Ordinal Deep Learning for Facial Age Estimation

Hao Liu, Jiwen Lu, *Senior Member, IEEE*, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— In this paper, we propose an ordinal deep learning approach for facial age estimation. Unlike conventional hand-crafted feature-based methods that require prior and expert knowledge, we propose an ordinal deep feature learning (ODFL) method to learn feature descriptors for face representation directly from raw pixels. Motivated by the fact that age labels are chronologically correlated and age estimation is an ordinal learning problem, our proposed ODFL enforces two criteria on the descriptors, which are learned at the top of the deep networks: 1) the topology-preserving ordinal relation is employed to exploit the order information in the learned feature space and 2) the age-difference cost information is leveraged to dynamically measure face pairs with different age value gaps. However, both the procedures of feature extraction and age estimation are learned independently in ODFL, which may lead to a sub-optimal problem. To address this, we further propose an end-to-end ordinal deep learning (ODL) framework, where the complementary information of both the procedures is exploited to reinforce our model. Extensive experimental results on five face aging datasets show that both our ODFL and ODL achieve superior performance in comparisons with most state-of-the-art methods.**

***Index Terms*— Facial age estimation, deep learning, feature learning, ordinal embedding.**

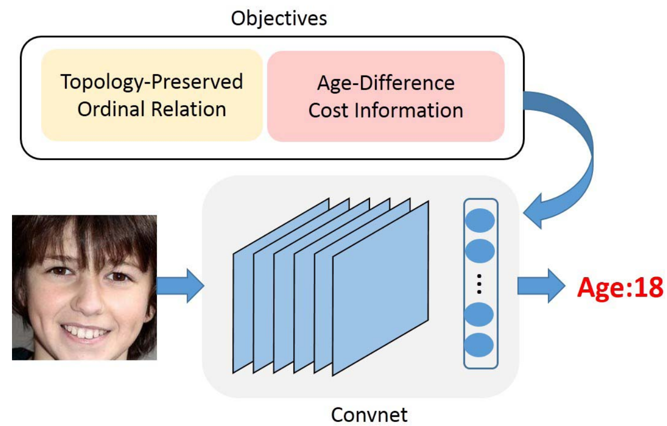


Fig. 1. The flowchart of the proposed approach. Specifically, we enforce two criterions on the face descriptors which are learned at the top of the deep Convnet. Moreover, we propose an end-to-end ordinal deep learning framework, where both tasks of learning face representation and age estimator are jointly optimized under a unified architecture. The network parameters are optimized by back-propagation.

I. Introduction

FACIAL age estimation attempts to predict exact age values for given facial images, which plays an important role in the human-computer interaction, visual advertisements and bio-metrics [1]–[5]. While extensive efforts have been devoted to, facial age estimation still remains a challenging problem, which is because face images usually captured in wild conditions, which undergoes large variations of lighting, facial expressions, appearance and cluttered background.

Existing facial age estimation systems are roughly divided into two key components: face representation [2], [3], [6] and age estimator learning [7]–[9]. However, most features employed in previous methods are ad hoc, which requires strong prior knowledge by hand. To address this, learning-based feature representation methods [7], [10], [11] have been proposed to learn discriminative feature representation directly from the image pixels. For example, Fu *et al.* [10] proposed a holistic feature learning method by using a discriminative manifold learning technique. Lu *et al.* [11] addressed the cost-sensitive problem for age estimation by learning local binary codes for face representation. However, their methods utilize linear feature filters so that they are not powerful enough to exploit the complex and nonlinear relationship between face samples and age labels. To address this nonlinear issue, deep learning techniques [12]–[16] have been applied to model the relationship between face features and age labels by a series of nonlinear transformations. For example, Yi *et al.* [12] proposed multi-scale features by leveraging deep convolutional neural networks [17], with additionally considering the gender and ethnicity attributes. Niu *et al.* [16] developed an ordinal regression method with multiple output via deep convolutional neural networks to perform age predicting. While promising performance has been obtained, these methods cannot explicitly model the structural and high-order relationships of face samples, which is useful to preserve the ordinal relation for age labels.

In this paper, we propose an ordinal deep learning approach for facial age estimation. Fig. 1 illustrates the flowchart of the proposed approach. Unlike existing facial age estimation methods which cannot explicitly exploit the structural order

Manuscript received April 21, 2017; revised July 15, 2017 and November 4, 2017; accepted December 8, 2017. Date of publication December 12, 2017; date of current version February 5, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61572271, and Grant 61527808, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Ministry of Education of China under Grant 20120002110033, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170438636, and in part by the Tsinghua University Initiative Scientific Research Program. This paper was recommended by Associate Editor A. Savakis. This work was presented in part at the 12th IEEE International Conference on Automatic Face and Gesture Recognition, in 2017. *(Corresponding author: Jiwen Lu.)*

The authors are with the Department of Automation, Tsinghua University, Beijing 100084, China, and also with the State Key Lab of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China (e-mail: h-liu14@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2782709

relationships such as the quadruplet and triplet-based comparisons among face samples, we propose an ordinal deep feature learning (ODFL) method to learn the high-order ordinal relation based on the mini-batched data during training process. To achieve this, our ODFL enforces two important criterions at the top of the deep network: 1) Topology-Preserving Ordinal Relation: for each sampled quadruplet, the topology structure towards ordinal relation is embedded in the learned feature space, and 2) Age-Difference Cost Information: the similarity of face pairs is smoothly measured based on the age difference values. However, the procedures of learning face representation and age estimator are optimized separately, which may be sub-optimal for this task. To address this, we elaborately develop an ordinal deep learning (ODL) framework for exact age prediction, where both the feature extraction and age estimation procedures are globally optimized in an end-to-end deep architecture. To achieve this, we firstly encode the age labels as the consistent binary outputs which aim to preserve the order information for age labels. Then we define two ordinal regression loss functions, *e.g.*, Square Loss and Cross-Entropy Loss, which minimize the mis-classified errors of assigning the true age labels for given face samples. The parameters of the whole deep networks are optimized by the standard back-propagation method. Hence, the ordered consistency can be passed backward to the whole network to promote the discriminativeness of the learned face representations. To verify the effectiveness of our proposed approach, we conduct experiments on five face aging datasets. Experimental results show significant performance compared with the state-of-the-art facial age estimation methods.

This work is an extension to our conference paper [18]. The newly incorporated work is described below:

1) We have designed an end-to-end ordinal deep learning (ODL) framework by including two ordinal regression loss functions, *e.g.*, Square loss and Cross-Entropy loss. Both losses optimize the whole networks containing both face representation mapping and age estimation procedures in a joint learning manner. Extensive experiments have been conducted to demonstrate the effectiveness of the proposed ODL.
2) We have conducted experiments to evaluate the importance of the proposed topology-preserving ordinal relation and age-difference cost information in our ODFL. The results show that our ODFL achieves exploiting the complementary information for both quadruplet and triplet-based comparisons of face samples, which simultaneously improves the age prediction performance.
3) We have compared our ODL and ODFL with various state-of-the-art approaches on five face aging datasets. The empirical results have clearly shown that both proposed methods achieve superior performance in comparisons with the state-of-the-arts.

The rest of this paper is organized as follows: Section II briefly reviews some related work. Section III describes the proposed ordinal deep learning approach for facial age estimation in details. Section IV reports experimental results and analysis, and Section V concludes the paper.

II. Related Work

In this section, we reviews the related works for facial age estimation methods and deep learning approaches, respectively.

A. Facial Age Estimation

Numerous facial age estimation methods [8], [9], [19]–[25] have been proposed over the past two decades. For example, Lanitis *et al.* [19] applied an age regression method to address the face aging problem. Zhang and Yueng [20] proposed an age estimation method by using a multi-task Gaussian process (MTWGP). Chang *et al.* [9] presented an ordinal hyperplane ranking (OHRanker) method which divided the age estimation problem as a series of sub-problems of binary classifications. Geng *et al.* [21], [26] proposed a label distribution learning (LDL) approach to model the relationship between face images and age labels. However, these methods usually employ hand-crafted features such as the holistic subspace feature [7], [27], local binary pattern (LBP) [2] and the bio-inspired feature (BIF) [3] for face representation, which require strong expert knowledge by hand. To address this, several attempts have been made to learn discriminative face descriptors by using advanced feature learning approaches [3], [11], [22]. For example, Guo *et al.* [28] proposed a holistic feature learning approach by utilizing a manifold learning technique. Lu *et al.* [11] proposed a local binary feature learning method (CS-LBFL) to learn a face descriptor which is robust to local illumination. However, these methods aim to seek simple feature filters, so that they are not powerful enough to exploit the nonlinear relationship of face samples in such cases that facial images are exposed to large variances of diverse facial expressions and cluttered backgrounds.

B. Deep Learning

Recently, deep learning methods, *i.e.*, deep convolutional neural networks (CNN), have been applied to many facial analysis tasks including face detection [29], face alignment [30] and face recognition [31], [32]. For example, Zhang *et al.* [30] utilized stacked auto-encoder networks to estimate facial landmarks in a coarse-to-fine manner. Sun *et al.* [31] developed a DeepID2 network to reduce the personalized inter-covariance jointly by using the identification and verification signals jointly. Parkhi *et al.* [32] proposed a VGG Face Net with a very deep architecture, which was pretrained by a large scale face dataset for face recognition. Inspired by the aforementioned works which learn task-adaptive face feature representation, deep learning has been also used to learn a set of nonlinear feature transformations for facial age estimation [13], [16], [33]–[38]. For example, Levi *et al.* [35] proposed a Multi-task deep CNN framework to jointly address the age and gender classification in a unified deep learning framework. Yang *et al.* [39] employed deep scattering transform networks (DeepRank) to predict ages via category-wise rankers. Niu *et al.* [16] developed an ordinal regression CNN-based (OR-CNN) method with multiple binary outputs for age estimation. While significant performance can be obtained, they ignored to take advantages

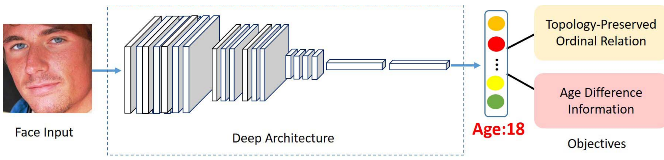


Fig. 2. The framework of the proposed ODFL. During the training stage, we enforce two objectives on learning age-related face descriptors, which aims to exploit both the topology-preserving ordinal relation and age-difference information at the top layer of the designed deep networks, *e.g.*, AlexNet [17], ResNet [40], VGG [32], etc. The network parameters are optimized via back-propagation. During the testing stage, we feed the face image to the networks and then predict the exact age value by a learned age ranker.

of the quadruplet-based ordinal relation during batch-wise training procedure in deep learning, which makes the learned features less accurate for age predicting.

III. Proposed Approach

In this section, we describe our proposed ODFL and ODL in details, respectively. Moreover, we present the difference between our approach compared with some related work.

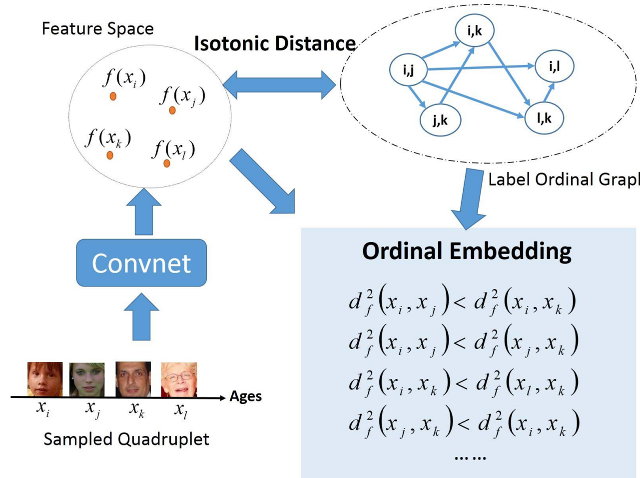


Fig. 3. Topology-Preserving Ordinal Relation. Given a quadruplet of face samples and age labels from a training batch, we construct a directed unweighted topology as the label ordinal graph towards ordinal embedding. Our ODFL aims to learn a deep Convnet, where the topology-preserving ordinal relation within the label ordinal graph has isotonic distance to that in the learned feature space. As a result, the topology-preserving ordinal relation is preserved in the transformed feature space.

A. ODFL

Fig. 2 shows the framework of the proposed ODFL. Let $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ denote the training set which contains N samples, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the ith face of d pixels. Our ODFL learns to compute feature representation $f(\mathbf{x}_i)$ for the ith face image $\mathbf{x}_i$ under the deep CNN architecture. Specifically, we feed the face image to the designed CNN and obtain the immediate feature representation, which is formulated as follows:

$$f(\mathbf{x}_i) = \mathbf{h}_i^{(m)} = \text{pool}\left(\text{ReLU}(\mathbf{W}^{(m)} \otimes \mathbf{x}_i + \mathbf{b}^{(m)})\right), \quad (1)$$

where $\otimes$ denotes the convolution operation, pool$(\cdot)$ denotes the max pooling operation, ReLU$(\cdot)$ denotes the nonlinear *ReLU* function and $m = \{1, 2, \cdots, M-2\}$ represents the mth layer, respectively.

The face descriptor at the top layer is computed as follows:

$$f(\mathbf{x}_i) = \mathbf{h}_i^{(M)} = \sigma(\mathbf{W}^{(M)}\mathbf{x}_i + \mathbf{b}^{(M)}), \quad (2)$$

where $\mathbf{W}^{(M)}$ and $\mathbf{b}^{(M)}$ denote the weights and bias of the top layer and $\sigma(\cdot)$ denotes the nonlinear function, respectively.

To sum up the total weights, we collect $m = \{1, 2, \cdots, M\}$ to train the whole CNN based on the dissimilarity on the face pair of $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, which is computed as follows:

$$d_f^2(\mathbf{x}_i, \mathbf{x}_j) = \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2, \quad (3)$$

where $\|\cdot\|_2$ denotes the Euclidean distance in the learned feature space and $f(\cdot)$ denotes the deep feature embedding based on the deep CNN architecture.

Therefore, how to learn the deep feature embedding $f(\cdot)$ is the crucial part in our ODFL. To learn efficient face descriptors for facial age estimation, the key design lies on preserving the ordinal relation among training samples in the transformed feature space. To this end, we propose two criterions including *topology-preserving ordinal relation* and *age-difference cost information* at the top of the deep network. Then the whole parameters of the deep network are optimized by back-propagation. In the following parts, we detail both proposed criterions accordingly.

1) Topology-Preserving Ordinal Relation: Unlike conventional facial age estimation methods [8], [11], [16], [41] which attempt on learning age rankers based on pairwise comparisons, we construct a label ordinal graph based on sets of quadruplets from training batches. Note that the label graph is embedded according to the smoothing degree of pairs of age labels [42]. Based on the label graph, the defined objective aims to enforce that the ordinal relation in the learned feature space should be *isotonic* to that in the label space [43], [44]. In other words, the compared relationships among face samples should be equal to those in the label space. To achieve this, our ODFL learns to map the face samples to a latent common space, where the topology-preserving ordinal relation is preserved in the learned face descriptors according to the smoothness of age labels.

As illustrated in Fig. 3, suppose we sample a quadruplet (i, j, k, l) from the training batch $\mathscr{B}$ with the knowing age labels (y_i, y_j, y_k, y_l). Based on the age labels, we encode such a quadruplet with a particular subset of ordinal constraints as follows:

$$\delta\left(y_i, y_j\right) < \delta\left(y_k, y_l\right), \quad \forall(i, j, k, l) \subseteq \mathscr{B}, \quad (4)$$

where $\delta(\cdot, \cdot)$ denotes the smooth function, which is viewed as a dissimilarity degree between a pair of age labels and is defined by the Gaussian function as follows:

$$\delta\left(y_i, y_j\right) = \delta_{ij} = \exp^{\frac{-(y_i - y_j)^2}{H^2}}, \quad (5)$$

where H denotes the label difference threshold to determine the variance of age label distribution.

To model the topo-structure for the quadruplet of age labels, we construct a label graph $G = (V, E) = [n]^4$, where each node $\delta_{ij} \in V$ represents the age dissimilarity degree between the ith and jth samples and meanwhile each directed edge $e_{(i,j,k,l)\subseteq\mathscr{B}} \subseteq E$ represents an ordinal relation of $\delta_{ij} < \delta_{kl}$. To achieve the topology-preserving ordinal relation, our ODFL aims to encode items in $\mathscr{B}$ as the projected feature representation, so that the ordinal constraints are preserved by the isotonic distance which is defined as follows:

$$\delta_{ij} < \delta_{kl} \Longrightarrow d_f^2(\mathbf{x}_i, \mathbf{x}_j) < d_f^2(\mathbf{x}_k, \mathbf{x}_l), \tag{6}$$

which means that the topology-preserving ordinal relation within the label ordinal graph has the isotonic distance with that in the learned feature space (refer to more details in Fig. 3). There are two common situations for (6), i.e., quadruplet ordinal relation where $(i, j, k, l) \subseteq \mathscr{B} \subseteq [n]^4$ and $(i, j, i, k) \subseteq \mathscr{B} \subseteq [n]^3$. Hence, the objective takes advantages of the fully structural ordinal relation of training batches, so that the high-order quadruplet and triplet based comparisons can be taken into account in the feature space simultaneously. Therefore, the distance of the face pair of the ith and jth samples should be smaller than that with the face pair of the kth and lth samples.

To involve the label information, we leverage the constructed ordinal label graph G to train the designed network in a globally supervised manner. For the ordinal relation of $e_{(i,j,k,l)\subseteq\mathscr{B}} \subseteq E$ in the batch $\mathscr{B}$, we expect the relation of age dissimilarity degree should be preserved by the learned feature space constrained by (6). To achieve this, we leverage Hinge Loss to optimize the violates of unsatisfied quadruplet comparisons. Hence, the objective J_1 is formulated as follows:

$$\sum_{v_{ij}, v_{kl} \in G} \zeta(v_{ij}, v_{kl}) \cdot \max[0, \alpha - d_f^2(\mathbf{x}_i, \mathbf{x}_j) + d_f^2(\mathbf{x}_k, \mathbf{x}_l)], \tag{7}$$

where $\zeta(v_{ij}, v_{kl})$ indicates 1 if there is a vertex v_{ij} to v_{kl}, and 0 vice versa. Note that α in (7) denotes a thresholding margin which was assigned to 1 in our experiments.

2) Age-Difference Cost Information: Since the traditional weighting functions in [45]–[47] were determined by a stochastic sampling technique during training process, which cannot be directly applied to exploit the smoothness of the real-world aging pattern. To better improve the discriminativeness of the face descriptors, we introduce a weighted ranking approximation method to smoothly consider the age difference information by a carefully designed weighting function. To this end, we define an objective function to measure the age-difference information in a ranking-preserving manner.

As is illustrated in Fig. 4, given a triplet of an anchor sample and other two samples, based on this anchor sample, the age-difference constraints aims to enforce that the difference of a pair with a small age gap should be smaller than that of a pair with a large age gap in the learned feature space. To this end, the age-difference information is weighted dynamically in the embedded feature space according to different age gaps, and the ranking weights are computed to show how they exploit different relation for different age gaps. Therefore, our goal of J_2 is to minimize the following objective function:

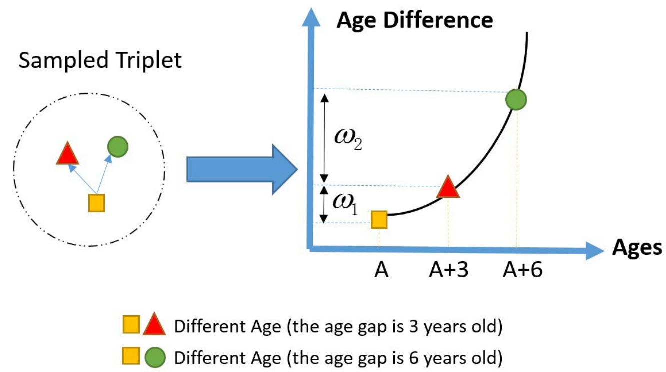


Fig. 4. Age-Difference Cost Information. Suppose there are three face samples from the training set and let the yellow square denote the anchor sample. Based on the anchor sample, the red triangle represents the face sample with an age gap of 3 years old and the green circle denotes that with a larger age gap of 6 years old. Our ODFL aims to learn a set of nonlinear feature transformations, where a face pair with a larger age gap has a larger ranking weight ω_2 than the ranking weight ω_1 with a smaller age gap. As a result, the ranking-preserving age-difference information can be exploited in the learned feature space.

$$\sum_{p}^{P} \left(1 - \ell_{p1,p2}(\tau - d_f^2(\mathbf{x}_{p1}, \mathbf{x}_{p2})) \cdot \omega_{y_{p1},y_{p2}}\right), \tag{8}$$

where $(p1, p2)$ denotes the face pair with different age value gaps according to the anchored face sample p. τ denotes the pre-defined threshold to enforce that the distance of the face pair $(p, p1)$ with a smaller age difference should be smaller than the threshold and meanwhile the distance of the face pair $(p, p2)$ with a larger age difference are larger than the threshold (typically, the value of τ was assigned to 1 in our experiments). $\ell(p1, p2)$ denotes the indicator which is set to 1 if the face pair belongs to the same age labels, and is set to -1, vice versa. y_{p1} and y_{p2} represent the age gaps computed based on the ground-truth, and $\omega_{y_{p1},y_{p2}}$ denotes the smoothness weighting function. The weighting function specifically measures the aging smoothness, which is defined as follows:

$$\omega_{y_{p1},y_{p2}} = \begin{cases} (|y_{p1} - y_{p2}| + 1)^{\eta}, & \text{if} \quad y_{p1} \neq y_{p2}. \\ 1, & \text{otherwise.} \end{cases} \tag{9}$$

where η is a constant parameter that describes the tolerance level of varying age relationship.

With the defined age-difference specific objective, the ranking weights are preserved by the smooth function instead of treating all pairs with different age gaps equally, so that the chronological aging process can be well measured in the embedded feature space. Moreover, the age-difference cost information is exploited in the transformed feature space by preserving age rankings.

3) Formulation: Based on the proposed two objectives including topology-preserving ordinal relation and age-difference cost information, we formulate our ODFL by combining (7) and (8) as minimizing the following optimization problem:

$$\min_{\{\mathbf{W},\mathbf{b}\}} J = J_1 + \lambda_1 J_2 + \lambda_2 J_3$$
$$= \sum_{\upsilon_{ij},\upsilon_{kl}\in G} \zeta(\upsilon_{ij},\upsilon_{kl}) \cdot \max[0, \alpha - d_f^2(\mathbf{x}_i,\mathbf{x}_j) + d_f^2(\mathbf{x}_k,\mathbf{x}_l)]$$
$$+ \lambda_1 \sum_{p}^{P} \Big(1 - \ell_{p1,p2}(\tau - d_f^2(\mathbf{x}_{p1},\mathbf{x}_{p2})) \cdot \omega_{y_{p1},y_{p2}}\Big)$$
$$+ \lambda_2 \sum_{m=1}^{M} (\|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2), \tag{10}$$

where hyperparameter λ_1 balances the proposed two criterions J_1 and J_2, hyperparameter λ_2 is utilized to control the penalty term to enhance the model generation, $\|\mathbf{W}^{(m)}\|_F^2$ denotes the Frobenius norm of matrix $\mathbf{W}^{(m)}$ to prevent the parameters of deep network from overfitting, respectively.

There are three objectives for (10):

1) The first term J_1 in (10) is to preserve the *topology-preserving ordinal relation* for each sampled quadruplet. Moreover, the fully order relationship of both quadruplet and triplet ranking comparisons are preserved simultaneously in the learned feature space in a purely supervised way.
2) The second term J_2 in (10) attempts to dynamically assign the ranking-preserving weights to achieve the *age-difference cost information* for the anchored triplets according to age value gaps, where the age difference is exploited in the transformed feature space to reinforce the age-related face representations.
3) The third term J_3 enforces the regularization on network parameters to reduce the model complexity, avoiding overfitting for very deep architecture.

4) Optimization: To optimize J_1 in (10), we present a landmark-based ordinal embedding method (LOE) [48], which considers the triplet comparisons from any training samples to the landmark. In this way, the number of ordinal constraints reduces from n^4 to $n \cdot L^2$, where L denotes the landmark number. Note that the subset (batch-size was assigned to 60 in our experiments) is already sufficient to guarantee the uniqueness of the ordinal relation of the learned feature descriptors [44]. Moreover, we apply a logistic loss function to relax the maximum non-convex function $\max[0, \Psi]$ that is not easy to optimize by $g(\Psi) = \frac{1}{\beta}\log(1+\exp(\beta\Psi))$, where β is a sharpness parameter. Based on the relaxation, J_1 in (10) is rewritten as follows:

$$J_1 = \sum_{i=1}^{n}\sum_{j,k=1}^{L} \zeta(\upsilon_{ij},\upsilon_{ik}) \cdot g(\alpha - d_f^2(\mathbf{x}_i,\mathbf{x}_j) + d_f^2(\mathbf{x}_k,\mathbf{x}_l)). \tag{11}$$

To solve the relaxed optimization problem of both (10) and (11), we leverage the stochastic gradient descent scheme to compute the parameters $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}$, where $m = \{1, 2, ..., M\}$. Specifically, the gradients of the objective J with respect to the parameters $\{\mathbf{W}^{(m)}\}$ and $\{\mathbf{b}^{(m)}\}$ can be computed accordingly as follows:

$$\frac{\partial J}{\partial \mathbf{W}^{(m)}} = \sum_{i=1}^{n}\sum_{j,k=1}^{L} \zeta(\upsilon_{ij},\upsilon_{ik}) \cdot g'(\Psi)\Theta_1^{(m)}$$
$$+ \lambda_1 J_2 \Theta_2^{(m)} \cdot \omega_{y_{p1},y_{p2}} + \lambda_2 \mathbf{W}^{(m)}, \tag{12}$$
$$\frac{\partial J}{\partial \mathbf{b}^{(m)}} = \sum_{i=1}^{n}\sum_{j,k=1}^{L} \zeta(\upsilon_{ij},\upsilon_{ik}) \cdot g'(\Psi)$$
$$\times [(\mathbf{L}_{ij}^{(m)} + \mathbf{L}_{ji}^{(m)}) - (\mathbf{L}_{kl}^{(m)} + \mathbf{L}_{lk}^{(m)})]$$
$$+ \lambda_1 J_2 (\mathbf{L}_{p1,p2}^{(m)} + \mathbf{L}_{p2,p1}^{(m)}) \cdot \omega_{y_{p1},y_{p2}}$$
$$+ \lambda_2 \mathbf{b}^{(m)}, \tag{13}$$

where the updating equations are computed as follows:

$$\Theta_1^{(m)} = [(\mathbf{L}_{ij}^{(m)}\mathbf{h}_i^{(m-1)^T} + \mathbf{L}_{ji}^{(m)}\mathbf{h}_j^{(m-1)^T})$$
$$- (\mathbf{L}_{kl}^{(m)}\mathbf{h}_k^{(m-1)^T} + \mathbf{L}_{lk}^{(m)}\mathbf{h}_l^{(m-1)^T})],$$
$$\Theta_2^{(m)} = (\mathbf{L}_{p1,p2}^{(m)}\mathbf{h}_{p1}^{(m-1)^T} + \mathbf{L}_{p2,p1}^{(m)}\mathbf{h}_{p2}^{(m-1)^T}),$$

where

$$\mathbf{L}_{ij}^{(M)} = (\mathbf{h}_i^{(M)} - \mathbf{h}_j^{(M)}) \odot \varphi'(\mathbf{z}_i^{(M)}),$$
$$\mathbf{L}_{ji}^{(M)} = (\mathbf{h}_j^{(M)} - \mathbf{h}_i^{(M)}) \odot \varphi'(\mathbf{z}_j^{(M)}),$$
$$\mathbf{L}_{kl}^{(M)} = (\mathbf{h}_k^{(M)} - \mathbf{h}_l^{(M)}) \odot \varphi'(\mathbf{z}_k^{(M)}),$$
$$\mathbf{L}_{lk}^{(M)} = (\mathbf{h}_l^{(M)} - \mathbf{h}_k^{(M)}) \odot \varphi'(\mathbf{z}_l^{(M)}),$$
$$\mathbf{L}_{1p,2p}^{(M)} = \ell_{1p,2p}(\mathbf{h}_{1p}^{(M)} - \mathbf{h}_{2p}^{(M)}) \odot \varphi'(\mathbf{z}_{1p}^{(M)}),$$
$$\mathbf{L}_{2p,1p}^{(M)} = \ell_{1p,2p}(\mathbf{h}_{2p}^{(M)} - \mathbf{h}_{1p}^{(M)}) \odot \varphi'(\mathbf{z}_{2p}^{(M)}),$$
$$\mathbf{L}_{ij}^{(m)} = (\mathbf{W}^{(m+1)^T}\mathbf{L}_{ij}^{(m+1)}) \odot \varphi'(\mathbf{z}_i^{(m)}),$$
$$\mathbf{L}_{ji}^{(m)} = (\mathbf{W}^{(m+1)^T}\mathbf{L}_{ji}^{(m+1)}) \odot \varphi'(\mathbf{z}_j^{(m)}),$$
$$\mathbf{L}_{kl}^{(m)} = (\mathbf{W}^{(m+1)^T}\mathbf{L}_{kl}^{(m+1)}) \odot \varphi'(\mathbf{z}_k^{(m)}),$$
$$\mathbf{L}_{lk}^{(m)} = (\mathbf{W}^{(m+1)^T}\mathbf{L}_{lk}^{(m+1)}) \odot \varphi'(\mathbf{z}_l^{(m)}),$$
$$\mathbf{L}_{1p,2p}^{(m)} = (\mathbf{W}^{(m+1)^T}\mathbf{L}_{1p,2p}^{(m+1)}) \odot \varphi'(\mathbf{z}_{1p}^{(m)}),$$
$$\mathbf{L}_{2p,1p}^{(m)} = (\mathbf{W}^{(m+1)^T}\mathbf{L}_{2p,1p}^{(m+1)}) \odot \varphi'(\mathbf{z}_{2p}^{(m)}),$$
$$\mathbf{z}_i^{(m)} = \mathbf{W}^{(m)}\mathbf{h}_i^{(m-1)} + \mathbf{b}^{(m)},$$

where $m = 1, 2, ..., M-1$ and $\odot$ denotes the element-wise multiplication.

Having obtained the gradients, parameters $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ are updated by using the gradient-decent algorithm as follows until convergence:

$$\mathbf{W}^{(m)} = \mathbf{W}^{(m)} - \rho\frac{\partial J}{\partial \mathbf{W}^{(m)}}, \tag{14}$$
$$\mathbf{b}^{(m)} = \mathbf{b}^{(m)} - \rho\frac{\partial J}{\partial \mathbf{b}^{(m)}}, \tag{15}$$

where ρ is the learning rate, which controls the convergence speed of the objective function J.

Algorithm 1 shows the optimization procedure of the proposed ODFL.

Algorithm 1: ODFL

Input: Training set: $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, learning rate ρ and iteration number T.
Output: The network parameters $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{i=1}^{M}$.
Step 1 (Parameters Intialization): Initialize the parameters $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{i=1}^{M}$ by the pretrained networks.
Step 2 (Optimization via Back-Propagation):
repeat
- 2.1 Randomly select an quadruplet (i, j, k, l) from a training batch $\mathscr{B}$, and then construct the label ordinal graph G by using the label quadruplet (y_i, y_j, y_k, y_l) according to (4) .
- 2.2 Perform forward propagation and map G to a landmark-based graph based on LOE [48].
- 2.3 Perform backward propagation and compute the gradients according to (12) and (13).
- 2.4 Update the parameters according to (14) and (15).

until *convergence or reaching the maximum iteration number T*;
Return: $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{i=1}^{M}$.

B. ODL

The proposed criterions including the topology-preserving ordinal relation and age-difference cost information mainly focus on embedding ordinal relation in feature space. Having obtained the face representation, we directly feed it to a learned age estimator, *e.g.*, OHRanker [9], for age value predicting. In this way, both procedures of feature extraction and age estimation are learned in a separated way, which may lead to local optimal during training process. Inspired by recent successes of the end-to-end deep learning architecture [17], [30], [49], [50], we propose an ordinal deep learning (ODL) framework, where both tasks of learning face representation and age estimator are jointly optimized in an end-to-end deep learning architecture. To achieve this goal, we elaborately design two ordinal regression loss functions, *e.g. Square Loss* and *Cross-Entropy Loss*, and then deploy them at the top of the deep network, which aims to directly map the raw face images to the exact age values in a joint learning manner. Specifically, we firstly embed the age labels as the consistent binary outputs to take the aging process into account. With the binary outputs, the age labels are encoded as the cumulative attribute for the aging progression in practice. Having obtained the consistent binary outputs, our ODL aims to regress deep feature embedding to the consistent binary outputs by leveraging the deep regression, dubbed ordinal regression in this work. Next, we describe the *consistent binary output* and *ordinal regression* in the following subsection.

1) Consistent Binary Outputs: Given nth training data point, our ODL first encodes age labels into a scalar vector $\mathbf{t}_n$, as illustrated in Fig. 5. The dimension of the scalar vector $\mathbf{t}_n$ contains $\aleph$ elements, *e.g.*, $\aleph = 60$ for a certain age dataset, where the maximal age label is 60 years old. Suppose we have

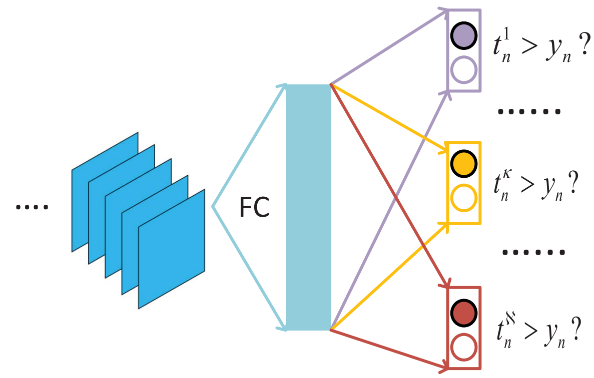


Fig. 5. The framework of the proposed ODL. Having obtained the latent feature representation from the deep Convnet fully connected (FC) layers, the basic idea of our ODL is to map the latent representation to the consistent binary outputs which performs ordinal decompositions for age labels. Let $\mathbf{t}_n^\kappa$ denote the κth element for the nth sample, the exact value is binary depend the order between the κ and the correct age label y_n, typically 1 if κ is bigger than y_n, and 0 otherwise. Hence, the age labels can be embedded as consistent binary outputs to better model the aging pattern, which improves the performance of facial age estimation.

N training set $\{(\mathbf{x}_n, y_n)\}$, the κth element of the scalar vector is computed as follows:

$$t_n^\kappa = \begin{cases} 1, & \text{when} \quad \kappa \le y_n, \\ 0, & \text{when} \quad \kappa > y_n, \end{cases} \tag{16}$$

where $\kappa = 1, 2, \cdots, \aleph$. For the scalar vector $\mathbf{t}_n$, the first y_n elements are all "ones" and the rest $\aleph - y_n$ elements are all "zeros".

To obtain exact age values, we collect the predicted consistent binary outputs and sum them up. The final age value for a given testing sample $\mathbf{x}'$ is predicted as follows:

$$\hat{y} = 1 + \sum_{\kappa=1}^{\aleph-1} f^\kappa(\mathbf{x}'), \tag{17}$$

where $f^\kappa(\mathbf{x}') \in \{0, 1\}$ is the predicted outputting result of the κth element for the sample $\mathbf{x}'$ (i.e., the κth output of our proposed deep networks. Ideally, these $f^\kappa(\mathbf{x}')$ should be consistent.

2) Ordinal Regression: For the training samples of $\mathbf{x}_i$ and $\mathbf{t}_i$ that depends on y_i for the ith face image, the basic idea of ordinal regression is to map the given deep feature embedding for face representation from deep Convnet FC layer to the consistent binary outputs for age labels. Hence, the objective function for κth element is formulated as follows (ignoring $\mathbf{b}$ for simplicity):

$$\min_{\{\mathbf{W}\}} \mathcal{O} = \sum_{n=1}^{N} \sum_{\kappa=1}^{\aleph} \text{loss}(t_n^\kappa, f^\kappa(\mathbf{x}_n)), \tag{18}$$

where $f^\kappa(\mathbf{x}_n) = \mathbf{w}^\kappa \mathbf{x}_i$, and loss$(\cdot)$ denotes the defined loss function, which aims to minimize the errors caused by the mis-classifed age labels for given face samples. To implement these losses, we propose two types of the loss functions, typically, *Square Loss* and *Cross-Entropy Loss*, which specifically achieve promising performance on a volume of visual analysis tasks [17], [30] by utilizing the deep learning architecture.

To optimize the parameters of the deep networks, we leverage the back-propagation method to compute and update the gradients w.r.t. the defined objectives in a layer-wise manner.

a) Square loss: The goal of this loss aims to minimize the Euclidean distance between the immediate representation from fully connected layers (FC) and the embedded binary outputs for age labels, which is formulated as follows:

$$\mathcal{O} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{\kappa=1}^{\aleph} \| t_n^{\kappa} - f^{\kappa}(\mathbf{x}_n) \|_2^2, \tag{19}$$

where $\|\cdot\|$ denotes the Euqlidean distance of the residual error for the ground-truth and prediction.

The gradients of the parameters $\mathbf{W}$ with respect to the objective $\mathcal{O}$ are performed as follows (ignoring the bias $\mathbf{b}$ for simplicity):

$$\frac{\partial \mathcal{O}}{\partial \mathbf{W}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{\kappa=1}^{\aleph} | t_n^{\kappa} - f^{\kappa}(\mathbf{x}_n) |, \tag{20}$$

b) Cross-entropy loss: The main objective of the cross-entropy loss is to maximize the cross-entropy energy (mutual information) between the feature representation and the corresponding ground-truth age labels, which is written as follows:

$$\mathcal{O} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{\kappa=1}^{\aleph} \mathbb{1}[o_n^{\kappa} = t_n^{\kappa}] \log(p(o_n^{\kappa} | \mathbf{x}_n, \mathbf{W})), \tag{21}$$

where $\mathbb{1}[\cdot]$ denotes a test function, where the result is 1 when the condition is ture, and 0 vice versa.

The gradients of the parameters $\mathbf{W}$ with respect to the objective $\mathcal{O}$ are performed as follows:

$$\frac{\partial \mathcal{O}}{\partial \mathbf{W}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{\kappa=1}^{\aleph} o_n^{\kappa} \cdot \Delta(\kappa), \tag{22}$$

where the updating equation is computed as:

$$\Delta(\kappa) = \mathbb{1}[o_n^{\kappa} = y_n^{\kappa}] - \log(p(o_n^{\kappa} | \mathbf{x}_n, \mathbf{W})).$$

Algorithm 2 shows the optimization procedure of the proposed ODL.

C. Discussions

In this subsection, we briefly discuss the main differences between our proposed approach and other deep learning-based facial age estimation methods.

1) Differences With Our Earlier Work [51]: Compared to our earlier work GA-DFL [51], the proposed ODL differs in two aspects: 1) Since the training set in face aging datasets usually undergo biases, GA-DFL [51] manually divides the whole age progression to a series of discrete age groups . This hand-crafted grouping strategy ignores the feature similarity of face pairs within the same age group in such cases when the appearance of face samples is quite different for neighouring ages. Differently, our ODL approach aims to simultaneously exploit the topology-preserving ordinal relation for age labels and age difference information in the transformed feature space. 2) The face descriptor and OHRanker [9] in GA-DFL [51] are learned separately, so that the optimization procedure may lead to local optima due the two-stage manner. In contrast to GA-DFL [51], we propose an end-to-end ODL method by including two ordinal regression loss functions, which specifically optimize both tasks of learning face representation and age estimator under a unified deep learning paradigm.

Algorithm 2: ODL

Input: Training set: $\mathbf{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, Testing set: $\mathbf{X}' = \{\mathbf{x}'_j\}_{j=1}^{N'}$,.

Output: the predicted age values for testing images $\{\mathbf{x}'_j\}_{j=1}^{N'}$.

Step 1: Pretraining parameters $\mathbf{W}$ by employing ODFL.

Step 2: Build consistent binary outputs for trainset $\mathbf{X}$.

Step 3: Optimization via Back-Propagation

repeat

3.1 Perform forward propagation.

3.2 Perform backward propagation and compute the gradients with respects to the losses $\mathcal{O}$.

3.3 Update the parameters according to (14) and (15).

until *convergence*;

Step 4: For each testing sample $\mathbf{x}' \in \mathbf{X}'$, forward $\mathbf{x}'$ to the learned networks and perform the final age predicting according to (17).

2) Differences With Deep Learning-Based Approaches [12], [15], [16], [52]–[54]: Although the facial age estimation methods [12], [15], [16], [52]–[54] also leveraged deep learning architectures in their models, our models differs in two-fold: 1) Unlike these deep learning-based methods such as [12], [15], [52], [53] which exploit little information of label correlation for ages, our proposed approach explicitly considers the label correlation by taking full access to the ordinal relations of quadruplets and triplets for each batch. 2) In contrast to [16], [39], [54] which cannot explicitly model the structural and high-order relationship for face aging data, our models simultaneously exploit the topology-preserving ordinal relation and the age-difference cost information, making full access to the order relationships of face pairs via both quadruplet-based and triplet-based comparisons. In particular, in contrast to [36], our models simultaneously exploit the topology-preserving ordinal relation and the age-difference cost information, making full access to the order relationships of face pairs via both quadruplet-based and triplet-based comparisons, other than modeling the age-difference information.[1] As a result, the ordinal uniqueness of age information is exploited in the learned feature embedding. Moreover, our model lies that our method, regarding as a feature learning method, is complementary to other facial age estimation methods.

[1]At the time writing, we have not made an access to the results of [36] for performance comparisons before the submission.

IV. Experiments

In the section, we present the employed datasets, evaluation protocols, evaluation settings, experimental results and analysis, respectively.

A. Evaluation Datasets

We evaluated our proposed ODFL and ODL on five widely used face aging datasets including MORPH (Album2) [55], FG-NET [19], FACES [56], LIFESPAN [57] and the apparent facial age estimation [58] datasets. In particular, the FACES [56] and LIFESPAN [57] datasets are exposed to diverse facial expressions which lead to large variances in face aging appearance. Moreover, since face samples in the apparent facial age estimation dataset were captured in the unconstrained conditions, these samples undergo diverse changes due to large poses, make-up appearance and partial occlusions.

1) MORPH (Album 2) [55]: This dataset consists of 55608 face images from about 13000 subjects. The age range lies from 16 to 77 years old and there exists averaging 4 samples per subject.

2) FG-NET [19]: This dataset has 1002 images of 82 persons and there exists averaging 12 samples for each person. The age range covers from 0 to 69. The dataset encounters large variations in pose, illumination and expression.

3) FACES [56]: The dataset contains 2052 face images from 171 persons. The age range covers from 19 to 80 years old. For each person, there are six expressions including neutral, sad, disgust, fear, angry and happy.

4) LIFESPAN [57]: The dataset contains 844 face images from 590 subjects. The age range covers from 18 to 94 years old. The face images of the same person from the LIFESPAN dataset were captured by two expressions: neural and happy. Each person has neural expression and some among them have happy expressions.

5) The Apparent Age Estimation Dataset [58]: This dataset contains 4112 images for training and 1500 images for validation. The age range covers from 0 to 100 years old, which were collected from social networks. The face images suffer from large variations of diverse facial expressions, poses and partial occlusions. Since the ground-truth age labels of testing datasets are not available, we performed age estimation by utilizing the validation set for testing.

B. Experimental Setting and Implementational Details

Before evaluation, we firstly detected the face bounding boxes on the original images based on the open source computer vision library DLIB [59]. We enlarged the detected size by 20% and rescale the detected faces to the size of 256×256×3 with RGB color channels. For each face image to be evaluated, we detected three landmarks including two centers of eyes and the nose base to align the face into the canonical coordinate system by using alternative affine transformation. It is valuable to notify that all face images were augmented by horizontal flipping and random cropping. In our experiments, we mainly leveraged the pre-trained parameters of VGG-16 Face Net [32]. After cropping, the VGG-16 Face Net [32] employed took the cropping size of 224×224×3 patches from 256× 256×3 images during each training epoch.

For the parameters employed in our ODFL and ODL, we set $H = 5$, $\eta = 0.5$, $\lambda_1 = 0.3$ and $\lambda_2 = 0.001$ by cross-validation. For feature comparisons in our ODFL, we adopted a new fully connected layer in the dimension of 4096-50 instead of substituting the last fully connected layer, where the dimension of each feature is reduced to 50. For the end-to-end age prediction in our ODL, we adopted a new fully connected layer in the dimension of 4096-$\mathcal{A}$, where $\mathcal{A}$ denotes the number of age labels on each evaluation dataset. In our experiments, we leveraged the uniform distribution [60] to initialize the parameter of the last layer, and we initialized the parameters of the remaining layers by using the pre-trained model such as VGG Face Net [32]. For the hyper-parameters of the network, we specified the values of the weight decay, moment empirically to 0.0001, 0.9, respectively. The whole training procedure converged until the validation error remained minimized and unchanged. It is valuable to notified that we randomly oversampled all face images during training process by horizontal flipping and shuffling to generate more training samples to reinforce the feature discriminativeness. The whole training procedure converged at around 2k iterations based on the VGG-16 Face Net [32].

Since our ODFL aims to learn face representation for ages, the aligned faces were fed to the designed networks to compute the face descriptors. Having obtained the face representation, we trained an age estimator OHRanker [9] and obtained exact age values during testing procedure. For the end-to-end framework ODL, we directly fed the testing facial images to the trained networks and obtained the final age values.

C. Evaluation Metrics

1) Mean Absolute Error: For the evaluation metrics, we utilized the mean absolute error (MAE) [1], [7], [16], [39] to measure the error between the predicted age and the ground-truth, which is computed as follows:

$$\epsilon = \frac{\sum_{i=1}^{N} \left\| \hat{y}_i - y_i^* \right\|_2}{N} \tag{23}$$

where $\hat{y}$ and y^* denote predicted and ground-truth age value, respectively, and N denotes the number of the testing samples.

2) Cumulative Score Curve: We also applied the cumulative score (CS) [20], [22], [27], [39] curve to quantitatively evaluate the performance of age estimation methods. The cumulative prediction accuracy at the error ϵ is computed as:

$$CS(\theta) = \frac{N_{\epsilon \leq \theta}}{N} \times 100\% \tag{24}$$

where $N_{\epsilon \leq \theta}$ is the number of images on which the error θ is no less than ϵ.

D. Comparisons With State-of-the-Art

To show the superiority of the proposed approach, we compared our ODFL and ODL with the state-of-the-art facial age estimation methods on the MORPH and FG-NET datasets.

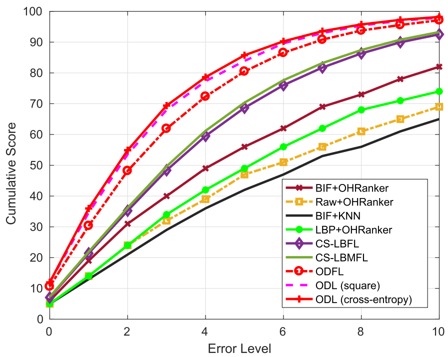


Fig. 6. The CS curves of our ODFL and ODL compared with different facial age estimation methods on the MORPH dataset.

Specifically, we firstly created baseline methods by utilizing the raw pixels, local binary patter (LBP) [2] and bio-inspired feature (BIF) [3] features, and carefully implemented several state-of-the-art methods including OHRanker [9], CS-LBFL [11] and CS-LBMFL [11] by following the details from the original papers. Furthermore, we compared of our approach with several different deep learning-based approaches including DeepRank [39], DeepRank+ [39] and OR-CNN [16], where the experimental results are directly cropped from the related papers.

For evaluation on the MORPH dataset, we performed 10-folds cross-validation for evaluation by following the settings in [11]. Specifically, we divided the whole dataset into ten folds and each fold has the nearly equal size. We used nine folds as the training set, and the remaining one was used for the testing set. We repeated this procedure 10 times and computed the average results as the final age estimation performance. Table I tabulates the MAEs of our methods compared with different facial age estimation methods, and Fig. 6 shows the CS curves of our approach compared with the state-of-the-arts, respectively. According to these results, we see that our methods outperform the hand-crafted features like BIF [3], OHRanker [9] and CS-LBFL [11]. This is because our approach aims to learn deep representation directly from raw pixels and exploits complex and nonlinear relationship between face representation and age labels. Moreover, our approach outperforms deep learning models including DeepRank [39], DeepRank+ [39] and OR-CNN [16], which is because the ordinal relation of quadruplet and triplet comparisons are fully taken into account in both the learned feature representation and age estimation procedures. Besides, our method outperforms [35] and obtain comparable performance with [36], [37] both of which involve external training face aging data in their models. The achievements of our method indicate that we make full use of ordinal relation for age labels in age estimation. However, the achievements of [36], [37] mainly benefit from external training data and the auxiliary attributes including facial race and gender. Thus, our method is complementary to any deep networks and we consider that our model will achieve a big improvement after employing a large scale of face aging data as well as facial attributes during training process.

TABLE I

COMPARISON OF MAEs WITH DIFFERENT STATE-OF-THE-ART APPROACHES ON THE MORPH DATASET (BEST PERFORMANCE IN BOLD, TOP THREE PERFORMANCE IN ITALIC)

Hand-Crafted Methods	**MAE**
BIF+KNN	9.64
AGES [1]	8.83
Raw+OHRanker [9]	7.34
LBP+OHRanker [9]	6.88
BIF+OHRanker [9]	6.49
MTWGP [21]	6.28
LDL [22]	5.69
CPNN [22]	5.67
CA-SVR [63]	4.87
MFOR [64]	5.88
BIF+OLPP [65]	4.20
CS-LDA [66]	6.03
CS-LBFL [12]	4.52
CS-LBMFL [12]	4.37
rKCCA + SVM [67]	3.91
CSOHR [68]	3.74
Deep Learning-Based Methods	**MAE**
DeepRank [41]	3.57
DeepRank+ [41]	3.49
Deep Reg	3.83
OR-CNN [17]	3.27
GA-DFL [53]	3.25
Age-Gender CNN [37]†,‡	3.06
Best from [38]†	**2.78**
Best from [39] †,‡	*2.96*
ODFL + OHRanker	3.12
ODL (Square Loss)	*3.01*
ODL (Cross-Entropy Loss)	*2.92*

†- Using external training data, *e.g.* , CASIA [69], AdienceFace [70], etc.
‡- Using auxiliary attributes such as race and gender

TABLE II

COMPARISON OF MAEs COMPARED WITH STATE-OF-THE-ART APPROACHES ON THE FG-NET DATASET

Hand-Crafted Methods	**MAE**
BIF+KNN	8.24
Raw+OHRanker [9]	6.25
LBP+OHRanker [9]	4.92
BIF+OHRanker [9]	4.48
MLP [22]	6.95
RUN [71]	5.78
AGES [1]	6.77
LARR [28]	5.07
PFA [72]	4.97
KAGES [73]	6.18
MSA [74]	5.36
SSE [75]	5.21
mKNN [76]	5.21
MTWGP [21]	4.83
RED-SVM [8]	5.21
PLO [77]	4.82
LDL [22]	5.77
CA-SVR [63]	4.67
CSOHR [68]	4.70
CS-LBFL [12]	4.43
CS-LBMFL [12]	4.36
CPNN [22]	4.76
Deep Learning-Based Methods	**MAE**
Deep Reg	4.88
GA-DFL [53]	3.93
ODFL + OHRanker	3.89
ODL (Cross-Entropy)	**3.71**

To conduct experiments on FG-NET dataset, we employed the widely used leave-one-person-out (LOPO) for evaluation protocol. Specifically, we randomly selected face images from

TABLE III

COMPARISON OF MAES WITH DIFFERENT STATE-OF-THE-ART APPROACHES ON THE FACES DATASET. FROM THE RESULTS, WE OBSERVE THAT OUR PROPOSED APPROACH EXHIBITS ROBUST TO VARIOUS FACIAL EXPRESSIONS

	Method	Neutral	Happy	Disgust	Fearful	Sad	Angry
Hand-Crafted	BIF [78]	9.50	10.70	13.26	12.65	10.78	13.26
	BIF+MFA [78]	8.14	10.32	12.24	10.73	10.66	10.96
	CS-LDA [66]	5.97	7.52	9.20	8.63	8.48	9.16
	BIF+OHRanker	5.16	7.64	8.31	7.00	6.87	7.87
	LBP+OHRanker	6.36	8.88	9.20	7.30	9.09	8.86
	CS-LBFL [12]	5.06	6.53	7.15	6.32	6.27	6.94
	CS-LBMFL [12]	4.84	5.85	5.70	6.10	4.98	5.50
Deep Learning	DeepRank [41]	5.99	7.12	8.15	6.35	7.77	6.68
	DeepRanker+ [41]	5.86	7.87	7.80	6.66	7.49	6.59
	ODFL + OHRanker	3.48	3.52	4.41	4.52	**3.96**	3.87
	ODL (Cross-Entropy)	**3.37**	**3.49**	**4.32**	**4.40**	4.00	**3.81**

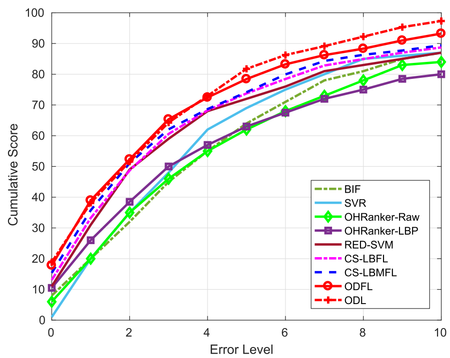


Fig. 7. The CS curves of our ODFL and ODL compared with different facial age estimation methods on the FG-NET dataset.

TABLE IV

COMPARISON OF MAES WITH DIFFERENT STATE-OF-THE-ART APPROACHES ON LIFESPAN DATASET

Method	Neutral	Happy
BIF [78]	8.93	10.75
BIF+MFA [78]	6.05	7.36
CS-LDA [66]	8.18	9.35
LBP+ OHRanker [9]	9.29	10.01
SIFT + OHRanker [9]	9.56	10.00
CS-LBFL [12]	5.79	5.84
CS-LBMFL [12]	5.26	5.84
DeepRank [41]	5.01	2.72
DeepRank+ [41]	5.64	4.18
ODFL + OHRanker	4.70	4.13
ODL (Cross-Entropy)	**4.51**	**3.99**

one person as testing images, and the faces of the remaining persons were used for training. In this way, the whole procedure were performed 82 folds for evaluation. Lastly, we averaged the 82 folds results as the final age estimation results. Table II and Fig. 7 shows the MAEs and the CS curves of our ODFL and ODL compared with the state-of-the-arts, respectively. From the results, we see that our proposed approach outperforms the state-of-the-arts facial age estimation approaches. The performed improvements show the effectiveness of our designed ordinal constraints by utilizing quadruplets and triplets within each batch.

E. Evaluation Regarding With Unbalanced Data

To evaluate our methods regarding with unbalanced training data, we conducted experiments based on our proposed approach when the training data become more and more sparse and unbalanced on both Morph [55] and FG-NET [19] datasets. To achieve this, we removed the data of certain age labels to make the data more and more sparse and unbalanced. Specifically, we randomly selected a fixed number of age groups (0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60+), each time to remove and then trained our models. We created the deep regression (dubbed *Deep Reg.*) with VGG-16 Face Net [32] which was finetuned by the L_2 loss function as the baseline method. Fig. 8 demonstrates facial age estimation performance regarding with the sparse and unbalanced data measured using MAEs on the FG-NET and MORPH datasets, respectively. From these results, we observe that our proposed ODFL and ODL achieve the robustness to the bias training set where the face samples of age groups were randomly removed. This is because our models focus on the ordinal relation of face aging data, more than directly mapping face images to the age targets by taking little label correlation into account.

F. Evaluation Regarding With Various Expressions

In our experimental setting, we conducted the experiments under the same expression on the FACES dataset. Fig. 9 shows the CS curves of our ODFL and ODL compared with different facial age estimation methods and Table III tabulates the MAEs, respectively. According to the results, we see our ODFL and ODL obtains significant performance compared with any other state-of-the-art methods. This is because our method achieves the age-related information across different facial expressions based on the VGG-16 Face Net, which contributes to the improvements for facial age estimation dataset where the face samples even undergo various expressions. Moreover, we conducted age estimation performance under the same expression on the LIFESPAN dataset. We performed five cross-validation for each expression set and computed the averaging MAEs for final results. Table IV demonstrates the experimental performance and Fig. 10 shows the CS curves of our methods on happy expression compared with several facial age estimation methods, respectively. According to these results, our methods significantly improve the performance of facial age estimation, which shows the robustness of our approach regarding with diverse expressions.

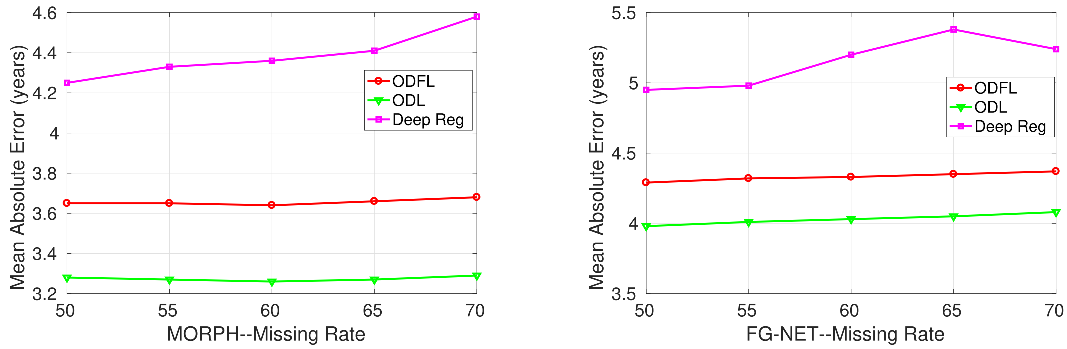


Fig. 8. Age estimation performance with sparse and unbalanced data measured using MAE (the lower the better) on FG-NET and MORPH datasets, respectively. We see that our methods slightly degrade while a subset of samples belonging to some age groups were removed during training procedure, which shows the robustness of our proposed methods to the sparse and unbalanced data.

TABLE V

COMPARISON OF MAEs AND GAUSSIAN ERRORS WITH DIFFERENT FACIAL AGE ESTIMATION APPROACHES ON THE APPARENT AGE ESTIMATION DATASET

Method	MAE	Gaussian Error
BIF+KNN	7.19	0.620
CS-LBFL	5.12	0.422
Deep Reg	5.05	0.456
Single Label	4.58	0.416
Gaussian Label	4.31	0.363
GA-DFL [53]	4.21	0.369
Best from [40]	**3.85**	0.33
ODFL + OHRanker	4.12	0.339
ODL (Cross-Entropy)	3.95	**0.312**

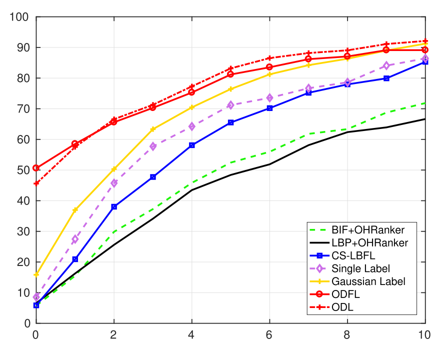


Fig. 9. The CS curves of our ODFL and ODL compared with different facial age estimation methods for Happy Expression on the FACES dataset.

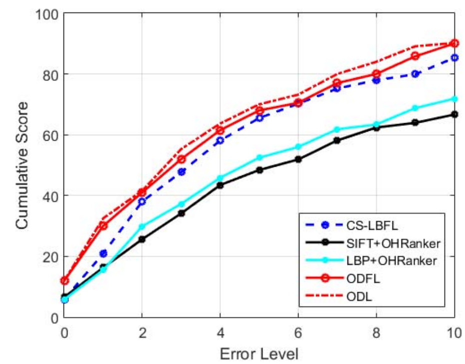


Fig. 10. The CS curves compared with our ODFL and ODL different facial age estimation methods for Neutral Expression on the LIFESPAN dataset.

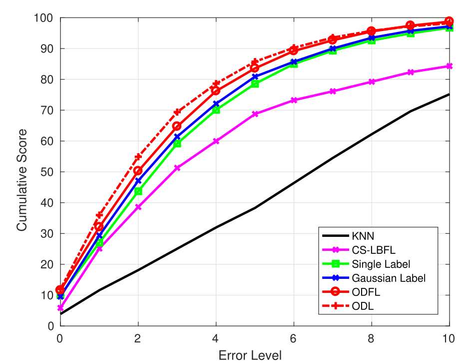


Fig. 11. The CS curves of our ODFL and ODL compared with different facial age estimation methods on the apparent facial age estimation dataset.

G. Evaluation on Unconstrained Dataset

To conduct the experiments of our ODFL and ODL on the apparent facial age estimation that were captured in the wild conditions, we created the single label and Gaussian label methods with the VGG-16 Face Net. Table V tabulates the MAEs and Gaussian errors [58], and Fig. 11 shows the CS curves, respectively. From these results, we see that our methods perform better than other deep learning methods without any additional labeled face aging data. This benefit from three aspects: 1) the learned deep representation can explicitly exploit the complexly nonlinear relationship between face samples and age labels, 2) the proposed criterions in our ODFL model the order information which is helpful for age estimation, and 3) our ODL jointly tuned the parameters of the deep face net by the ordinal regression losses for age predicting, which contributes the improvements for age estimation performance. In addition, we illustrated some resulting samples in Fig. 12, where the age prediction errors are below one year old. From these sampled examples, we see that our model achieves robustness to large variations caused by varying facial expressions, large poses, etc. We also provided some failure examples in Fig. 13 and these results indicates that these failures are mainly generated from extreme challenging cases including diverse mark-up, low resolution and intense illumination.

Fig. 12. The selected examples from the apparent age estimation dataset, where the age prediction errors are below one years old. According to these resulting samples, we see that our approach is robust to large variances of facial wearing glasses, poses and expressions.

Fig. 13. The example faces from the apparent facial age estimation are selected where the predicted errors are larger than 5 years old.

TABLE VI
COMPARISON OF MAEs OF OUR PROPOSED METHODS COMPARED WITH DIFFERENT DEEP NETWORKS ARCHITECTURES ON THE MORPH DATASET

Method	Cropping Size	MAE
AlexNet [18]	227× 227 ×3	3.72
ResNet [42]	224× 224 ×3	3.47
GoogleNet [79]	224× 224 ×3	3.49
ResNet for Face [80]	224× 224 ×3	3.00
Lightened CNN for Face [81]	128× 128 ×1	3.97
VGG-16 Face Net [34] (ODFL)	224× 224 ×3	3.12
VGG-16 Face Net [34] (ODFL+ODL[1])	224× 224 ×3	3.01
VGG-16 Face Net [34] (ODFL+ODL[2])	224× 224 ×3	**2.92**

1. Square Loss 2. Cross-Entropy Loss

H. Parameter Selections

In this part, we investigated the performance effects of different network architectures and tuning parameters employed in our approach.

1) Comparisons With Existing Networks: We compared the performance of our ODFL and ODL with existing deep networks such as AlexNet [17], ResNet-101 [40] and GoogleNet [77] which were pretrained by ImageNet images and the deep architectures including VGG-16 Face Net [32], ResNet for Face [78] and Lightened CNN for Face [79] which were pretrained by face data. Specifically, we directly deployed our proposed objectives of ODFL and ODL to finetune the deep networks. Note that the AlexNet was fed with the color facial images in the size of 227×227. For the remaining deep models, we used gray images of 128×128 for the Lightened CNN, and color facial images of 224×224 for the others. Table VI tabulates the results of our ODFL compared with existing networks. From these results, we see that our approach with the VGG-16 Face Net obtains the best performance. The reason is that the VGG-16 Face Net were pretrained by a large amount of face images for 2622 person identities, which learns to capture more facial patterns than those of any other networks, in order to improve the discriminativeness of learned deep face representation.

TABLE VII
COMPARISON OF MAE WITH DIFFERENT AGE ESTIMATORS ON THE FG-NET DATASET

Method	MAE
VGG-16 Face Net [34] + KNN	4.88
ODFL + SVR	4.47
VGG-16 Face Net [34] + Single Label	3.63
VGG-16 Face Net [34] + Gaussian Label	3.44
VGG-16 [34] features + OHRanker	5.89
ODL + Square Loss	3.31
ODL + Cross-Entropy	3.24
ODFL + OHRanker	3.12
ODFL + ODL + Square Loss	3.01
ODFL + ODL + Cross-Entropy	**2.92**

2) Comparisons With Different Age Estimators: We investigated the effectiveness of different facial age estimators with our learned features. To be specific, we first employed the pretrained VGG-16 Face Net [32] without the fine-tuning training as the feature extractor. We created a baseline method with the unsupervised VGG-16 features and KNN classifier. Then, we deployed the softmax loss [17] as the single label method, and the deep label distribution learning [14] as the Gaussian label methods at the top of the VGG-16 Face Net and finetuned these networks. Moreover, we compared with support vector regression (SVR) [80] and OHRanker, and then computed the MAEs for final performance. As the results are demonstrated in Table VII, we see our ODFL with OHRanker performs better than deep learning based age estimators. The reason is that the structural ordinal relation is exploited by our model in the learned face feature representation, which take advantages of the fully order relationship of quadruplet comparisons. Moreover, our ODL jointly optimized the exacting feature representation and age estimation in an end-to-end manner, so that the complementary information from both phases is exploited to improve facial age estimation performance.

3) Performance Effects of Different Learning Strategies: To address the importance of the proposed two criterions J_1 and J_2, and the regularization term J_3 with the parameter selection of λ_1 and λ_2, we investigated the contributions of different terms in our ODFL model on the MORPH dataset. We defined the following five alternative baselines to investigate the importance of different terms in our deep feature learning model:

- ODFL-1: learning age net only from J_1.
- ODFL-2: learning age net only from J_2.
- ODFL-3: learning age net from J_1 and J_2, where λ_1 was specified to 0.8 ($\lambda_2 = 0$).
- ODFL-4: learning age net from J_1, J_2 and J_3, where λ_1 and λ_2 were specified to 0.8 and 0.001, respectively.
- ODFL-5: learning age net from J_1, J_2 and J_3, where λ_1 and λ_2 were specified to 0.3 and 0.001, respectively.

Accordingly, ODFL-1 and ODFL-2 aim to learn the parameters of the proposed deep CNN architecture by employing J_1 and J_2 separately, ODFL-3 performs the optimization procedure without the regularization term J_3, and ODFL-4 and ODFL-5 perform Algorithm 1 by specifying the

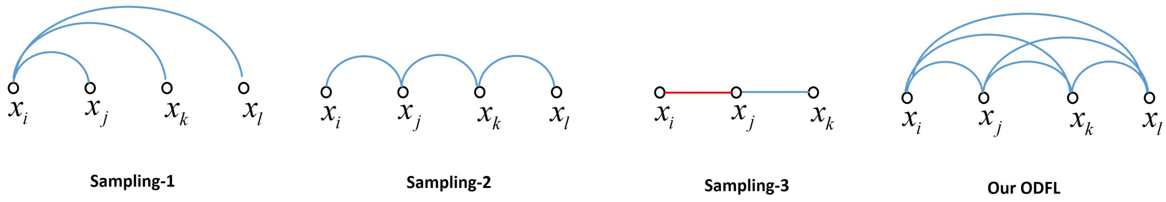


Fig. 14. An illustration of different quadruplets or anchor triplets in batches. Specifically, *Sampling-1* performs the anchored quadruplets which were similar with the anchored triplets; *Sampling-2* performs the quadruplets for only neighbouring pairs; *Sampling-3* performs the triplet method without the weighting function to smooth the distances with age differences (red line denotes the positive pair while blue line denotes the negative pair). In contrast, our ODFL explicitly takes into account total pairwise edges by all quadruplets and triplets within the mini-batch.

TABLE VIII

COMPARISON OF MAEs AND CED VALUES OF OUR METHOD FOR THE GIVEN $\theta = \{1, 5\}$ WITH DIFFERENT LEARNING STRATEGIES ON THE MORPH DATASET

Method	MAE	CDE$_{\theta<1}$	CDE$_{\theta<5}$
ODFL-1	3.45	26.2%	72.8%
ODFL-2	3.51	23.3%	69.5%
ODFL-3	3.24	28.5%	74.3%
ODFL-4	3.19	30.8%	76.9%
ODFL-5	**3.12**	**31.4%**	**80.2%**

parameters λ_1 and λ_2 to 0.8 ,0.3 and 0.001, respectively. It is notified that we utilized ODFL-5 as the final experimental settings. The following table tabulates the mean absolute errors (MAE, years old) and the cumulative scores (CS) for evaluation of ODFL and other four variations on the MORPH dataset.

Table VIII tabulates the performance effects of different learning strategies. According to these results, we see that both criterions J_1 and J_2 in our proposed method achieve discriminative information in our learned face descriptor, and J_1 contributes more than J_2 by exploiting the ordinal information. In terms of the penalty term J_3, λ_2 was set to 0.001 empirically and our approach is not sensitive to it. Moreover, the highest performance can be obtained when all three terms are used to learn the face descriptor, where the complementary information of both the ordinal relation and age-difference information for the chronological age labels is explicitly exploited, simultaneously.

4) Comparisons With Different Sampling Strategies: To further investigate the performance effects of our ODFL regarding with different quadruplets and anchor triplets, we created three baseline methods according to various sampling strategies as follows (refer to the illustration of different methods based on quadruplets and triplets in Fig. 14):

- *Sampling-1:* Within the quadruplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$, suppose we have an anchored sample $\mathbf{x}_i$ and formed comparisons with other samples $\mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l$.
- *Sampling-2:* Within the quadruplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$, we only paired the neighbouring samples such that the constraint compares for the distances of neighbouring face samples.
- *Sampling-3:* Within the triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ anchored by $\mathbf{x}_i$, we sampled the positive face pair $\mathbf{x}_i$ and $\mathbf{x}_j$ with the same age and the negative face pair $\mathbf{x}_i$ and $\mathbf{x}_k$ with the different age values.

Note that our ODFL considers the full pairing comparisons within the quadruplets of both neighbouring and high-order ordinal relationships, as well as the triplets of age difference information. Table IX shows the results of our ODFL regarding with different quadruplets and triplets on the MORPH dataset. From the results, we see that our ODFL with both quadruplet and triple-based relationships achieves the best performance compared with *Sampling-1* and *Sampling-2*, which benefits from the complementary information of both the topology-preserving ordinal relation and age-difference information. Another reason lies on that our model takes full access to the face pairs and meanwhile exploits the high-order relation among face pairs. Moreover, compared with the baseline method *Sampling-3*, our ODFL with J_2 performs better results which demonstrates the importance of the age-difference information exploited in the feature subspace.

TABLE IX

PERFORMANCE OF OUR ODFL REGARDING WITH DIFFERENT SAMPLING STRATEGIES OF QUADRUPLETS AND TRIPLETS ON THE MORPH DATASET. NOTE THAT THE EMPLOYED DEEP NETWORK WAS VGG-16 FACE NET [32] AND WE LEVERAGED THE OHRANKER [9] AS THE AGE ESTIMATOR FOR EVALUATION

Method	Sampling Strategy	MAE
Sampling-1	Quadruplet	3.67
Sampling-2	Quadruplet	3.58
Sampling-3	Triplet	3.73
our ODFL with J_1	Quadruplet	3.45
our ODFL with J_2	Triplet	3.51
our ODFL with J_1 and J_2	Quadruplet & Triplet	**3.12**

TABLE X

COMPUTATION TIME (SECOND) COMPARISONS OF OUR METHODS WITH DIFFERENT FEATURE LEARNING-BASED APPROACHES ON THE MORPH DATASET. NOTE THAT THESE SHALLOW FEATURE LEARNING-BASED MODELS WERE TESTED ON WITH A CPU, WHILE OUR MODELS USED WERE EVALUATED WITH A GPU COMPUTATION CARD

Method	Testing Time (imgs/s)
DFD [83]	2
LQP [84]	10
RICA [85]	3.5
CS-LBFL [12]	20
AlexNet [18]	2425.3
ResNet-101 [42]	256.8
ResNet for Face [80]	256.8
Lightened CNN for Face [81]	2173.2
GoogleNet [79]	346.2
VGG-16 Face Net	143.2

I. Computational Time

Our approach was implemented by the open source Caffe [84] deep learning toolbox, and we trained our model

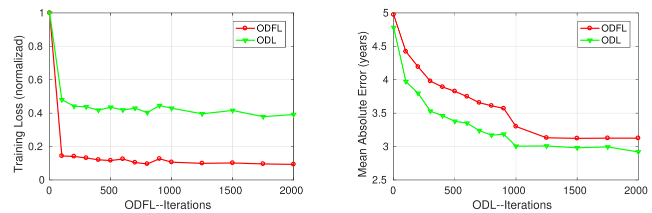


Fig. 15. Loss and Testing MAEs across iterations of both our ODFL and ODL evaluated on the MORPH dataset. Note that we decreased the learning rate by 0.1 after the 1000-th iteration.

with a speed-up parallel computing technique by using single GPU with NVIDIA GTX 1080. Our models converged at about 2000 iterations by monitoring the convergence rate versus the testing performance in Fig. 15. Moreover, we compared our models with several shallow facial age estimation approaches such as DFD [81], LQP [82], RICA [83] and CS-LBFL [11] with a CPU. We also reported the computational time under the GPU parallel computing card compared with different deep architectures. Table X tabulates the comparisons of the computational time during the testing phase. From these results, we see that the deep architectures achieve the real-time age estimation with a GPU platform. Besides, the OHRanker employed in our experiments takes 0.04 seconds by using an Intel i7-CPU@3.40GHz PC, which satisfies the real-time requirement.

J. Discussion

The above experimental results suggest the following three key observations:

1) Compared with facial age estimation methods which employ hand-crafted features [9], [19]–[21] and linear feature filters [3], [10], [11], our ODFL and ODL achieve the best performance than the state-of-the-art approaches on five facial age estimation datasets. This is because our approach automatically learns feature representation directly from raw pixels, which achieves strong robustness to diverse facial expressions, aspect ratios and cluttered background. Moreover, our model learns to exploit the nonlinear relationship between face samples and age labels, which at the same time embeds the ordinal relation for aging pattern in the learned feature space. Hence, higher age estimation performance is obtained.
2) Compared with the age estimation methods which utilize deep learning techniques [14], [16], [17], [39], each of the proposed criterions in our feature learning method ODFL is effective to exploit the order information for age labels. Hence, the best age estimation performance is obtained when all these terms are used together for ordinal feature representation learning.
3) Our proposed ODL outperforms most of the state-of-the-art approaches. This is because our ODL leverages the ordinal regression losses for end-to-end age predicting, so that the complementary information of both feature extraction and age predicting phases are exploited to reinforce our model.

V. Conclusions and Future Work

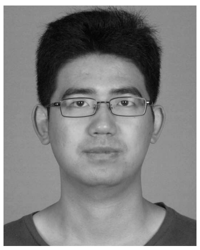
We have proposed an ordinal deep learning approach for facial age estimation. We have developed a feature learning method named ODFL by enforcing two defined criterions, which aims to learn face descriptors directly from raw pixels. Furthermore, we have proposed an end-to-end deep learning framework ODL, so that both procedures of extracting facial features and predicting age values are jointly optimized in a unified deep learning framework. Experimental results on five face aging datasets show the effectiveness of the proposed methods. Since our method is complementary to any deep networks, we believe that our model can achieve a big improvement after introducing a large scale of face aging data, as well as auxiliary facial attributes. It is desirable to address facial age estimation with the feed-back deep networks [49], [50] to further exploit with the complementary information for the personalized aging pattern. Moreover, how to exploit the order information for face aging problem which might help to promote performance of the age-invariant face recognition is an interesting work in the future.

References

[1] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.
[2] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[3] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. CVPR*, Jun. 2009, pp. 112–119.
[4] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proc. ICCV*, 2015, pp. 3970–3978.
[5] W. Wang *et al.*, "Recurrent face aging," in *Proc. CVPR*, 2016, pp. 2378–2386.
[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[7] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.
[8] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A ranking approach for human ages estimation based on face images," in *Proc. ICPR*, Aug. 2010, pp. 3396–3399.
[9] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. CVPR*, Jun. 2011, pp. 585–592.
[10] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.
[11] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, Dec. 2015.
[12] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. ACCV*, 2014, pp. 144–158.
[13] X. Liu *et al.*, "AgeNet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. ICCVW*, 2015, pp. 16–24.
[14] X. Yang *et al.*, "Deep label distribution learning for apparent age estimation," in *Proc. ICCVW*, 2015, pp. 102–108.
[15] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. WACV*, 2015, pp. 534–541.
[16] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. CVPR*, 2016, pp. 4920–4928.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[18] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep feature learning for facial age estimation," in *Proc. FG*, Jun. 2017, pp. 157–164.

[19] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.
[20] Y. Zhang and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. CVPR*, 2010, pp. 2622–2629.
[21] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[22] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. CVPR*, 2011, pp. 657–664.
[23] Z. Lou, F. Alnajar, J. Alvarez, N. Hu, and T. Gevers, "Expression-invariant age estimation using structured learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 365–375, Feb. 2018
[24] S. Feng, C. Lang, J. Feng, T. Wang, and J. Luo, "Human facial age estimation by cost-sensitive label ranking and trace norm regularization," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 136–148, Jan. 2017.
[25] H. Dibeklioğlu, F. Alnajar, A. A. Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1928–1943, Jun. 2015.
[26] G. Guodong and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 761–770, 2014.
[27] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.
[28] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. CVPR*, 2009, pp. 112–119.
[29] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. ICCV*, 2015, pp. 3676–3684.
[30] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in *Proc. ECCV*, 2014, pp. 1–16.
[31] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.
[32] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, p. 6.
[33] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," *Neurocomputing*, vol. 187, pp. 4–10, Apr. 2016.
[34] Z. Kuang, C. Huang, and W. Zhang, "Deeply learned rich coding for cross-dataset facial age estimation," in *Proc. ICCVW*, 2015, pp. 338–343.
[35] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. CVPRW*, 2015, pp. 34–42.
[36] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3087–3097, Jul. 2017.
[37] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognit.*, vol. 66, pp. 106–116, Jun. 2017.
[38] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, "Kernel ELM and CNN based facial age estimation," in *Proc. CVPRW*, 2016, pp. 785–791.
[39] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen, "Automatic age estimation from face images via deep ranking," in *Proc. BMVC*, 2015, pp. 55.1–55.11.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[41] C. Li, Q. Liu, J. Liu, and H. Lu, "Ordinal distance metric learning for image ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1551–1559, Jul. 2015.
[42] K. H. Huang and H. T. Lin. (2016). "Cost-sensitive label embedding for multi-label classification." [Online]. Available: https://arxiv.org/abs/1603.09048
[43] M. Kleindessner and L. U. von, "Uniqueness of ordinal embedding," in *Proc. COLT*, 2014, pp. 40–67.
[44] Y. Terada and U. Luxburg, "Local ordinal embedding," in *Proc. PMLR*, 2014, pp. 847–855.
[45] J. Weston, S. Bengio, and N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. IJCAI*, 2011, pp. 2764–2770.
[46] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, 2014, pp. 1386–1393.
[47] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. CVPR*, 2015, pp. 1556–1564.
[48] E. Arias-Castro. (2015). "Some theory for ordinal embedding." [Online]. Available: https://arxiv.org/abs/1501.02861
[49] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. CVPR*, 2015, pp. 1110–1118.
[50] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with LSTM recurrent neural networks," in *Proc. CVPR*, 2015, pp. 3547–3555.
[51] H. Liu, J. Lu, J. Feng, and J. Zhou, "Group-aware deep feature learning for facial age estimation," *Pattern Recognit.*, vol. 66, pp. 82–94, Jun. 2017.
[52] R. Ranjan *et al.*, "Unconstrained age estimation with deep convolutional neural networks," in *Proc. ICCVW*, 2015, pp. 351–359.
[53] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," in *Proc. BTAS*, 2016, pp. 1–8.
[54] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.
[55] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. FG*, 2006, pp. 341–345.
[56] N. C. Ebner, M. Riediger, and U. Lindenberger, "FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behav. Res. Methods*, vol. 42, no. 1, pp. 351–362, 2010.
[57] M. Minear and D. C. Park, "A lifespan database of adult facial stimuli," *Behav. Res. Methods*, vol. 36, no. 4, pp. 630–633, 2004.
[58] S. Escalera, "ChaLearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. ICCVW*, 2015, pp. 243–251.
[59] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.
[60] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Aistats*, vol. 9. 2010, pp. 249–256.
[61] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. CVPR*, 2013, pp. 2467–2474.
[62] R. Weng, J. Lu, G. Yang, and Y.-P. Tan, "Multi-feature ordinal ranking for facial age estimation," in *Proc. FG*, 2013, pp. 1–6.
[63] G. Guo and G. Mu, "Human age estimation: What is the influence across race and gender?" in *Proc. CVPR*, 2010, pp. 71–78.
[64] J. Lu and Y. P. Tan, "Cost-sensitive subspace learning for human age estimation," in *Proc. ICIP*, 2010, pp. 1593–1596.
[65] G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 761–770, 2014.
[66] 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.*, vol. 24, no. 3, pp. 785–798, Mar. 2015.
[67] D. Yi, Z. Lei, S. Liao, and S. Z. Li. (2014). "Learning face representation from scratch." [Online]. Available: https://arxiv.org/abs/1411.7923
[68] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.
[69] S. Yan, H. Wang, X. Tang, and T. S. Huang, "Learning auto-structured regressor from uncertain nonnegative labels," in *Proc. ICCV*, 2007, pp. 1–8.
[70] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "A probabilistic fusion approach to human age prediction," in *Proc. CVPRW*, 2008, pp. 1–6.
[71] X. Geng, K. Smith-Miles, and Z. H. Zhou, "Facial age estimation by nonlinear aging pattern subspace," in *Proc. ACM MM*, 2008, pp. 721–724.
[72] X. Geng and K. Smith-Miles, "Facial age estimation by multilinear subspace analysis," in *Proc. ICASSP*, 2009, pp. 865–868.
[73] S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, and T. S. Huang, "Synchronized submanifold embedding for person-independent pose estimation and beyond," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 202–210, Jan. 2009.
[74] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning distance metric regression for facial age estimation," in *Proc. ICPR*, 2012, pp. 2327–2330.
[75] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *Proc. CVPR*, 2012, pp. 2570–2577.
[76] G. Guo and X. Wang, "A study on human age estimation under facial expression changes," in *Proc. CVPR*, 2012, pp. 2547–2553.
[77] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

[78] I. Masi, A. T. Trân, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Proc. ECCV*, 2016, pp. 579–596.
[79] X. Wu, R. He, Z. Sun, and T. Tan, "A lightened CNN for deep face representation." [Online]. Available: https://arxiv.org/abs/1511.02683
[80] A. Smola and V. Vapnik, "Support vector regression machines," in *Proc. NIPS*, 1997, pp. 155–161.
[81] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 289–302, Feb. 2014.
[82] S. Ul Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. BMVC*, 2012, p. 11.
[83] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. NIPS*, 2011, pp. 1017–1025.
[84] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093



Hao Liu received the B.S. degree in software engineering from Sichuan University, China, in 2011, and the M.E. degree in computer technology from the University of Chinese Academy of Sciences, China in 2014. He is currently pursuing the Ph.D. degree with the Department of Automation, Tsinghua University. His research interests include face alignment, facial age estimation, and deep learning.

Jiwen Lu (M'11–SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. From 2011 to 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored or co-authored over 180 scientific papers in these areas, including 52 IEEE Transactions papers. He is currently a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society, respectively. He was a recipient of the National 1000 Young Talents Plan Program. He serves as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, *Pattern Recognition Letters*, the *Journal of Visual Communication and Image Representation*, *Neurocomputing*, and IEEE ACCESS.

Jianjiang Feng (M'10) received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007, respectively. From 2008 to 2009, he was a Post-Doctoral Researcher with the PRIP Laboratory, Michigan State University, East Lansing, MI, USA. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing. His research interests include fingerprint recognition and computer vision. He is an Associate Editor of *Image and Vision Computing*.

Jie Zhou (M'01–SM'04) received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 60 papers have been published in top journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and Conference on Computer Vision and Pattern Recognition. His current research interests include computer vision, pattern recognition, and image processing. He was a recipient of the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *International Journal of Robotics and Automation*, and two other journals.