Journal of Engineering Sciences and Information Technology (JESIT) • Vol 8, Issue 2 (2024) • P: 50 - 38
https://journals.ajsrp.com/index.php/jesit
ISSN: 2522-3321(Online) • ISSN: 2522-3321(Print)

Enhancing Cross-Age Facial Recognition with T2T-ViT Networks and Multi-Scale Attention Decomposition

Zakarya Mutahar Al-Haeer*, Li Mengxia
Yangtze University | 1 Nanhuan Road, Jingzhou | Hubei Province | 434023 | China

Received: 16/03/2024
Revised: 27/03/2024
Accepted: 08/04/2024
Published: 30/06/2024

* Corresponding author: alhaeer.zakarya@qq.com

Citation: Al-Haeer, Z. M., & Mengxia, L. (2024). Enhancing Cross-Age Facial Recognition with T2T-ViT Networks and Multi-Scale Attention Decomposition. *Journal of engineering sciences and information technology, 8(2)*, 38 – 50. https://doi.org/10.26389/AJSRP.Z160324

Abstract: This paper presents a cross-age facial recognition model that integrates Convolutional Neural Networks (CNN) with Transformers. The model first utilizes a depth-separable T2T-ViT network to extract rich facial features. Subsequently, it employs a multi-scale attention decomposition module to nonlinearly decouple age and identity features. The feature decomposition is jointly constrained by mutual information minimization, cross-entropy, and the Arcface function. The model achieves accuracy rates of 94.97%, 99.51%, and 95.81% on three benchmark datasets: FG-NET, CACD_VS, and CALFW, respectively, matching or surpassing the state-of-the-art (SOTA) performance. These results indicate that the proposed model can extract robust facial information and efficiently decouple features, achieving advanced recognition performance.

Keywords: Cross-Age; Transformer; network; Multi-Scale Attention

تعزيز التعرف على الوجه عبر الأعمار باستخدام شبكات T2T-ViT وتحليل الانتباه متعدد الأحجام

زكريا مطهر الحائر*، لي منغ شيا
جامعة اليانغتسي | 1 طريق نانهوان | جينغتشو ، مقاطعة هوبي | 434023 | الصين

المستخلص: يقدم هذا البحث نموذجًا للتعرف على الوجوه عبر الأعمار يدمج شبكات الالتفاف العصبية (CNN) مع المحولات (Transformers). يستخدم النموذج في البداية شبكة T2T-ViT القابلة للفصل بالعمق لاستخراج ميزات الوجه الغنية. بعد ذلك، يستخدم وحدة تحليل الانتباه متعددة الأحجام لفك الترابط اللاخطي بين ميزات العمر والهوية. يتم تقييد تحليل الميزات مشتركًا بواسطة تقليل المعلومات المتبادلة، والتقاطع الإنتروبي، ووظيفة Arcface. يحقق النموذج معدلات دقة تبلغ 94.97%، و99.51%، و95.81% على ثلاث مجموعات بيانات معيارية: FG-NET، وCACD_VS، وCALFW، على التوالي، متطابقة أو تفوق أداء الحالة الفنية (SOTA). تشير هذه النتائج إلى أن النموذج المقترح يمكنه استخراج معلومات وجهية قوية وفك الترابط بين الميزات بكفاءة، محققًا أداء تعرف متقدمًا.

الكلمات المفتاحية: عبر الأعمار؛ المحولات؛ شبكة T2T-ViT؛ الانتباه متعدد الأحجام.

1.Introduction

Facial recognition technology has become a widely adopted method for identity verification and security assurance in modern society. However, age-related factors remain a significant bottleneck in the field of facial recognition. As individuals age, their facial characteristics undergo nonlinear changes, leading to high intra-class variability and inter-class similarity, which poses challenges to the accuracy and stability of facial recognition technologies. Deep learning currently stands as one of the main approaches for cross-age facial recognition. The Transformer, as an emerging deep learning model, possesses rapid inference capabilities and powerful feature extraction capacities, making it adept at capturing key features across different identities for use in cross-age facial recognition. However, Transformer-based models for cross-age facial recognition still face challenges in adequately expressing local, low-level features and in thoroughly decomposing features.

To address the Transformer's deficiency in capturing local, low-level feature information, this paper integrates depth wise separable convolution (DSC) into the T2T-ViT Transformer architecture, creating an efficient and straightforward depth wise separable T2T-ViT network (DST2T-ViT). This model combines the strengths of Convolutional Neural Networks (CNNs) in extracting low-level features and enhancing locality with the Transformer's capability in establishing long-range dependencies. This integration aims to enrich the extraction of low-level features at a minimal additional computational cost.

Regarding the issue of incomplete decomposition of identity and age features, inspired by the attention mechanism's ability to adaptively focus on features relevant to the target task while suppressing irrelevant information, this paper proposes constructing a multi-scale attention decomposition module (MSADM) by concatenating improved channel and spatial attentions. This module employs multi-scale attention in both channel and spatial dimensions, allowing the network to selectively focus on age-related features to promote efficient feature decomposition. The mutual information (MI) metric is used to quantitatively measure the degree of decoupling between these features. By minimizing the MI between identity and age features, the decomposed features are constrained in terms of their relevance, capturing complete identity information.

1.2 Research problem

1- How can a facial recognition model effectively integrate Convolutional Neural Networks (CNNs) and Transformers to accurately recognize faces across different ages?
2- What methodologies can be utilized within a depth-separable T2T-ViT network to extract rich facial features essential for cross-age facial recognition?
3- How can a multi-scale attention decomposition module be employed to non-linearly decouple age and identity features in facial recognition models?
4- In what ways can mutual information minimization, cross-entropy, and the Arcface function jointly constrain feature decomposition to enhance the accuracy of cross-age facial recognition?
5- What are the performance metrics of the proposed model on benchmark datasets such as FG-NET, CACD_VS, and CALFW, and how do they compare to the state-of-the-art (SOTA) models in terms of accuracy rates?
6- How does the proposed model manage to achieve robust extraction and efficient decoupling of facial features to attain advanced recognition performance across a wide age range?

1.3 Research hypotheses

1. **Hypothesis 1:** A facial recognition model that integrates Convolutional Neural Networks (CNNs) with Transformers, specifically through a depth-separable T2T-ViT network, will significantly enhance the extraction of rich facial features, leading to improved accuracy in cross-age facial recognition.
2. **Hypothesis 2:** Employing a multi-scale attention decomposition module within the facial recognition model will effectively decouple age and identity features, thereby reducing the impact of age variation on recognition accuracy.
3. **Hypothesis 3:** The joint constraint of feature decomposition by mutual information minimization, cross-entropy, and the Arcface function will lead to a more precise separation of age-related and identity-related facial features, resulting in higher recognition performance across diverse age groups.

4. **Hypothesis 4:** The proposed facial recognition model, by integrating advanced neural network architectures and feature decomposition techniques, will achieve or surpass state-of-the-art (SOTA) performance on benchmark datasets (FG-NET, CACD_VS, and CALFW), demonstrating its efficacy in handling the variability of facial features across different ages.
5. **Hypothesis 5**: The advanced recognition performance of the proposed model, as indicated by high accuracy rates on benchmark datasets, suggests that it can robustly extract and efficiently decouple facial features, making it a superior solution for cross-age facial recognition challenges.

1.4 Research objectives

1. **Objective 1:** To develop a cross-age facial recognition model that effectively integrates Convolutional Neural Networks (CNNs) with Transformers to address the challenges of recognizing faces across a wide range of ages.
2. **Objective 2:** To utilize a depth-separable T2T-ViT network within the model to enhance the extraction of rich facial features critical for improving recognition accuracy across age variations.
3. **Objective 3:** To implement a multi-scale attention decomposition module in the model for the nonlinear decoupling of age and identity features, aiming to minimize the impact of age on recognition accuracy.
4. **Objective 4:** To apply a joint constraint mechanism involving mutual information minimization, cross-entropy, and the Arcface function for the effective separation of age-related and identity-related features within the facial recognition model.
5. **Objective 5:** To evaluate the model's performance on benchmark datasets (FG-NET, CACD_VS, and CALFW) to compare its accuracy rates with those of state-of-the-art (SOTA) models, aiming to match or surpass their performance.
6. **Objective 6:** To demonstrate that the proposed model can robustly extract facial information and efficiently decouple features, thereby achieving advanced recognition performance suitable for practical applications in cross-age facial recognition scenarios.

1.5 Importance of research

1. **Advancement in Facial Recognition Technology:** By integrating Convolutional Neural Networks (CNNs) with Transformers, the research represents a significant step forward in facial recognition technology, particularly in addressing the challenge of cross-age recognition. This advancement is crucial for applications where accurate identification is required regardless of age changes, such as security systems, personalized services, and age-invariant access controls.
2. **Innovation in Feature Extraction and Decomposition:** The use of a depth-separable T2T-ViT network for extracting rich facial features and a multi-scale attention decomposition module for decoupling age and identity features introduces innovative methods for handling the complex variability in facial characteristics across different ages. This approach enhances the model's ability to focus on identity-specific features while minimizing the influence of age-related changes.
3. **High Accuracy Across Diverse Datasets:** Achieving accuracy rates of 94.97%, 99.51%, and 95.81% on benchmark datasets FG-NET, CACD_VS, and CALFW, respectively, demonstrates the model's effectiveness across various datasets, showcasing its potential for widespread applicability. Matching or surpassing state-of-the-art performance validates the model's superiority and its significance in pushing the boundaries of what's achievable in facial recognition.
4. **Contribution to the Body of Knowledge:** The research contributes to the body of knowledge in the field of computer vision and artificial intelligence by proposing a novel integration of technologies and methodologies. This contribution not only provides a new tool for researchers and practitioners but also opens up avenues for further exploration in age-invariant facial recognition and beyond.
5. **Potential for Real-world Applications:** The model's ability to robustly extract facial information and efficiently decouple features has significant implications for real-world applications. From enhancing security and surveillance systems to improving user experience in consumer electronics and supporting forensic investigations, the implications of achieving advanced recognition performance are vast and varied.
6. **ddressing a Key Challenge in Biometrics:** Finally, the research addresses a fundamental challenge in biometrics: accurately recognizing individuals across different ages. This challenge has been a significant hurdle in deploying facial recognition systems that can maintain high accuracy over time. The proposed model's ability to overcome this hurdle is of paramount importance for the future development and deployment of reliable, age-invariant facial recognition systems.

2 Related Work

In this paper, the cross-age facial recognition is achieved through the DST2T-ViT network, which efficiently captures a wealth of initial facial features, combined with the use of the MSADM and MI minimization regularization algorithm to obtain robust identity characteristics. Therefore, the related work will be discussed from two aspects: facial feature extraction and facial feature decoupling.

2.1 Facial Feature Extraction

In recent years, research on cross-age facial recognition has made significant progress through the use of CNN-based models. Document [2] utilizes the ResNet network as both encoder and decoder, proposing a hybrid network capable of learning stable identity features while achieving realistic facial synthesis. Document [3] adds a pyramid feature fusion module to the ResNet network to learn effective features from multiple scales, aiming for robust feature extraction. Methods based on CNNs typically involve a high number of parameters and multiply-accumulate operations (MACs). Moreover, CNNs focus on modeling the relationships between adjacent pixels, which can leave gaps in the capture of global facial information. Alexey et al. [4] introduced the Transformer into computer vision tasks, proposing the vision Transformer (ViT) network model. Document [5] applies the T2T-ViT model to cross-age facial recognition tasks, addressing the issues of high complexity and computational cost associated with CNNs. While this approach shows promising performance in global information modeling, its effectiveness in extracting local information needs improvement. Therefore, this paper embeds CNN into the T2T-ViT model, utilizing depthwise separable convolution (DSC) to capture local information and T2T-ViT to grasp global facial features, thereby extracting a rich set of facial information.

2.2 Facial Feature Decoupling

To learn discriminative identity features, document [6] introduces a latent factor analysis algorithm that represents initial facial features as a linear combination of age components, identity components, and noise. This approach mitigates the impact of age factors on recognition. Document [7] employs two parallel fully connected layers to learn identity and age features from deep features and introduces a direct sum module to eliminate redundant features in the age and identity subspaces. Document [8] utilizes a linear canonical correlation analysis module to obtain age features and introduces a decorrelation adversarial learning algorithm to reduce the correlation between them. Considering the nonlinear relationships between feature vectors, document [9] uses a channel attention block to nonlinearly decompose facial features in the high-level semantic feature space, aiming to learn robust identity features. Documents [10-11] incorporate spatial attention mechanisms into the feature decomposition module, assigning different attention weights at spatial and channel levels to enhance the representation of age features. Spatial attention maps, calculated by compressing channels, tend to distribute spatial attention weights uniformly across channels, resulting in the extraction of identity features that include age characteristics. This paper employs multi-scale depthwise separable convolutions to construct spatial attention, calculating spatial attention maps for each channel individually. This promotes a dynamic distribution of attention weights in the spatial dimension, thereby learning robust identity features.

3. Overall Framework

The primary task during cross-age face recognition is to extract identity features that are not affected by age factors and are complete. The overall framework proposed in this paper is illustrated in Figure 1, comprising three main components: the DST2T-ViT network, MSADM, and a multi-task training module consisting of MI estimator, identity discriminator, and age discriminator. Through continual training and parameter optimization, the identity and age features are thoroughly decoupled. Finally, the optimized model is utilized to extract identity features, which are then compared with feature vectors in the database using cosine distance calculation to achieve cross-age face recognition.

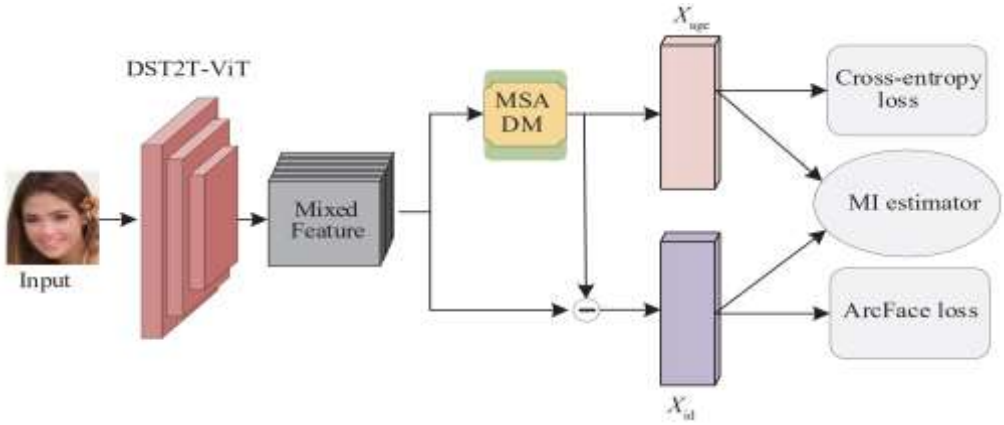


Fig. 1 Overall framework diagram

3.1 DST2T-ViT Feature Extraction Network

In this study, the fusion of DSC and T2T-ViT is employed to design the DST2T-ViT feature extraction network, as illustrated in Figure 2. This network primarily comprises the patches embedding module (Patch embedding), Transformer layers, and Depthwise Separable Convolution (DSC) Block. The patches embedding module is constructed with convolutional and pooling layers, leveraging the advantages of CNN in extracting low-level features. It extracts patches from the feature map, overcoming the Transformer's limitation in modeling local information. Specifically, a convolutional layer with a kernel size of 7 and a stride of 2 is utilized to extract shallow local facial features, generating feature maps with 32 channels. Subsequently, a BatchNorm layer stabilizes the model training, followed by a max-pooling layer with a kernel size of 3 and a stride of 2 to compress the feature maps, resulting in feature maps that are four times smaller than the input image, facilitating the model to learn more detailed features.

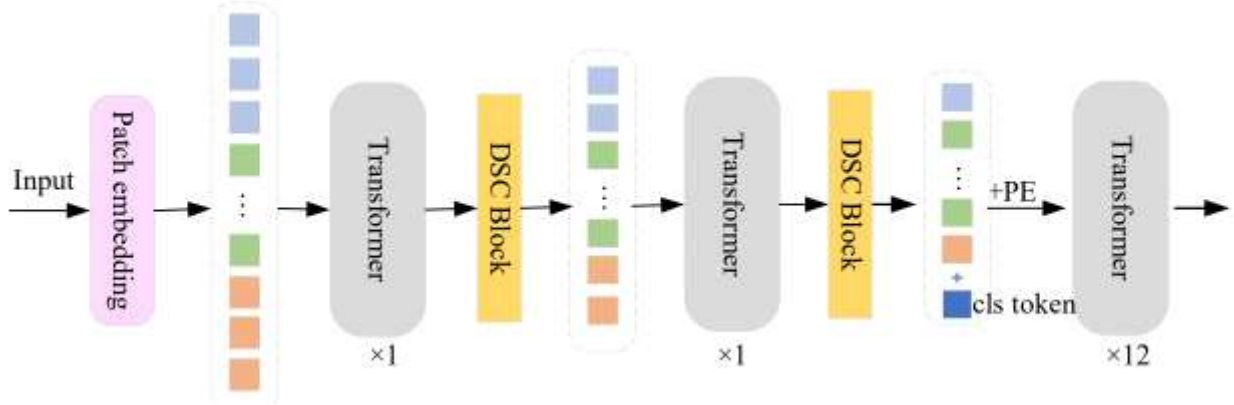


Fig. 2 Diagram of the DST2T-ViT network architecture

The Transformer layer, a fundamental unit of the Vision Transformer (ViT) model, comprises two sub-layers: the multi-head attention (MHA) mechanism and the multilayer perceptron (MLP). It employs residual connections around each sub-layer and is utilized for modeling the contextual information of faces. The multi-head attention sub-layer leverages multiple sets of attention weights to learn various semantic information. For an attention sub-layer with h h heads, the input features are transformed into Query, Key, and Value vectors through linear transformations. The computation formula for the attention mechanism is as follows:

$$q^{(l,i)} = W_Q^{(l,i)} L_{\mathrm{N}}\left(x^{(l-1)}\right) \in R^{D_k}, \quad (1)$$

$$k^{(l,i)} = W_K^{(l,i)} L_{\mathrm{N}}\left(x^{(l-1)}\right) \in R^{D_h}, \quad (2)$$

$$v^{(l,i)} = W_V^{(l,i)} L_{\mathrm{N}}\left(x^{(l-1)}\right) \in R^{D_h}, \quad (3)$$

Among these, $l \in \{1, \cdots, L\}$ represents the number of Transformer layers, $i \in \{1, \cdots, h\}$ represents the number of heads, L_{N} denotes a linear transformation, different l have different weight parameters, $D_h = \frac{D}{h)}$ represents the dimension of each

attention head. Then, for different heads $q^{(l,i)}, k^{(l,i)}, v^{(l,i)}$, scaled dot-product attention is computed in parallel. Finally, the results of the scaled dot-product attention are concatenated and projected again as the final output. The computation process is as follows:

$$\text{head } i = \text{Attention}\left(\boldsymbol{q}^{(l,i)}, \boldsymbol{k}^{(l,i)}, \boldsymbol{v}^{(l,i)}\right) = \sigma\left(\frac{\boldsymbol{q}^{(l,i)}\boldsymbol{k}^{(l,i)}}{\sqrt{D_h}}\right)\boldsymbol{v}^{(l,i)}, \quad (4)$$

Here, σ denotes the activation function, enhancing the nonlinear relationships between features. The MLP (multilayer perceptron) layer maps the data to different dimensional spaces through two fully connected layers and the GeLU (Gaussian Error Linear Unit) activation function, learning more abstract features of the face. Residual connections are used around both sublayers to prevent information loss.

The depthwise separable convolution module is an efficient convolution operation that decomposes traditional convolution operations into channel and spatial dimensions. It consists of depthwise convolution and pointwise convolution. In both DSC (Depthwise Separable Convolution) Blocks, depthwise convolutions with a kernel size of 3 and stride of 2 aggregate local spatial information on the channels, scaling the feature map size down to half. Then, multiple 1×1 convolutions perform pointwise linear combinations on the feature maps of each channel, fusing information between channels. This module effectively reduces the length of the token sequence while expanding the channel capacity.

3.2 Multi-Scale Attention Decomposition Module

To minimize the loss of identity features during the feature decomposition process, the Multi-Scale Attention Decomposition Module (MSADM) was constructed to nonlinearly decompose mixed facial features in high-level semantic spaces. MSADM is primarily divided into the Improved Channel Attention (ICA) and Multi-Scale Spatial Attention (MSSA) modules. ICA allows the network to selectively focus on age-related objects, while MSSA concentrates on important spatial regions. Through dynamically distributed attention weights in both dimensions, MSADM learns age features to promote efficient feature decomposition, as illustrated in Figure 3.

3.2.1 Improved Channel Attention

The ICA module combines global average pooling and max pooling in parallel, using average pooling to maintain the invariance of global information and max pooling to emphasize attention to key channels. It introduces a learnable parameter \(\alpha \) to weight the features on the pooling channels, enhancing the selection of effective features on the channels. To overcome the issue of partial information loss during channel interactions, ICA incorporates a one-dimensional fast convolution to facilitate cross-channel local information exchange, thereby strengthening the representational capability of the feature map. The output expression of a feature map \(X_{in} \) after passing through the ICA module is as follows:

$$\boldsymbol{X}_{\text{age}} = \boldsymbol{X}_{\text{in}} \otimes \sigma\{\text{Conv}\,1[F_{\text{GAP}}(\boldsymbol{X}_{\text{in}})] \times \alpha + \text{Conv}\,2[F_{\text{GMP}}(\boldsymbol{X}_{\text{in}})] \times (1-\alpha)\}, \quad (5)$$

Where X_{age} represents the age features, and X_{in} represents the initial facial features. F_{GMP} and F_{GAP} respectively denote global max pooling and global average pooling. $Conv1$ and $Conv2$ refer to two 1D convolutions with a kernel size of 5. α represents a learnable parameter, and $\otimes$ signifies the element-wise multiplication of tensors.

3.2.2 Multi-Scale Spatial Attention (MSSA):

The MSSA module assigns different weights to each spatial position based on the contribution of different spatial regions to the age classification outcome, highlighting key age structure features in the feature map as a supplement to channel attention. To dynamically adjust the weights of spatial attention on each channel dimension, this paper utilizes depthwise convolution to capture spatial relationships between features, ensuring the preservation of inter-channel relationships while reducing computational complexity. A multi-scale structure enhances the convolution operation's ability to capture spatial relationships. Channel mixing is performed by 1×1 convolution, thus generating more refined attention maps. The output expression of the MSSA module is as follows.

$$\boldsymbol{X}'_{\text{age}} = \boldsymbol{X}_{\text{age}} \otimes \text{Conv}_{1\times1}\left(\sum_{i=0}^{3} \text{Branch}_i\left(\text{DConv}\left(\boldsymbol{X}_{\text{age}}\right)\right)\right) \quad (6)$$

Where $\boldsymbol{X}'_{\text{age}}$ represents the age features output by the MSSA module, and (X_{age}) represents the age features output by the ICA module. $DConv$ denotes depthwise convolution, and $Branch_i$ represents the (i) branch. In each branch, two depthwise strip convolutions are used to approximate a standard depthwise convolution with a large kernel. The kernel size for each channel varies to capture multi-scale information. This paper cascades this module with the ICA module, forming the multi-scale attention decomposition module, thereby promoting the effective selection of age features in a high-level semantic space.

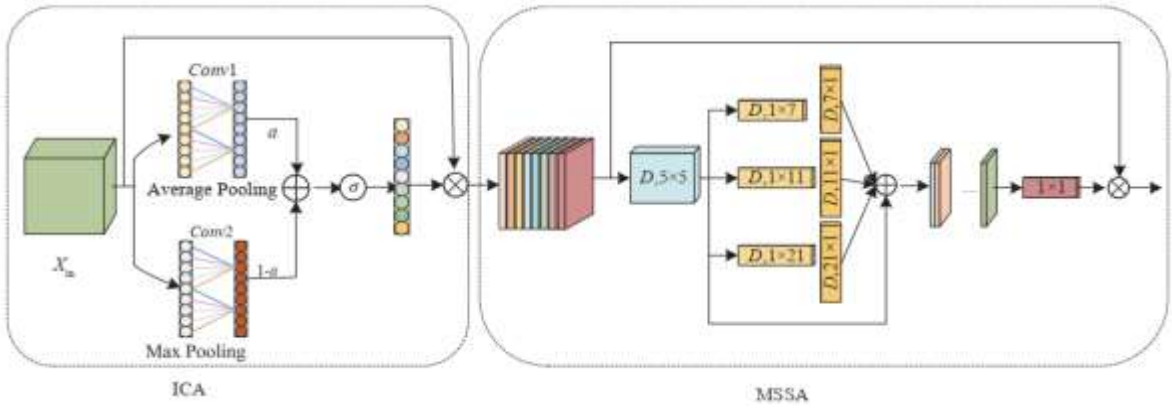


Fig. 3 Diagram of MSADM structure

3.3 Multi-Task Training

This paper employs multi-task training to constrain feature learning, incorporating three fundamental constraint modules: an identity discriminator, an age discriminator, and an MI (Mutual Information) estimator. For identity feature discrimination, the ArcFace function [12] is used to supervise the learning of identity features X_{id}. The ArcFace function is defined as

$$L_{\text{id}} = -\log \frac{\exp\left(s\cos\left(\theta_{y_i} + m\right)\right)}{\exp\left(s\cos\left(\theta_{y_i} + m\right)\right) + \sum_{j \neq y_i}^{n} \exp\left(s\cos\theta_j\right)} \tag{7}$$

Where n represents the number of individuals, s denotes the scaling factor, m is a constant margin term controlling the angle, y_i is the identity label for the i sample, and $\cos\theta_j$ represents the cosine of the angle between the i feature X_{id} and the weight vector of label y_j.

For the age discriminator, given that age labels come with inherent noise, following document [5], the age labels are divided into 8 non-overlapping age groups, treating these as age categories. The cross-entropy function is used to assess the discrepancy between the predicted age group and the true age group. The cross-entropy function is defined as

$$L_{\text{age}} = -\log \frac{e^{z_j}}{\sum_{j=1}^{N} e^{z_j}} = -z_i + \log\left(\sum_{j=1}^{N} e^{z_j}\right) \tag{8}$$

Where N represents the number of age groups, and z_i denotes the age group label corresponding to sample i.

The MI (Mutual Information) estimator is utilized to reduce the correlation between age features X_{age} and identity features X_{id}. The mutual information $I\left(X_{\text{age}}; X_{\text{id}}\right)$ between the given vectors X_{age} and X_{id} is defined as

$$I\left(\boldsymbol{X}_{\text{age}}; \boldsymbol{X}_{\text{id}}\right) = E_{p\left(X_{\text{age}}, X_{\text{id}}\right)}\left[\log \frac{p\left(\boldsymbol{X}_{\text{age}}, \boldsymbol{X}_{\text{id}}\right)}{p\left(\boldsymbol{X}_{\text{age}}\right) p\left(\boldsymbol{X}_{\text{id}}\right)}\right], \tag{9}$$

By minimizing $I\left(X_{\text{age}}; X_{\text{id}}\right)$, the network is trained to generate identity features that are insensitive to age. In the case of facial feature decomposition, the conditional distribution $p\left(X_{\text{age}} \middle| X_{\text{id}}\right)$ is not accessible, so $q_\psi\left(X_{\text{age}} \middle| X_{\text{id}}\right)$ is used to approximate $p\left(X_{\text{age}} \middle| X_{\text{id}}\right)$. For a given $sample\left(X_{\text{age}}, X_{\text{id}}\right)$, the MI minimization objective function [14] is defined as

$$L_{\mathrm{mi}} = \hat{I}_{\mathrm{CLUB}}(\boldsymbol{X}_{\mathrm{age}}; \boldsymbol{X}_{\mathrm{id}}) = \frac{1}{N}\sum_{i=1}^{N}\left[\log q_{\psi}(\boldsymbol{X}_{\mathrm{age}}^{i} \mid \boldsymbol{X}_{\mathrm{id}}^{i}) - \frac{1}{N}\sum_{j=1}^{N}\log q_{\psi}(\boldsymbol{X}_{\mathrm{age}}^{j} \mid \boldsymbol{X}_{\mathrm{id}}^{j})\right], \quad (10)$$

Where N represents the number of training samples. To make the upper bound more closely approximate the true value, it is constrained by maximizing the corresponding log-likelihood function, which is defined as

$$L_{\mathrm{ma}} = \frac{1}{N}\sum_{j=1}^{N}\log q_{\psi}(\boldsymbol{X}_{\mathrm{age}}^{i} \mid \boldsymbol{X}_{\mathrm{id}}^{i}) \quad (11)$$

Combining equations (7) and equations (8)-(10), the overall multi-task training objective function for the entire network is defined as

$$L = L_{\mathrm{id}} + \lambda_1 L_{\mathrm{age}} + \lambda_2 L_{\mathrm{mio}} \quad (12)$$

Where λ_1 and λ_2 represent the proportionality coefficients that balance the three loss functions.

4 Experiments

4.1 Experimental Setup

4.1.1 Data Preprocessing

Multi-task Cascaded Convolutional Networks (MTCNN) [15] are utilized for detecting facial regions and key points within face images. Similarity transformation is applied to five facial key points, and the input face images are cropped to 112×112 RGB images. Finally, the pixel values of the cropped face images are normalized by subtracting 127.5 and dividing by 128, as shown in Figure 4.

Fig. 4 Face alignment effect

4.1.2 Network Architecture

The backbone of this study employs a network structure similar to T2T-ViT-14, adopting a deep narrow structure with fewer hidden dimensions but more layers. Convolutional operations are used to scale down the feature maps three times in spatial dimensions, reducing them to 1/4, 1/8, and 1/16 of the original size, thereby modeling structural information while decreasing the length of the token sequence. The first two Transformer layers use only one Transformer layer to capture global information in the shallow features, with hidden size and MLP size both set to 64. The depth of the final Transformer layer is set to 14, with hidden size and MLP size being 384 and 1,152 respectively. This deep narrow structure design can reduce model complexity and enhance feature representation capability.

4.1.3 Training Details

A large-scale face dataset, Faces Emore [16], is employed to pre-train the network model, which is further fine-tuned on the cross-age face dataset CACD to achieve efficient training of the entire network. A pre-trained age estimation model [17] is used to estimate the age information of faces in the training dataset, resulting in 85,742 individuals with age-labeled data, totaling 5,774,205 face images. Age information is divided into eight groups: 0-12, >12-18, >18-25, >25-35, >35-45, >45-55, >55-65, >65. During model

pre-training, the hardware setup includes a single NVIDIA GeForce RTX 3090 card, and the model training is implemented using PyTorch version 1.8.1. The batch size is set to 512, the number of iterations is 25, and the stochastic gradient descent (SGD) method is utilized for optimizing model parameters, with an initial learning rate of 0.01. At iteration rounds 14, 18, and 22, the learning rate decays by a factor of 0.1 compared to the previous round, while the momentum factor is set to 0.9. The hyperparameters s and m in Equation (7) are set to 64 and 0.5 respectively. Through iterative experimentation and comparison, the balance coefficients $\lambda 1$ and $\lambda 2$ in Equation (12) are determined to be optimal at 0.1 and 0.01 respectively for the best recognition performance. The initial learning rate of the MI estimator is set to 1×10−5, and during training, the encoder forwards propagate once while the MI estimator is optimized five times.

4.2 Experimental Results Analysis

4.2.1 Analysis of FG-NET Dataset Experiment Results

FG-NET is one of the most popular face aging datasets for cross-age face recognition, comprising 1,002 mixed-color and grayscale facial images of 82 individuals collected by scanning photos of individuals aged 0 to 69 years. Following the protocols established in literature [6, 11], cross-validation is conducted using a leave-one-out strategy. Specifically, one image is chosen as the test data while the remaining 1,001 facial images are used for model fine-tuning. This process is repeated 1,002 times, and the average rank-1 recognition rate is reported. This evaluation strategy effectively reflects the performance of the recognition model since each participant in the dataset has multiple facial images at different ages.

Tab. 1 Comparison results of different methods on FGNET dataset

Methods	Acc/%
HFA[61	69
LF-CNN[181	88.1
AA-CNN[151	89.34
AD-CNN[111	90
DAL[81	94.5
MTLFace[21	94.78
Methods of this article	94.97

The comparison of the proposed method with existing cross-age face recognition methods on the FG-NET dataset is shown in Table 1. From Table 1, it can be observed that the proposed method achieves higher accuracy compared to other methods, with a recognition accuracy of 94.97%, surpassing the current state-of-the-art method by 0.19%.

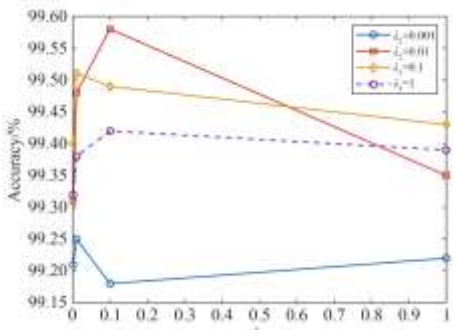
Fig. 5 Incorrectly retrieved face images

Visualization of facial images with retrieval failures is presented in Figure 5. The failed retrieval images mainly occur in infants and children aged 0-12 years. Since the pre-training dataset, Faces Emore, contains a relatively small proportion of underage facial images, and even the CACD dataset used for fine-tuning the model does not include facial images of individuals aged 0-12, there are limitations in attempting to learn the potential distribution of this specific age group through data-driven methods.

4.2.2 Analysis of CACD-VS Dataset Experiment Results

CACD-VS consists of 163,446 facial images of 2,000 celebrities aged 16 to 62 years, sourced from various lighting conditions, different poses, and makeup effects on the internet. It effectively reflects the robustness of cross-age face recognition algorithms. CACD-

VS is a subset of CACD, containing 4,000 pairs of face images for face verification, including 2,000 pairs of positive samples and 2,000 pairs of negative samples. This study strictly follows the experimental setup in literature [18] for evaluation on the CACD-VS dataset. Considering that the hyperparameters λ1 and λ2 in Equation (12) affect model performance, λ1 and λ2 are set to {1, 0.1, 0.01, 0.001} for validation on the CACD-VS dataset to explore their reasonable values. Figure 6 illustrates the accuracy of face verification for different



The accuracy curve of face verification with different λ1 and λ2 values

values, indicating that the model achieves optimal performance when λ1=0.1 and λ2=0.01.

Tab. 2 Comparison results of different methods on CACD-VS dataset

Methods	Acc/ %	AUC/%
HFA[61	84.4	88.8
LF-CNN[181	98.5	99.3
DAL[151	98.95	99.6
FSDS-CNN[61	99.2	99.7
T2T-DAL[51	99.35	—
Methods of this article	99.51	99.7

Table 2 presents the comparison of the proposed method with existing methods in terms of Acc and AUC. From Table 2, it can be observed that the proposed method is not inferior to existing models in both evaluation metrics, with an accuracy of 99.51%, surpassing the highest existing model by 0.16%. This demonstrates the superiority of the proposed method in terms of robustness.

4.2.3 Analysis of CALFW Dataset Experiment Results

The CALFW dataset is specifically designed for unconstrained face verification with significant age differences, comprising 12,176 facial images of 4,025 individuals. Each individual has at least two images, with 600 pairs of positive sample images with the same age difference and 600 pairs of negative sample images with the same gender and different races selected. The performance of the proposed method is evaluated using the accuracy (Acc) and Equal Error Rate (EER) indicators.

Tab. 3 Comparison results of different methods on CALFW dataset

Methods	Acc/%	EER/%
LF-CNN[191	90.7	—
ArcFace[121	95.45	—
DAL[81	95.48	12.17
MTLFace[21	95.62	12.05
FSDS-CNN[61	—	10.3
Methods of this article	95.81	11.02

As shown in Table 3, the proposed method achieves a recognition accuracy of 95.81% on the CALFW dataset, setting a new record for the CALFW dataset. Since this dataset lacks age information and the model's training and fine-tuning processes do not involve its parameters, experimental evaluation is conducted on this dataset, fully verifying the superior generalization ability of the proposed method.

4.2.4 Analysis of Ablation Experiment Results

To demonstrate the effectiveness of the proposed modules, four sets of comparative models are designed on the FG-NET, CACD-VS, and CALFW datasets, following the parameter settings outlined earlier.

1. Baseline1: Directly utilizes the T2T-ViT network to extract initial facial features, with the ArcFace function employed as the identity discriminator for supervised training.
2. Baseline2: Incorporates DSC into the T2T-ViT network to extract features.
3. Baseline3: Improves the T2T-ViT network by introducing MSADM, which nonlinearly decomposes initial facial features in a high-level semantic space, with age features constrained by the cross-entropy loss function.
4. Our: The model proposed in this paper, built upon Baseline3, incorporates the MI regularization algorithm to remove correlations between identity and age. The MI estimator and identity/age discriminators are simultaneously trained.

Tab. 4 Accuracy of different module fusion on cross-age face dataset

Model	FG-NET	CACD-VS		CALFW	
	Acc/%	Acc/ %	AUC/%	Acc/%	EER/%
Baseline1	**92.56**	**98.53**	**98.2**	**94.21**	**13.10**
Baseline2	**93.21**	**99.02**	**98.4**	**94.92**	**12.63**
Baseline3	**94.91**	**99.47**	**99.34**	**95.73**	**11.38**
Our	**94.97**	**99.51**	**99.70**	**95.81**	**11.02**

As shown in Table 4, Baseline1 simply employs the traditional T2T-ViT network to extract identity features for recognition, resulting in poor recognition performance across the three datasets. By embedding the DSC module into the T2T-ViT network, the recognition accuracy on the three datasets improves by 0.65%, 0.49%, and 0.71% respectively, indicating that DSC can compensate for the deficiency of the Transformer model in representing low-level local features. Baseline3, with the addition of the feature decomposition module and age loss function constraint, shows certain improvement in recognition performance, validating that the MSADM module can highlight age-related information and effectively reduce the interference of age factors on identity recognition. The proposed method in this paper, built upon Baseline3 and incorporating the MI discriminator to constrain identity and age feature decomposition, significantly improves model performance, demonstrating the strong robustness of our method to age variations.

5.Conclusion

In this study, a method based on multi-task learning is proposed, utilizing the DST2T-ViT network to extract facial features. This network embeds the DSC module into the T2T-ViT network to capture more local low-level feature information. To capture comprehensive identity information, MSADM is designed to nonlinearly decouple facial features in a high-level semantic space, while the MI minimization algorithm constrains the relationship between age and identity features to achieve efficient and precise feature decomposition. The excellent experimental results on three benchmark datasets demonstrate the model's advancement in recognition performance.

However, through experimentation, it was found that the lack of underage facial images in publicly available benchmark datasets hinders the model from fully learning and representing the unique features of underage faces. This results in a decrease in accuracy when it comes to underage face recognition. Therefore, the learning of underage facial features will be a focus of future research efforts.

Recommendations

Here we offer some recommendations to Enhancing Cross-Age Facial Recognition

1. Expanding Dataset Diversity for Underage Facial Features

Collect and integrate a more diverse set of facial images for individuals aged 0-12 years into the training datasets. This could involve creating a specialized dataset focusing on underage individuals to improve the model's learning and representation of this specific age group's unique facial features.

2. Investigating Additional Attention Mechanisms

Explore the integration of newer or different attention mechanisms within the Transformer architecture to further refine the model's ability to decouple age and identity features. Techniques such as self-attention or cross-attention might offer additional benefits in focusing on relevant features and suppressing irrelevant ones.

3. Enhancing Local Feature Extraction Capabilities

Develop and test additional methods or modules for enhancing the extraction of local, low-level features. This might involve experimenting with alternative convolutional structures or incorporating recent advancements in neural network architectures that focus on capturing finer details.

4. Exploring Generative Models for Data Augmentation

Utilize generative adversarial networks (GANs) or variational autoencoders (VAEs) to generate synthetic facial images across a wide range of ages. This could help address data scarcity issues, particularly for underrepresented age groups, and potentially improve the robustness of the model to age-related variations.

5. Integrating Ethical and Bias Mitigation Considerations

Implement a systematic evaluation of the model's performance across different demographics beyond age, such as ethnicity, gender, and facial expressions. This involves developing metrics and methodologies for assessing and mitigating biases in the model to ensure fair and equitable performance across all user groups.

6. Cross-Modal Feature Extraction

Investigate the feasibility of incorporating cross-modal data, such as voice or textual descriptions, to aid in the facial recognition process. This approach could leverage additional context that may be relevant for identifying individuals across significant age gaps.

7. Real-World Deployment and Feedback Loop

Pilot the model in real-world applications, such as security systems, personal identification, and age-invariant access controls. Collect feedback and performance data to continuously refine and adapt the model based on practical challenges and operational requirements.

8. Computational Efficiency and Deployment

Focus on optimizing the model for computational efficiency and ease of deployment on a variety of platforms, including mobile devices and cloud-based systems. This may involve quantization, pruning, and other model compression techniques to reduce computational resource requirements while maintaining high accuracy.

9. Exploring Fusion with Other Biometric Modalities

Examine the integration of facial recognition with other biometric modalities, such as fingerprint or iris recognition, to enhance overall identification accuracy and security. This multimodal approach could be particularly effective in scenarios where facial features alone may not provide sufficient identification certainty.

10. Continuous Learning and Update Mechanism

Develop a mechanism for continuous learning where the model can be periodically updated with new data without needing complete retraining. This approach would allow the model to adapt over time to changes in appearance and new data trends, maintaining its relevance and accuracy.

Implementing these recommendations could significantly advance the state of cross-age facial recognition technology, addressing current limitations and opening new avenues for research and application.

References:

[1] YUAN L, CHEN Y, WANG T, et al. Tokens-to-token vit : Training vision transformers from scratch on ImageNet [C] //2021 IEEE/CVF International Conference on Computer Vision (ICCV) . Los Angeles : IEEE Computer Society, 2021 : 538-547.

[2] HUANG Z, ZHANG J, SHAN H. When age-invariant face recognition meets face age synthesis : A multi-task learning framework J] . arXiv preprint arXiv, 2021 : 2103. 01520.

[3] HUANG Y, HU H. A parallel architecture of age adver-sarial convolutional neural network for cross-age face recognitio [J] .IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31 (1) : 148-159.

[4] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words：Transformers for image recognition at scale［J］. arXiv preprint arXiv, 2020：2010. 11929.

[5] 刘成，曹良才，靳业，等．一种基于 Transformer 的跨年龄人脸识别方法［J］．激光与光电子进展，2022，60（10）：210-215.

[6] GONG D, LI Z, LIN D, et al. Hidden factor analysis for age invariant face recognition［C］//2013 IEEE International Conference on Computer Vision（ICCV）. Los Angeles：IEEE Computer Society, 2013：2872-2879.

[7] 叶继华，郭祺玥，江爱文，等．基于特征子空间直和的跨年龄人脸识别方法［J］．郑州大学学报（工学版），2021，42（5）：7-12.

[8] WANG H, GONG D, LI Z, et al. Decorrelated adversarial learning for age-invariant face recognition［J］. arXiv preprint arXiv, 2019：1904. 04972.

[9] LI S, LEE H J. Effective attention-based feature decomposition for cross-age face recognition［J］. Applied Sciences, 2022, 12（10）：4816.

[10] 孙文斌，王荣，孙连烛，等．基于深度学习的跨年龄人脸识别［J］．激光与光电子学进展，2022，59（2）：340-349.

[11] 何星辰，郭勇，李奇龙，等．基于深度学习的抗年龄干扰人脸识别［J］．自动化学报，2022，48（3）：877-886.

[12] DENG J, GUO J, YANG J, et al. Arcface：Additive angular margin loss for deep face recognition［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44（10）：5962-5979.

[13] CHENG P, HAO W, DAI S, et al. Club：A contrastive log-ratio upper bound of mutual information［J］. arXiv preprint arXiv, 2020：2006. 12013.

[14] HOU X, LI Y, WANG S. Disentangled representation for age-invariant face recognition：A mutual information minimization perspective［C］//2021 IEEE/CVF International Conference on Computer Vision（ICCV）. Los Angeles：IEEE Computer Society, 2021：3672-3681.

[15] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks［J］. IEEE signal processing letters, 2016, 23（10）：1499-1503.

[16] KEMELMACHER-SHLIZERMAN I, SEITZ S, MILLER D, et al. The megaFace benchmark：1 million faces for recognition at scale［J］. arXiv preprint arXiv, 2015：1512. 00596.

[17] ROTHE R, TIMOFTE R, VAN GOOL L. Dex：Deep expectation of apparent age from a single image［C］//2015 IEEE International Conference on Computer Vision Workshop（ICCVW）. Los Angeles：IEEE Computer Society, 2015：252-257.

[18] LI H, ZOU H, HU H. Modified hidden factor analysis for cross-age face recognition［J］. IEEE Signal Processing Letters, 2017, 24（4）：465-469.