PML: Progressive Margin Loss for Long-tailed Age Classification

Zongyong Deng[1], *Hao Liu*[1,2,*], *Yaoxing Wang*[1], *Chenyang Wang*[1], *Zekuan Yu*[3], *Xuehong Sun*[1,2]
[1]School of Information Engineering, Ningxia University, Yinchuan, 750021, China
[2]Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence
Co-founded by Ningxia Municipality and Ministry of Education, Yinchuan, 750021, China
[3]Academy for Engineering and Technology, Fudan University, Shanghai, 200433, China
zongyongdeng_nxu@outlook.com; liuhao@nxu.edu.cn; yaoxing.wang_nxu@outlook.com
chenyang.wang_nxu@outlook.com; yzk@fudan.edu.cn; sunxh@nxu.edu.cn

Abstract

In this paper, we propose a progressive margin loss (PML) approach for unconstrained facial age classification. Conventional methods make strong assumption on that each class owns adequate instances to outline its data distribution, likely leading to bias prediction where the training samples are sparse across age classes. Instead, our PML aims to adaptively refine the age label pattern by enforcing a couple of margins, which fully takes in the in-between discrepancy of the intra-class variance, inter-class variance and class center. Our PML typically incorporates with the ordinal margin and the variational margin, simultaneously plugging in the globally-tuned deep neural network paradigm. More specifically, the ordinal margin learns to exploit the correlated relationship of the real-world age labels. Accordingly, the variational margin is leveraged to minimize the influence of head classes that misleads the prediction of tailed samples. Moreover, our optimization carefully seeks a series of indicator curricula to achieve robust and efficient model training. Extensive experimental results on three face aging datasets demonstrate that our PML achieves compelling performance compared to state of the art. Code will be made publicly.

1. Introduction

Facial age classification (*a.k.a.*, facial age estimation) aims to predict the exact biological ages from given facial images, which has a lot of potential computer vision applications such as human-computer interaction [53, 15] and facial attribute analysis [39, 4]. While numerous works have been devoted recently [17, 27, 18, 50, 43], the performance still remains limited in wild conditions, which is mainly due to that the datasets often undergo long-tailed distribution with many minority classes (tail) and a few common classes (head). When learning with the long-tailed age data, a common problem is that the head classes usually dominate the training convergence. Therefore, the learned age classification model tends to perform better on head classes, whereas the performance degrades in tail classes. This quite motivates us to develop a robust facial age classification approach versus imbalanced age data. In the left of Fig. 1, we visualize some failure cases caused by existing age classification methods.

Facial age classification approaches could be roughly divided into the single label learning (SLL)-based [17, 18, 43, 11] and the label distribution learning (LDL)-based [50, 31, 51, 52, 36]. SLL-based methods typically classify one single age for a given facial image, which treats each age independently. However, they ignore human face changes gradually with progressive ages, thus the facial appearance is usually indiscriminative at adjacent age classes. To further model the age correlation, Geng *et al*. [19] proposed an LDL method to map the real-valued ground-truth to a Gaussian label distribution. However, the performance degrades in such long-tailed case where the feature representation of minority is suppressed by the majority classes.

To address the long-tailed data issue, we propose a progressive margin loss (PML) approach for age classification, which aims to leverage semantic margins to reduce intra-class variance and enlarge inter-class variance simultaneously. As shown in Fig. 2, we carefully develop a progressive margin loss at the top of deep neural networks with preserving the age-difference cost information. Technically, our proposed PML is composed of two crucial branches including an ordinal margin learning and a variational margin learning. The ordinal margin attempts to extract discriminative features while maintaining the relation of the age order. For efficient optimization, we develop a series of indicators by following the curriculum-learning method. To validate the effectiveness of our proposed method, we per-

*Corresponding author.

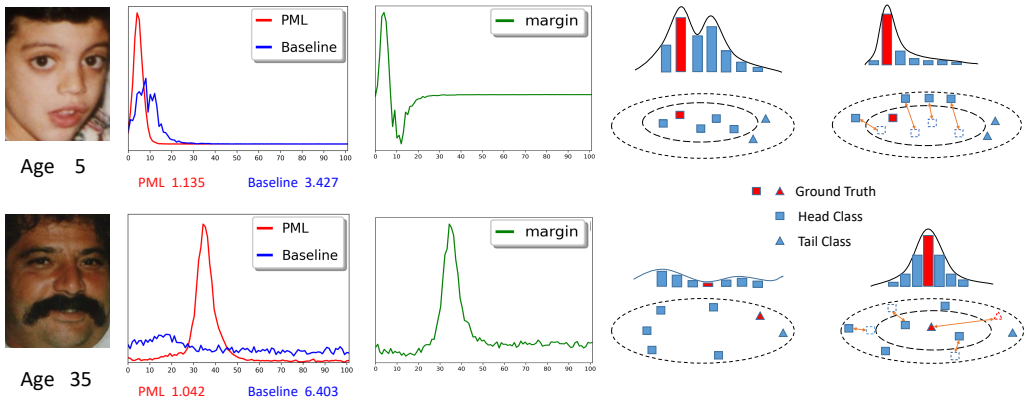


Figure 1. Our approach versus existing label distribution learning approaches. We expect the ground truth (red square or red triangle) to be at the center of the label prediction, where one sample in the head category (blue square) and another sample in the tail category (blue triangle) should be enforced by a margin from the real age. It is valuable to be notified that the dotted frame represents the position before adjustment. Top: The baseline method reasons the multi-modal distribution, because the head classes dominate the tail classes. Our proposed PML addresses this error by preventing the tail class from disturbance of the head. Bottom: The baseline method could hardly find effective features in tail categories and limited to output the uniform distribution. Our PML achieves robust feature representation by integrating the relation of adjacent age classes. (Best viewed in color PDF file.)

form extensive experiments on three widely-used face aging datasets, where each dataset undergoes varying degrees of the imbalance. From the results, we achieve superior performance compared with the state-of-the-art methods especially only with fewer samples. For example, without using any external datasets, we decrease the MAE by 1.56 compared with the recently reported benchmark with only sparse and limited samples.

2. Related Work

In this section, we briefly review the related works on facial age estimation and imbalanced classification, respectively.

Facial Age Classification. Conventional age classification methods could be roughly divided into two types: feature representation [2, 9, 12] and age prediction [16, 8]. Feature representation-based methods aim to exploit discriminative feature patterns from the facial images. Respectively, age prediction-based methods learn to classify the age labels with the extracted features. However, both types are optimized in a two-stage manner, which likely leads to local solution. To circumvent this limitation, deep learning has been applied to jointly optimize both procedures of feature representation and age prediction. For example, Rothe *et al.* [50] formulated the age estimation as the expectation-based classification problem, where the prediction is accomplished by maximizing the expectation of outputting logits. Nevertheless, these methods hardly exploit the full chronological relationship of practical ages. To introduce the age label correlation to the model, Liu *et al.* [31] proposed an ordinal deep feature learning (ODFL) method, which enforces both the topology ordinal relation and the age-difference information in the learned feature space. Furthermore, Li *et al.* [27] developed the BridgeNet to model the ordinal relation of age labels via gated local regressors. To further alleviate the problem of label ambiguity, Geng *et al.* [19] designed a distribution learning approach to transform the single scalar label to a vector. Nevertheless, the LDL schema gives rise to bias in predicting minority classes, where the samples within each age class are variant in appearance. We cope with this issue by a progressive margin loss framework, which elaborately adjusts the learned age patterns by fully considering the distributed property of neighboring age classes.

Imbalanced Classification. With the remarkable success achieved by data-driven CNNs [26, 32, 40, 19], deep models have witnessed that the generation capacity is limited especially for imbalanced and distributed data [6, 25]. Existing imbalanced classification methods are coarsely divided into re-sampling [3, 6, 48] and cost-sensitive loss function [28, 30, 37]. Accordingly, re-sampling-based methods aim to balance the scalability of the head classes and tail classes, but such schema easily prones to overfitting in the tail classes. The major reason is that the training model memorizes irrelevant noise when utilizing the tailed data repeatedly [29]. Cost-sensitive methods are developed to improve the influence of minority classes by assigning higher misclassification costs to the minority class than to the majority ones. In addition, both works [28, 34] were proposed by the focal loss to mine hard-negative in-

stances online and adaptive margin softmax to adjust the margins for different classes adaptively. However, these methods ignore the original relationship within samples *w.r.t.* neighboring age classes. As far as one can tell from the literature, few works of imbalanced classification have been visited yet in facial age classification.

3. Approach

In this work, we claim that *margin matters* for robust age label distribution learning. To achieve the proposed progressive margin loss, we enforce our model to integrate with the practical age progression in the learned distribution, which is semantic and interpretable. Fig. 2 demonstrates the overall architecture of our proposed method. In detail, the PML contains three components: a backbone feature extractor $f^E(\cdot)$, an ordinal margin learning branch $f^O(\cdot)$ and a variational margin learning branch $f^V(\cdot)$. For an input image I , the feature denoted by $\boldsymbol{x}$ is extracted by the layer-4 of the backbone ResNet-34 network [23]. Then we define the class center $\boldsymbol{c}$, the inter-variance ϕ and the intra-variance $\boldsymbol{\psi}$, which will be updated according to the recursive formula for calculating the mean and the variance. Moreover, our approach learns both the ordinal and variational margins by taking $\boldsymbol{c}$, ϕ and $\boldsymbol{\psi}$ as the inputs to the $f^O(\cdot)$ and $f^V(\cdot)$. Finally, we introduce a curriculum learning protocol [24, 20] to smoothly simulate data distribution from being balanced to imbalanced. To clarify the notations, Table 1 tabulates the detailed descriptions of all employed variables and functions in this work.

3.1. Problem Formulation

Let $y \in \{0, ..., 100\}$ denote the ground-truth age for each input I. Based on the property of label ambiguity, a facial image feature responds to different similarities across ages and the similarity roughly obeys the Gaussian distribution [17]. Our goal is to transform the scalar age value y to an adaptive label distribution $\boldsymbol{y} \in \mathbb{R}^{101}$ as follows.

$$y_k = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(k-y)^2}{2\sigma^2}), \tag{1}$$

where $k \in [0, 100]$, σ is the variance of label distribution. y_k is the k-th element of $\boldsymbol{y}$ which represents the probability that the true age is k years old, respectively.

To classify the progressive ages with the long-tailed data, we propose a progressive margin loss to reason out robust label distributions. To this end, our approach maintains the ordinal age correlation and suppresses the noise of majority classes on the minority ones in the learned feature space. Moreover, our approach leverages Kullback-Leibler(KL) divergence to measure the distance between the ground-truth distribution and the predicted one.

Table 1. Detailed description of the variables.

Symbol	Definition
$I \in \mathbb{R}^{W \times H}$	Raw face image with $W \times H$ pixels
$\boldsymbol{x} \in \mathbb{R}^{D}$	Extracted feature with D-dimension
$\boldsymbol{y} \in \mathbb{R}^{c}$	Age label distribution consisting of c Age scalar values y
$\boldsymbol{c} \in \mathbb{R}^{c \times D}$	Class centers with D-dimension
$\phi \in \mathbb{R}^{c \times 1}$	Intra-class variances
$\boldsymbol{\psi} \in \mathbb{R}^{c \times c}$	Inter-class variances
$V \in \mathbb{R}^{c \times (D+1+c)}$	Concatenation of the $\boldsymbol{c}$, ϕ and $\boldsymbol{\psi}$
$s(\cdot)$	Dot similarity measure function
$d(\cdot)$	Cosine distance measure function
$f^O(\cdot)$	Function of ordinal margin learning
$f^V(\cdot)$	Function of variational margin learning
$M_o \in \mathbb{R}^{c \times 2}$	Ordinal margins including a tuple of mean and variance
$M_v \in \mathbb{R}^{c \times c}$	Variational margins computing by one-vs.-all (OvA) schema

In this way, the optimal parameter θ^* is determined by

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i log \frac{\boldsymbol{y}_i}{\hat{\boldsymbol{y}}_i} \\ &= \arg\min_{\theta} -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i log \hat{\boldsymbol{y}}_i, \end{aligned} \tag{2}$$

Actually, Equ.2 is the softmax cross-entropy loss function, which was widely used in the margin-based metric learning [22, 7, 30]. The main insight of these methods is to enforce the intra-class concentrations and inter-class diversity by introducing margins to the softmax loss. However, these methods only consider single labels independently, thus ignoring correlated information of neighboring ages. Hence, the fixed positive margin is inflexible to exploit the real-world age distribution. To address the aforementioned problem, our PML suits the distribution learning framework by the newly-learned margins and moreover can be optimized by the standard back-propagation algorithm. The PML is formulated as follows.

$$\mathcal{L}_{m_p} = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i log \hat{\boldsymbol{y}}_i^*, \tag{3}$$

$$\begin{aligned} \hat{\boldsymbol{y}}_i^* = [&\frac{exp(s(\boldsymbol{x}_i, W_1) - m_{p1})}{exp(s(\boldsymbol{x}_i, W_1) - m_{p1}) + \sum_{t \neq 1} exp(s(\boldsymbol{x}_i, W_t))}, \\ ..., &\frac{exp(s(\boldsymbol{x}_i, W_c) - m_{pc})}{exp(s(\boldsymbol{x}_i, W_c) - m_{pc}) + \sum_{t \neq c} exp(s(\boldsymbol{x}_i, W_t))}]^{\mathrm{T}}, \end{aligned} \tag{4}$$

where $s(\cdot)$ denotes the similarity function, *e.g.* dot product similarity [35], and m denotes the parameters for our learned margins of the k-th class, respectively.

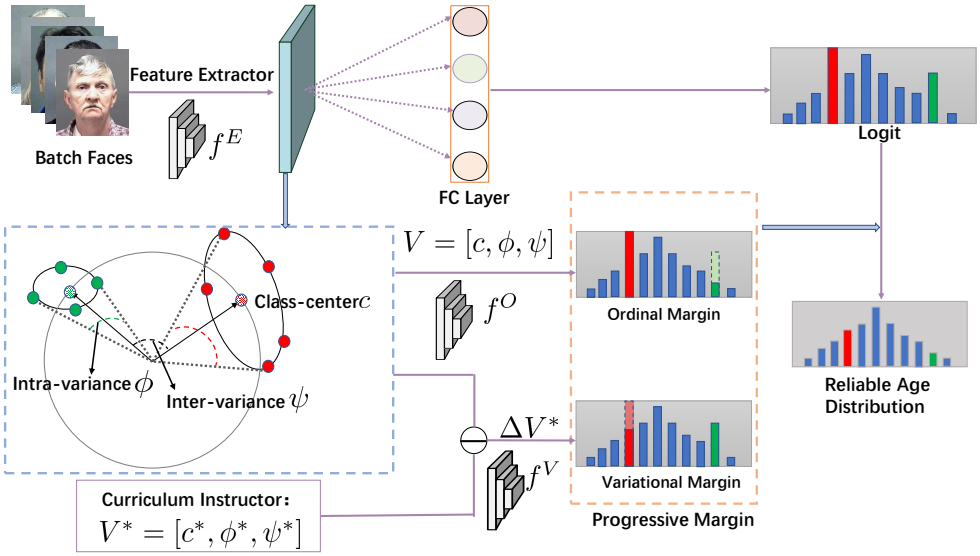


Figure 2. Flowchart of Our PML. Our architecture starts with the input faces, feeding to the feature extractor network $f^E(\cdot)$. Having obtained these deep features, we first compute class center $\boldsymbol{c}$, intra-class variance ϕ and inter-class variance $\boldsymbol{\psi}$ with each class as is shown in blue dotted rectangle. Then we reason out ordinal margin based on the concatenation of all variables (*i.e.* V) mentioned above. Meanwhile, we perform the residual ΔV^* by subtracting the preserved prior curriculum instructor variable V^* from the concatenated variable V. Based on the residual, the variational margin is deduced. Finally, we fuse with the both types of progressive margins versus imbalanced age classes.

Obviously, how to learn the appropriate and interpretable margins is a crucial part in our PML. Since long-tailed age classification is determined by the chronological relation and imbalanced degree of data simultaneously, our proposed PML takes both factors into account in our learned margins, which could be optimized in a globally-tuned manner.

3.2. Progressive Margin Loss

The proposed progressive margin loss framework mainly includes the ordinal margin learning module and the variational margin learning module. To be specific, the ordinal margin aims at making the deep feature more discriminative and simultaneously preserving the ordinal correlation. We assume that the class center is the mean of its samples, which responses the holistic property of one class in the embedding space. In other words, it not only indicates the feature discriminability, but also includes the discrepancy between the majority and minority class. Since a high-level feature $\boldsymbol{x}$ embeds abundant semantic information of the input sample, we take $\boldsymbol{x}$ to represent this sample. For each $\boldsymbol{x}$ *w.r.t* one class, the class center is performed as follows.

$$\boldsymbol{c}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \boldsymbol{x}_i, \tag{5}$$

where $\boldsymbol{c_j}$ and N_j denote the center of j-th class and the number of samples in the j-th class, and $\boldsymbol{x}_i$ is the feature which belongs to this class, respectively. However, Equ.5 requires N_j instances of j-th class (*i.e.* the whole instances belong to j-th class), which cannot be directly applied to mini-batch iterative training. Based on the recursive formula for calculating the mean, our PML computes the class center as follows.

$$\boldsymbol{c}_j^t = \boldsymbol{c}_j^{t-1} + I(y_i = j) \frac{I(y_i = j)\boldsymbol{x}_i - \boldsymbol{c}_j^{t-1}}{N_j^{t-1} + 1}, \tag{6}$$

where t denotes the training iterations and $I(\cdot) \in \{0, 1\}$ is the indicator function. $I(\cdot)$ outputs 1 only if the conditions in brackets are true, vice versa. According to the class center $\mathbf{c}$, the inter-class variance is performed as follows.

$$\boldsymbol{\psi}_j^t = [d(\boldsymbol{c}_j^t, \boldsymbol{c}_0), ..., d(\boldsymbol{c}_j^t, \boldsymbol{c}_c)], \tag{7}$$

where $d(\cdot)$ denotes the cosine distance measure function [41].

To compute the intra-class, we reformulate the recursive formula for calculating the variance as follows.

$$\phi_j^t = \phi_j^{t-1} + I(y_i = j) d(\boldsymbol{x}_i, \boldsymbol{c}_j^{t-1}) d(\boldsymbol{x}_i - \boldsymbol{c}_j^t), \tag{8}$$

where $\boldsymbol{c}_j$, ϕ_j and $\boldsymbol{\psi}_j$ response holistic feature representation, intra-class variance and inter-class variance of j-th class, respectively.

Our proposed PML concats all of them as inputs to $f^O(\cdot)$ to get ordinal margin. For simplicity, we assume that M_o

obeys the Gaussian distribution. Having enforced the constraint to our ordinal margin network, the margins are generated as below.

$$M_o = f^O([\boldsymbol{c}, \phi, \boldsymbol{\psi}]), M_o \in \mathbb{R}^{c\times 2}, \tag{9}$$

where $[\boldsymbol{c}, \phi, \boldsymbol{\psi}] \in \mathbb{R}^{c\times(D+1+c)}$ denotes the concatenation of these variables. Note that M_o is composed of the computed mean and variance, which transforms to $M_o^* \in \mathbb{R}^{c\times c}$ by discretely sampling from the range of 0 to c. After combing Equ.9 with Equ.3, the ordinal margin can be optimized in a unified framework.

M_o enhances the feature discriminativeness by considering the age ordinal relation. However, it may fall into a suboptimal solution for adjacent age classes with imbalanced training samples. In such case, the minority age samples are likely misclassified to the majority age labels. Instead, our proposed variational margin is progressively suppressed the majority classes with its own influence. We performed the residual of class center, inter-class variance and intra-class variance between to adjacent iterations as.

$$\begin{aligned} \Delta V &= [\boldsymbol{c}^t, \phi^t, \boldsymbol{\psi}^t] - [\boldsymbol{c}^{t-1}, \phi^{t-1}, \boldsymbol{\psi}^{t-1}], \\ M_v &= f^V(\Delta V), M_v \in \mathbb{R}^c. \end{aligned} \tag{10}$$

Noticing that $M_o^* \in \mathbb{R}^{c\times c}$ exploits the relation about the one-vs.-all (OvA) mechanism [5], which pays attention to local examples. $M_v \in \mathbb{R}^c$ is complementary to M_o^* by enhancing the learned feature of each class especially for the minority class, which efficiently prevents the disturbance of other classes. We formulate this sense as follows.

$$M_{pj} = \lambda M_o^* + \beta M_v, \tag{11}$$

where M_{pj} denotes progressive margin, λ and β is used to balance M_o^* and M_v, respectively.

3.3. Optimization with Curricula

To further make the margin learning process more stable and fast, our proposed PML follows the insight of curriculum learning [24, 20]. Specifically, we divide the training data into five curricula to optimize the network parameters, where each curriculum consists of varying degrees of data imbalance. In this way, the proposed PML is learned by gradually including samples distribution from being balanced to imbalanced. Unlike classic curriculum learning mechanism where each curriculum contains non-crossing label fields, we design a sampling method to model the consistency of label fields. Our protocol of curriculum learning is defined as

$$\begin{aligned} & \mathcal{D}_1 \subset \mathcal{D}_2 \subset \mathcal{D}_3 \subset \mathcal{D}_4 \subset \mathcal{D}_5, \quad \mathcal{D}_5 = \mathcal{D}_{all}, \\ & \mathcal{D}_i = \{X_i, Y_i\}, \\ s.t. \quad & X_i = x^{(0,\delta_i)} \cup \boldsymbol{\rho}\left(x^{(\delta_i+1,c)}\right). \end{aligned} \tag{12}$$

Algorithm 1: Training Procedure of Our PML

Input: Training set: $\mathcal{D} = \{I_i\}^{i=1:n}$, maximal iteration T.
Output: Parameters of $f^E(\cdot)$, $f^O(\cdot)$ and $f^V(\cdot)$.

for $t < T$ **do**
- /*Extracting the feature of i-th face image.*/
- $\boldsymbol{x}_i = f^E(I_i)$;
- /*Assuming $\boldsymbol{x}_i$ belongs to j-th class and updating the class of each class by Equ.5.*/
- $\boldsymbol{c}_j^t = \boldsymbol{x}_i$, /*For iteration-1*/;
- $\boldsymbol{c}_j^t = \boldsymbol{c}_j^{t-1} + \frac{\boldsymbol{x}_i - \boldsymbol{c}_j^{t-1}}{N_j^{t-1}+1}$;
- /*Updating the inter-class variance by Equ.7.*/
- $\boldsymbol{\psi}_j^t = [d(\boldsymbol{c}_j^t, \boldsymbol{c}_0), ..., d(\boldsymbol{c}_j^t, \boldsymbol{c}_c)]$;
- /*Updating the intra-class variance by Equ.8.*/
- $\phi_j^t = \phi_j^{t-1} + d(\boldsymbol{x}_i - \boldsymbol{c}_j^{t-1})d(\boldsymbol{x}_i - \boldsymbol{c}_j^t)$;
- /*Learning the ordinal margin by Equ.9.*/
- $M_o = f^O([\boldsymbol{c}, \phi, \boldsymbol{\psi}])$;
- /*Optimization with Curricula.*/
- $\Delta V = [\boldsymbol{c}^t, \phi^t, \boldsymbol{\psi}^t] - [\boldsymbol{c}^*, \phi^*, \boldsymbol{\psi}^*]$;
- /*Learning the variational margin by Equ10.*/
- $M_v = f^V(\Delta V), M_v \in \mathbb{R}^c$;
- /*Optimizing our PML by Equ.11 and Equ.3.*/
- $M_{pj} = \lambda M_o^* + \beta M_v$;
- $\mathcal{L}_{m_p} = -\boldsymbol{y}_i log\hat{\boldsymbol{y}}_i^*$;

end
Return: $f_\theta^E(\cdot)$, $f_\theta^O(\cdot)$ and $f_\theta^V(\cdot)$.

For splitting, we firstly sort each class by its owned instances in an ascending order, where δ_i denotes the dividing line of the i-th course. Then, the function of $\boldsymbol{\rho}(x^{(a,b)})$ represents a sampling operation, which draws the same number of instances as the $(a-1)$-th class from the range of a to b.

More specifically, our PML takes data from D_1 to D_5 as the inputs to train the deep convolution neural networks, until it converges in each curriculum. Based on the class property of previous curriculum $V_{pre} = [\boldsymbol{c}^*, \phi^*, \boldsymbol{\psi}^*]$, the variational margin between the adjacent curricula can be obtained as

$$\begin{aligned} \Delta V^* &= [\boldsymbol{c}^t, \phi^t, \boldsymbol{\psi}^t] - V_{pre}, \\ M_v^* &= f^V(\Delta V^*), M_v \in \mathbb{R}^c. \end{aligned} \tag{13}$$

Since the class property V_{pre} is acquired in a further balanced course than current ones, the learned V_{pre} is the unbiased representation towards each class. By referencing this unbiased instructor, the learning procedure of M_p becomes stable. Through this curriculum learning fashion, the optimization process is slightly affected by data imbalance. Experimentally, we find out that this learning schema can achieve comparable performance with the state-of-the-art methods by using fewer training samples. Note that we

only enforce theses margins in the training procedure for achieving discriminative feature representation.

Algorithm 1 shows the optimization procedure of the proposed PML.

4. Experiments

To evaluate the effectiveness of the proposed method, we conducted experiments on three widely-used datasets for uncontrolled age classification. We conducted experiments our method on Morph II [49], FG-NET [44] and ChaLearn LAP 2015 [13]. For fair comparisons, we only used additional IMDB-WIKI dataset [51] for pre-training to evaluate ChaLearn LAP 2015.

4.1. Evaluation Datasets and Metrics

Evaluation Datasets. *Morph II*. This database is the widely-used benchmark for age estimation, which consists of 55,134 face images of 13,617 subjects. The age range of this database covers from 16 to 77 years old. In our experiments, we used two types of testing protocols in our evaluations. **Setting I.** The dataset was randomly divided to training part (80%) and testing part (20%). **Setting II.** A subset of 5,493 face images from Caucasian descent followed the work [54].

FG-NET. The FG-NET dataset contains 1,002 face images of 82 subjects and the age ranges from 0 to 69. We followed the previous methods [31, 43] to use leave-one-out (LOPO) setting for evaluation.

ChaLearn LAP 2015. This dataset was released in 2015 at the ChaLearn LAP challenge, which collects 4,691 images. The ChaLearn LAP was labeled with the apparent age, and each label was set as an average of at least 10 people. This dataset contains training, validation and testing subsets with 2476, 1136 and 1079 images, respectively.

IMDB-WIKI. The IMDB-WIKI consists of 523,051 images in total and the range is from 0 to 100. To follow the common setting, we selected about 300,000 images for training, where all non-face and severely occluded images were removed.

Evaluation Metrics. In the experiments we leveraged Mean Absolute Error (MAE) to calculate the discrepancy between estimated age and the ground-truth. Obviously, the lower the MAE value, the better performance it achieves. According to previous work [36], we also used the ϵ-error to measure the performance on the ChaLearn dataset. In particular, this standard testing protocol is defined as follows.

$$\epsilon = 1 - \sum_{i=1}^{n} exp(-\frac{(y_i - y_i^*)^2}{2\sigma_i^{*2}}),$$

where y_i^* is the ground-truth age value, σ^* is the annotated standard deviation, respectively.

4.2. Implementation Details

For each input image, we first detected the whole face with MTCNN [58]. Then we aligned it based on the detected facial landmarks. For IMDB-WIKI, we straightly removed invalid images. In the training stage, we augmented all images randomly with horizontal flipping, scaling, rotation and translation. Moreover, we adopted ResNet-34 [23] as our backbone network and this network was pretrained on ImageNet [10]. For all experiments, we employed the Adam optimizer and SGD optimizer [46]. The weight decay and the momentum were set to 0.0005 and 0.9, respectively. The initial learning rate was set to 0.0001 and we leveraged two methods for learning rate adjustment. λ and β were tuned by cross validations. In the Adam optimization method, we used CosineAnnealingLR [38] to adjust learning rates. Meanwhile, we used ExponentialLR for the SGD optimizer. For parallel acceleration, we trained our model with PyTorch [45] on 4 Tesla V100 GPUS.

4.3. Results and Analysis

Comparisons on Morph II. Table 2 and Table 3 show the MAEs of our approach on Morph II dataset with different settings. Noticing that we did not use IMDB-WIKI for pretraining in this dataset. According to the results, our model achieves 2.150 and 2.307 under the Setting I and Setting II, respectively. More specifically, in Setting I, our method achieves the best performance among all models except AVDL, but this model was pretrained on IMDB-WIKI. In Setting II, our model achieves the best performance among all state-of-the-art methods regardless of using the external datasets. From the results, we made two-fold conclusions: (1) Compared label distribution learning methods such as DLDL-V2 [18] and M-V Loss[43] leverages a fixed pattern to learned feature. Such schema ignores the issue of age imbalance, which likely hurts the discriminativeness of minority features. (2) Particularly from the results on Setting II, we see that our PML outperforms most state of the arts with sparse training data. This achievement is due to that the learned margin enlarges the inter-class variance by preserving the age-related semantic information.

Comparisons on FG-NET. As shown in Table 4, we compared our model with the state-of-the-art models on FG-NET. Our method PML achieves the lowest MAE of 2.17. Moreover, compared with AVDL that was pretrained by IMDB-WIKI, our PML decreases the MAE by 0.17 with our progressive margin loss. Compared with the state-of-the-art DHAA that was trained from scratch, our PML decreases the MAE by a large margin. Obviously, the results show that our method significantly works well on few-shot dataset.

Comparisons on ChaLearn LAP 2015. We further compared our model with the state-of-the-art models on the ChaLearn LAP 2015. As shown in Table 5, our method

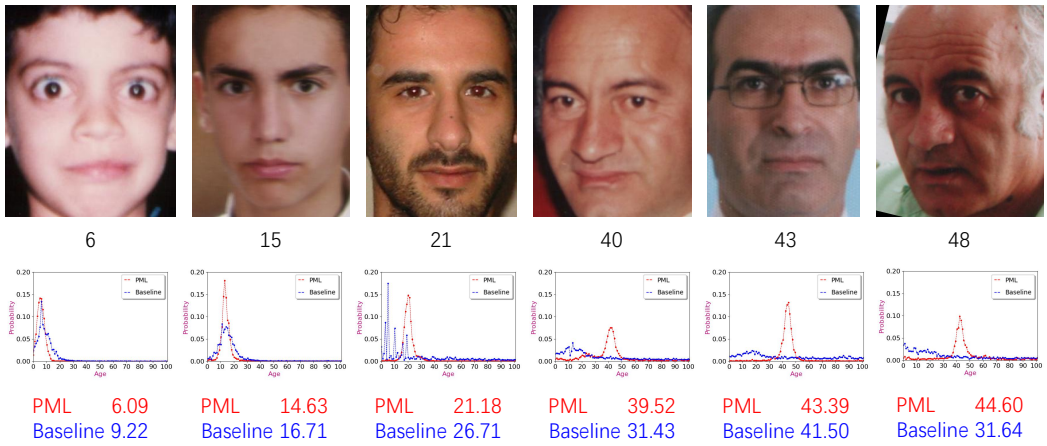


Figure 3. Comparisons of predicted distributions on FG-NET. The first row shows six aligned faces and their corresponding ground-truths. The second row shows the predicted distributions of the baseline and our PML approach. Seeing from these distributions, in our method, the predictions is more accurate and reliable than baseline method.

Table 2. Comparisons of MAEs of our approach compared with different state-of-the-art methods on Morph II under Setting I. Bold indicates the best (* indicates the model was pre-trained on the IMDB-WIKI dataset and † indicates the model was pre-trained on the MS-Celeb-1M, respectively. We annotated the 2nd performance as the italic type.)

Method	Morph	Year
OR-CNN [42]	3.34	2016
ODFL [31]	3.12	2017
ARN [1]	3.00	2017
CasCNN [56]	3.30	2018
M-V Loss[43]	2.41/2.16*	2018
DRFs [52]	2.17	2018
DLDL-V2 [18]	1.97†	2018
SADAL [33]	2.75	2019
BridgeNet	2.38*	2019
AVDL [57]	**1.94***	2020
PML	*2.15*	**-**

Table 3. Comparisons of MAEs of our approach compared with different state-of-the-art methods on Morph II under Setting II.

Method	Morph	Year
DEX [50]	3.25/2.68*	2018
AgeED [54]	2.93/2.52*	2018
DRFss [52]	2.91	2018
DHAA [55]	2.49*	2019
AVDL [57]	*2.37**	2020
PML	**2.31**	**-**

achieves 2.915 MAE which was pretrained on IMDB-WIKI and surpasses the state-of-the-art performance. The results prove that our PML deals with the samples with large variance, while the progressive margin learning achieves to filter noisy instance.

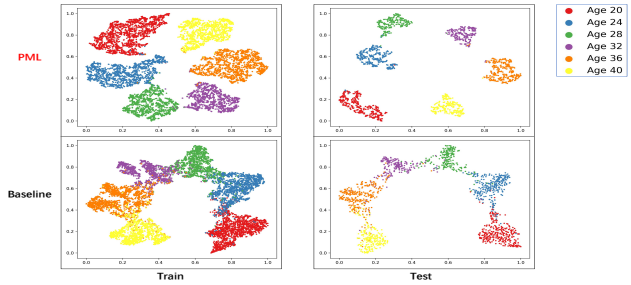
Table 4. Comparisons of MAEs of our approach compared with different state-of-the-art methods on the FG-NET dataset.

Method	FG-NET	Year
DEX [50]	4.63/3.09*	2018
DRFs [52]	3.85	2018
M-V Loss [43]	4.10/2.68*	2018
AgeED [54]	4.34/2.96*	2018
C3AE [57]	2.95	2019
BridgeNet [27]	2.56*	2019
DHAA [55]	3.72/2.59*	2019
AVDL [57]	*2.32**	2020
NRLD [11]	2.55*	2020
PML	**2.16**	**-**

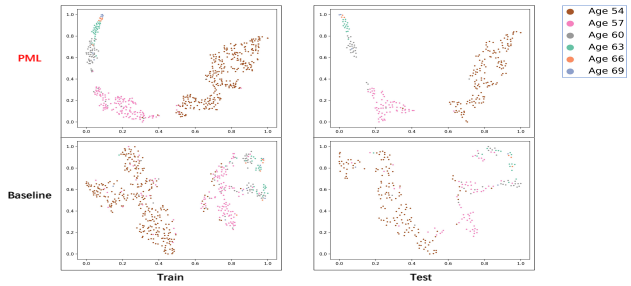
Table 5. Comparisons of MAEs of our approach compared with different state-of-the-art methods on ChaLearn LAP 2015 dataset.

Method	ChaLearn	ϵ-error	Year
ARN [1]	3.153*	-	2017
TinyAgeNet [18]	3.427†	0.301†	2018
CVL_ETHZ [51]	3.252*	0.282*	2018
AgeED [54]	3.210*	0.280*	2018
ThinAgeNet [18]	3.135†	0.272†	2018
ODL [31]	3.950	0.312	2019
DHAA [55]	*3.052**	*0.265**	2019
PML	**3.455**	**0.293**	**-**
PML*	**2.915***	**0.243***	**-**

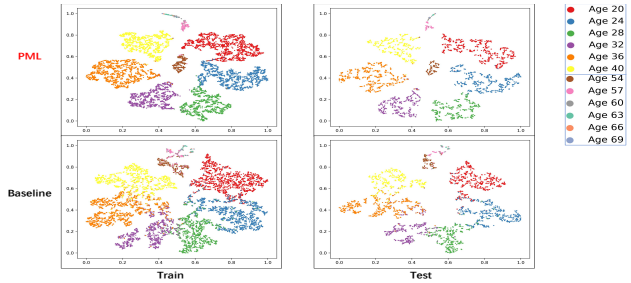
Qualitative Results. To better demonstrate the effectiveness of our PML intuitively, we visualized the predicted distributions and the learned features with versus without the PML framework. For fair comparisons, we created the baseline model, which has the same architectures as our PML except using the standard KL loss. Fig. 3 shows the six resulting examples from young to old on FG-NET. From



(a) Majority Classes



(b) Minority Classes



(c) Majority and Minority Classes

Figure 4. The visualization of learned feature $\boldsymbol{x}$ with t-SNE. We conducted both experiments of training and testing splits on Morph II, compared with the baseline method without the progressive margin. (a) The visualization of features from 6 majority classes. Seeing from these results, each head class is distinguished by our PML. (b) The visualization of our embedded features from 6 minority classes. With these minority classes, our approach learns more discriminative feature than the baseline. (c) Learned by both the majority and minority classes, the spanned feature space of minority classes is narrowed and disturbed by the majority classes in the baseline method. Fortunately, our PML framework teaches each class to characterize their manifolds by constraining the margin to all classes. (Zoomed in for better visualization.)

the Fig. 3, we observe that the learned label distributions of our PML significantly suit real-world age correlation than the baseline model. Fig. 4 shows the learned feature with t-SNE [47]. We see that the proposed progressive margin loss effectively guarantees the boundary of each class in the learned embedding space.

Table 6. Comparisons of MAEs of our approach on Morph II and ChaLearn LAP 2015 dataset under different curriculum learning protocols. (WITHOUT pre-training on IMDB-WIKI)

Dataset	Groups	Imbalance Ratio	Sample	MAE
Morph II	$\mathcal{D}_1(20\%)$	27/1	1,382	3.751
	$\mathcal{D}_2(40\%)$	430/1	17,611	2.828
	$\mathcal{D}_3(60\%)$	1054/1	37,342	2.503
	$\mathcal{D}_4(80\%)$	1335/1	42,438	**2.314**
ChaLearn	$\mathcal{D}_1(20\%)$	6/1	199	6.720
	$\mathcal{D}_2(40\%)$	16/1	840	5.306
	$\mathcal{D}_3(60\%)$	28/1	1,303	4.622
	$\mathcal{D}_4(80\%)$	60/1	2,009	**3.878**

Analysis. To further investigate the effects of our PML regrading with different quantity of training samples, we conducted comparisons on both Morph II and ChaLearn with various courses. For simplicity, we set the dividing line $\{\delta_1, \delta_2, \delta_3, \delta_4\}$ of dataset to $\{20\%, 40\%, 60\%, 80\%\}$ respectively. By following Equ.12, a series of curricula from balance to imbalance could be achieved gradually. As the 3*rd* and 4*th* columns of Table 6 show, we see that with the quantity of samples increases, the imbalance ratio increases. As the Table 6 shows, we see that our PML decreases the MAEs from curriculum $\mathcal{D}_1$ to $\mathcal{D}_4$ on both datasets. More specifically, in course $\mathcal{D}_4$, we achieve comparable results with the state-of-the-art methods while training with less samples, *i.e.*, 80% of the entire dataset. It mainly benefits from the learning instructor of previous curriculum, these instructors assign balanced initial spaces for all classes. Hence, this reduces the probability from trapping into sub-optima.

5. Conclusions

In this paper, we have proposed a progressive margin loss framework (PML) for unconstrained facial age classification. The proposed PML has progressively learned the age label pattern by taking both real-world age relations and critical property of the class center into account. Experiments on three datasets have demonstrated the effectiveness of proposed approach. In future works, we will focus on self-supervised margin learning in a contrastiveA manner [21, 14] by including fewer labels.

6. Acknowledge

This work was supported in part by the National Science Foundation of China under Grant 61806104 and 62076142, in part by the West Light Talent Program of the Chinese Academy of Sciences under Grant XAB2018AW05, and in part by the Youth Science and Technology Talents Enrollment Projects of Ningxia under Grant TJGC2018028.

References

[1] Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Anchored regression networks applied to age estimation and super resolution. In *ICCV*, pages 1652–1661, 2017. 7

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face description with local binary patterns: Application to face recognition. *PAMI*, 28(12):2037–2041, 2006. 2

[3] Shin Ando and Chun Yuan Huang. Deep over-sampling framework for classifying imbalanced data. In *ECML PKDD*, pages 770–785, 2017. 2

[4] Raphael Angulu, Jules-Raymond Tapamo, and Aderemi Oluyinka Adewumi. Age estimation via face images: a survey. *EJIVP*, 2018:42, 2018. 1

[5] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006. 5

[6] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 2

[7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NIPS*, pages 1567–1578, 2019. 3

[8] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, pages 585–592, 2011. 2

[9] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *PAMI*, 23(6):681–685, 2001. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[11] Zongyong Deng, Mo Zhao, Hao Liu, Zhenhua Yu, and Feng Feng. Learning neighborhood-reasoning label distribution (nrld) for facial age estimation. In *ICME*, pages 1–6, 2020. 1, 7

[12] Mohamed Y. Eldib and Motaz El-Saban. Human age estimation using enhanced bio-inspired features (EBIF). In *ICIP*, pages 1589–1592, 2010. 2

[13] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *ICCV-W*, pages 243–251, 2015. 6

[14] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *CVPR*, pages 10364–10374, 2019. 8

[15] Nickolaos F. Fragopanagos and John G. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005. 1

[16] Yun Fu and Thomas S. Huang. Human age estimation with regression on discriminative aging manifold. *TMM*, 10(4):578–584, 2008. 2

[17] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *TIP*, 26(6):2825–2838, 2017. 1, 3

[18] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *IJCAI*, pages 712–718, 2018. 1, 6, 7

[19] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. 1, 2

[20] Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *ICML*, pages 1311–1320, 2017. 3, 5

[21] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742, 2006. 8

[22] Munawar Hayat, Salman H. Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *ICCV*, pages 6468–6478, 2019. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6

[24] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. Self-paced curriculum learning. In *AAAI*, pages 2694–2700, 2015. 3, 5

[25] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019. 2

[26] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334, 2015. 2

[27] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for estimation. In *CVPR*, pages 1145–1154, 2019. 1, 2, 7

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 2

[29] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, pages 438–455, 2020. 2

[30] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: margin-aware reinforcement learning for deep face recognition. In *ICCV*, pages 10052–10061, 2019. 2, 3

[31] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Ordinal deep feature learning for facial age estimation. In *FG*, pages 157–164, 2017. 1, 2, 6, 7

[32] Hao Liu, Jiwen Lu, Minghao Guo, Suping Wu, and Jie Zhou. Learning reasoning-decision networks for robust face alignment. *PAMI*, 42(3):679–693, 2020. 2

[33] Hao Liu, Penghui Sun, Jiaqiang Zhang, Suping Wu, Zhenhua Yu, and Xuehong Sun. Similarity-aware and variational deep adversarial learning for robust facial age estimation. *TMM*, 22(7):1808–1822, 2020. 7

[34] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *CVPR*, pages 11947–11956, 2019. 2

[35] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, pages 507–516, 2016. 3

[36] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *ICCV-W*, pages 258–266, 2015. 1, 6

[37] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019. 2

[38] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[39] Stéphane Mérillou and Djamchid Ghazanfarpour. A survey of aging and weathering phenomena in computer graphics. *Comput. Graph.*, 32(2):159–174, 2008. 1

[40] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *PAMI*, 31(4):607–626, 2008. 2

[41] Hieu V. Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *ACCV*, pages 709–720, 2010. 4

[42] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *CVPR*, pages 4920–4928, 2016. 7

[43] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, pages 5285–5294, 2018. 1, 6, 7

[44] Gabriel Panis, Andreas Lanitis, Nicholas Tsapatsoulis, and Timothy F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, 5(2):37–46, 2016. 6

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[46] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. 6

[47] Paulo E. Rauber, Alexandre X. Falcão, and Alexandru C. Telea. Visualizing time-dependent data using dynamic t-sne. In Enrico Bertini, Niklas Elmqvist, and Thomas Wischgoll, editors, *Eurographics Conference on Visualization*, pages 73–77, 2016. 8

[48] Fulong Ren, Peng Cao, Wei Li, Dazhe Zhao, and Osmar Zaiane. Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. *Computerized Medical Imaging and Graphics*, 55:54–67, 2017. 2

[49] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *FG*, 2006. 6

[50] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV-W*, pages 10–15, 2015. 1, 2, 7

[51] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2016. 1, 6, 7

[52] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L. Yuille. Deep regression forests for age estimation. In *CVPR*, pages 2304–2313, 2018. 1, 7

[53] Xiangbo Shu, Jinhui Tang, Zechao Li, Hanjiang Lai, Liyan Zhang, and Shuicheng Yan. Personalized age progression with bi-level aging dictionary learning. *PAMI*, 40(4):905–917, 2018. 1

[54] Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan Z. Li. Efficient group-n encoding and decoding for facial age estimation. *PAMI*, 40(11):2610–2623, 2018. 6, 7

[55] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z. Li. Deeply-learned hybrid representations for facial age estimation. In *IJCAI*, pages 3548–3554, 2019. 7

[56] Jun Wan, Zichang Tan, Zhen Lei, Guodong Guo, and Stan Z. Li. Auxiliary demographic information assisted age estimation with cascaded structure. *TCYB*, 48(9):2531–2541, 2018. 7

[57] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3AE: exploring the limits of compact model for age estimation. In *CVPR*, pages 12587–12596, 2019. 7

[58] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016. 6