GroupFace: Imbalanced Age Estimation Based on Multi-Hop Attention Graph Convolutional Network and Group-Aware Margin Optimization

Yiping Zhang, Yuntao Shou, *Student Member, IEEE*, Wei Ai, Tao Meng, and Keqin Li, *Fellow, IEEE*

***Abstract*—With the recent advances in computer vision, age estimation has significantly improved in overall accuracy. However, owing to the most common methods do not take into account the class imbalance problem in age estimation datasets, they suffer from a large bias in recognizing long-tailed groups. To achieve high-quality imbalanced learning in long-tailed groups, the dominant solution lies in that the feature extractor learns the discriminative features of different groups and the classifier is able to provide appropriate and unbiased margins for different groups by the discriminative features. Therefore, in this novel, we propose an innovative collaborative learning framework (GroupFace) that integrates a multi-hop attention graph convolutional network and a dynamic group-aware margin strategy based on reinforcement learning. Specifically, to extract the discriminative features of different groups, we design an enhanced multi-hop attention graph convolutional network. This network is capable of capturing the interactions of neighboring nodes at different distances, fusing local and global information to model facial deep aging, and exploring diverse representations of different groups. In addition, to further address the class imbalance problem, we design a dynamic group-aware margin strategy based on reinforcement learning to provide appropriate and unbiased margins for different groups. The strategy divides the sample into four age groups and considers identifying the optimum margins for various age groups by employing a Markov decision process. Under the guidance of the agent, the feature representation bias and the classification margin deviation between different groups can be reduced simultaneously, balancing inter-class separability and intra-class proximity. After joint optimization, our architecture achieves excellent performance on several age estimation benchmark datasets. It not only achieves large improvements in overall estimation accuracy but also gains balanced performance in long-tailed group estimation.**

***Index Terms*—Age estimation, imbalanced learning, graph convolutional network, reinforcement learning.**

I. Introduction

AGE is one of the most important biometric traits in faces, and the general precision in age estimation has seen notable advancements in recent years. It encompasses a broad spectrum of application scenarios, including social media, visual surveillance, image retrieval, marketing, and public safety [1], [2], [3].

Driven by the rapid development of deep learning, age estimation based on Convolutional Neural Networks (CNN) achieves promising performance. Rothe et al. [1] utilized a pre-trained VGG-16 network for face representation learning, then used the classification probability multiply the corresponding labels to get a regression result, which is much better than the single classification or regression methods. Chen et al. [4] proposed Ranking-CNN, which converted the age estimation task into a ranking challenge, and obtained age prediction results by summing the binary classification results. With the advent of Vision Transformer (ViT) [5], its powerful ability to globally model dependencies was put into use for face representation learning. Kuprashevich et al. [6] proposed a unified dual model Multi Input VOLO (MiVOLO), which integrated the gender and age estimation tasks in the field based on the Vision Transformer. Qin et al. [7] proposed Transformer-based SwinFace, which achieved multi-task face feature extraction such as age estimation through a shared backbone and a sub-net for each related task. Moreover, benefits from the flexibility of Graph Convolutional Neural Networks (GCN) in processing complex and irregular objects, GCN achieves performance similar to or even surpassing that of Transformer in feature extraction. Shou et al. [8] proposed Masked Contrastive Graph Representation Learning (MCGRL) to capture the rich structural face information more flexibly and with low redundancy, which outperformed most of the CNN and ViT-based methods on age estimation task. These frontier age estimation researches mainly focus on designing feature extraction networks for more robust face representation learning and optimizing age estimation strategies for more accurate prediction performance, which have achieved great improvement in the overall age estimation performance. However, a prominent issue is that these approaches overlook the issue of imbalanced data distribution, fail to maintain balanced performance on long-tailed group recognition and show significant degradation when encountering specific scenarios.

Received 17 June 2024; revised 15 November 2024; accepted 14 December 2024. Date of publication 18 December 2024; date of current version 7 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 61802444 and in part by the Research Foundation of the Education Bureau of Hunan Province of China under Grant 22B0275. The associate editor coordinating the review of this article and approving it for publication was Prof. Zhen Lei. *(Corresponding author: Tao Meng.)*

Yiping Zhang, Yuntao Shou, Wei Ai, and Tao Meng are with the College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, Hunan 410004, China (e-mail: yipingzhang@csuft.edu.cn; shouyuntao@stu.xjtu.edu.cn; aiwei@hnu.edu.cn; mengtao@hnu.edu.cn).

Keqin Li is with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/TIFS.2024.3520020

Due to different difficulties in collecting samples from different age groups, the age estimation datasets imbalance problem is so severe that the prediction accuracy for some groups or specific application scenarios is poor. For example, in the long-span dataset, MIVIA [9], the number of images for certain age groups (0.6% for "children", 4.4% for "teenager", and 6.3% for "senior") is much lower than for the head age group (88.7% for "adult"), and this imbalanced distribution leads to biased identification of long-tailed groups. In many cases, although the tail categories are numerically small, ignoring them will be costly. Application scenarios such as grading systems in social media, anti-addiction of minors in games, searching for lost children in public safety, and fraud prevention for the elderly. Seldom studies have focused on long-tailed recognition in age estimation. Bao et al. [10] designed a two-stage framework that decouples the learning process into representation learning and classification, which utilized a balanced sampling dataset to train a balanced classifier. Deng et al. [11] utilized globally-tuned ResNet-34 for feature extraction and proposed the variational margin to minimize the effect of misleading tail sample predictions for head class. Wang et al. [3] proposed Meta-Set Learning (MSL) based on RestNet-34 and created an unfair filtering network to identify and filter the noisy samples, and then obtained a balanced meta-set for meta-weighting to alleviate the unfairness between age estimation. Though these techniques mitigate the class imbalance issue to some extent, they still have some shortcomings: **i) Weak in discriminative feature learning.** The CNN-based local feature extraction [3], [11] or Transformer-based global feature extraction [6], [7] tends to learn the consistency of most classes, which is difficult to effectively learn the personalized information of different samples. In particular, the lack of discriminative sample mining leads to the over-fitting of minority classes. **ii) Poor in identifying appropriate and unbiased margins for different groups.** Some methods [8], [10] decoupled the two-stage learning process, which is no explicit control over the distribution of learned features and hard to provide optimal margins for different groups from staged extracted features. Other methods [3], [12] introduced meta-learning for learning the distribution of imbalanced samples, but meta-learning is extremely sensitive to noise, making it difficult to flexibly and stably provide appropriate margins for different groups.

To overcome the above shortcomings for improving the generalization performance of age estimation, it is necessary to focus on both discriminative feature extraction of different groups and an imbalanced margin learning strategy for long-tailed classes. In this novel, an innovative collaborative learning framework (GroupFace) is proposed that integrates a multi-hop attention graph convolutional network and dynamic group-aware margin strategy based on reinforcement learning. To extract the discriminative features among different age groups, we design an enhanced multi-hop attention graph convolutional network. It has the capability to capture the interactions of neighboring nodes at different distances, effectively extending the receptive field of the graph model. It also models facial deep aging by fusing local and global information, thus capturing diverse representations of different

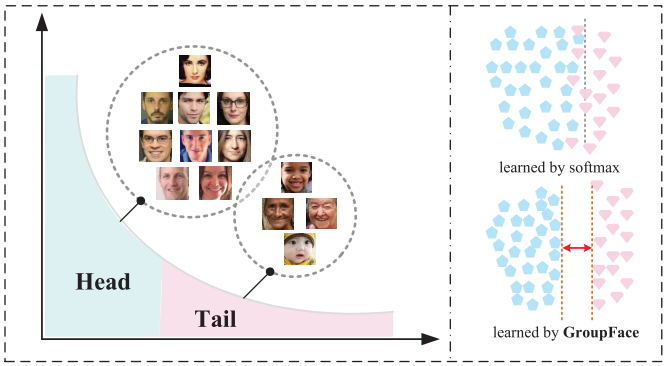

Fig. 1. The illustration of age estimation with class imbalanced learning. Most face datasets have an imbalanced distribution of race and age groups, such that the recognition bias is high for the long-tailed groups. Our GroupFace can achieve a balanced generalization capability for different age groups by discriminative feature extraction and group-aware margin optimization.

groups. Specifically, to maintain message diversity for deeper mining of discriminative features, we design an adaptive decay strategy during graph diffusion to adaptively assign learnable weights based on different hop distances, randomly drop messages during message propagation to prevent over-fitting and incorporate residual blocks in graph convolution to prevent over-smoothing. Moreover, to address the computational inefficiency of higher-order graph models, we utilize a power iteration method to approximate the accelerated inverse matrix.

At the same time, to further address the class imbalance problem, we design a dynamic group-aware margin strategy based on reinforcement learning. We roughly categorize ages into four groups ("children", "teenager", "adult", and "senior"), and design a more flexible and stable group-aware margin loss function. While considering employing a Markov decision process to identify the optimum margins for various age groups, we utilize deep q-learning to acquire a strategy for selecting appropriate group margins. As shown in Fig. 1, under the guidance of the agent, the representation bias in the feature space and the margin deviation in the classification space between different groups can be reduced simultaneously, while the inter-class separability and intra-class proximity be improved, then achieving a balanced generalization ability. The resulting adaptive and unbiased margins for different age groups are more conducive to subsequent accurate age classification and regression from coarse to fine.

Finally, through joint optimization, the group-aware margin policy will guide and facilitate our whole age estimation architecture GroupFace. The contributions to this novel are concluded as follows:

- We propose an innovative collaborative learning framework (GroupFace) that integrates a multi-hop attention graph convolutional network and a group-aware margin strategy, which focuses on both discriminative feature extraction of different groups and an imbalanced margin learning strategy for long-tailed classes.
- To achieve more discriminative representation learning, we propose an enhanced multi-hop attention graph convolutional network fusing local and global information to model aging changes in faces and design adaptive decay

diffusion, random message dropping, and power iteration methods to enhance the graph model.
- To overcome the imbalanced distribution problem, we propose a dynamic group-aware margin strategy based on reinforcement learning with a more flexible and stable margin loss function. The balanced generalization ability is achieved by facilitating the search for optimal margins for different age groups through reinforcement learning.
- Extensive experiments have shown that our architecture not only provides a significant improvement in overall estimation accuracy but also balances performance in long-tailed groups.

II. Related Work

In this section, we first present an overview of the frontier research on Graph Neural Networks. Then we review some studies in the general and imbalanced age estimation. Finally, we provide a brief description of reinforcement learning.

A. Graph Neural Network

Graph Neural Networks and their variants extend deep networks from regular grids to non-Euclidean graph-structured data, showing great potential in areas such as action recognition [13], social media [14], traffic prediction [15] and computer vision [16].

Existing models mainly adhere to a message-passing structure, aggregating information from adjacent nodes in a direct connection. Graph Convolutional Neural Network (GCN) [17] aggregated the representations of their one-hop neighbor nodes in a recursive manner and in the process normalized the weights of the edges using the Laplace matrix. On the other hand, Graph Attention Network (GAT) [18] introduced a multi-head self-attention mechanism that dynamically discerned the significance of various neighboring nodes, thus discarding the fixed neighbor weight settings in traditional methods. To address the challenges of processing large graph data, GraphSAGE [19] proposed an efficient batch training algorithm that ensures constant computational and memory complexity regardless of the graph size, which significantly improved the scalability of large graphs. However, these methods restrict information extraction to the local neighborhood, limiting deep feature extraction and representation of the graph model.

Since the use of multiple single-hop message-passing layers might lead to a decline in model efficacy due to the effects of Laplace smoothing, multi-hop graph networks have been put forth to address this issue by capturing information from the k-hop neighborhood vicinity. Simplified Graph Convolutional Network (SGC) [20] utilized powers of the adjacency matrix to generate multi-hop neighbor representations. Abu-El-Haija et al. [21] designed MixHop, a mixed-hop GNN that broadens the receptive field by reiterating the feature representations of neighbors at varying distances. Wang et al. [22] proposed Multi-hop Attention Graph Neural Network (MAGNA) to deal with the over-smoothing problem by graph attention and diffusion methods, but fixed decay weights make it inflexible for different nodes at the same distance. In this work, we pay attention to the optimization of the multi-hop model to obtain more discriminative face features.

B. General Age Estimation

Age estimation becomes a challenging task due to many internal and external factors, and the mainstream works for general age estimation have focused on two directions: designing more robust feature extraction models [8], [23] and optimizing more accurate age prediction strategies [24], [25]. Zhang et al. [23] proposed C3AE to train multi-scale images using cascade networks to make full use of contextual information. Shin et al. [24] developed a new sequential regression technique named Moving Window Regression (MWR), integrating the notion of relative rank and establishing both local and global relative regressors to attain the rho-rank in the whole and specific rank ranges. Chen et al. [25] presented the Delta Age AdaIN (DAA), comprising components such as a facial encoder, DAA operation, binary code mapping, and age decoder module. Shou et al. [8] proposed the new representation framework Masked Contrastive Graph Representation Learning (MCGRL), which can learn face structure and semantics flexibly.

These techniques designed for general age estimation tasks have enhanced the overall precision. However, overlooking the long-tailed distribution of age datasets, their performance deteriorates severely when it comes to younger and older individuals.

C. Imbalanced Age Estimation

The imbalanced distribution of data is widespread in real life. There has been quite a bit of groundwork in the visual imbalanced learning species to address such problems, which can be categorized into two main types: data level [10], [26] and algorithm level [27], [28]. The data level balances the class distribution by resampling the training data, but this tends to destroy the original expression. The algorithm level improves the importance of minority classes by improving existing algorithms, including cost-sensitive learning, ensemble learning, and other methods.

However, studies dealing with imbalanced age estimation are still scarce, and ignoring the marginalized is disastrous and costly in some scenarios. Bao et al. [10] decoupled face representation learning and age classification at the data level by training a balanced classifier separately from a balanced dataset obtained by class-balanced sampling. Deng et al. [11] proposed variational margins at the algorithm level to mitigate the misleading tail-sample prediction in the head class. Wang et al. [29] carefully created an unfair filtering network under a meta-learning paradigm that utilizes meta-re-weighting interventions to reduce training bias caused by category imbalance. Bao et al. [30] proposed Pixel-level Auxiliary learning (PA) and Feature Rearrangement (FR) to better utilize the facial features, while Adaptive Routing (AR) was devised to select the appropriate classifiers to improve the long-tailed recognition.

In this work, we consider both the discriminative feature extraction and the imbalanced learning to achieve a balanced

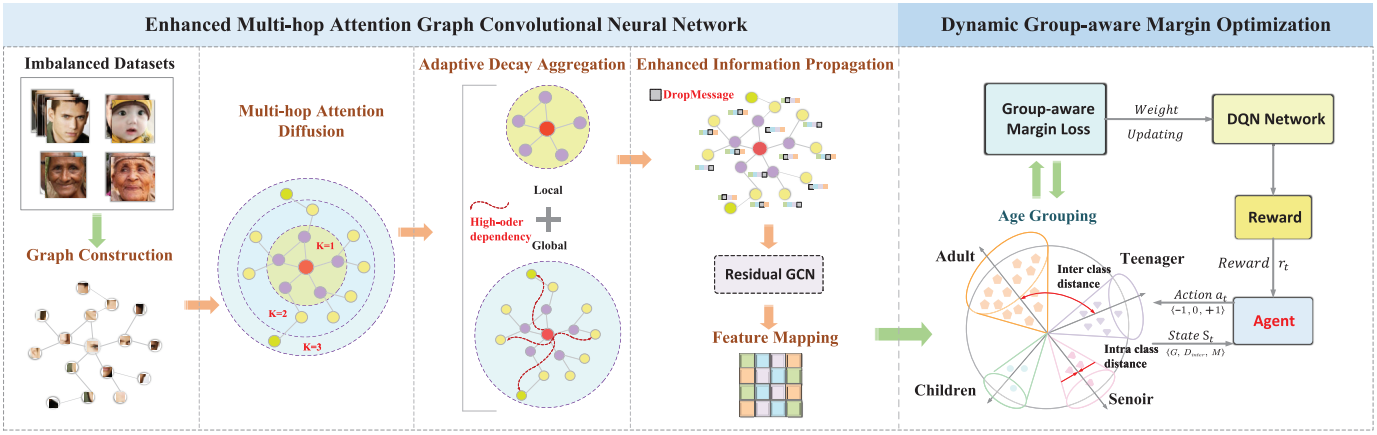

Fig. 2. **The overall framework** of our proposed imbalanced learning method **GroupFace**. Face images are segmented into patches as nodes, and then a multi-hop attention graph convolutional network will fuse global and local information to model deep facial aging for capturing the discriminative features of different groups. Through joint optimization, the dynamic group-aware margin strategy based on reinforcement learning will identify the optimum margins for various age groups to mitigate the bias of imbalanced learning.

performance for the long-tailed groups while improving the overall accuracy.

D. Reinforcement Learning

Reinforcement learning mimics the human decision-making process training agents to learn trial-and-error-based strategies by maximizing cumulative rewards in dynamic environments. In addition to applications such as robot control and gaming, reinforcement learning has recently been successfully applied to several visual recognition tasks. Lin et al. [31] modeled a sequential decision-making process to classify the images, and devised more rewards for the minority category. Liu et al. [32] presented fair loss, which is a margin-aware reinforcement learning-based loss function to learn an adaptive margin. Wang et al. [29] proposed a reinforcement learning-based Racial Balancing Network (RL-RBN) to find the most suitable margin for non-Europeans and can reduce the skewness of feature dispersion among races. In this work, a dynamic group-aware margin strategy based on reinforcement learning is designed with a more flexible and stable margin loss function for imbalanced age estimation.

III. Preliminaries

In this section, we will elaborate on our proposed imbalanced learning framework GroupFace consisting of the Enhanced Multi-hop Attention Graph Convolutional Network and Dynamic Group-aware Margin Optimization. The overall pipeline is shown in Fig. 2.

A. Enhanced Multi-Hop Attention Graph Convolutional Network

Aging changes between facial key points profoundly affect the accuracy of age estimation. However, common GCNs only focus on one-hop nodes to aggregate information from local domains, and simply deepening the network easily leads to over-smoothing issues. To improve the model for capturing

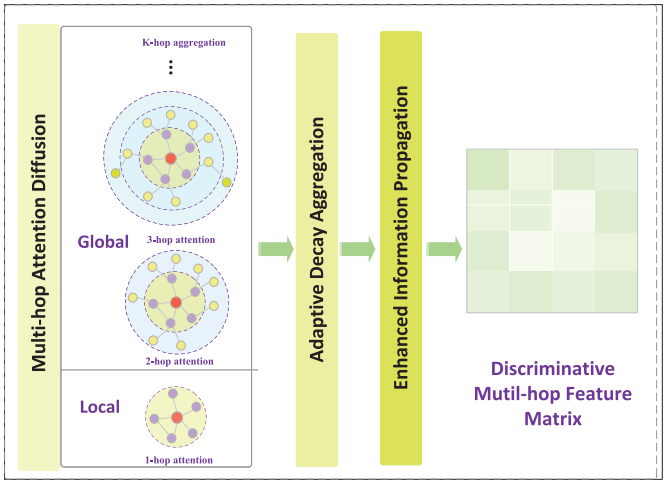

Fig. 3. The illustration of the main designs of EMAGCN to capture discriminative features fusing global and local information.

the long dependencies between distant nodes, we introduce a multi-hop approach to expand the graph model receptive field.

Specifically, following [22], we address some of the shortcomings of multi-hop networks with enhancements. The main design of our Enhanced Multi-hop Attention Graph Convolutional Network (EMAGCN) is shown in Fig. 3. Since previous work aggregating different distance information with fixed weights tends to introduce higher-order noise and degrade the performance, EMAGCN employs an adaptive decay strategy to adaptively learn useful information. We also randomly drop messages during message propagation to prevent over-fitting, and incorporate residual blocks in graph convolution to prevent over-smoothing. Finally, to address the problem of inefficient computation of dense matrices for higher-order graph models, we utilize a power iteration method to approximate the inverse matrix for acceleration.

1) Graph Construction: Formally, consider that an input face image of shape $H \times W \times 3$, we segment it into N equal size patches and each patch is transformed into feature embedding

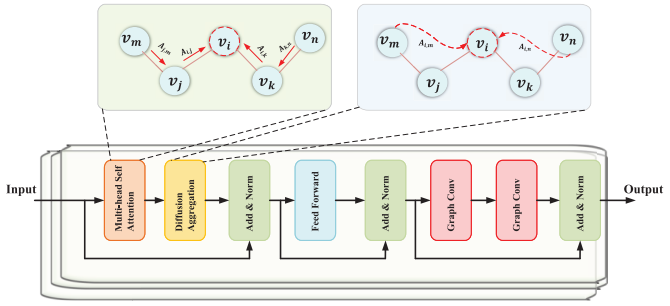

Fig. 4. The illustration of EMAGCN blocks. In the schematic above, the relations between multi-hops are obtained by attention diffusion.

$x_i \in \mathbb{R}^d$ utilizing the convolution-based patch embedding methods [16], where d denotes the feature dimensions. Regarding these patches representations $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ as nodes $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, we employ the K-Nearest algorithm [33] to calculate the neighbors $\mathcal{N}(v_i)$ of each node to build the edge. Then we obtain a graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{M})$, where $\mathcal{E}$ is the set of all edges between nodes, $\mathcal{A}$ denotes the adjacency matrix represented relation, $\mathcal{M}$ is the message matrix of the message-passing GNNs. The A_{ij} is initialized to 1 when nodes i to j have edges, otherwise to 0, while $D = \sum_j A_{ij} = \{d_1, \ldots, d_N\}$ denotes the degree matrix, d_i denotes the edge weight sum of node v_i. The $\mathcal{M} = \{m_1, \ldots, m_N\}$ are first initialized based on each node's own feature vector, and during each layer of message passing, updates the message matrix based on the features and edges of its neighboring nodes. Specifically, The message matrix propagate from node i to node j at l-th layer can be formulated as $M_{ij}^{(l)} = AGG_{j\in N(i)}(h_i^{(l)}, h_j^{(l)}, e_{ij})$, where h_i is the hidden representation of node v_i.

2) Multi-Hop Attention Diffusion: Similar to Graph Attention Network (GAT) [18], the first step is computing the attention values for all edges, the attention value of nodes i to j can be calculated as:

$$\alpha_{ij} = \sigma\left(a_{(l)}^T \tanh\left(W_i^{(l)} h_i^{(l)} \parallel W_j^{(l)} h_j^{(l)}\right)\right) \tag{1}$$

where $h_i^{(l)}$ denotes the node i embedding at l-th layer, and $h_i^{(0)} = x_i$. $a_{(l)}^T$, $W_i^{(l)}$ and $W_j^{(l)}$ are the learnable weights shared by l-th layer, $\sigma(\cdot)$ is the *LeakyReLU* activation function and $\parallel$ denotes the concatenation operation.

For all edges of l, we calculate 1-hop correlation by Eq. 1, then obtain an attention value matrix:

$$S_{ij} = \begin{cases} \alpha_{ij}, & if\ A_{ij} = 1 \\ -\infty, & otherwise \end{cases} \tag{2}$$

We further apply *softmax* operation on S_{ij} to acquire the attention matrix at l-th layer:

$$A_{ij}^{(l)} = sofmax\left(S_{ij}\right) \tag{3}$$

3) Adaptive Decay Aggregation: Subsequently, we utilize the graph diffusion method to compute the attention among nodes that lack a direct connection. As shown in the upper two schematics in Fig. 4, computing multi-hop attention allows creating attention shortcuts between nodes without explicit connection, utilizing the attention dependent on both their previous layer representation and the path relations between the nodes, thus effectively capturing long-distance interactions. However, as the distance increases, the correlation between nodes becomes weaker and weaker, which tends to introduce useless information or noise. To overcome this weakness, we employ an adaptive decay strategy that separates semantic correlations of the network from different hops and assigns adaptive learnable decay weights, the power of the K-hop attention matrix is formulated as:

$$\mathcal{A} = \sum_{k=0}^{K} \bar{A}^k \odot \delta_k \tag{4}$$

where $\bar{A}^k$ is the power of attention matrix, such as $\bar{A}_{ij}^k$ represents the relational path number from node i to j of maximum length k, which can increase the receptive domain of attention. The $\delta_k = \{\delta_1, \ldots, \delta_k\}$ denotes the adaptive learnable attention decay factor adjusted by a *Sigmoid* function $\delta_k = \frac{1}{1+e^{-\omega_k}}$, where ω_k is the weight parameter learned by the model during the training process. Since the attention decay factor is adaptive rather than fixed, the learned weights are different for different hops and paths with different levels of importance, thus providing more flexibility in learning useful information while suppressing noisy information. Then the attention diffusion $AD(\cdot)$ to updated graph representation can be defined as $AD(\cdot) = \mathcal{A}H^{(l)}$.

Furthermore, the attention diffusion equation for each individual head i is computed distinctly as follows:

$$\begin{aligned} \tilde{H}^{(l)} &= MSA\left(\hat{H}^{(l)}, \mathcal{G}\right) = \left(\|_{i=1}^{M} head_i\right) W_h \\ head_i &= AD\left(\mathcal{G}, \tilde{H}^{(l)}, \Theta_i\right), \ \hat{H}^{(l)} = LN\left(H^{(l)}\right) \end{aligned} \tag{5}$$

where $MSA(\cdot)$ is the multi-head self attention, $AD(\cdot)$ denotes the attention diffusion, $LN(\cdot)$ is the layer normalization for stabilizing the computation procedure, W_h is a parameter matrix and Θ_i is the parameter of the i-th head.

With attention diffusion, the receptive domain of the graph model is enlarged and global and local information are effectively captured, while the decay strategy reduces redundant information from distant nodes.

4) Enhanced Information Propagation: During the node information propagation process, we randomly drop some messages [34] for enhancement to maintain the diversity of topological information as well as to prevent the training of long-tailed samples from over-fitting. Unlike the previous drop methods, DropMessage directly drops the message matrix M. By dropping the message matrix with drop ratio ϱ, the process can be elaborated as:

$$\tilde{M}_{ij} = \frac{1}{1-\varrho} \epsilon_{ij} M_{ij} \tag{6}$$

where $\epsilon_{ij} \sim Bernoulli(1-\varrho)$ denotes an independent mask determining whether a reservation will be made. Thus node representation is updated based on node feature information and messages from multi-hop neighbors, which can be represented as:

$$h_i^{(l+1)} = h_i^{(l)} + \tilde{M}_i \tag{7}$$

Moreover, considering the significant increase in the computation of dense matrices generated by the multi-hop graph

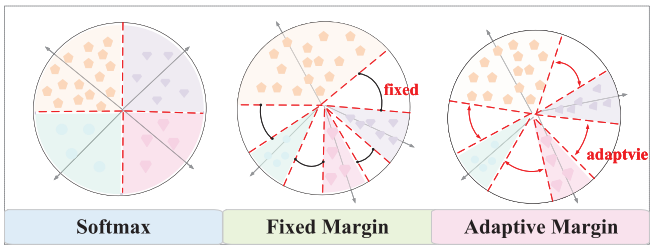

Fig. 5. The illustration of different margin losses. GroupFace employs the dynamic group-aware margin loss, which can adaptively provide suitable and unbiased margins for different age groups.

model, we introduce a power iteration method to estimate the inverse matrix, thereby attaining linear complexity accelerating training, and reducing computational overhead. The specific computational procedure can be calculated as:

$$\begin{aligned} H^{(n)} &= Q^T H^{(n-1)} + H^{(1)} \\ &= Q^n X + Q^{n-1} X + \ldots + Q^2 X + Q^T X \end{aligned} \tag{8}$$

where $Q = \tilde{A}\tilde{D}^{-1}$ is a row normalization instead of standard Laplacian smoothing normalization and n is the iterations. When n tends to infinity, the matrix hierarchy $(I - Q)^{-1}$ converges to $I + M + M^2 + \ldots + M^{n-1} + M^n$.

Finally, EMAGCN propagates all information into the face features through the process of graph convolution, as shown in Fig. 4, which contains a fully connected feed-forward sublayer, add layer normalization and residual connections to achieve a more expressive information propagation process:

$$\begin{aligned} \tilde{H}^{(l+1)} &= \tilde{H}^{(l)} + H^{(0)}, \\ H^{(l+1)} &= \sigma\left(LN\left(\tilde{H}^{(l+1)}\right) W_1^{(l)}\right) W_2^{(l)} + \tilde{H}^{(l+1)} \end{aligned} \tag{9}$$

where $\sigma(\cdot)$ denotes the *LeakyReLU* activation function, $LN(\cdot)$ is the layer normalization and $W_1^{(l)}, W_2^{(l)}$ are the different trainable weight.

B. Dynamic Group-Aware Margin Loss

Large margin loss functions based on softmax are often used to train feature extractors to make learned features more discriminative. A unified form is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{f(\theta_{y_i}, m)}}{e^{f(\theta_{y_i}, m)} + \sum_{j \neq y_i}^{n} e^{s\cos\theta_j}} \tag{10}$$

where y_i is the label index, m denotes the margin, and θ_{y_i} denotes the angle between the weight and feature vector of the j-th classifier.

However, most margins [35], [36] use fixed values, which are not flexible enough for real-life complex classification and have a large bias for category imbalance, so many studies [37], [38] have begun to explore adaptive margins.

Considering the poor generalization ability of long-tailed groups, we introduce the idea of dynamic adaptive into imbalanced age estimation to improve the large margin loss, which can be shown in Fig. 5. Our goal is to explore the adaptive margins between different classes, which can be guided by subsequent reinforcement learning, and the angular margins can be adjusted by the loss function automatically. By balancing the margins between different classes, it is realized that the majority class will not converge with too large a gap, while the minority class will converge to the majority class with a smaller gap. Following the [38], we design a dynamic group-aware margin loss to balance the additional margins between different classes, which can be formulated as:

$$\mathcal{L}_{DGM} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{a_i(t)\left(\theta_j - h_i(t)\right)^2 + k_i(t)}}{e^{a_i(t)\left(\theta_j - h_i(t)\right)^2 + k_i(t)} + \sum_{j \neq y_i}^{n} e^{s\cos\theta_j}} \tag{11}$$

where $a_i(t)$, $h_i(t)$ and $k_i(t)$ are the adaptive margin parameters associated with different groups, which can be determined in the reinforcement learning training stage.

By converting the cosine function to a quadratic function, model over-fitting can be avoided and it helps to reduce the computation overhead. Three learnable parameters make the margins more adaptive between different groups, thus enhancing the inter-class discrimination and intra-class compactness of age estimation features.

C. RL-Based Dynamic Group-Aware Margin Optimization

Inspired by [29] and [32], we conceptualize the problem of identifying suitable adaptive margins within the framework of a Markov Decision Process (MDP). With state s_t as input, the Q-value $Q(s_t, a)$ is estimated by the deep Q-learning network (DQN), where the agent will be trained to use different actions a_t to adapt the margins for each state. In turn, the environment will give the action a reward $r(s_t, a_t)$, and then update the state by s_{t+1}. The training objective of deep Q-learning is to seek the optimal function Q^*, which indicates that the agent can obtain the highest cumulative reward value by adjusting the margins of each iteration with policy π.

1) State: The dynamic group-aware margin is adaptively adjusted to the number of images in different groups, inter-class, and intra-class distance, which is defined by the triple $\{G,\ D_{inter},\ M\}$. We divide the age groups into four categories $G = \{0, 1, 2, 3\}$ where *Children* (group 0), *Teenager* (group 1), *Adult* (group 2) and *Senior* (group 3). M is equivalent to the dynamic group-aware margin. Since the number of the *Adult* age group in most age datasets is the largest, we keep the margin of *Adult* as the anchor and achieve imbalanced learning by adjusting the angular skewness of the long-tailed age groups with respect to the head class *Adult*. D_{inter} denotes the deviation of the inter-class distance between age group i of the long-tailed classes and the head class *Adult*, which can be expressed as:

$$\begin{aligned} D_{inter} &= |d^i_{inter} - d^2_{inter}| \\ d^i_{inter} &= \frac{1}{N_i}\sum_{i=1}^{N_i} \max_{k=1:N_i} \cos\left(x_k, x_i\right) \end{aligned} \tag{12}$$

where d^i_{inter} is the inter-class distance of i-th age group, d^2_{inter} denotes the inter-class distance of the head class *Adult*. Moreover, $\cos(\cdot,\cdot)$ represents the cosine distance function, N_i

is the i-th age group's number and x_i denotes the feature center of i-th age group.

Assuming that different age groups have different requirements for margins, these requirements may vary depending on their D_{inter}. When the bias is large, long-tailed groups may require larger margins to improve their generalization ability. In addition, to make the space of states discrete, we let D_{inter} and M identified to discrete space $\mathcal{D}$ and $\mathcal{M}$, where $\mathcal{D} = \{d_1, d_2, \ldots, d_{nD}\}$ and $\mathcal{M} = \{m_1, m_2, \ldots, m_{nM}\}$.

2) Action: Depending on the different state, there are three types of action spaces $\mathcal{A} = \{-1, O, +1\}$, where O means keep the same, -1 means shrink to a constant value κ and $+1$ means expand to a constant value κ.

We train the agent to make better decisions while taking action a obeying the cumulative reward $\pi^*(s, a) = arg\max_a Q(s, a)$. For instance, if at time step t the agent opts to execute action -1 based on the Q value and the state $s_t = \{3,\ d_1,\ m_1\}$, the margin of the Senior will be updated to $m_2 = m_1 - \kappa$.

3) Reward: The reward function $r(s_t, a_t)$ is set to motivate the agent to take a better action a_t at state s_t. We expect that the long-tailed classes have close generalization ability of the head class *Adult*, then the age estimation biases of the different groups are balanced, and so we utilize the deviation of the inter-class and intra-class distances to form the rewards. The deviation of intra-class distances between age group i of the long-tailed classes and the head class *Adult* can be expressed as:

$$\begin{aligned} D_{intra} &= |d^i_{intra} - d^2_{intra}| \\ d^i_{intra} &= \frac{1}{N_i}\sum_{i=1}^{N_i} \cos\left(x_k, x_i\right) \end{aligned} \tag{13}$$

where d^i_{intra} denotes the intra-class distance of i-th age group, d^2_{intra} denotes the intra-class distance of the head class *Adult*. N_i is the number of i-th age group, and x_i denotes the feature center of i-th age group. Combining the deviation of the inter-class and intra-class distances together, the reward to adjust the margin can be formulated as:

$$\begin{aligned} r\left(s_t, a_t\right) &= \mathcal{R}_{t+1} - \mathcal{R}_t \\ \mathcal{R} &= -\left(D_{intra} + D_{inter}\right) \end{aligned} \tag{14}$$

4) Deep Q-Learning: It is employed to seeks the optimal function Q^* to guide the agent acquiring the highest cumulative reward value. During the training process, we iteratively change the Q-function to update the model by minimizing the loss:

$$\begin{aligned} \mathcal{L}_{DQN} &= \mathbb{E}_{s_t, a_t}\|y_t - Q\left(s_t, a_t\right)\|^2 \\ y_t &= \mathbb{E}_{\pi}\left(r_t + \gamma \max Q\left(s_{t+1}, a_{t+1} | s_t, a_t\right)\right) \end{aligned} \tag{15}$$

where y_t is the target Q value, γ is the discount factor, $y_t - Q(s_t, a_t)$ denotes the deviation error at t-th step and a future reward $\gamma \max Q(s_{t+1}, a_t | s_t, a_t)$.

5) Training Network: The samples are used as inputs to train the DQN, and the agent dynamically adapts the margin through actions for different groups. After traversing all the states through continuous iterative updating, the optimal group-aware margin policy will be obtained.

Then the group-aware margin policy will guide and optimize our overall age estimation network, including robust extraction of face features and accurate classification of age groups. Our age estimation method starts with age classification at smaller intervals through the four age groups, and the classification results are further regressed to obtain predicted values.

In the multi-classification process, for the i-th element output z_i, softmax is employed for age grouping to g categories with $G_i = Soft\max(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{g} e^{z_j}}$. After the softmax operation, the predicted age with label y_i is calculated by the expectation $\hat{y}_i = E = \sum_{i=0}^{|g-1|} y_i.G_i$. The joint two-stage estimation loss consists of cross-entropy loss and average absolute loss, which is formulated as:

$$L_{GroupFace} = \lambda L_{CE} + (1 - \lambda) L_{MAE} \tag{16}$$

IV. Experiment

In this section, we present our experiments in detail, which will be organized around the following questions:

RQ1: *How effective and robust is our method under general age estimation?*

RQ2: *Can our method achieve balanced generalization performance in long-tailed age estimation?*

RQ3: *How do the key designs and hyperparameters in the architecture affect the performance?*

A. Datasets

1) Morph-Ii: The dataset [39] is a widely used age estimation benchmark comprising 55,134 facial images from 13,617 individuals, ages spanning from 16 to 77. It also provides demographic details like gender, race, and glasses usage. In our experiments, we used two protocols in our evaluation. **Setting I** [40]: The entire dataset is randomly segmented into two distinct sections, with 80% allocated for training and 20% reserved for testing. **Setting II** [41]: Based on a subset of 5493 facial images of European ethnicity, the subset was randomly segmented into two sections: 80% and 20% for training and testing.

2) UTK-Face: The dataset [42] is a large-scale, long age-span (from 0 to 116 years) face dataset. Collected in an unconstrained setting, it includes over 20,000 images labeled with gender, age, and ethnicity, capturing a broad spectrum of variations. In our work, we randomly select 80% and 20% for training and testing.

3) Chalearn LAP 2015: The dataset [43] is a competitive dataset released at the ChaLearn LAP Challenge 2015 and contains 4699 face images with annotations averaged from at least 10 users. It is divided into training (2476 images), validation (1136 images), and testing (1079 images) subsets.

4) CACD: The dataset [44] is a vast dataset from 2,000 celebrities with 163,446 images aged 14 to 62. Due to estimated age annotations, it has more noise. In our experiments, we cleansed the noise and utilized 1800 celebrities for training, and 80 and 120 for validation and testing.

TABLE I
THE SUMMARY OF WIDELY-USED AGE ESTIMATION DATASETS. SOME DETAILED INFORMATION IS FROM [45]

Dataset	#Images	#Subjects	Age range	Label type	In-the-wild	Class imbalance
MORPH II [39]	55,134	13,618	16-77	Real age	No	Yes
UTK-Face [42]	23,708	-	0-116	Apparent age	Yes	Yes
CLAP15 [43]	4,691	-	3-85	Apparent age	Yes	Yes
CACD [44]	163,336	2000	14-62	Apparent age	Yes	Yes
MIVIA [9]	575,073	-	1-81	Apparent age	Yes	Yes

5) MIVIA: The dataset [9], derived from the CAIP competition and sourced from VGGFace2, contains 575,073 images labeled with ages from 1 to 81 using "knowledge distillation." With the test set pending release, MIVIA was split into 458,752 and 114,688 images for training and validation.

B. Evaluation Metrics

1) MAE (Mean Absolute Error): It is defined as the average distance between the actual and predicted age, which is formulated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{17}$$

where y_i and $\hat{y}_i$ are the true and predicted age values of the i-th sample, respectively, and N is the amount of test images.

2) ϵ-Error (Normal Score): The age in the CLAP-2015 dataset is labeled as the mean of different people, and the true age in the data contains two attributes, mean and variance. Thus considering these factors can be a more accurate measure, which is calculated as:

$$\epsilon = 1 - \sum_{i=1}^{N} exp\left(-\frac{(y_i - \hat{y}_i)^2}{2\sigma_i^2}\right) \tag{18}$$

where y_i and $\hat{y}_i$ denote the true and predicted age values of the i-th sample, respectively, N is the amount of test images, and σ_i^2 is the labeled standard deviation. The smaller the ϵ-error, the more accurate the age estimate.

3) AAR (Age Accuracy and Regularity): To further evaluate the performance of the model under imbalanced or long-tailed distributions, we introduce the protocol of [9] and [10]. This metric can assume values between 0 and 10, weighted with 70% MAE (accuracy) and 30% σ (regularity) contributions, which can be formulated as:

$$\begin{aligned} AAR &= \max(0; 7 - MAE) + \max(0; 3 - \sigma), \\ \sigma &= \sqrt{\frac{\sum_{j=1}^{n}\left(MAE^j - MAE\right)^2}{n}}. \end{aligned} \tag{19}$$

where n is the number of age groups, MAE is the mean absolute error of the entire test set, σ denotes the standard deviation of different age groups, and MAE^j is the MAE computed for the sample whose actual age is in the j-th age group.

C. Implementation Details

We first utilize the face landmark algorithm MTCNN [46] to detect and align each face image, and then crop them to 224 * 224. During the training process, the samples are augmented with translations, color dithering, and random rotations. The starting learning rate was established at 0.0001 for the entirety of the experiments and attenuated using a cosine annealing strategy. We used the Adam optimizer [47] with parameters for weight decay and momentum configured to 0.0005 and 0.9, respectively. Each model underwent training on NVIDIA RTX 3090 GPUs using PyTorch.

D. Comparisons With The State-of-the-Art Methods (RQ1)

To validate the efficacy of our proposed GroupFace, detailed experiments are conducted on three face image datasets. For the characteristics of different datasets, we utilize appropriate experimental settings and evaluation criteria to compare them with state-of-the-art methods (SOTAs).

1) Comparisons on Morph II: On the most popular restricted datasets, TABLE II lists the avenue year, backbone network, and number of parameters for the SOTAs. Our method achieves the MAEs of 2.09 and 1.86 (with IMDB-WIKI dataset pre-trained weights) under Setting I, which outperforms PML [11], DCT [48], and MSL [3] that also consider imbalanced learning. It is inferior to GLAE [30] pre-trained on extra datasets using MS-CELEB-1M, TAA-GCN [49], MetaAge [50], and MWR [24] without using extra datasets but is able to achieve similar performance to the SOTAs with the IMDB-WIKI dataset weights. It is worth noting that compared to the commonly used age estimation pre-training dataset IMDB-WIKI, MS-CELEB-1M is far superior to it in terms of the number and quality of images. So a high-quality pre-training dataset and three excellent network designs enabled GLAE [30] to achieve a performance that far exceeded the best of others. Under Setting II, our method achieves the second-best MAE 2.01 with the IMDB-WIKI dataset weights, which approaches the GLAE. In addition, our network has 8.6 M parameters, which is smaller than these methods with the highest prediction accuracy.

2) Comparisons on UTK-Face: We test the performance of GroupFace on an unconstrained large-scale dataset spanning ages 0 to 116. TABLE III shows that our GroupFace achieves the second-best MAE of 4.32 and performs far better than previous methods with a smaller number of parameters of 8.6M. Note that MSL [3] uses a deeper ResNet-34, and MWR [24] uses a larger VGG-16, both of which effectively reduce the MAE. Andrey Savchenko [55] uses a more compact network MobileNet-v2, which has the lowest number of parameters but lacks in performance.

3) Comparisons on ChaLearn LAP 2015: We further compare our method with the SOTAs on the unrestricted

TABLE II

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON MORPH II. (* INDICATES USED THE IMDB-WIKI DATASET FOR PRE-TRAINING,† INDICATES USED THE MS-CELEB-1M DATASET FOR PRE-TRAINING, AND '↓' INDICATES THE SMALLER IS BETTER)

Method	Venue Year	Backbone Network	MAE↓		Param.↓
			Setting I	Setting II	
DEX [1]	IJCV 2016	VGG-16	-	3.15/2.68*	138M
Ranking-CNN [4]	CVPR 2017	Binary CNNs	2.96*	-	500M
DLDLF [51]	NIPS 2017	VGG-16	2.24	-	138M
MV [52]	CVPR 2018	VGG-16	2.79 /2.16*	-	138M
SSR-NET [53]	IJCAI 2018	SSR-NET	3.16*	-	40.9K
C3AE [23]	CVPR 2019	C3AE	2.78*	2.95*	**39.7K**
BridgeNet [54]	CVPR 2019	VGG-16	2.38*	2.35*	138M
PML [11]	CVPR 2021	ResNet-34	2.15	2.31	21M
MWR [24]	CVPR 2022	VGG-16	2.00*	2.13*	138M
MetaAge [50]	TIP 2022	VGG-16	1.81*	2.23*	138M
DAA [25]	CVPR 2023	ResNet-18	2.25/2.06*	-	11M
TAA-GCN [49]	PR 2023	TAA-GCN	1.69	-	-
DCT [48]	TIFS 2023	ResNet-50	2.28/2.17*	-	23M
MSL [3]	TIFS 2023	ResNet-34	2.10	2.03	21M
GLAE [30]	TIP 2023	ResNet-50	**1.14**†	**2.00**†	23M
GroupFace (Ours)	**-**	**EMAGCN**	2.09/1.86*	2.27/2.01*	8.6M

TABLE III

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON UTK-FACE. (* INDICATES USED THE IMDB-WIKI DATASET FOR PRE-TRAINING, AND '↓' INDICATES THE SMALLER IS BETTER)

Method	Backbone	MAE↓	Param.↓
Andrey Savchenko [55]	MobileNet-v2	5.44*	**3.4M**
CORAL [56]	ResNet-50	5.47*	25.6M
DCDL [57]	VGG-16	4.48*	138M
MWR [24]	VGG-16	4.37	138M
MSL [3]	ResNet-34	**4.31***	21M
GroupFace (Ours)	**EMAGCN**	4.32*	8.6M

TABLE IV

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON CHALEARN LAP 2015. (* INDICATES USED THE IMDB-WIKI DATASET FOR PRE-TRAINING, † INDICATES USED THE MS-CELEB-1M DATASET FOR PRE-TRAINING, AND '↓' INDICATES THE SMALLER IS BETTER)

Method	MAE↓	ϵ-error↓	Param.↓
DEX [1]	3.25*	0.282	138M
DHAA [58]	3.05*	0.265	100M
BridgeNet [54]	2.98*	0.267	138M
DLDL [59]	3.51	0.310	138M
AGEn [41]	3.21	0.280	138M
PML [11]	2.91*	0.243	21M
MWR [24]	2.95*	0.262	138M
DCT [48]	2.87†	0.242	11M
GLAE [30]	**2.852**†	0.242	11M
GroupFace (Ours)	2.91*	**0.239**	**8.6M**

competitive dataset CLAP 2015. TABLE IV shows our results, where we achieve the third best MAE of 2.91 after GLAE [30] and DCT [48], and the best ϵ-error of 0.239, while the parameter count of 8.6M is the smallest among these methods. GLAE designed Adaptive Routing (AR) to select suitable classifier for improving the long-tailed recognition while maintaining the head class. However, our method achieves similar performance with the help of the IMDB-WIKI dataset while GLAE with the help of MS-CELEB-1M, indicating that GroupFace is also effective in handling samples with large variance.

TABLE V

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON CACD. (* INDICATES USED THE IMDB-WIKI DATASET FOR PRE-TRAINING, † INDICATES USED THE MS-CELEB-1M DATASET FOR PRE-TRAINING, AND '↓' INDICATES THE SMALLER IS BETTER)

Method	MAE↓	Param.↓
DEX [1]	4.79*	138M
DHAA [58]	4.35*	100M
CR-MTk [60]	4.48*	67M
MWR [24]	4.41	138M
DCT [48]	4.09†	23M
TAA-GCN [49]	4.09	-
MSL [3]	4.126*	21M
GLAE [30]	4.09†	23M
GroupFace (Ours)	**4.07***	**8.6M**

4) Comparisons on CACD: We also compare our GroupFace with the SOTAs on the large dataset CACD, which originates from web crawling with a lot of noisy data and large variations in face background and illumination. As shown in TABLE V, our framework achieves an optimal performance of 4.07 MAE with IMDB-WIKI dataset pre-trained weights. Compared with GLAE [30], MSL [3] and DCT [48], the amount of parameters is nearly halved thanks to the flexibility of the graph model, although our model only reduces 0.02 MAE. Clearly, the results show that our GroupFace is capable of unconstrained age estimation and robust feature extraction in samples containing more noise.

TABLE VI

THE LONG-TAILED GENERALIZATION ANALYSIS ON THREE LONG-RANGE AND IMBALANCED AGE ESTIMATION DATASETS. ('↓' INDICATES THE SMALLER IS BETTER, WHILE '↑' INDICATES THE LARGER THE BETTER)

Method	MORPH II (Setting I)							MORPH II (Setting II)						
	Group MAE				Overall			Group MAE				Overall		
	MAE0	MAE1	MAE2	MAE3	MAE↓	$\sigma\downarrow$	AAR↑	MAE0	MAE1	MAE2	MAE3	MAE↓	$\sigma\downarrow$	AAR↑
Baseline	-	3.21	2.16	5.23	2.42	1.69	5.89	-	3.14	2.29	4.97	2.50	1.47	6.03
GroupFace (Ours)	**-**	**2.17**	**2.03**	**4.27**	**2.09**	**1.25**	**6.66**	**-**	**2.59**	**2.14**	**4.29**	**2.27**	**1.18**	**6.55**
Method	**UTK-Face**							**MIVIA**						
	Group MAE				Overall			Group MAE				Overall		
	MAE0	MAE1	MAE2	MAE3	MAE↓	$\sigma\downarrow$	AAR↑	MAE0	MAE1	MAE2	MAE3	MAE↓	$\sigma\downarrow$	AAR↑
Baseline	5.24	8.45	4.11	6.22	4.78	2.01	3.21	6.86	3.23	1.62	3.81	1.86	2.77	5.37
GroupFace (Ours)	**4.49**	**5.91**	**4.08**	**4.63**	**4.32**	**0.82**	**4.86**	**3.78**	**2.31**	**1.58**	**2.56**	**1.68**	**1.17**	**7.15**

Our GroupFace strikes a great improvement which probably owing to: i) GroupFace effectively integrates Enhanced Multi-hop Graph Convolutional Networks (EMAGCN) and reinforcement learning-based group-aware margin strategy, focusing on both discriminative feature extraction for different groups and imbalanced learning for the long-tailed class. ii) The graph model is more flexible in dealing with complex irregular objects that effectively model aging changes in the face, and reduces the learning of excessive redundant information thus making the model more compact.

E. Long-Tailed Generalization Analysis (RQ2)

To extra evaluate the generalization performance of our architecture GroupFace in different age groups, we consider evaluating the long-tailed age estimation on three imbalanced datasets. Our experiments categorize the samples into four groups: children (0-12), teenager (13-17), adult (18-65), and senior (66+). In most datasets, adult is treated as the head class, while children, teenager, and senior are treated as the tail classes.

Specifically, as shown in TABLE I, MORPH II with an age span of 16-77, lacks the children group and has a teenager group of about 20%, with the tail class concentrated in senior at about 1.7%. For CACD ranging 16-62 years, it missing the children and senior groups, so these datasets are not used in the long-tailed experiment. In addition, the Chalearn LAP 2015 dataset is also not used in that experiment considering that the data distribution imbalance is relatively insignificant and contains only 7591 images. In contrast, both UTK-Face and MIVIA are large-scale datasets with long age spans and significant long-tailed distributions, which are appropriate for evaluating the imbalanced learning performance of the models. And we perform the generalization analysis on the Morph II dataset without the help of external dateset, while the analysis on the UTK-Face and MIVIA datasets with the help of pre-training on the IMDB-WIKI dataset.

Following the long-tailed recognition [10], we apply both MAE and AAR to evaluate the performance of different age groups. TABLE VI shows the long-tailed recognition performance of GroupFace with three different datasets, where Baseline utilizes one-hop GCN and sofmax for classification without margin optimization. It is obvious that our GroupFace not only has a large improvement in the overall MAE but also has a significant improvement in the group MAE. Under two settings of MORPH II, the group MAE gap between our different classes narrows and σ decreases significantly, with the AAR reaching as high as 6.64. Meanwhile, our long-tailed recognition performance improvement is more obvious in UTK-Face and MIVIA, where the age span and imbalance are more severe. The lowest σ of 0.77 and the highest AAR of 7.54 are achieved in MIVIA. This is due to the effective integration of EMAGCN and dynamic group-aware margin optimization in GroupFace, which greatly improves the model's balanced generalization performance for long-tailed recognition.

TABLE VII

THE LONG-TAILED AGE ESTIMATION COMPARISONS ON MIVIA DATASET. ('↓'INDICATES THE SMALLER IS BETTER, WHILE '↑'INDICATES THE LARGER THE BETTER)

Method	GroupMAE			Overall		
	0-17	18-65	66-100	MAE↓	$\sigma\downarrow$	AAR↑
GLAE	3.41	1.68	**2.31**	1.73	**0.69**	**7.58**
GroupFace	**2.73**	**1.58**	2.56	**1.68**	1.71	7.15

Meanwhile, we compare our GroupFace on the MIVIA dataset using the same evaluation metrics with the imbalanced age estimation benchmark GLAE [30]. Because of the different age grouping strategies, we follow GLAE to combine children and teenager (0-18) in the comparison. As shown in the TABLE VII, our method underperforms GLAE in the comprehensive metric AAR, but achieves a lower overall MAE of 1.68. GLAE utilized the smaller ResNet-18 (11M) or the better ResNet-50 (23M) as the backbone network, and designed Adaptive Routing (AR) to select the appropriate classifiers, which achieved both general and imbalanced age estimation with excellent results. Our GroupFace utilizes a more flexible multi-hop graph model (8M) to extract local and global features, and designs dynamic Group-aware margin optimization for imbalanced learning, which also achieves impressive performance.

TABLE VIII

THE EFFECT OF DIFFERENT KEY COMPONENTS ON TREE IMBALANCED AGE ESTIMATION DATASETS. ('↓' INDICATES THE SMALLER IS BETTER, WHILE '↑' INDICATES THE LARGER THE BETTER)

Components		MORPH II (Setting I)			MORPH II (Setting II)			UTK-Face			MIVIA		
EMAGCN	DGMO	MAE	$\sigma\downarrow$	AAR↑	MAE↓	$\sigma\downarrow$	AAR↑	MAE↓	$\sigma\downarrow$	AAR↑	MAE↓	$\sigma\downarrow$	AAR↑
-	-	2.42	1.69	5.89	2.50	1.47	6.03	4.78	2.01	3.21	1.86	2.77	5.37
✓	-	2.18	1.57	6.25	2.29	1.32	6.39	4.61	2.14	3.25	1.72	1.95	6.33
-	✓	2.24	1.38	6.38	2.37	1.26	6.37	4.69	1.58	3.73	1.81	1.21	6.98
✓	✓	**2.09**	**1.25**	**6.66**	**2.27**	**1.18**	**6.55**	**4.32**	**0.82**	**4.86**	**1.68**	**1.17**	**7.15**

F. Ablation Studies (RQ3)

To validate the effectiveness and robustness of our architecture GroupFace, as well as to explore the performance enhancement of the architecture by different components, we perform a series of ablation studies across multiple datasets. And the evaluation of the Morph II dataset is conducted without the help of external dateset, while the evaluation of UTK-Face and MIVIA datasets is with the help of pre-training on the IMDB-WIKI dataset. First, We roughly divide the architecture into two main components: the Enhanced Multi-hop Attention GCN (EMAGCN) and the Dynamic Group-aware Margin Optimization (DGMO), and conduct initial ablation experiments on these two components. Then the different modules and strategies in these two components are further explored in more detail one by one.

1) Effect of Key Components: To verify the effectiveness of the two key components on the entire age estimation architecture, we set the baseline as a one-hop GCN and sofmax for classification without margin optimization, and then compare the gains after adding different components. As shown in TABLE VIII, the results show that the improvement of overall MAE using EMAGCN is obvious, while DGMO is able to significantly reduce σ and improve AAR. Especially in the case of long age-span imbalanced datasets UTK-Face and MIVIA, combining the two improves the AAR by 1.65 and 1.78, respectively. This manifests that EMAGCN can achieve more robust facial feature extraction and capture the discriminative features of the different groups. DGMO reduces the skewness of feature representations and achieves a balanced performance across different groups. Our GroupFace integrating both can attain joint enhancements in both general and long-tailed age estimation tasks.

2) Effect of Enhanced Multi-Hop Attention GCN: To achieve stronger learning of discriminative representations, we design an enhanced multi-hop graph convolutional network to model the aging changes of the face and also design Adaptive Decay strategy (AD), DropMessage (DM), and Residual connection (RC) to enhance the graph model. Before exploring the effects of these designs for deepening the graph model, we adjust the hop count K and the decay on the UTK-Face dataset to find the optimal initial hop count. As shown in Fig. 6, the MAE is significantly reduced when multi-hop neighbors ($K > 1$) are adopted. The adaptive decay model reaches the lowest MAE of 4.32 when $K = 4$, while the fixed decay model reaches the lowest MAE of 4.44 when $K = 3$. Besides, the

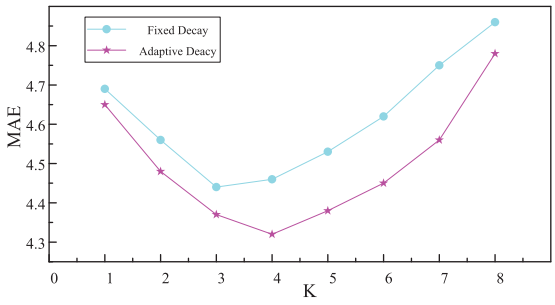

Fig. 6. The effect of multi-hop K and adaptive decay for EMAGCN on UTK-Face dataset.

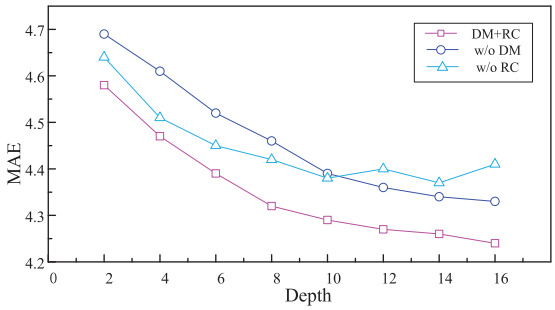

Fig. 7. The effect of the designed strategy for EMAGCN and the depth of EMAGCN on the UTK-Face dataset.

MAE of adaptive decay is far smaller than fixed decay in most multi-hops, which demonstrates that the adaptive decay strategy can flexibly regulate the degree of attention decay at different distances and improve the effectiveness of multi-hop attention diffusion by reducing noise from long-range distance. However, the performance gradually goes down as K increases after reaching the minimum. This may be due to the fact that when the number of hops is too large, aggregating information about neighbors that are too far away introduces more noise than useful information. Therefore, we set the multi-hop count K to 4 and utilize adaptive decay in our experiments.

Based on the optimal hop count $K = 4$, we continue to investigate how much the designed method and strategy improve the multi-hop attention GCN. The comparisons of DM+RC, w/o DM, and w/o RC at different network layers were obtained by removing some components separately. The results are shown in Fig. 7, DropMessage (DM) helps to keep the diversity of the message delivery while preventing over-fitting, and Residual connection (RC) can maintain initial information while preventing over-smoothing. The combination of the two is more beneficial to keep the performance of

TABLE IX
THE EFFECT OF DIFFERENT MARGIN LOSSES ON THE UTK-FACE DATASET. ('↓' INDICATES THE SMALLER IS BETTER, WHILE '↑' INDICATES THE LARGER THE BETTER)

Type	Group MAE				Overall		
	MAE0	MAE1	MAE2	MAE3	MAE↓	$\sigma\downarrow$	AAR↑
Sofmax	4.64	8.78	4.07	5.42	4.61	2.14	3.25
CosFace [35]	4.51	7.12	4.08	5.39	4.48	1.41	4.11
ArcFace [36]	4.48	7.09	4.07	5.41	4.47	1.40	4.12
ElasticFace-Cos [37]	**4.24**	6.43	4.08	5.37	4.39	1.14	4.47
ElasticFace-Arc [37]	4.26	6.38	4.06	5.32	4.37	1.12	4.51
GroupFace (Ours)	4.49	**5.81**	**4.05**	**4.93**	**4.32**	**0.82**	**4.86**

the deep network, while with a maximum gap of 0.14 or 0.17 MAE without DM or RC, respectively.

3) Effect of Margin Loss Function: To investigate the efficacy of our adaptive margin loss on imbalanced age estimation, we compare the experiments using different margin losses on the long-span UTK-Face dataset. As shown in TABLE IX, the large margin loss functions CosFace and ArcFace with fixed values can effectively lower the overall MAE compared to the ordinary sofmax. ElasticFace, which can flexibly change the margins, shows significant improvement in the overall performance as well as the performance of different groups. And benefit from the excellent adaptive margin loss function and the dynamic optimization of reinforcement learning, our GroupFace, which can recognize different groups very well, improves the overall recognition accuracy and also effectively achieves a balanced generalization performance. We achieve the lowest overall MAE of 4.32 and the highest AAR of 4.86.

4) Effect of RL-Based Group-Aware Margin Optimization: In dynamic group-aware margin strategy, the number of samples of different age groups is used as the classification of head or tail classes, and the head class Adult is regarded as the anchor point. The rest of the classes are regarded as long-tailed classes, which are categorized into Minor Class, Sub-minor Class, and Moderate Class according to the number of samples from smallest to largest. The above categorization varies according to the distribution of age groups in different datasets. In the UTK-Face dataset, the head class is Adult, and the long-tailed classes are Minor Class (Teenager), Sub-minor Class (Senior), and Moderate Class (Children) respectively. The margin of the head class Adult is kept constant after selecting the optimal margin, while the other classes are guided by reinforcement learning to dynamically adjust the margin to minimize the angular skewness between the long-tailed class and the head class. As shown in Fig. 8, a part of the strategy for different age groups of the trained agent is demonstrated. We can find that the long-tailed classes tend to increase the margin by the head class adult, where the Minor Class (Teenager) increases the largest margin, and the Sub-minor Class (Senior) and Moderate Class (Children) increase the similar margin. Having a larger inter-class deviation leads to an increase in the margin, since a larger intra-class distance usually reflects an imbalanced performance in recognizing the group, and thus an increased margin is essential to bolster the group's generalization capability. This demonstrates the adaptability and reliability of our strategy in identifying long-tailed groups, which can dynamically find appropriate margins for different age groups, reducing the skewness of feature representations between different groups and balancing intra-class proximity and inter-class separability.

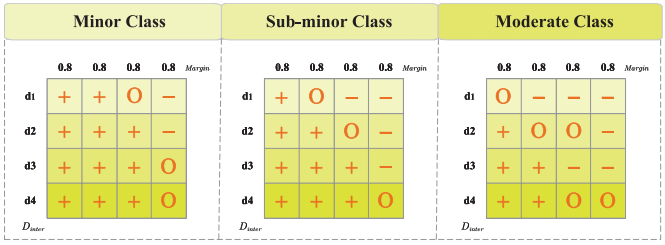

Fig. 8. The examples of RL-based group-aware margin strategy from trained agents on the UTK-Face dataset. For the state $s_t = \{G, D_{inter}, M\}$, each grid denotes an action $a_t = \{-1, O, +1\}$ to adjust group margins. D_{inter} is represented as $d_1 < d_2 < d_3 < d_4$, while M is corresponding to discrete spaces $\{0.2, 0.4, 0.6, 0.8\}$.

G. Qualitative Results

We select the samples of different age groups on the long-span, large-scale UTK-Face dataset for example demonstration of results. As shown in Fig. 9, our GroupFace performs well in all four age groups and achieves balanced generalization performance in long-tailed recognition. The blue numbers show that our method improves significantly in children, teenager and senior, which is attributed to the mining of discriminative facial features from different age groups, as well as the group-aware margin optimization. The red numbers show some of the failed samples, which may be caused by exaggerated facial expressions, heavy makeup, and so on.

H. Feature Visualization

We further compare the learned feature distributions from Baseline and our GroupFace on UTK-Face dataset by t-SNE visualization. For a fair comparison, Baseline uses a normal one-hop GCN and sofmax for classification without margin optimization. From Fig. 10(a), it can be seen that the features learned by the Baseline are more dispersed, and the margin overlaps, which is not sufficiently distinguishable for different age groups. While from Fig. 10(b), it can be observed that our GroupFace has a more compact feature distribution

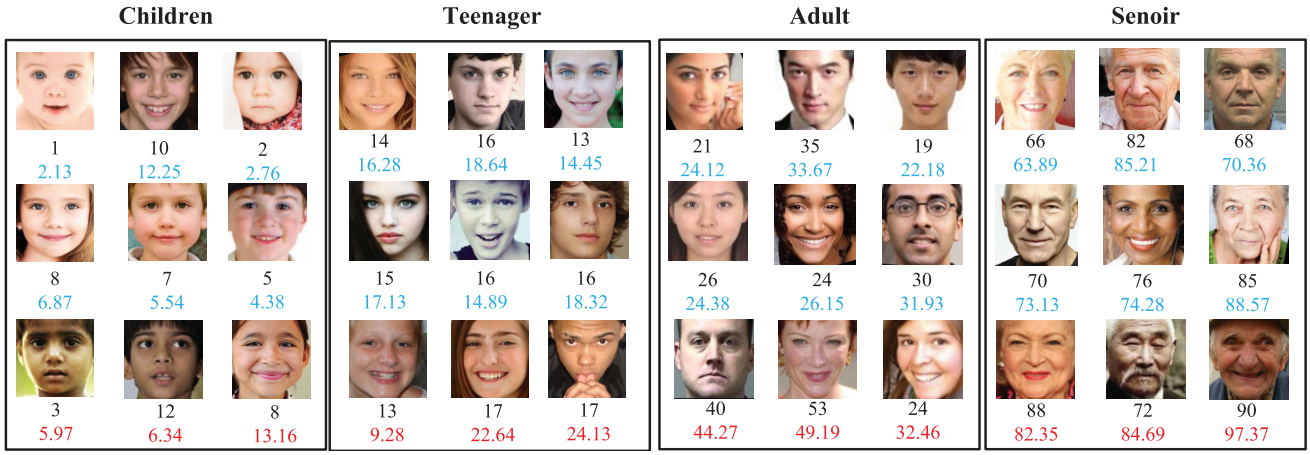

Fig. 9. The examples of age estimation results using our GroupFace on UTK-Face dataset. The ground truth label is the black number, the reliable estimation results are shown in the blue number, and the poor estimation results are shown in the red number.

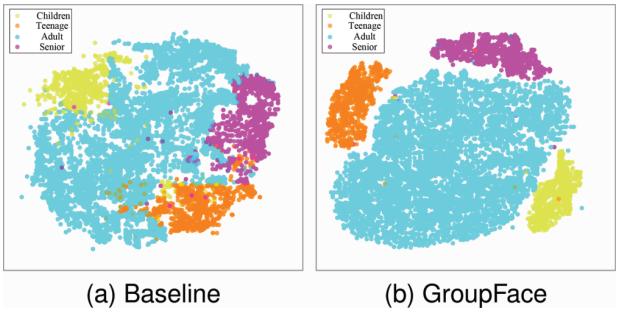

Fig. 10. The t-SNE visualization of the Baseline and our architecture GroupFace.

compared to Baseline, and the margins of different age groups are clearly distinguished. This indicates that GroupFace is effective in extracting the discriminative features of different age groups and providing appropriate and unbiased margins for different groups, balancing inter-class separability and intra-class proximity, which also demonstrates that GroupFace is very effective for imbalanced age estimation.

V. Conclusion

In this novel, we have presented an innovative collaborative learning framework GroupFace, which integrates an Enhanced Multi-hop Attention Graph Convolutional Network (EMAGCN) and a dynamic group-aware margin strategy based on reinforcement learning. The EMAGCN fuses local and global information to model aging changes in faces, which can achieve more discriminative representation learning. In addition, the dynamic group-aware margin strategy based on reinforcement learning provides appropriate and unbiased margins for different groups, which can balance inter-class separability and intra-class proximity. Extensive experiments have shown that our GroupFace not only provides a significant improvement in overall estimation accuracy but also balances performance in long-tailed groups. For future work, we are interested in further improving the effectiveness and generalizability of imbalanced learning with the help of the language-image pre-training method.

Acknowledgment

The author's deepest gratitude goes to the anonymous reviewers and AE for their careful work and thoughtful suggestions that have helped improve this article substantially.

References

[1] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep EXpectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 10–15.
[2] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul.2021.
[3] C. Wang, Z. Li, X. Mo, X. Tang, and H. Liu, "Exploiting unfairness with meta-set learning for chronological age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 5678–5690, 2023.
[4] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5183–5192.
[5] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[6] M. Kuprashevich and I. Tolstykh, "MiVOLO: Multi-input transformer for age and gender estimation," in *Proc. Int. Conf. Anal. Images, Social Netw. Texts*. Cham, Switzerland: Springer, Jan. 2024, pp. 212–226.
[7] L. Qin et al., "SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2223–2234, Apr.2024.
[8] Y. Shou, X. Cao, H. Liu, and D. Meng, "Masked contrastive graph representation learning for age estimation," *Pattern Recognit.*, vol. 158, Feb.2025, Art. no. 110974.
[9] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "Effective training of convolutional neural networks for age estimation based on knowledge distillation," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21449–21464, Dec.2022.
[10] Z. Bao et al., "LAE: Long-tailed age estimation," in *Proc. 19th Int. Conf. Comput. Anal. Images Patterns*. Cham, Switzerland: Springer, 2021, pp. 308–316.
[11] Z. Deng, H. Liu, Y. Wang, C. Wang, Z. Yu, and X. Sun, "PML: Progressive margin loss for long-tailed age classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10498–10507.
[12] W. Cai, X. Dong, and H. Liu, "Meta descent learning for class imbalanced age estimation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.
[13] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 128–145, Apr.2021.
[14] S. Yan, C. Li, H. Wang, B. Lin, and Y. Yuan, "Feature interactive graph neural network for KG-based recommendation," *Expert Syst. Appl.*, vol. 237, Mar.2024, Art. no. 121411.

[15] F. Li et al., "Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution," *ACM Trans. Knowl. Discovery Data*, vol. 17, no. 1, pp. 1–21, Feb.2023.
[16] Y. Zhang, Y. Shou, T. Meng, W. Ai, and K. Li, "A multi-view mask contrastive learning graph convolutional neural network for age estimation," *Knowl. Inf. Syst.*, vol. 66, no. 11, pp. 7137–7162, Nov.2024.
[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
[19] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 1–11.
[20] F. Wu, A. H. Souza, T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6861–6871.
[21] S. Abu-El-Haija et al., "MixHop: Higher-order graph convolutional architectures via sparsified neighborhood mixing," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 21–29.
[22] G. Wang, R. Ying, J. Huang, and J. Leskovec, "Multi-hop attention graph neural network," 2020, *arXiv:2009.14332*.
[23] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3AE: Exploring the limits of compact model for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12587–12596.
[24] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Moving window regression: A novel approach to ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18760–18769.
[25] P. Chen et al., "DAA: A delta age AdaIN operation for age estimation via binary code transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15836–15845.
[26] H. Xiong and A. Yao, "Deep imbalanced regression via hierarchical classification adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 23721–23730.
[27] J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo, "Nested collaborative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6939–6948.
[28] J. S. Baik, I. Y. Yoon, and J. W. Choi, "DBN-mix: Training dual branch network using bilateral mixup augmentation for long-tailed visual recognition," *Pattern Recognit.*, vol. 147, Mar.2024, Art. no. 110107.
[29] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9319–9328.
[30] Z. Bao, Z. Tan, J. Li, J. Wan, X. Ma, and Z. Lei, "General vs. long-tailed age estimation: An approach to kill two birds with one stone," *IEEE Trans. Image Process.*, vol. 32, pp. 6155–6167, 2023.
[31] E. Lin, Q. Chen, and X. Qi, "Deep reinforcement learning for imbalanced classification," *Appl. Intell.*, vol. 50, no. 8, pp. 2488–2502, Aug.2020.
[32] B. Liu et al., "Fair loss: Margin-aware reinforcement learning for deep face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10051–10060.
[33] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision GNN: An image is worth graph of nodes," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2022, pp. 8291–8303.
[34] T. Fang, Z. Xiao, C. Wang, J. Xu, X. Yang, and Y. Yang, "DropMessage: Unifying random dropping for graph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 4, pp. 4267–4275.
[35] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.
[36] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[37] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 1578–1587.
[38] J. Xu et al., "X2-softmax: Margin adaptive loss function for face recognition," *Expert Syst. Appl.*, vol. 249, Sep.2024, Art. no. 123791.
[39] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, Apr. 2006, pp. 341–345.
[40] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 712–718.
[41] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, and S. Z. Li, "Efficient group-n encoding and decoding for facial age estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2610–2623, Nov.2018.
[42] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5810–5818.
[43] S. Escalera et al., "ChaLearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 243–251.
[44] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, Jun.2015.
[45] O. Agbo-Ajala and S. Viriri, "Deep learning approach for facial age classification: A survey of the state-of-the-art," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 179–213, Jan.2021.
[46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct.2016.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[48] Z. Bao, Z. Tan, J. Wan, X. Ma, G. Guo, and Z. Lei, "Divergence-driven consistency training for semi-supervised facial age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 221–232, 2023.
[49] M. Korban, P. Youngs, and S. T. Acton, "TAA-GCN: A temporally aware adaptive graph convolutional network for age estimation," *Pattern Recognit.*, vol. 134, Feb.2023, Art. no. 109066.
[50] W. Li, J. Lu, A. Wuerkaixi, J. Feng, and J. Zhou, "MetaAge: Meta-learning personalized age estimators," *IEEE Trans. Image Process.*, vol. 31, pp. 4761–4775, 2022.
[51] W. Shen, K. Zhao, Y. Guo, and A. Yuille, "Label distribution learning forests," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Feb. 2017, pp. 834–843.
[52] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5285–5294.
[53] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "SSR-Net: A compact soft stagewise regression network for age estimation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1078–1084.
[54] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "BridgeNet: A continuity-aware probabilistic network for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1145–1154.
[55] A. V. Savchenko, "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet," *PeerJ Comput. Sci.*, vol. 5, p. e197, Jun.2019.
[56] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec.2020.
[57] H. Sun, H. Pan, H. Han, and S. Shan, "Deep conditional distribution learning for age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4679–4690, 2021.
[58] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Deeply-learned hybrid representations for facial age estimation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3548–3554.
[59] 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.
[60] N. Liu, F. Zhang, and F. Duan, "Facial age estimation using a multi-task network combining classification and regression," *IEEE Access*, vol. 8, pp. 92441–92451, 2020.

Yiping Zhang is currently pursuing the Graduate degree with the College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China. His research interests include computer vision.

Yuntao Shou (Student Member, IEEE) is currently pursuing the Graduate degree with the College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China. His research interests include multimodal data mining.

Wei Ai received the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. She is currently an Assistant Professor with the Central South University of Forestry and Technology. Her research interests include data mining, big data, cloud computing, and parallel computing.

Tao Meng received the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. He is currently an Assistant Professor with the Central South University of Forestry and Technology. His research interests include data mining, network analysis, and deep learning.

Keqin Li (Fellow, IEEE) received the B.S. degree in computer science from Tsinghua University in 1985 and the Ph.D. degree in computer science from the University of Houston in 1990. He is a SUNY Distinguished Professor with the State University of New York; and the National Distinguished Professor with Hunan University, China. He has authored or co-authored more than 990 journal articles, book chapters, and refereed conference papers. He holds nearly 75 patents announced or authorized by the Chinese National Intellectual Property Administration. He is a member of the SUNY Distinguished Academy. He is also an AAAS Fellow, an AAIA Fellow, and an ACIS Founding Fellow. He is an Academician Member of the International Artificial Intelligence Industry Alliance. He is a member of Academia Europaea (Academician of the Academy of Europe). He received several Best Paper Awards from international conferences, including PDPTA-1996, NAECON-1997, IPDPS-2000, ISPA-2016, NPC-2019, ISPA-2019, and CPSCom-2022. He was a recipient of the 2017 Albert Nelson Marquis Lifetime Achievement Award for being listed in Marquis Who's Who in Science and Engineering, Who's Who in America, Who's Who in the World, and Who's Who in American Education for over 20 consecutive years. He also received the Distinguished Alumnus Award from the Computer Science Department, University of Houston, in 2018; the IEEE TCCLD Research Impact Award from the IEEE CS Technical Committee on Cloud Computing in 2022; the IEEE TCSVC Research Innovation Award from the IEEE CS Technical Community on Services Computing in 2023; and the IEEE Region 1 Technological Innovation Award (Academic) in 2023. He is among the world's top five most influential scientists in parallel and distributed computing in terms of single-year and career-long impacts based on a composite indicator of the Scopus citation database.