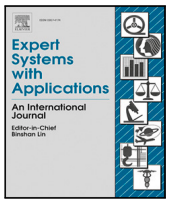
Expert Systems With Applications 273 (2025) 126819

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



LRA-GNN: Latent Relation-Aware Graph Neural Network with initial and Dynamic Residual for facial age estimation ☆

Yiping Zhang [a], Yuntao Shou [a], Wei Ai [a], Tao Meng [a,*], Keqin Li [b]

[a] College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, Hunan 410004, China
[b] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Keywords:
Age estimation
Latent relation
Deep residual graph convolution
Reinforcement learning

ABSTRACT

Face information is mainly concentrated among facial key points, and frontier research has begun to use graph neural networks to segment faces into patches as nodes to model complex face representations. However, these methods construct node-to-node relations based on similarity thresholds, so there is a problem that some latent relations are missing. These latent relations are crucial for deep semantic representation of face aging. In this novel, we propose a new Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (LRA-GNN) to achieve robust and comprehensive facial representation. Specifically, we first construct an initial graph utilizing facial key points as prior knowledge, and then a random walk strategy is employed to the initial graph for obtaining the global structure, both of which together guide the subsequent effective exploration and comprehensive representation. Then LRA-GNN leverages the multi-attention mechanism to capture the latent relations and generates a set of fully connected graphs containing rich facial information and complete structure based on the aforementioned guidance. To avoid over-smoothing issues for deep feature extraction on the fully connected graphs, the deep residual graph convolutional networks are carefully designed, which fuse adaptive initial residuals and dynamic developmental residuals to ensure the consistency and diversity of information. Finally, to improve the estimation accuracy and generalization ability, progressive reinforcement learning is proposed to optimize the ensemble classification regressor. Our proposed framework surpasses the state-of-the-art baselines on several age estimation benchmarks, demonstrating its strength and effectiveness.

1. Introduction

With the continuous development of deep learning, the field of face recognition shows an increasingly prosperous trend. As one of the important attributes of the face, age is also an important topic in current face research. The potential value of age has been gradually discovered and applied in numerous fields, such as human–computer interaction (Yang, Huang, Lin, Hsiu, & Chuang, 2018), social media (Duan, Li, & Li, 2017), and video surveillance (Rothe, Timofte, & Van Gool, 2015).

However, age estimation tasks continue to be challenging due to the complicated internal and external factors (Agbo-Ajala & Viriri, 2021; Zhang, Liu, Yuan et al., 2019). Early age estimation methods were mainly based on manual feature extractors and machine learning algorithms, which were difficult to deal with the increase in image complexity and data size. With the increase in computational power, image feature extraction based on Convolutional Neural Network (CNN) has shown outstanding performance. Rothe et al. (2015) utilized a pre-trained VGG-16 network for facial representation and obtained an age regression result through classification probabilities multiplied by the corresponding labels. Zhang, Liu, Yuan et al. (2019) integrated LSTM units with the residual networks (ResNets) to extract local age-sensitive features and used the Deep EXpectation algorithm (DEX) for age regression. Moreover, with the success of Transformer, many efforts have been based on the Vision Transformer (ViT) (Dosovitskiy et al., 2020) for long-range dependency modeling. Kuprashevich and Maksim (2023) proposed Multi Input VOLO (MiVOLO) utilizing the newest vision transformer for age and gender estimation in the wild. Qin et al. (2023) proposed SwinFace based on ViT for multi-task facial feature extraction including age estimation.

Despite the CNN and Transformer methods have outstanding capabilities in image processing, they cannot be applied to data in non-Euclidean spaces. In particular, facial attributes are mainly defined

☆ Our code is publicly available at https://github.com/Bottle010/LRA-GNN.

* Corresponding author.
E-mail addresses: yipingzhang@csuft.edu.cn (Y. Zhang), shouyuntao@stu.xjtu.edu.cn (Y. Shou), weiai@csuft.edu.cn (W. Ai), mengtao@hnu.edu.cn (T. Meng), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.eswa.2025.126819
Received 13 August 2024; Received in revised form 26 December 2024; Accepted 7 February 2025
Available online 15 February 2025

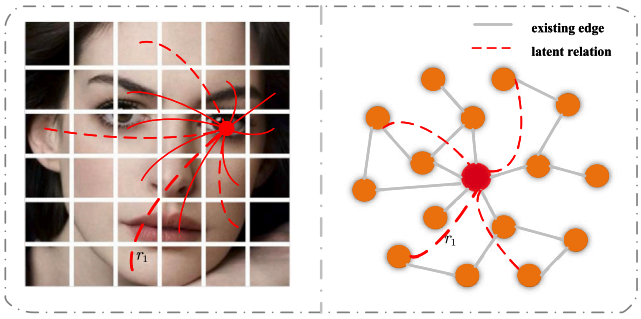


Fig. 1. The illustration of the latent relation (e.g., the dashed part). Existing GNN-based methods utilize the similarity threshold method and may ignore relations between facial key points and wrinkles such as r_1, which are not similar enough but important to attain more accurate estimation.

around the specific facial key points, so these modeling approaches are inflexible for complex and irregular human faces. Therefore, remarkable works have utilized Graph Neural Networks (GNN) to model facial structural information by directly learning potential embedded nodes based on their neighbors and relationships. Korban, Youngs, and Acton (2023) proposed a Time-Aware Adaptive Graph Convolutional Network (TAA-GCN) to compute the temporal age dependence using Temporal Memory Module (TMM). Shou, Cao, Liu, and Meng (2025) proposed a Masked Contrastive Graph Representation Learning (MCGRL) to capture the rich structural information of the face, which outperforms the CNN and Transformer based methods. Ge, Jose, Xu, Liu, and Han (2024) proposed facial action units detection network MGRR-Net with multi-level graph feature learning and relational reasoning.

All of the above methods have achieved excellent performance, but they still have some deficiencies. (1) **Insufficient in capturing latent relations.** GNN-based methods consider explicit modeling relationships based on similarity thresholding, but they cannot capture the latent relations of facial key points. The latent relations are schematically shown in Fig. 1, which are indispensable for the deep semantic representation of face aging. Recent work has begun to focus on this point, Jiang et al. (2023) utilizes the dilated K-nearest neighbors algorithm to learn the latent connections between face graph nodes. However, this approach is limited and incomplete for capturing latent relations. (2) **Inadequate to co-optimization of multi-stage age estimation.** Existing age estimation often utilizes hybrid algorithms, but the optimization of different stages is learned independently. Duan et al. (2017) introduced an extreme learning machine (ELM) to optimize age estimation, but the ELM classifier and ELM regressor were trained separately, which is easy to fall into the suboptimal problem.

To settle the two problems above together, we propose a new Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (LRA-GNN) for robust and comprehensive facial representation, as well as design the progressive reinforcement learning to synergistically optimize multi-stage age estimation. Specifically, face images are segmented into patches of the same size as graph nodes, and we utilize facial key points as prior knowledge to construct an initial graph. Then we employ a random walk strategy on the initial graph to obtain the global structure, which can collect more information with fewer paths in the search space. Under the guidance of the aforementioned strategies, LRA-GNN leverages the multi-attention mechanism to capture the latent relations and generates a set of fully connected graphs containing rich facial information and complete structure. Unlike the commonly used similarity computing (Jang et al., 2024; Korban et al., 2023) and K-nearest neighbor algorithms (Jiang et al., 2023; Shou et al., 2025), our mechanism can cover all relations without omission. In addition, to avoid over-smoothing issues for deep feature extraction on the fully connected graphs, we design the deep residual graph convolution networks to convolve the fully connected graph. These networks fuse adaptive initial residuals and dynamic developmental residuals to ensure the consistency and diversity of information, which are powerful for deep representation learning.

After fed the robust and comprehensive facial features into age estimation, to improve the accuracy and generalization of our model, we propose the progressive reinforcement learning to synergistically optimize the subsequent age group classification and final regression. We define age estimation by classification and then regression as a walk-to-the-end problem on a grid, which emphasize the continuity and correlation of ages unlike common categorical regression. In addition, to bootstrap agents to be generalizable for age estimation, we carefully design reward functions which takes into account the age distribution of different age groups for more robust age grouping and more accurate regression prediction. When the agent possesses the optimum behavior to maximize the cumulative reward through interaction with the environment, it can perform step-wise age estimation by classification and then regress the samples as correctly as possible.

1.1. Our contributions

Extensive experiments have demonstrated that the model we proposed is flexible, effective, and comprehensive. Our contributions to this paper are summarized as follows:

- We propose a new Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (LRA-GNN) for obtaining robust and comprehensive facial representations.
- To effectively explore the semantic and structural information of faces, we utilize the face key points as prior knowledge to construct an initial graph and apply a random walk strategy to the initial graph to obtain global structure.
- We carefully design a multi-attention mechanism to capture latent relations and a deep residual graph convolution network that fuses adaptive initial residuals and dynamic developmental residuals to ensure the consistency and diversity of information.
- To improve the generalization ability of our architecture, we propose progressive reinforcement learning to synergistically optimize the age group classification and final regression. The robust reward function and loss function are designed together to guide our age estimation.

2. Related work

In this section, we first introduce the frontier research of graph neural networks. Then we review some related works in deep learning-based age estimation. Finally, we briefly describe the reinforcement learning.

2.1. Graph neural networks

Graph Neural Networks (GNNs) have achieved impressive results in the field of processing non-Euclidean data and become a popular research. They have been widely applied in several tasks such as recommendation systems (Cui, Yu, Guo, Cao, & Wang, 2024), image retrieval (Qin, Li, Pang, & Hao, 2024), action recognition (Qiu & Hou, 2024), and multi-modal emotion recognition (Meng, Shou, Ai, Yin, & Li, 2024).

There are several variants developed from GNNs, such as Graph Convolutional Neural Networks (GCNs) (Kipf & Welling, 2016), Graph Attention Networks (GATs) (Veličković et al., 2017), GraphSAGE (Hamilton, Ying, & Leskovec, 2017) and so on. For the most commonly used GCNs, existing work mainly focuses on two streams: spectral-based (Henaff, Bruna, & LeCun, 2015) and spatial-based (Atwood & Towsley, 2016; Niepert, Ahmed, & Kutzkov, 2016). However, they utilized the shallow networks that limit their representation capabilities, while deep GCNs are prone to over-smoothing which results in a rapid

decrease in node differentiation. To get over this weakness, the mainstream attempts are two-fold: Li, Muller, Thabet and Ghanem (2019) proposed DeepGCN similar to DeepCNN utilizing dilated convolutions, dense or residual connections to improve the expressive power. Abu-El-Haija et al. (2019) proposed MixHop which involved discerning neighborhood connections at different distances by iteratively mixing the feature embedding.

Following the former, we integrate a multi-head attention mechanism with a deep residual graph convolutional network fusing adaptive initial residuals and dynamic developmental residuals to obtain robust and comprehensive facial representation.

2.2. Age estimation

Age estimation stands as a crucial and difficult task within the realm of computer vision. Current age estimation methods usually work on designing more robust face representation networks (Kuprashevich & Maksim, 2023; Shou et al., 2025; Zhang, Shou, Ai, Meng and Li, 2024) or more efficient age estimation techniques (Chen et al., 2023; Shin, Lee, & Kim, 2022; Wang, Li, Mo, Tang, & Liu, 2023).

Shin et al. (2022) proposed moving window regression (MWR), which formed a search window through two reference instances and then estimated the rho-rank. Chen et al. (2023) proposed the Delta Age AdaIN (DAA), utilizing binary code mapping and age encoder–decoder. Kuprashevich and Maksim (2023) proposed Multi Input VOLO (MiVOLO) utilizing the newest vision transformer for age and gender estimation in the wild. Wang et al. (2023) utilized meta-learning paradigm to built an unfair filtering network that reduce category bias in age estimation. Zhang, Shou, Ai et al. (2024) proposed GroupFace, integrating a multi-hop attention GCN with a group-aware margin strategy, which is effective in imbalanced age estimation. Shou et al. (2025) proposed a Masked Contrastive Graph Representation Learning (MCGRL) to capture the rich structural information of the face, which outperforms the CNN and Transformer based methods. However, these methods is sufficient in capturing latent relations.

In this novel, we design the multi-head attention mechanism with a deep GCN to capturing latent relations effectively and efficiently.

2.3. Reinforcement learning

Reinforcement learning (RL) has become a new paradigm in artificial intelligence technology, which has also gained high speed in recent years. RL has shown a broad application prospect in the fields of finance, gaming, automation, robotics, and so on. The RL aims to train intelligence to make autonomous decisions by maximizing future cumulative rewards, which can effectively optimize image classification and regression. Lin, Chen, and Qi (2020) considered imbalanced data classification as a Markov decision process, where samples are states, classification is actions, and rewards are based on the match between predicted and true values. Wen and Wu (2021) modeled the decision tree generation as a Markov decision process and achieved significant results by guiding tree construction through reinforcement learning. Yang et al. (2023) further optimized the original reinforcement learning algorithm to avoid overestimation of values and improve the training effectiveness.

In this novel, the Double Deep Q Network (DDQN) is introduced to reinforcement learning for achieving better classification and prediction results, which identifies high-level features using a reward function that distinguishes between different classes, i.e., punish minorities more harshly or reward them more generously.

3. Methodology

In this section, we are set to furnish an exhaustive delineation of our proposed approach the Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (LRA-GNN) for feature extraction and the Progressive Reinforcement Learning-based Age Estimation. The overall pipeline is shown in Fig. 2.

3.1. Overview of our framework

The main design of our framework is organized around the following questions:

Q1: *What should pay attention to get robust and comprehensive facial representation?*

Q2: *How to capture latent relations effectively and efficiently?*

Q3: *How to extract deep features based on the obtained fully connected graphs?*

Q4: *What are the strategies to synergistically optimize age estimation for higher accuracy and better generalization?*

In feature extraction, the face image is first segmented into equal-sized patches as graph nodes. **Q1:** As shown in Algorithm 1, to get robust and comprehensive facial representation, we propose a new Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (LRA-GNN) for capturing the latent relations. **Q2:** Specifically, we first construct an initial graph utilizing facial key points as prior knowledge, and then a random walk strategy is employed to the initial graph for obtaining the global structure, both of which together guide the subsequent effective exploration and comprehensive representation. Then LRA-GNN leverages the multi-attention mechanism to capture the latent relations and generates a set of fully connected graphs containing rich facial information and complete structure based on the aforementioned guidance. **Q3:** In addition, to avoid over-smoothing issues for deep feature extraction on the fully connected graphs, we design the deep residual graph convolution networks to convolve the fully connected graph. These networks fuse adaptive initial residuals and dynamic developmental residuals to ensure the consistency and diversity of information, which are powerful for deep representation learning.

After feeding the comprehensive and robust facial features into age estimation, **Q4:** to improve the generalization and accuracy of our framework, we propose progressive reinforcement learning to optimize the subsequent age group classification and final regression synergistically. We define age estimation by classification and then regression as a walk-to-the-end problem on a grid. With a well-designed reward function, a trained agent can perform step-wise age estimation as correctly as possible.

3.2. Initial graph construction

3.2.1. Graph construction

Face attributes are mainly concentrated at facial key points, and we first employ a facial landmark algorithm to select the most information-rich representations. Existing face landmark detection algorithms can easily give the $2D$ coordinates of the facial key points, and we utilize the algorithm (Korban et al., 2023) to select the key points that are susceptible to age changes and insensitive to facial expressions. Each face image has n keypoints as $U_i = \{u_1, u_2, \dots, u_n\}$, each keypoint contains coordinate information $u_i = (x, y)$.

Then the original face image can be partitioned into multiple same-size patches, where the similar pixel blocks around the keypoints will be stitched in the same patch as much as possible based on the coordinate information. Formally, assume that the input image I has shape $H \times W \times 3$, which is segmented to N patches corresponding to the key points, and the patches are embedded into M feature dimensions forming a feature vector. Each latent vector is treated as a node $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$, yielding us a graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where E is the edges set, A denotes the adjacency matrix with initializing to a 0–1 matrix by the similarity computing method. Moreover, let $\mathcal{X} \in \mathbb{R}^{N \times M}$ be the node features and $\mathcal{E} \in \mathbb{R}^{N \times M}$ be the edge features, where x_i denotes the node embedding and e_{ij} denotes the edge embedding.

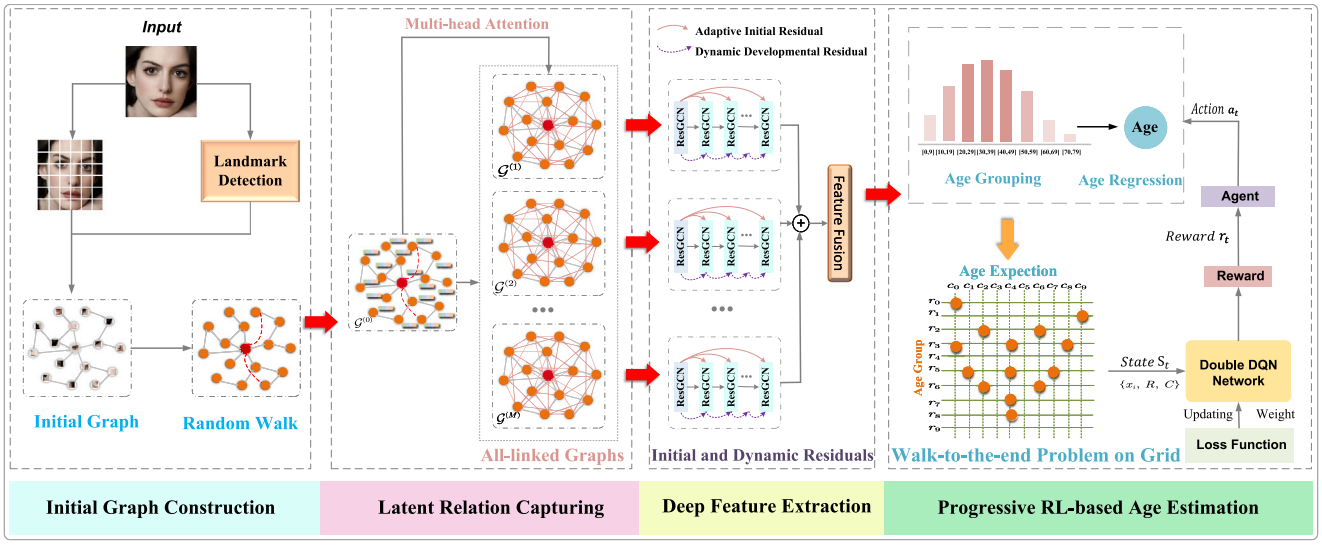


Fig. 2. The Overall Pipeline of our Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (**LRA-GNN**). First, utilizing facial key points as prior knowledge, face images are segmented into patches as graph nodes to construct an initial graph. Then LRA-GNN leverages the multi-attention mechanism to capture the latent relations and generates a set of fully connected graphs. Finally, this set of fully connected graphs is fed into deep residual graph convolutional networks for feature extraction and through progressive reinforcement learning to achieve robust and accurate age estimation.

3.2.2. Random walk updating

The key to graph representation learning is the aggregation and propagation of domain node information, which is closely related to paths. Therefore, to initially obtain the global structure, the random walk strategy (Grover & Leskovec, 2016; Zhou, Wang, & Zhang, 2024) is employed in the initial graph to collect more information with fewer paths in the search space.

Given a random walk that uniformly samples a random vertex v_i, which has visited edge e_{ik} to reach node v_k, assume that the next walk will visit v_j, and the non-normalized transfer probability formula for the walk is:

$$\mathcal{P}\left(v_j \mid v_k,\ v_i\right) = \begin{cases} 1/p, & if\ d_{ij} = 0 \\ 1, & if\ d_{ij} = 1 \\ 1/q, & if\ d_{ij} = 2 \\ 0, & otherwise \end{cases} \tag{1}$$

where m, n are the adjustable parameters and d_{ij} is the shortest path length of node v_i to v_j. The random walk employs Depth First Search (DFS) when $p > 1$ and $q < 1$ while employing Breadth First Search (BFS) when $p < 1$ and $q > 1$.

By optimizing the problem of random walk, it updates the adjacency matrix by different walk paths. Calculating the cosine similarity of higher-order structural information can determine whether there are stronger connections between nodes in the global structure and the update formula can be expressed as:

$$\begin{aligned} &\tilde{A}_{ij} = A_{ij} + f\left(i,j\right), \\ &f\left(i,j\right) = \begin{cases} 1, & if\ \cos\left(x_i, x_j\right) \geq \tau \\ 0, & otherwise \end{cases} \end{aligned} \tag{2}$$

where $\cos\left(\cdot\right)$ denotes the cosine similarity calculation and τ is the threshold different from initial similarity.

As shown in Fig. 3, after utilizing the random walk strategy, we update the graph representation and obtain a new optimized graph embedding $\tilde{\mathcal{G}} = \left(\tilde{\mathcal{V}},\ \tilde{\mathcal{E}},\ \tilde{\mathcal{A}}\right)$. With different similarity thresholds, the updated graph embeds a few more explicit relations. This not only initially captures the global structural features in the graph, but also significantly reduces the number of nodes and edges that need to be taken into account, reducing the burden of the subsequent graph convolution and speeding up the computational process.

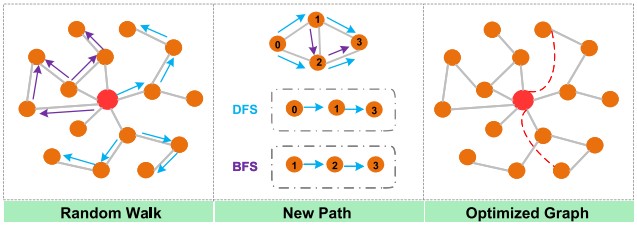


Fig. 3. The illustration of Random Walk Updating. BFS tends to visit the immediate neighbors of the source node, and DFS tends to explore nodes that are further and further away from the source node. The combination of the two can effectively capture the local and global relations between nodes.

3.3. Latent relation capturing

After explicit exploration of semantic and structural facial information using random walk, to obtain more comprehensive facial representation, the multi-head attention (Ren, Huang, Li, Song, & Nie, 2021) is then applied on the optimized initial graph embedding $\tilde{\mathcal{G}}$ to explore latent relations between facial keypoints. Specifically, the multi-head self-attention is performed to construct a set of fully connected graphs that integrates the correlations between vertices. Generating m-indexed fully connected graphs can be formulated as:

$$\tilde{A}^{(m)} = soft\max\left(\frac{\tilde{X}_i W_m^i \times \left(\tilde{X}_j W_M^j\right)^T}{\sqrt{M}}\right)\tilde{A}^{(0)} \tag{3}$$

where $\tilde{A}^{(0)}$ is the initial adjacency matrix of optimized graph $\tilde{\mathcal{G}}$, $\tilde{X}_i$ and $\tilde{X}_j$ denotes the node embedding from the node set $\tilde{\mathcal{V}} = \left\{v_1,\ \ldots\ ,\ v_i,\ \ldots\ ,\ v_j,\ \ldots\ v_N\right\}$. M is feature vector's dimension, while W_m^i and W_m^j are the pairwise transfer matrices for $\tilde{X}_i$ and $\tilde{X}_j$, respectively. From this mechanism, aggregation by fully connected graphs can capture latent relations between facial key points more comprehensively, increase the diversity of correlations, and address the weakness of common GNNs missing some important information.

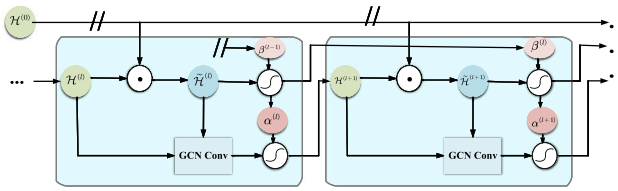


Fig. 4. The illustration of deep feature extraction with the adaptive initial residuals and dynamic developmental residuals. The adaptive initial residuals obtain the personalized characteristics from initial embedding $H^{(0)}$ and hidden embedding $H^{(l)}$. The dynamic developmental residuals gain the developmental pattern $\alpha^{(l)}$ from the residual embedding $\tilde{H}^{(l)}$.

3.4. Deep feature extraction

To extract deep and rich facial feature information, we perform the graph convolution operation on the obtained m different fully connected graphs.

DeepGCNs can expand the receptive field and improve the graph model representation performance, which can be formulated as:

$$H^{(l+1)} = \mathrm{ReLU}\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{4}$$

where $\tilde{D} = \sum_j \tilde{A}_{ij}$ denotes the degree matrix, $\tilde{A} = A + I$, I is the unit matrix, $H^{(l)}$ is the lth layer embedding of the graph nodes and $W^{(l)}$ is the learnable weight.

However, GCNs stack too many layers tending to over-smoothing, where the representations of the nodes converge and become indistinguishable. For the fully connected graphs generated by multi-head attention mechanisms, the over-smoothing problem caused by deepening the number of layers of GCNs is more severe and needs to be urgently addressed. Some works have begun to introduce residual connections to alleviate the over-smoothing problem and it is working well, ResGCN (Li, Muller et al., 2019) is expressed as:

$$H^{(l+1)} = \mathrm{ReLU}\left((1-\alpha)\,\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)} + \alpha H^{(l-1)}\right) \tag{5}$$

where $W^{(l)}$ is the residual connection weight parameter. And the initial residual connection replaces the node representation of the previous layer by combining $H^{(l-1)}$ with the initial representation $H^{(0)}$ (see Fig. 4).

Nevertheless, ResGCN uses fixed values to represent layer-to-layer node connections, ignoring the individualized characteristics of the initial nodes and the dynamically developmental correlations between layers. And learning the personalized characteristics of initial nodes and the related properties of dynamic development between layers can not only solve the over-smoothing problem, but also enhance the expressive power of GCNs. Inspired by the Zhang, Yan, He, Li, and Chu (2023), we design the deep residual GCNs to combine the adaptive initial residuals and dynamic developmental residuals. The adaptive initial residuals can adaptively derive information from the initial representation to prevent noise-mitigating over-smoothing problems, which can be formulated as:

$$\tilde{H}^{(l+1)} = \left(1-\beta^{(l)}\right)\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)} + \beta^{(l)}H^{(0)} \tag{6}$$

where $\beta^{(l)}$ denotes the proportion of adaptively connected initial features at layer l.

The dynamic developmental residuals capture the residual developmental state of the layers and avoid stacking too many non-linear mappings resulting in vanishing gradients, which can be formulated as:

$$H^{(l+1)} = \mathrm{ReLU}\left(\left(1-\alpha^{(l)}\right)\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\tilde{H}^{(l)}W^{(l)} + \alpha^{(l)}\tilde{H}^{(l-1)}\right) \tag{7}$$

where $\alpha^{(l)} = \Psi\left(\tilde{H}^{(l)},\ \tilde{H}^{(l-1)}\right)$ denotes the dynamic developmental factor to adjust the percentage of information retained from the previous layer, $\Psi(\cdot)$ is the developmental function.

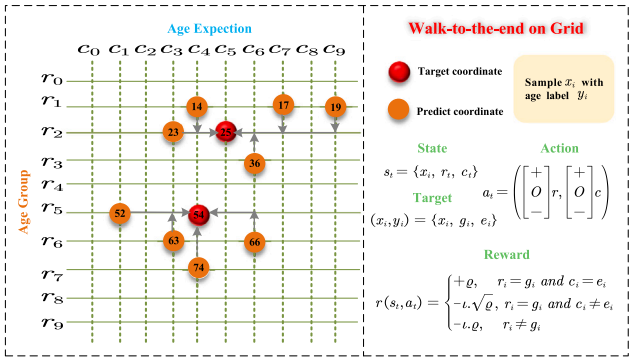


Fig. 5. The illustration of Progressive RL-based Age Estimation. The age estimation through classification and then regression is defined as a walk-to-the-end problem on a grid.

Algorithm 1 Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (LRA-GNN).

Input: Images $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$, multi-head number M;
Output: Comprehensive facial feature embedding $\mathcal{H}$.
for number of training epochs **do**
  for I_i **in** all images **do**
    // **Initial Graph Construction**
    Construct an initial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ using facial keypoints as priori knowledge.
    // **Random Walk Updating**
    Obtain $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{A}})$ on initial graph by Eq. (1), (2).
    // **Latent Relation Capturing**
    for v_i **in** $\tilde{\mathcal{V}}$ **do**
      Generate fully connected graphs by Eq. (3).
    end for
    // **Deep Feature Extraction**
    for v_i **in** $\tilde{\mathcal{V}}$ **do**
      Calculate adaptive initial residual by Eq. (6).
      Calculate dynamic developmental residual by Eq. (7).
      Fusing two residuals to obtain embedding $\mathcal{H}$.
    end for
  end for
end for
Return: The feature embedding $\mathcal{H}$.

3.5. Progressive RL-based age estimation

We define the age estimation through classification and then regression as a walk-to-the-end problem on a grid. As shown in Fig. 5, the grid is divided into rows and columns, with the rows representing the classified age groups and the columns representing the serialized age regression values within that age group. At each time interval, the agent is presented with a sample and places (classifies) it into the appropriate row (age group), subsequently positioning (regresses) it within one of the columns of that row (predicted value). Depending on the agent's different actions, the environment provides an immediate reward and the subsequent sample. When the agent moves the sample to the correct row and column, the environment will give the agent a positive reward, otherwise, it will give the agent a penalty. As the agent acquires the optimal conduct through its engagement with the environment to achieve the highest cumulative reward, it can make progressive age estimation through classification and then regression on the samples as correctly as possible.

We model the progressive RL-based age estimation as a Markov Decision Process (MDP) and describe it using a five-tuple $\{S, A, R, P, Y\}$. In the framework, the state space $s_t \in S$, action space $a_t \in A$, rewards $r_t \in R$, policy $\pi \in P$ and discount factor $\gamma \in (0, 1)$.

3.5.1. State

The state of the environment is determined by the trained age estimation samples, which can be denoted as $\{x_i,\ R,\ C\}$. Specifically, the sample x_i with real label y_i can be split into corresponding age groups $g_i \in G$ (rows $r_i \in R$) and values within groups $e_i \in E$ (columns $c_i \in C$). G and E contain 0–9, e.g. g_1 denotes the age group 10–19 years, while e_1 in g_1 denotes the second age value of this age group i.e. 11 years.

3.5.2. Action

Agent's actions are associated with the labels of the training datasets, which are targeted to move to the correct coordinates. The main actions include: row move d_r units, column move d_c units, or none of the row or column move. Represent the coordinates separately by binary groups i.e. $a_t = (+d_r, O)$ means the row increases by d_r units and the column stays the same, $a_t = (O, -d_c)$ means row stay the same and column decreases by d_c units.

3.5.3. Reward

Rewards are designed based on the engagement between the environment and the agent, which can measure the agent's performance after acting. Because the sample is categorized into the correct age group is a critical step, the reward given for whether the agent chooses the correct row is higher than for choosing the correct column. We also consider the continuity of age distribution to add the distance from the label value on the reward function. Meanwhile, to make age estimation more robust and generalizable, we design the imbalance ratio to guide the agent to more effectively learn actions within an unbalanced dataset. To enhance the identification of samples from the minority class, the algorithm is attuned to the minority class, delivering a greater reward or penalty upon encountering such a sample. The reward function is delineated as follows:

$$r(s_t, a_t) = \begin{cases} +\varrho, & r_i = g_i \ and \ c_i = e_i \\ -\imath.\sqrt{\varrho}, & r_i = g_i \ and \ c_i \neq e_i \\ -\imath.\varrho, & r_i \neq g_i \end{cases} \tag{8}$$

where $\imath = |r_i - g_i| + |c_i - e_i|$ denotes the distance from label value and $\varrho = \frac{N_{g_M}}{N_{g_i}}$ denotes the imbalance ratio. N_{g_M} is the amount of majority group and N_{g_i} the amount of ith group.

3.5.4. Policy

The policy π is a mapping that maps states to actions, denoted by $\pi(a \mid s)$. Depending on the currently observed state s, the π_θ will decide which action a the agent should perform. We employ a random policy to provide a probability distribution for choosing each action in state s. This randomness increases the exploratory ability of the agent and prevents it from falling into a local optimal solution. We define the age estimation problem as the strategy π^* that finds the optimal row and column coordinates to maximize the cumulative reward, and the strategy π_θ can be viewed as an ensemble classifier regressor with parameter Ω.

3.5.5. Deep Q-learning

Q-learning is a reinforcement learning approach rooted in value iteration, which pursues the optimum function π^*. This signifies that, by opting for the most favorable action a within the present state s, the agent is positioned to secure the greatest cumulative reward.

The mapping of the action-value function adheres to the *Bellman*'s equation, which can be articulated as:

$$Q^*(s, a) = E_\pi\left(r_t + \gamma Q\left(s_{t+1},\ a_{t+1}\right) | s_t = s,\ a_t = a\right) \tag{9}$$

where $\gamma \in (0, 1)$ falls within the open interval $(0, 1)$, embodying the discount factor that mirrors the influence of the current action's impact on the reward value. The action with the largest Q in state s is chosen by the optimal policy $\pi_*(s,\ a)$, i.e., the greedy policy, which can be formulated as:

$$\pi^*(s, a) = arg\max_a Q(s, a) \tag{10}$$

Since the Deep Q Network selects a greedy strategy for both action selection and action evaluation and uses the same neural network parameters, it is easy to cause an overestimation of the value function during the learning process, i.e., the projected value function exceeds the factual value, which will affect the final optimal policy. To address this problem, Van Hasselt, Guez, and Silver (2016) proposed the Double Q-learning algorithm, which decouples the selection of actions and action evaluation to solve the overestimation problem towards the traditional Deep Q Network (DQN).

The Double-DQN method initially identifies the action that corresponds to the highest Q value within the online network Q^θ. Subsequently, it determines the target Q value stemming from the chosen action a in the target network Q^τ, which can be formulated as:

$$Y_t^{Double\ Q} = r_t + \gamma Q\left(s_{t+1},\ arg\max_a Q^\theta\left(s_{t+1},\ a;\ \theta_o\right); \theta_t\right) \tag{11}$$

where $Y_t^{Double\ Q}$ is the target Q value, γ is the discount factor, θ_o is the online weight and θ_t is the target weight.

To ensure the Q-value prediction accuracy of the model, we introduce an ε-greedy strategy to guide the action selection process. This strategy selects the best action based on the current Q-value in most cases, i.e., with probability $1-\varepsilon$, to fully utilize the known information. However, to prevent the model from falling into a local optimal solution and to encourage it to explore the unknown state, we simultaneously retain a certain degree of randomness, i.e., randomly selecting actions with a probability of ε. This strategy of balancing exploration and utilization helps the model find the optimal balance between exploring new knowledge and utilizing existing knowledge. During the training process, we make full use of the objective function as well as the empirical playback mechanism, to optimize the performance of the Q-network.

Algorithm 2 Progressive RL-based Age Estimation(PRLAE).

Input: Training data $D = \{(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)\}$, episode number K.
Output: Optimized ensemble classification regressor with parameters Ω.
1: Initialize simulation environment E.
2: **for** episode $k = 1$ **to** K **do**
3: Shuffle the training data D.
4: Initialize the current state $s_t = \{x_t,\ r_t,\ c_t\}$.
5: Calculate the target coordinate (g_i, e_i).
6: **for** time $t = 1$ **to** T **do**
7: Choose an action based on ϵ-greedy policy: $a_t = \pi_\theta(a_t)$.
8: Calculate the reward $r(s_t, a_t)$ by Eq (8).
9: **end for**
10: Obtain next state $s_{t+1} = \{x_{t+1},\ r_{t+1},\ c_{t+1}\}$.
11: Calculate the target Q-value by Eq. (11).
12: Update the parameter Ω by computing the loss with Eq. (12).
13: **if** reach the maximum cumulative reward **then** break
14: **end for**
15: **Return:** The parameter Ω.

3.5.6. Training network

As shown in Algorithm 2, taking the face sample features as inputs to train Double DQN, the agent traverses all states by moving on the grid with progressive iterative updates while optimizing the ensemble classifier regressor. Inspired by Lin, Goyal, Girshick, He, and Dollár (2017), our progressive age estimation loss function consists of focal loss (FL) $\mathcal{L}_{FL} = -(1-p_i)^\tau \log(p_i)$ and average absolute loss (MAE), which can be formulated as:

$$\mathcal{L}_{PRLAE} = \eta \mathcal{L}_{FL} + (1-\eta)\mathcal{L}_{MAE} \tag{12}$$

where p_i represents the model's predicted probability and τ is an adjustable focusing parameter. Higher values of τ reduce the loss of easy samples, which allows the model to turn its attention to hard

samples. When $\tau = 0$, it becomes the standard cross-entropy loss. We set $\tau = 1.3$ in our study for better classification results.

The loss function $\mathcal{L}_{PRLAE}$ can dynamically consider both hard samples and is optimized together with the reward function that can distinguish between classes, which is committed to improving the generalization ability and accuracy of our age estimation network.

4. Experiment

In this section, we introduce the datasets and follow the evaluation criteria utilized in the experimental procedures. In addition, we present the specifics of our experiment and compare the results with state-of-the-art methods to validate the efficacy of our proposed technique. Finally, an ablation analysis is performed on the key elements within our approach to clarify how various aspects influence the overall performance.

4.1. Datasets

MORPH-II: The dataset (Ricanek & Tesafaye, 2006) is a large collection of cross-age facial images widely used in facial analysis studies. The dataset encompasses 55,134 facial photos featuring 13,000 distinct individuals, spanning an extensive age spectrum from 16 to 77 years. This experiment references two widely used evaluation schemes: **Setting I**: following Gao, Zhou, Wu, and Geng (2018), the dataset is randomly partitioned into two distinct segments, with the training set comprising 80% and the test set accounting for 20%. **Setting II**: following Tan et al. (2017), we select a subset of 5493 facial images from Caucasian ethnic groups, which is split into two separate portions, allocating 80% for training and 20% for testing.

FG-NET: The dataset (Lanitis, Taylor, & Cootes, 2002) encompasses 1002 facial photographs, representing 82 individuals with ages varying from 0 to 69 years, demonstrating a range of ages from children to the elderly. The dataset provides information on 68 key points of the face in each image, which facilitates when performing tasks such as facial aging simulation and facial expression analysis. We followed the setup of previous methods (Pan, Han, Shan, & Chen, 2018) using leave-one-person-out (LOPO) cross-validation.

ChaLearn LAP 2016: The dataset (Escalera et al., 2016) is used in a challenge contest organized by ChaLearn focusing on the estimation of the appearance age of faces. It contains normally distributed age labels based on the annotation results of at least 10 people, which is very due for age estimation tasks that need to deal with real-world conditions. This dataset can be divided into a training set of 4113 images, a validation set of 1500 images, and a test subset of 1978 images.

UTK-Face: The dataset (Zhang, Song, & Qi, 2017) encompasses a vast age range that spans from infancy at 0 years to elderly at 116 years. This extensive collection, comprising more than 20,000 photographs, is annotated with demographic information such as gender, age, and ethnicity, thereby encapsulating a diverse array of human features. In our work, we employ a stratified sampling approach, allocating 80% of the dataset for training purposes and reserving the remaining 20% for the testing of our models.

4.2. Evaluation criteria

4.2.1. MAE(Mean Absolute Error)

It refers to the mean absolute error between the estimated age and the actual label value, which can be expressed as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{13}$$

where y_i and $\hat{y}_i$ denote the actual and estimated age values of the ith sample, respectively, and N is the amount of test images. The lower the value of the MAE, the superior the model's prediction performance.

4.2.2. ϵ-error (normal score)

In the LAP-2016 dataset, the age is indicated as an average derived from various individuals' labels, and the true age in the data contains both mean and variance attributes. Therefore considering these factors can be a more accurate measure of the age estimation performance. A smaller ϵ-error indicates improved performance in age estimation, and it can be expressed as follows:

$$\epsilon = 1 - \sum_{i=1}^{N} exp\left(-\frac{\left(y_i - \hat{y}_i\right)^2}{2\sigma_i^2}\right) \tag{14}$$

where y_i and $\hat{y}_i$ are the actual and estimated age values of the ith sample, respectively, N is the amount of test images, and σ_i^2 is the labeled standard deviation.

4.2.3. CS(Cumulative Score)

It focuses on the accuracy of the prediction within a certain error, which can be calculated as:

$$CS\,(j) = \frac{N_{e\leq j}}{N} \times 100\% \tag{15}$$

where $N_{e\leq j}$ indicates the number of image tests where the absolute error of the estimate does not exceed j. N denotes the number of test images. The larger its value, the method is more robust.

4.3. Implementation details

First, we utilize MTCNN (Zhang, Zhang, Li, & Qiao, 2016) for full face detection, align the image using the identified facial landmarks, and then crop and adjust it to a size of 224 * 224 pixels. Throughout the training phase, the image undergoes augmentation through random inversion, rotation, and translation. For the entirety of our experiments, we applied the Adam optimizer (Kingma & Ba, 2014), with decay and momentum parameters set at 0.0005 and 0.9, respectively. The learning rate was initialized at 0.001, with a decay strategy based on cosine annealing. Utilizing PyTorch on an NVIDIA RTX 3090 GPU, we trained our framework over 120 epochs with a batch size of 32.

Meanwhile, during the experimental tuning process, considering the model performance and parameter overhead, we set the number of LRA-GNN layers L to 12, the number of multi-head M to 8, the loss function hyperparameter η to 0.5, the initial graph construction threshold to 0.936, and random walk updating threshold to 0.824.

4.4. Comparisons with the state-of-the-art methods

To demonstrate the efficacy of our LRA-GNN, we perform comprehensive experiments across three facial image datasets. For the characteristics of different datasets, we utilize appropriate experimental settings and evaluation criteria to compare them with the state-of-the-art (SOTA) methods.

4.4.1. Comparisons on Morph II

On the most widely used Morph II dataset, we list the SOTA works year by year and give the backbone network used for that work as well as the number of parameters. We follow different protocols for comparison from the point of view of MAEs and the number of parameters. Table 1 shows the detailed results, where our method achieves an excellent performance of 1.79 (Setting I) and 1.94 (Setting II) with pre-training on an external dataset. Under Setting I, we only underperform HR (Hiba & Keller, 2023), GLAE (Bao et al., 2023) and TAA-GCN (Korban et al., 2023). HR proposed a hierarchical model integrating discrete age predictions with a collection of specialized regressors, each fine-tuned to enhance the probability estimation within specific age brackets. GLAE proposed Feature Rearrangement (FR) and Pixel-level Auxiliary learning (PA) to take advantage of the facial features. TAA-GCN designed the Temporal Memory Module (TMM) and Adaptive Graph Convolutional Layer (AGCL) to model age aging and temporal dependence. Under Setting II, we achieved the best results.

Table 1
Comparison with the state-of-the-art methods on Morph II dataset (see Bao et al., 2022; Chen, Zhang, Dong, Le, & Rao, 2017; Li, Lu et al., 2019; Li, Lu, Wuerkaixi, Feng, & Zhou, 2022; Shen, Zhao, Guo, & Yuille, 2017; Zhang, Liu, Xu and Zhu, 2019; Zhao, Liu, & Wei, 2024).

Method	Venue year	Backbone network	MAE		Param.
			Setting I	Setting II	
DEX (Rothe et al., 2015)	IJCV 2016	VGG-16	–	3.15/2.68[a]	138M
Ranking-CNN (Chen et al., 2017)	CVPR 2017	Binary CNNs	2.96[a]	–	500M
DLDLF (Shen et al., 2017)	NIPS 2017	VGG-16	2.24	–	138M
MV (Pan et al., 2018)	CVPR 2018	VGG-16	2.79 /2.16[a]	–	138M
SSR-NET (Yang et al., 2018)	IJCAI 2018	SSR-NET	3.16[a]	–	40.9K
C3AE (Zhang, Liu, Xu et al., 2019)	CVPR 2019	C3AE	2.78[a]	2.95[a]	**39.7K**
BrigeNet (Li, Lu et al., 2019)	CVPR 2019	VGG-16	2.38[a]	2.35[a]	138M
PML (Deng et al., 2021)	CVPR 2021	ResNet-34	2.15	2.31	21M
MWR (Shin et al., 2022)	CVPR 2022	VGG16	2.00	2.13	138M
MetaAge (Li et al., 2022)	TIP 2022	VGG16	1.81	2.23	138M
DAA (Chen et al., 2023)	CVPR 2023	ResNet-18	2.25/2.06[a]	–	11M
TAA-GCN (Korban et al., 2023)	PR 2023	TAA-GCN	1.69	–	–
DCT (Bao et al., 2022)	TIFS 2023	ResNet-50	2.28/2.17[a]	–	23M
MSL (Wang et al., 2023)	TIFS 2023	ResNet-34	2.10	2.03	21M
HR (Hiba & Keller, 2023)	TPAMI 2023	VGG16	**1.13**[a]	2.53[a]	138M
GLAE (Bao et al., 2023)	TIP 2023	ResNet-50	1.14[a]	2.00[a]	23M
GroupFace (Zhang, Shou, Ai et al., 2024)	TIFS 2024	EMGCN	2.09/1.86[a]	2.27/2.01[a]	8.6M
Zhao et al. (2024)	ISCI 2024	ResNet34(1/4)	1.85[a]	2.42[a]	–
LRA-GNN (Ours)	**–**	**LRA-GNN**	2.02/1.79[a]	2.21/**1.94**[a]	13M

[a] Indicates used the extra datasets for pre-training.

Table 2
Comparison with the state-of-the-art methods on FG-NET dataset (see Deng et al., 2021; Shen et al., 2018).

Method	MAE	CS (%)	Param.
DEX (Rothe et al., 2015)	4.63/3.09[a]	72.4	138M
DRFs (Shen et al., 2018)	3.85	80.6	138M
MV (Pan et al., 2018)	4.10/2.68[a]	–	138M
C3AE (Zhang, Liu, Xu et al., 2019)	2.95[a]	–	**40.9K**
BrigeNet (Li, Lu et al., 2019)	2.56	86.0	138M
PML (Deng et al., 2021)	2.16[a]	–	16M
MWR (Shin et al., 2022)	2.23	91.1	–
DAA (Chen et al., 2023)	2.19[a]	–	11M
TAA-GCN (Korban et al., 2023)	3.58	–	–
MCGRL (Shou et al., 2025)	2.86	88.0	–
LRA-GNN (Ours)	**2.14**[a]	**91.6**	13M

[a] Indicates used the extra datasets for pre-training.

Table 3
Comparison with the state-of-the-art methods on ChaLearn LAP 2016 dataset (see Duan, Li, Yang, & Li, 2018; Sun, Pan, Han, & Shan, 2021).

Method	MAE	ϵ-error	Param.
AGEn (Tan et al., 2017)	3.82	0.3100	138M
MV (Pan et al., 2018)	3.14	0.287	138M
DLDL-v2 (Gao et al., 2018)	3.45	0.267	**3.7M**
RAGN (Duan et al., 2018)	–	0.367	–
DCDL (Sun et al., 2021)	3.33	–	138M
MetaAge (Li et al., 2022)	3.49	0.265	138M
MSL (Wang et al., 2023)	3.25	–	21M
LRA-GNN (Ours)	**3.11**[a]	**0.258**	13M

[a] Indicates used the extra datasets for pre-training.

Thanks to our comprehensive capturing of latent relations, our LRA-GNN achieves performance similar to or even surpassing the SOTAs. At the same time, while achieving notable performance, we have fewer network parameters, which is attributed to the effectiveness and low redundancy of GCN in modeling facial key points.

4.4.2. Comparisons on FG-NET

We use the Mean Absolute Error (MAE) and Cumulative Score (CS) metrics on this few-shot dataset to compare with the SOTAs. CS represents the proportion of images where the absolute error does not exceed a threshold of j, following previous work with a setting of $j = 5$. As shown in Table 2, our work earns the lowest MAE of 2.14 and the highest CS of 91.6%, better than the similar GNN-based method MCGRL (Shou et al., 2025) that not capture the latent relations. This may be due to the fact that we have designed a series of graph enhancement strategies such as random walk and latent relation capturing, which make our model capable of learning discriminative features even in the face of fewer datasets.

4.4.3. Comparisons on ChaLearn LAP 2016

To enhance the assessment of our approach's efficacy in unconstrained conditions, we compared it against SOTAs on the ChaLearn LAP 2016 dataset. CLAP2016 is a challenging dataset done by multiple annotators, with the presence of manually subjectively labeled mean and variance, and thus we evaluate the model performance in conjunction with the ϵ-error. Table 3 demonstrates the final results, where our method attains the minimal MAE at 3.11 and the minimal ϵ-error at 0.258 for a moderate number of parameters. Compared to past models based on CNN architectures, our LRA-GNN achieves a more comprehensive feature extraction in a more flexible way, demonstrating the superiority of our method.

4.4.4. Comparisons on UTK-face

We evaluated the performance of LRA-GNN on a large-scale, unconstrained dataset spanning ages from 0 to 116 years. As seen in Table 4, GroupFace achieved the lowest Mean Absolute Error (MAE) of 4.22, significantly outperforming previous methods with a relatively modest number of parameters, totaling 13M. It is worth noting that MSL (Wang et al., 2023) employed a deeper ResNet-34, and MWR (Shin et al., 2022) utilized a larger VGG16 architecture, both effectively reducing the MAE. Conversely, Andrey Savchenko (Savchenko, 2019) utilized the more compact MobileNet-v2 with the fewest parameters but lagged in performance. These comparisons collectively demonstrate the effectiveness and reliability of our network across various types of facial datasets.

4.5. Ablation studies

To demonstrate the efficacy of our method and strategy, we split the different components for a series of ablation experiments on several datasets without loading the eternal dataset pre-trained weights. We

Table 4
Comparison with the state-of-the-art methods on UTK-Face dataset (see Cao, Mirjalili, & Raschka, 2020).

Method	Backbone	MAE	Param.
Savchenko (2019)	MobileNet-v2	5.44[a]	**3.4M**
CORAL (Cao et al., 2020)	ResNet-50	5.47[a]	25.6M
DCDL (Sun et al., 2021)	VGG16	4.48[a]	138M
MWR (Shin et al., 2022)	VGG16	4.37[a]	138M
MSL (Wang et al., 2023)	ResNet-34	4.31	21M
MIVOLO (Kuprashevich & Maksim, 2023)	VOLO-D1	4.23[a]	25.8M
LRA-GNN (Ours)	**LRA-GNN**	**4.22**	13M

[a] Indicates used the extra datasets for pre-training.

divide our method into three main constituent components: Latent Relation Capturing (LRC), Deep Feature Extraction (DFE), and Progressive RL-based Age Estimation (PRLAE). We first explore the effectiveness of these three components as a whole, and then further conduct more detailed ablation experiments on each of the three components to demonstrate the performance improvement of different modules.

4.5.1. Impact of different components

To test the impact of different components for age estimation accuracy, we set baseline as common DeepGCN, and then add components one by one for the experiment. As shown in Table 5, the results give the MAEs of common GCN, and GCN with capturing latent relations, partial LRA-GNN, and full LRA-GNN on the three datasets. It can be seen that Latent Relation Capturing (LRC) contributes the most to enhancing the accuracy of age estimation, reducing the MAE by 0.28 in CLAP 2016. This may be due to the fact that unrestricted face samples are more complex and more necessary to capture the latent relations of facial key points. Secondly, Progressive RL-based Age Estimation (PRLAE) is also critical to performance improvement, directly reducing the MAE by 0.25 under Setting II in Morph II. Thirdly, Deep Feature Extraction (DFE) provides a significant gain to the model as well, directly reducing the MAE by 0.29 in the FG-NET dataset, while the combination of DFE and LRC enables a more comprehensive and deeper facial representation.

4.5.2. Impact of latent relation capturing

In Latent Relation Capturing, we utilize facial key points as prior knowledge (FK) and Random Walk Strategy (RW) to jointly guide the effective capturing and comprehensive representation of latent relations. To analyze the gain of these two strategies on latent relation capturing and to discuss the impact of the quantity of M heads in the mechanism of multi-head attention, we conduct experiments on the FG-NET dataset, which is evaluated using the cumulative score (CS).

As shown in Fig. 6, both the utilization of face key points as prior knowledge and the random walk strategy are effective in improving the performance of age estimation. Without the joint guidance, as the number of fully connected graphs M increases, more redundant information is easily generated, making it difficult to effectively capture latent relations. We observe that the most performance of 91.6% is achieved when the quantity of multi-head is 8, so we choose M to 8.

We also explore the effect of specific strategy types in random walk updating on our graph model. As shown in Table 6, considering both the Breadth First Search (BFS) and Depth First Search (DFS) can effectively reduce the age prediction error. This is because BFS tends to explore local neighbors and is difficult to capture long-distance dependencies, while DFS tends to explore more deeply but lacks a description of the exact nature of these dependencies. LRA-GNN flexibly combines the two which can better capture global-local relations.

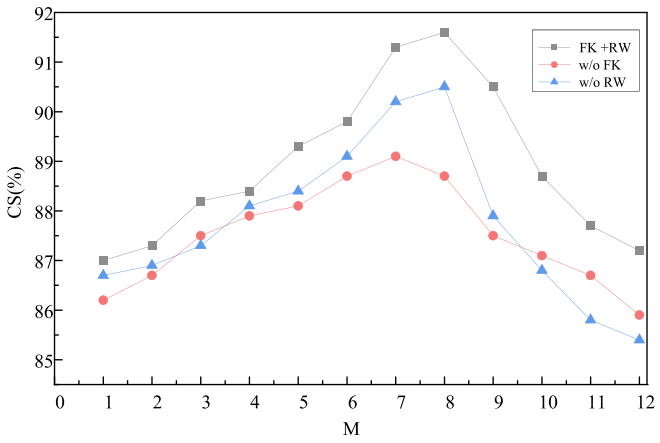


Fig. 6. The ablation study of Latent Relation Capturing on the FG-NET dataset.

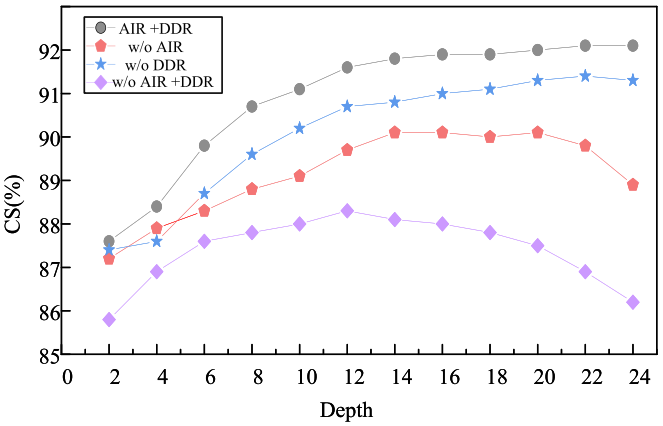


Fig. 7. The ablation study of Deep Feature Extraction on the FG-NET dataset.

4.5.3. Impact of deep feature extraction

We continue to evaluate the effectiveness of our Deep Feature Extraction. To quantify the contribution of the adaptive initial residual (AIR) and dynamic developmental residual (DDR) of our design, keeping the other components unchanged, four objects were set up: AIR+DDR, w/o AIR, w/o DDR and w/o AIR+DDR. As shown in Fig. 7, the performance without the adaptive initial residual (AIR) and dynamic developmental residual (DDR) grows with the number of layers starts to degrade after a shallower growth, which could be the introduction of noise leading to over-smoothing problems. With AIR or DDR mitigates the over-smoothing problem somewhat, where with AIR performance starts to decline after reaching the optimum, and the upper limit of the performance without DDR is difficult to break through and improves slowly. Our Deep Feature Extraction fusing AIR and DDR can better alleviate the GCN over-smoothing problem and ensure the consistency and diversity of the information, thus effectively improving the age estimation performance. Considering the balance between performance and parameter count, the performance improvement is less but the number of parameters increases more after $L > 12$, so we set the number of layers to 12.

4.5.4. Impact of progressive RL-based age estimation

In Progressive RL-based Age Estimation, we define age estimation by classification as a walk-to-the-end problem on a grid, where the grid is divided into rows and columns. As shown in Table 7, our age estimation method allows for simultaneous optimization of classification and regression, achieving a considerable improvement over the decoupling method. Besides, our well-designed reward function not

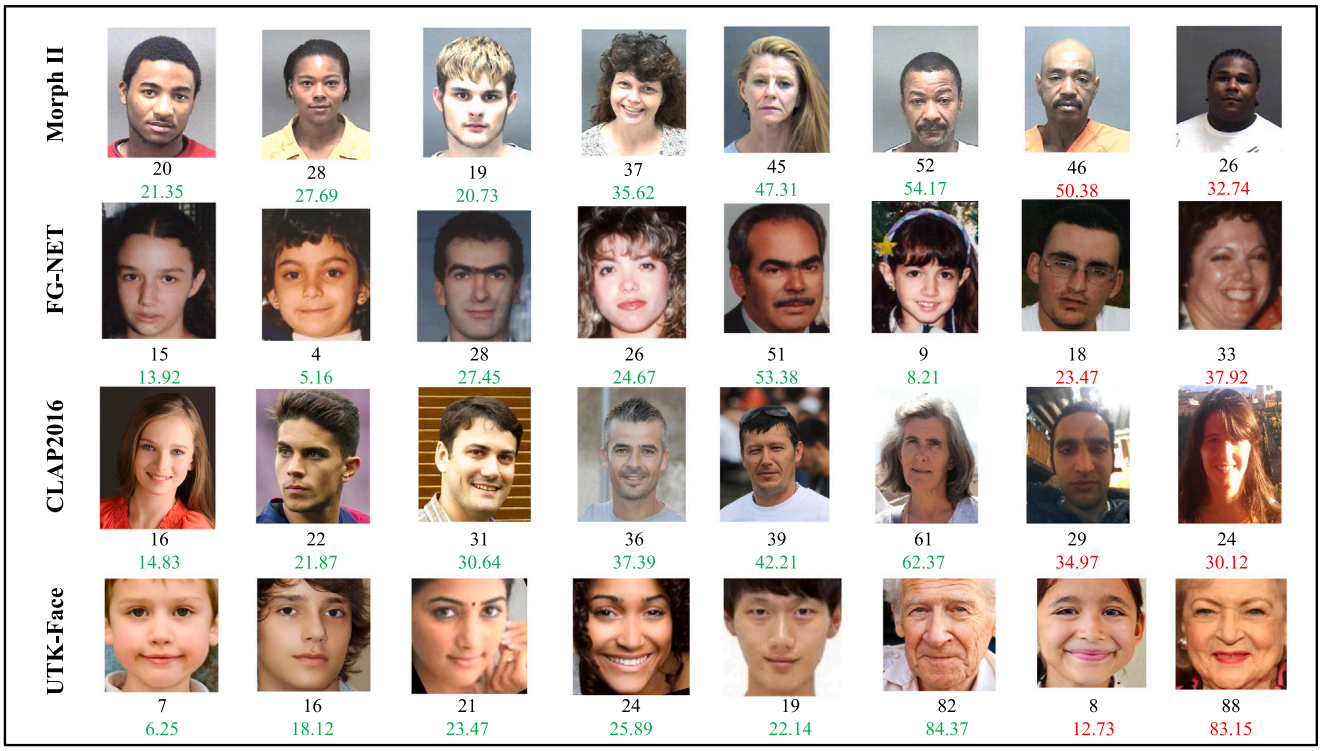


Fig. 8. The examples of age estimation results of our LRA-GNN on three facial datasets. The factual label is the black number, the reliable estimation results are shown in the green number, and the poor estimation results are shown in the red number.

Table 5
The impact of different key components.

Components			Morph II		FG-NET	CLAP 2016	UTK-Face
LRC	DFE	PRLAE	Setting I	Setting II			
–	–	–	2.47	2.88	2.68	3.54	4.62
✓	–	–	2.29	2.64	2.41	3.26	4.38
–	✓	–	2.35	2.73	2.39	3.32	4.45
✓	✓	–	2.18	2.46	2.32	3.21	4.34
–	✓	✓	2.21	2.43	2.28	3.24	4.41
✓	–	✓	2.15	2.39	2.24	3.18	4.28
✓	✓	✓	**2.02**	**2.21**	**2.14**	**3.11**	**4.22**

Table 6
The impact of different types of Random Walk Updating.

Types	Morph II		FG-NET	CLAP 2016	UTK-Face
	Setting I	Setting II			
BFS	2.13	2.32	2.28	3.23	4.31
DFS	2.07	2.26	2.21	3.19	4.28
BFS + DFS	**2.02**	**2.21**	**2.14**	**3.11**	**4.22**

only designs the imbalance ratio to guide the agent to better perform the behavior in the unbalanced dataset but also adds the distance from label value to take into account the continuity of the age distribution, which both significantly improves the generalization of age estimation. Under Setting II of Morph II, we get the maximum gain of 0.25 MAE.

In addition, we evaluate the impact of the parameter η in the loss function. As shown in Table 8, Morph II (Setting I) and FG-NET datasets achieved the best results at $\eta = 0.5$ and $\eta = 0.4$, respectively. This is probably because Focal Loss is a benefit for coping with hard samples and class distributions, and jointly optimizing the two loss functions for categorical regression with about the same weights can obtain excellent results.

4.6. Qualitative results

To better qualitative the effectiveness of our method, we first discuss the prediction accuracy of LRA-GNN for different age groups and analyze the possible reasons for the results. Then, we measure and compare the runtime and number of parameters of our architecture to analyze the efficiency. Finally, we randomly select some samples in the three datasets to demonstrate the prediction results and analyze the reasons for the success and failure of age estimation.

4.6.1. Accuracy analysis

We discuss the estimation accuracy of LRA-GNN on Morph II (Setting I) and FG-NET datasets for different age groups. Baseline utilizes a common DeepGCN without capturing latent relations. As shown in Fig. 9, the performance of each age group of our LRA-GNN on both datasets is somewhat improved compared to the Baseline. The most obvious improvement is in the g_1 (10–19 years old) age group of Morph II and g_6 (60–69 years old) of FG-NET, probably due to our mining of latent relations and the design of reward function improving the hard sample representation more.

4.6.2. Efficiency analysis

We first perform a theoretical analysis of the time complexity of the main design part of the LRA-GNN. For the construction of the graph, we

Table 7
The impact of progressive RL-based age estimation.

Designs	Morph II		FG-NET	CLAP 2016	UTK-Face
	Setting I	Setting II			
Decouple	2.18	2.46	2.32	3.21	4.34
Co-optimization	2.14	2.37	2.25	3.16	4.31
+Imbalance ratio	2.09	2.28	2.19	3.13	4.25
+Distribution distance	**2.02**	**2.21**	**2.14**	**3.11**	**4.22**

Table 8
The impact of the parameter η in the age estimation loss function.

η	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Morph II	2.21	2.15	2.12	2.06	**2.02**	2.05	2.13	2.18	2.25
FG-NET	2.29	2.25	2.21	**2.14**	2.16	2.23	2.27	2.31	2.33

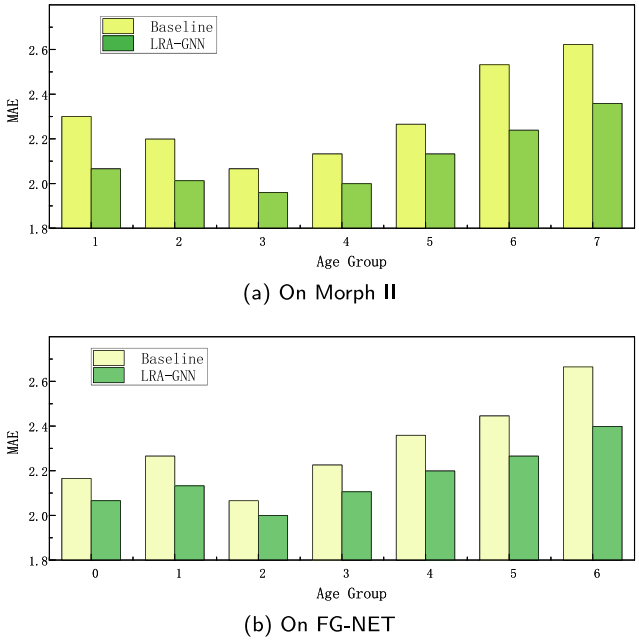


Fig. 9. The accuracy analysis of the baseline and our LRA-GNN on Morph II and FG-NET datasets.

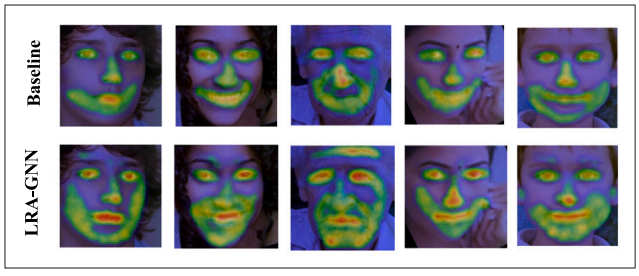


Fig. 10. The visualization examples of LRA-GNN for facial feature extraction utilizing attention heat map.

have N nodes (i.e., facial keypoints) with m features per node, which mainly depends on calculating the similarity between nodes, which takes $\mathcal{O}\left(N^2 \times m\right)$ time. The time complexity of a random walk depends on the number of steps t of the walk and the sparsity of the graph. For each node, we need to update the probability of its neighboring nodes, which takes $\mathcal{O}(t \times d)$ time, where d is the average degree of the graph. The time complexity of the multi-head attention mechanism mainly depends on the number of heads M and the amount of computation for each head. For each head, we need to compute the attention weights between N nodes, which takes $\mathcal{O}\left(N^2 \times M \times m\right)$ time. The time complexity of the deep residual graph convolutional network depends on the number of layers L and the amount of computation per layer. For each layer, we need to perform graph convolution operation which takes $\mathcal{O}\left(N^2 \times L \times m\right)$ time. The time complexity of the reinforcement learning part depends on the size of the state space S, the size of the action space A and the number of training iterations T, which is $\mathcal{O}(S \times A \times T)$. Thus the total time complexity can be expressed as $\mathcal{O}\left((M+L)\,N^2 \times m + t \times N \times d + S \times A \times T\right)$.

Table 9
The efficiency analysis on UTK-Face dataset.

Architectures	Runtime	Param	MAE
VGG-16	149.29 ms	138M	4.37
ResNet-34	**61.88 ms**	21M	4.31
VOLO-D1	194.74 ms	25.8M	4.23
ViG-S	76.43 ms	27.3M	4.28
LRA-GNN (Ours)	141.96 ms	**13M**	**4.22**

Then we perform the efficiency comparison tests using GPUs under the same experimental setup on UTK-Face dataset. We compare the reasoning runtime of CNN (Shin et al., 2022; Wang et al., 2023), Transformer (Kuprashevich & Maksim, 2023) and GNN (Zhang, Shou, Meng, Ai and Li, 2024). As shown in Table 9, compared to ViG-S, which does not capture latent relations, LRA-GNN increases on the fold of the computational overhead but achieves the lowest MAE and number of parameters. Moreover, the running time of LRA-GNN still outperforms VGG-16 and VOLO-D1, which shows that our architecture is worthwhile and reliable.

4.6.3. Visualization analysis

We selected samples from three different types of facial datasets as a presentation of the estimation results. As shown in Fig. 8, our LRA-GNN achieves excellent performance on both constrained and unconstrained datasets. The green numbers show that our method performs well in different datasets and age groups, which can be due to the capturing of latent relations, as well as the progressive reinforcement learning-based co-optimization for age estimation. However, the red numbers show some of the poor estimation results, which may be caused by severe occluded faces, poor backgrounds, and so on.

Furthermore, we employ the attention heat map for visualization to compare the face attention regions obtained from Baseline and LRA-GNN. As can be seen from Fig. 10, compared with Baseline, LRA-GNN covers a larger area of attention and provides better recognition of key points of the face such as eyes, mouth, and nose, and is more sensitive to the detection of some wrinkles. This proves that our network can capture more useful information for age estimation by utilizing the multi-head attention mechanism.

5. Conclusion

In this novel, we have presented a new Latent Relation-Aware Graph Neural Network with Initial and Dynamic Residual (LRA-GNN) to achieve robust and comprehensive facial representation. We first construct an initial graph utilizing facial key points as prior knowledge, and then a random walk strategy is employed on the initial graph

to obtain the global structure. The LRA-GNN leverages the multi-attention mechanism to capture the latent relations and generates a set of fully connected graphs. We also design the deep residual graph convolutional networks for deep feature extraction on the fully connected graphs, which fuse adaptive initial residuals and dynamic developmental residuals to ensure the consistency and diversity of information. Finally, we propose progressive reinforcement learning to co-optimize the ensemble classification regressor. Our proposed method outperforms the state-of-the-art methods on Morph II, FG-NET, and CLAP 2016 age estimation benchmarks, demonstrating its strength and effectiveness. In the future, we consider further optimization of the age estimation algorithm to achieve more robust results under unconstrained conditions.

CRediT authorship contribution statement

Yiping Zhang: Investigation, Conceptualization, Design of study, Acquisition of data, Software, Writing – original draft. **Yuntao Shou:** Methodology, Analysis and interpretation of results, Writing – review & editing. **Wei Ai:** Funding acquisition, Resources, Reviewing and editing. **Tao Meng:** Methodology Analysis and interpretation of results, Supervision, Investigation, Writing – review & editing. **Keqin Li:** Reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 69189338), Excellent Young Scholars of Hunan Province of China (Grant No. 22B0275).

Data availability

Data will be made available on request.

References

Abu-El-Haija, S., Perozzi, B., Kapoor, A., Alipourfard, N., Lerman, K., Harutyunyan, H., et al. (2019). Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *International conference on machine learning* (pp. 21–29). PMLR.

Agbo-Ajala, O., & Viriri, S. (2021). Deep learning approach for facial age classification: a survey of the state-of-the-art. *Artificial Intelligence Review*, *54*(1), 179–213.

Atwood, J., & Towsley, D. (2016). Diffusion-convolutional neural networks. *Advances in Neural Information Processing Systems*, *29*.

Bao, Z., Tan, Z., Li, J., Wan, J., Ma, X., & Lei, Z. (2023). General vs. Long-tailed age estimation: An approach to kill two birds with one stone. *IEEE Transactions on Image Processing*, *32*, 6155–6167.

Bao, Z., Tan, Z., Wan, J., Ma, X., Guo, G., & Lei, Z. (2022). Divergence-driven consistency training for semi-supervised facial age estimation. *IEEE Transactions on Information Forensics and Security*, *18*, 221–232.

Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, 325–331.

Chen, S., Zhang, C., Dong, M., Le, J., & Rao, M. (2017). Using ranking-CNN for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5183–5192).

Chen, P., Zhang, X., Li, Y., Tao, J., Xiao, B., Wang, B., et al. (2023). DAA: A delta age adain operation for age estimation via binary code transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15836–15845).

Cui, Y., Yu, H., Guo, X., Cao, H., & Wang, L. (2024). RAKCR: Reviews sentiment-aware based knowledge graph convolutional networks for personalized recommendation. *Expert Systems with Applications*, *248*, Article 123403.

Deng, Z., Liu, H., Wang, Y., Wang, C., Yu, Z., & Sun, X. (2021). Pml: Progressive margin loss for long-tailed age classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10503–10512).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Duan, M., Li, K., & Li, K. (2017). An ensemble CNN2elm for age estimation. *IEEE Transactions on Information Forensics and Security*, *13*(3), 758–772.

Duan, M., Li, K., Yang, C., & Li, K. (2018). A hybrid deep learning CNN–ELM for age and gender classification. *Neurocomputing*, *275*, 448–461.

Escalera, S., Torres Torres, M., Martinez, B., Baró, X., Jair Escalante, H., Guyon, I., et al. (2016). Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1–8).

Gao, B.-B., Zhou, H.-Y., Wu, J., & Geng, X. (2018). Age estimation using expectation of label distribution learning. *vol. 1*, In *IJCAI* (p. 3).

Ge, X., Jose, J. M., Xu, S., Liu, X., & Han, H. (2024). MGRR-Net: Multi-level graph relational reasoning network for facial action unit detection. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–20.

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 855–864).

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, *30*.

Henaff, M., Bruna, J., & LeCun, Y. (2015). Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163.

Hiba, S., & Keller, Y. (2023). Hierarchical attention-based age estimation and bias analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jang, S., Lee, H., Kim, W. J., Lee, J., Woo, S., & Lee, S. (2024). Multi-scale structural graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Jiang, F., Huang, Q., Mei, X., Guan, Q., Tu, Y., Luo, W., et al. (2023). Face2nodes: learning facial expression representations with relation-aware dynamic graph convolution networks. *Information Sciences*, *649*, Article 119640.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Korban, M., Youngs, P., & Acton, S. T. (2023). Taa-gcn: A temporally aware adaptive graph convolutional network for age estimation. *Pattern Recognition*, *134*, Article 109066.

Kuprashevich, & Maksim, T. I. (2023). Mivolo: Multi-input transformer for age and gender estimation. In *International conference on analysis of images, social networks and texts* (pp. 212–226). Springer.

Lanitis, A., Taylor, C. J., & Cootes, T. F. (2002). Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 442–455.

Li, W., Lu, J., Feng, J., Xu, C., Zhou, J., & Tian, Q. (2019). Bridgenet: A continuity-aware probabilistic network for age estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1145–1154).

Li, W., Lu, J., Wuerkaixi, A., Feng, J., & Zhou, J. (2022). Metaage: meta-learning personalized age estimators. *IEEE Transactions on Image Processing*, *31*, 4761–4775.

Li, G., Muller, M., Thabet, A., & Ghanem, B. (2019). Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9267–9276).

Lin, E., Chen, Q., & Qi, X. (2020). Deep reinforcement learning for imbalanced classification. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *50*(8), 2488–2502.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Meng, T., Shou, Y., Ai, W., Yin, N., & Li, K. (2024). Deep imbalanced learning for multimodal emotion recognition in conversations. *IEEE Transactions on Artificial Intelligence*.

Niepert, M., Ahmed, M., & Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In *International conference on machine learning* (pp. 2014–2023). PMLR.

Pan, H., Han, H., Shan, S., & Chen, X. (2018). Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5285–5294).

Qin, X., Li, L., Pang, G., & Hao, F. (2024). Heterogeneous graph fusion network for cross-modal image-text retrieval. *Expert Systems with Applications*, *249*, Article 123842.

Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., et al. (2023). Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Qiu, H., & Hou, B. (2024). Multi-grained clip focus for skeleton-based action recognition. *Pattern Recognition*, *148*, Article 110188.

Ren, M., Huang, X., Li, W., Song, D., & Nie, W. (2021). LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition. *IEEE Transactions on Multimedia*, *24*, 4422–4432.

Ricanek, K., & Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)* (pp. 341–345). IEEE.

Rothe, R., Timofte, R., & Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 10–15).

Savchenko, A. V. (2019). Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet. *PeerJ Computer Science*, *5*, Article e197.

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., & Yuille, A. L. (2018). Deep regression forests for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2304–2313).

Shen, W., Zhao, K., Guo, Y., & Yuille, A. L. (2017). Label distribution learning forests. *Advances in Neural Information Processing Systems*, *30*.

Shin, N.-H., Lee, S.-H., & Kim, C.-S. (2022). Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18760–18769).

Shou, Y., Cao, X., Liu, H., & Meng, D. (2025). Masked contrastive graph representation learning for age estimation. *Pattern Recognition*, *158*, Article 110974.

Sun, H., Pan, H., Han, H., & Shan, S. (2021). Deep conditional distribution learning for age estimation. *IEEE Transactions on Information Forensics and Security*, *16*, 4679–4690.

Tan, Z., Wan, J., Lei, Z., Zhi, R., Guo, G., & Li, S. Z. (2017). Efficient group-n encoding and decoding for facial age estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2610–2623.

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. *vol. 30*, In *Proceedings of the AAAI conference on artificial intelligence*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

Wang, C., Li, Z., Mo, X., Tang, X., & Liu, H. (2023). Exploiting unfairness with meta-set learning for chronological age estimation. *IEEE Transactions on Information Forensics and Security*.

Wen, G., & Wu, K. (2021). Building decision tree for imbalanced classification via deep reinforcement learning. In *Asian conference on machine learning* (pp. 1645–1659). PMLR.

Yang, J., El-Bouri, R., O'Donoghue, O., Lachapelle, A. S., Soltan, A. A., Eyre, D. W., et al. (2023). Deep reinforcement learning for multi-class imbalanced training: applications in healthcare. *Machine Learning*, 1–20.

Yang, T.-Y., Huang, Y.-H., Lin, Y.-Y., Hsiu, P.-C., & Chuang, Y.-Y. (2018). Ssr-net: A compact soft stagewise regression network for age estimation.. *vol. 5*, In *IJCAI* (p. 7).

Zhang, C., Liu, S., Xu, X., & Zhu, C. (2019). C3AE: Exploring the limits of compact model for age estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12587–12596).

Zhang, K., Liu, N., Yuan, X., Guo, X., Gao, C., Zhao, Z., et al. (2019). Fine-grained age estimation in the wild with attention LSTM networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(9), 3140–3152.

Zhang, Y., Shou, Y., Ai, W., Meng, T., & Li, K. (2024). GroupFace: Imbalanced age estimation based on multi-hop attention graph convolutional network and group-aware margin optimization. *IEEE Transactions on Information Forensics and Security*.

Zhang, Y., Shou, Y., Meng, T., Ai, W., & Li, K. (2024). A multi-view mask contrastive learning graph convolutional neural network for age estimation. *Knowledge and Information Systems*, *66*(11), 7137–7162.

Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5810–5818).

Zhang, L., Yan, X., He, J., Li, R., & Chu, W. (2023). Drgcn: Dynamic evolving initial residual for deep graph convolutional networks. *vol. 37*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11254–11261).

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.

Zhao, Q., Liu, J., & Wei, W. (2024). Mixture of deep networks for facial age estimation. *Information Sciences*, *679*, Article 121086.

Zhou, C., Wang, X., & Zhang, M. (2024). Facilitating graph neural networks with random walk on simplicial complexes. *Advances in Neural Information Processing Systems*, *36*.