED-SVM [3]	5.24	LOPO
OHRank [4]	4.48	LOPO
DEX [38]	4.63	LOPO
DEX*	3.09	LOPO
Proposed	**4.10**	LOPO
Proposed*	**2.68**	LOPO

Table 4. Comparisons of the age estimation MAEs by the proposed approach and the state-of-the-art methods on FG-NET. * The IMDB-WIKI dataset was used to pre-train the model.

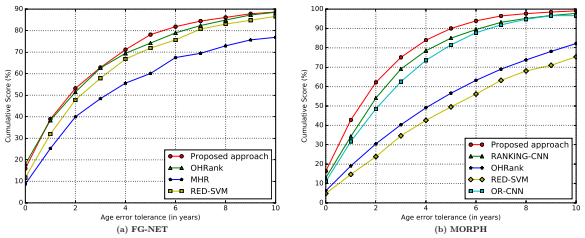

Figure 4. Comparisons of the age estimation cumulative scores by the proposed approach and the state-of-the-art methods on the FG-NET dataset with a LOPO protocol, and the MORPH II dataset with a random split (RS) protocol.

IMDB-WIKI have very similar performance using SE protocol. This suggests that our method has a good generalization ability into unseen scenarios. Figure 4 shows the entire CS curves on the MORPH II and FG-NET databases using the RS and LOPO protocols, respectively. We can see that the proposed approach performs consistently better than the state-of-the-art methods.

The age estimation results by our approach and a number of the state-of-the-art approaches [6] on CLAP2016 are reported in Table 5. The first-place method [1] in the CLAP2016 competition reported a lower error than our method, but they used a score-level fusion of multiple CNN models. Such a method is likely to have large memory and computational cost consumptions. In addition, a number

Rank	Team Name	ε-error	Single model?
1	OrangeLabs [1]	0.2411	NO
*	**Proposed**	0.2867	**YES**
2	palm_seu [23]	0.3214	NO
3	cmp+ETH	0.3361	NO
4	WYU_CVL	0.3405	NO
5	ITU_SiMiT	0.3668	NO
6	Bogazici	0.3740	NO
7	MIPAL_SNU	0.4565	NO
8	DeepAge	0.4573	YES

Table 5. Comparisons of the age estimation ε-errors by the proposed approach and the state-of-the-art methods on the CLAP2016 database. The results of the state-of-the-art methods are from [6].

Method	Cls	Reg	[26]	EMD [21]	Proposed
ρ	0.5923	0.6239	0.6782	0.6682	0.6647

Table 6. Comparisons between our approach and the state-of-the-art methods on the AADB dataset in terms of ρ value.

of children's face images were collected from the Internet in [1] to improve the corresponding age estimation accuracy. The second-place method [23] also used an age distribution learning method and achieved an ε error of 0.3214. However, our approach performs much better than [23], which suggests that the proposed mean-variance loss is very effective for the age estimation task.

Figures 5 and 6 show some examples of good and poor age estimation results by our approach on MORPH II, FG-NET, and CLAP2016. We can see that the proposed approach performs quite robust for young, middle-aged, and old subjects. The age estimation accuracy may decrease when the face images have very bad illumination, large (self-) occlusion, and blurring (see the bottom rows of Figures 5 and 6).

4.5. Image Aesthetics

We use AlexNet with the proposed mean-variance loss to perform image aesthetics assessment on AADB, and compare the results with two state-of-the-art methods [26, 21]. The results of individual approaches in terms of ρ value are reported in Table 6. From Table 6, we can see that even though [21] uses a much deeper network (VGG-16) and [26] utilizes more information, such as attributes, rank and content features, our approach achieves results comparable to these methods, which suggests that our approach can generalize well to image aesthetics regression task.

5. Conclusions

In this paper, we propose a mean-variance loss for robust age estimation via distribution learning. We show that the proposed loss is useful for obtaining a concentrated yet ac-

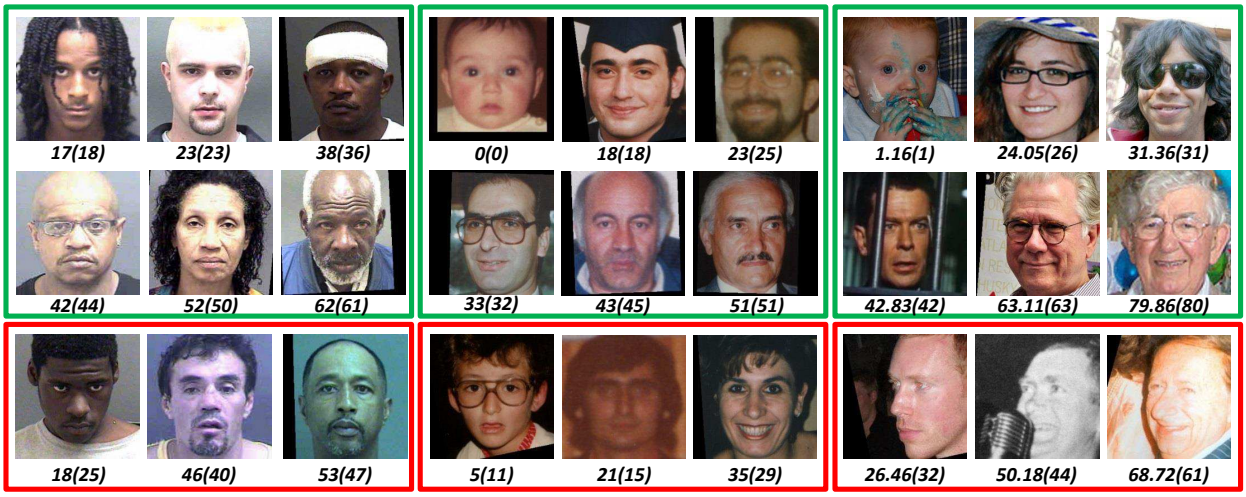

Figure 5. Examples of age estimation results by the proposed approach on the (a) MORPH II, (b) FG-NET and (c) CLAP2016. The top two rows show some good age estimation examples, and the third row shows some poor age estimation examples. The numbers below each image show the ground-truth age and estimated age of the subject, *i.e.*, ground-truth age (estimated age).

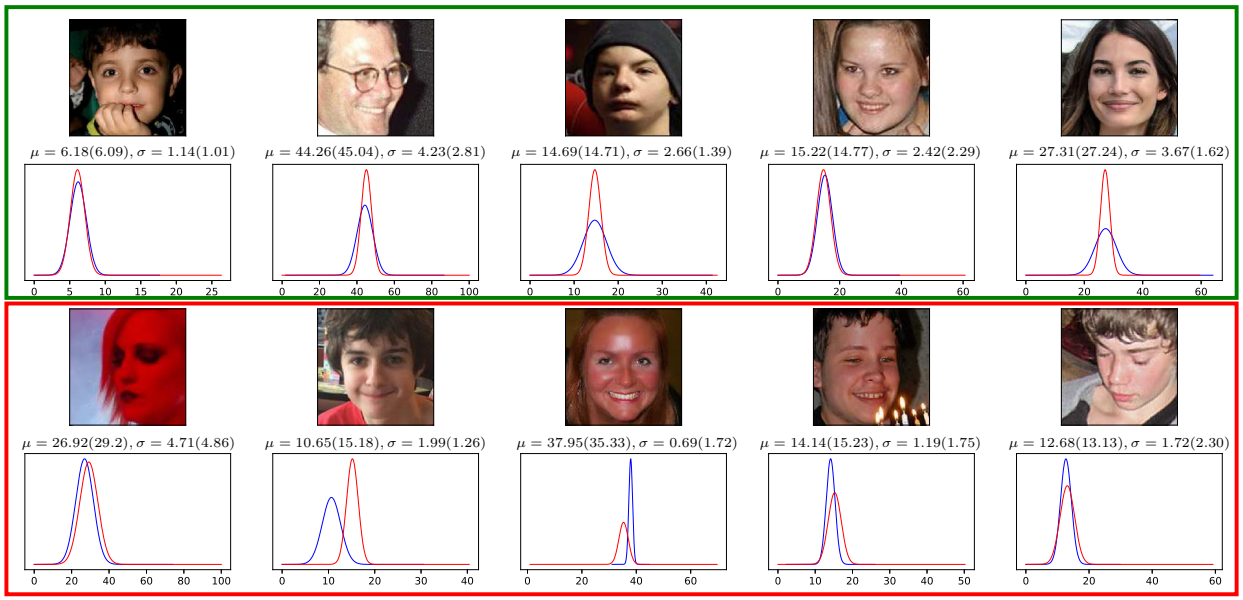

Figure 6. Examples of distributions estimated by our approach on the CLAP2016. The top two rows show good distribution results, and the bottom two rows show poor distribution results. The red and blue curves are the estimated and ground-truth distributions, respectively. The numbers below each image show the age and std of ground-truth and estimated of the subject, *i.e.*, μ = ground-truth age (estimated age), σ = ground-truth std (estimated std).

curate label distribution estimation during age estimation. Experiments on the MORPH II, FG-NET, CLAP2016, and AADB databases show that our approach performs better than the state-of-the-art methods, and generalizes well into image aesthetics assessment task. In our future work, we would like to study feature representations that are robust to large pose and illumination variations. In addition, we would like to investigate the effectiveness of the proposed loss in other learning tasks.

Acknowledgement This research was supported in part by the National Basic Research Program of China (grant 2015CB351802), Natural Science Foundation of China (grants 61732004 and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), Strategic Priority Research Program of CAS (grant XDB02070004), and Youth Innovation Promotion Association CAS (2018135).

References

[1] G. Antipov, M. Baccouche, S. Berrani, and J. Dugelay. Apparent age estimation from face images combining general and children-specialized deep learning models. In *IEEE CVPR Workshops*, pages 801–809, 2016.

[2] K. Y. Chang and C. S. Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Trans. Image Process.*, 24(3):785–798, Jan. 2015.

[3] K. Y. Chang, C. S. Chen, and Y. P. Hung. A ranking approach for human age estimation based on face images. In *ICPR*, pages 3396–3399, 2010.

[4] K. Y. Chang, C. S. Chen, and Y. P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *IEEE CVPR*, pages 585–592, 2016.

[5] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-CNN for age estimation. In *IEEE CVPR*, pages 5183–5192, 2017.

[6] S. Escalera, M. Torres, B. Martinez, X. Bar, H. J. Escalante, I. Guyon, I. Guyon, I. Guyon, M. Oliu, M. A. Bagheri, and M. A. Bagheri. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *IEEE CVPR Workshops*, pages 706–713, 2016.

[7] Y. Fu and T. Huang. Human age estimation with regression discriminative aging manifold. *IEEE Trans. Multimedia*, 10(4):578–584, May 2008.

[8] D. Gabor. Theory of communication. *J. Inst. Electr. Eng.*, 93(26):429–457, Nov. 1946.

[9] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *IEEE Trans. Image Process.*, 26(6):2825–2838, 2017.

[10] X. Geng. Label distribution learning. *IEEE Trans. Knowl. Data Eng.*, 28(7):1734–1748, July 2016.

[11] X. Geng and Y. Xia. Head pose estimation based on multivariate label distribution. In *IEEE CVPR*, pages 1837–1842, 2014.

[12] X. Geng, C. Yin, and Z. H. Zhou. Facial age estimation by learning from label distributions. *IEEE Trans. Pattern Anal. Mach. Intell*, 35(10):2401–2412, Oct. 2013.

[13] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Process.*, 17(7):1178–1188, May 2008.

[14] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. In *IEEE CVPR*, pages 112–119, 2009.

[15] G. Guo and X. Wang. A study on human age estimation under facial expression changes. In *IEEE CVPR*, pages 2547–2553, 2012.

[16] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, Aug. 2017.

[17] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *IEEE ICB*, pages 1–8, 2013.

[18] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):1148–1161, June 2015.

[19] K. He, G. Gkioxari, P. D. ar, and R. Girshick. Mask R-CNN. In *IEEE ICCV*, pages 2980–2988, 2017.

[20] Z. He, X. Li, Z. Zhang, F. Wu, X. Geng, Y. Zhang, M.-H. Yang, and Y. Zhuang. Data-dependent label distribution learning for age estimation. *IEEE Trans. Image Process.*, 26(8):3846–3858, 2017.

[21] L. Hou, C.-P. Yu, and D. Samaras. Squared earth movers distance loss for trainingdeep neural networks on ordered-classes. In *NIPS Workshop*, 2017.

[22] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Report 07-49, University of Massachusetts, Amherst*, 2007.

[23] Z. Huo, X. Yang, C. Xing, Y. Zhou, P. Hou, J. Lv, and X. Geng. Deep age distribution learning for apparent age estimation. In *IEEE CVPR Workshops*, pages 722–729, 2016.

[24] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducinginternal covariate shift. *arXiv:1502.03167v3*, 2015.

[25] K. R. Jr. and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *IEEE FG*, pages 341–345, 2006.

[26] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, pages 662–679, 2016.

[27] A. Krizhevsky, S. Ilya, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[28] Y. Kwon and N. Lobo. Age classification from facial images. In *IEEE CVPR*, pages 762–767, 1994.

[29] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. Syst. Man, Cybern. B, Cybern*, 34(1):621–628, Jan. 2004.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2015.

[32] D. Lowe. Object recognition from local scale-invariant features. In *IEEE ICCV*, pages 1150–1157, 1999.

[33] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output CNN for age estimation. In *IEEE CVPR*, pages 4920–4928, 2016.

[34] T. Ojala, M. Pietikinen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *ICPR*, pages 582–585, 1994.

[35] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *IET Biom*, 5(2):37–46, May 2015.

[36] O. M. Parkhi, A. Vedaldia, and A. Zisserman. Deep face recognition. In *BMVC*, pages 41.1–41.12, 2015.

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[38] R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.*, 126(2):1–14, Aug. 2016.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, Dec. 2015.

[40] M. Simon, E. Rodner, and J. Denzler. ImageNet pre-trained models with batch normalization. *arXiv preprint arXiv:1612.01452*, 2016.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–14, 2015.

[42] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, Sept. 2009.

[43] F. Wang, H. Han, S. Shan, and X. Chen. Multi-task learning for joint prediction of heterogeneous face attributes. In *IEEE FG*, pages 173–179, 2017.

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

[45] B. Xiao, X. Yang, H. Zha, Y. Xu, and T. Huang. Metric learning for regression problems and human age estimation. In *PCM*, pages 88–99, 2009.

[46] X. Yang, B. B. Gao, C. Xing, Z. W. Huo, X. S. Wei, Y. Zhou, J. Wu, and X. Geng. Deep label distribution learning for apparent age estimation. In *IEEE ICCV Workshops*, pages 102–108, 2015.

[47] X. Yang, X. Geng, and D. Zhou. Sparsity conditional energy label distribution learning for age estimation. In *IJCAI*, pages 2259–2265, 2016.

[48] Z. Yang and H. Ai. Demographic classification with local binary patterns. In *ICB*, pages 464–473, 2007.

[49] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *IEEE ACCV*, pages 144–158, 2015.

[50] Y. Zhou, H. Xue, and X. Geng. Emotion distribution recognition from facial expressions. In *ACM MM*, pages 1247–1250, 2015.