Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Facial Age Estimation Using Hybrid Architecture: Vision Transformers and ResNet50 with Mixup Data Augmentation

Ahmed Chaouki Chami[1*], Riadh Ajgou[1,2], Abdelmalik Taleb-Ahmed[3]

1.LGEERE Laboratory Faculty of TechnologyUniversity of El Oued,39000 El Oued, Algeria.

2.Department of Electronics and Automation University of Biskra, Algeria.

3.Institute of Electronics, Microelectronics and Nanotechnology (IEMN), Universite Polytechnique Hauts de France, Universit ´ e de Lille, ´ Centre National de la Recherche Scientifique (CNRS), Valenciennes, France,

* Corresponding Author:chami-ahmedchaouki@univ-eloued.dz.

ARTICLE INFO

Received: 10 May 2025

Accepted: 10 Feb 2026

Published: 01 May 2026

ABSTRACT

Facial age classification is an intrinsically hard problem in computer vision, due to the gradual, continuous nature of ageing, the huge inter-subject variability and the visual ambiguity at the boundary of adjacent age classes. Despite recent advances in deep learning, most existing methods are restricted to either convolutional neural networks or transformer-based architectures and cannot simultaneously capture fine-grained local facial texture cues and long-range global contextual dependencies, which are critical for accurate age-class discrimination. In this work, we propose a novel adaptive fusion architecture with dual branches to jointly integrate a pre-trained ResNet50 backbone for local feature extraction and Vision Transformer for global contextual modelling via an Adaptive Feature Fusion Module with a channel-wise attention mechanism. Our approach is motivated by the observation that age-related changes in faces exist at multiple scales, from fine local texture patterns such as wrinkles and skin degradation, to holistic structural changes across the whole face. To improve generalisation across ambiguous age class boundaries we further augment the training data with Mixup, a technique known to improve generalisation. We performed extensive experiments on three benchmark datasets, MORPH II, UTKFace, and UAGD, each with different characteristics. Our model achieves an MAE of 3.42 on MORPH II, 4.42 on UTKFace, and 5.63 on UAGD, consistently outperforming classical methods such as OR-CNN, CORAL, Ranking-CNN, and CSCS-Swin. The ablation studies verify that each component of the architecture is important to the final performance. These results show that the complementary integration of convolutional local descriptors and transformer global representations with adaptive attention-based fusion is a robust and generalisable solution for facial age classification on controlled, in-the-wild and uniformly distributed benchmarks.

Keywords: Facial age estimation, ResNet50, Vision Transformer, adaptive feature fusion, attention mechanism, Mixup augmentation, deep learning.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Introduction

Facial age estimation the computational task of inferring a subject's chronological age or age class from a single facial image has emerged as one of the most actively researched problems in computer vision and biometric analysis over the past two decades. Age estimation from facial images is an exciting and challenging task, where traits derived from facial images are used to determine age, gender, ethnic background, and emotional state. Its practical relevance spans an exceptionally broad range of real-world applications, including video surveillance, access control, forensic investigation, customer profiling, human-computer interaction, social media analysis, and clinical demographic statistics [1][2][3]. Despite sustained research efforts, accurate facial age estimation remains a fundamentally open problem due to the complex and non-stationary nature of the aging process, which is simultaneously influenced by genetic factors, gender, ethnicity, lifestyle, illumination conditions, image quality, facial expression, and pose variation [4]

Early approaches to facial age estimation relied on handcrafted feature representations derived from aging-pattern analysis, such as anthropometric measurements, active appearance models, and aging manifold learning techniques, which were subsequently fed into classical classifiers or regression models.[5]While these methods established the foundational understanding of age-related facial changes, their performance under real-world unconstrained conditions remained limited. With the rapid development of deep learning in recent years, deep learning-based facial age estimation methods have significantly improved the accuracy and robustness of face-based age estimation, especially under unconstrained conditions. Modern facial age estimation methods typically follow two main directions: improving the learning capacity of the neural network architecture itself, or incorporating auxiliary age-related features to support network training.

Within the deep learning paradigm, three dominant methodological families have emerged. Regression-based methods directly map facial features to a continuous age value, while classification-based approaches decompose the problem into discrete age group prediction. Ordinal regression methods, which treat age estimation as a series of binary classification subproblems exploiting the inherent ordering of age labels, have gained considerable traction as they are theoretically more consistent with the continuous and monotonic nature of aging than naive multi-class classification. Label distribution learning (LDL) methods, which assign a probability distribution over neighboring age labels rather than a single discrete label, have emerged as state-of-the-art approaches for handling the ambiguity inherent at age boundaries, where faces from adjacent ages of the same individual are often visually indistinguishable. Methods such as Ranking-CNN [6], DLDL [7], and BridgeNet[8] have achieved competitive performance on standard benchmarks by jointly exploiting ordinal and distributional information. However, as convolutional networks have continued to improve, the potential of CNN-based facial age estimation models has been increasingly exploited, and the growing number of network parameters has raised the cost of training while yielding diminishing returns.

Vision Transformer models have recently gained significant importance in computer vision tasks due to their self-attention mechanisms, while CNNs have dominated the field by achieving remarkable results across various applications. The introduction of the Vision Transformer (ViT)[9] marked a paradigm shift by demonstrating that pure attention-based architectures, without any convolutional inductive bias, can match or surpass CNNs on image recognition tasks when trained on sufficiently large datasets. Subsequent works such as Swin Transformer [10] have further validated the effectiveness of hierarchical transformer architectures for dense visual prediction tasks. In the context of facial age estimation, transformer-based models have shown promise in capturing long-range dependencies and holistic facial structural relationships that CNNs inherently struggle to model due to their local receptive field constraints [11]. Nevertheless, purely transformer-based approaches tend to overlook the fine-grained local texture cues such as wrinkles, skin degradation, and localized

Research Article

morphological changes that are critically important for distinguishing between neighboring age classes.

Despite the complementary nature of CNN local feature extraction and transformer global context modeling, the combined effectiveness of deep residual networks and Vision Transformer models in estimating human chronological age has not been thoroughly examined. Most existing hybrid methods either apply simple concatenation or fixed-weight fusion strategies that fail to adaptively emphasize the most discriminative feature dimensions, and few works have explicitly addressed the challenge of age class boundary ambiguity through augmentation-based regularization strategies during training. Furthermore, the majority of existing methods are evaluated exclusively on well-established and demographically skewed datasets such as MORPH II, where the natural over-representation of middle-aged subjects allows models to exploit distributional bias rather than learning genuinely robust age-discriminative representations. Evaluation on uniformly distributed and in-the-wild benchmarks, which more faithfully reflect the challenges of real-world deployment, remains rare and insufficiently explored in the literature.

To address these identified gaps, we propose a novel dual-branch deep learning architecture that jointly integrates a pre-trained ResNet50 backbone and a Vision Transformer through an Adaptive Feature Fusion Module (AFFM) equipped with a channel-wise attention mechanism, enabling the model to simultaneously capture fine-grained local facial cues and long-range global contextual dependencies for robust facial age classification. To further improve model generalization across inherently ambiguous age class boundaries, we incorporate the Mixup data augmentation strategy [12] during training, which has demonstrated consistent improvements in prediction robustness across diverse classification tasks. We evaluate our proposed framework on three benchmark datasets of markedly different characteristics: the controlled longitudinal MORPH II [13], the uniformly distributed UAGD [14], and the large-scale in-the-wild UTKFace[15], providing a comprehensive and multi-faceted assessment of model robustness and generalization ability.

The main contributions of this paper are as follows.

- We propose a dual-branch CNN-Transformer architecture that combines the complementary strengths of ResNet50 local feature extraction and ViT global context modeling through an AFFM with channel-wise attention-based adaptive weighting.
- We incorporate Mixup augmentation as a principled regularization strategy specifically tailored to the ordinal and ambiguous nature of age class boundaries.
- We conduct comprehensive experiments across three diverse benchmarks, including one of the first independent evaluations reported on the UAGD dataset, demonstrating consistent and competitive performance across controlled, uniformly distributed, and in-the-wild evaluation scenarios.
- We assess the performance of our proposed model by comparing it against single-branch baselines and state-of-the-art methods, demonstrating its robustness and generalizability for facial age classification across diverse real-world conditions.

The remainder of this paper is organized as follows. Section II describes the proposed dual-branch architecture and its key components in detail. Section III presents the experimental setup, datasets, and evaluation metrics. Section IV reports and analyzes the experimental results. Section V concludes the paper and outlines future research directions.

Proposed Methodology

In this section, we present a novel parallel dual-branch architecture designed to jointly capture fine-grained local facial features and long-range global contextual representations for robust facial age

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

group classification. Rather than relying on a single backbone, the proposed framework exploits the complementary strengths of two well-established deep learning architectures: ResNet-50[16], which excels at extracting hierarchical local texture and structural features through residual convolutional learning, and the Vision Transformer (ViT), which models long-range spatial dependencies across the entire facial region through multi-head self-attention over sequences of image patches. The representations produced by both branches are integrated through an Adaptive Feature Fusion Module (AFFM) equipped with a channel-wise attention mechanism that selectively weights the contribution of each feature source based on its relevance to the current input, yielding an interaction-aware fused descriptor that neither branch nor any non-interactive fusion strategy could produce in isolation. To further strengthen the generalisation capacity of the model and mitigate overfitting across the ambiguous and continuous age group boundaries, we incorporate Mixup augmentation into the training pipeline, generating synthetic samples by linearly interpolating between pairs of facial images and their corresponding age group labels and thereby encouraging the model to learn smoother and more calibrated decision boundaries across the age continuum. The proposed framework consists of four key components: (i) a ResNet-50 local feature extraction branch, (ii) a ViT global context modelling branch, (iii) an Adaptive Feature Fusion Module (AFFM) with channel-wise attention-based adaptive weighting, and (iv) a Mixup-regularised training strategy. The overall architecture is illustrated in Figure.1 and the specific design of each component is described in detail in the following subsections.

I.1. Overall architecture:

The overall structure of our proposed framework is shown in Figure. 1 and consists of three main stages. The input facial image of size 224×224×3 is first fed simultaneously into two parallel feature extraction branches: a pre-trained ResNet50 backbone that extracts a 2048-dimensional local feature vector
$F_R \in R^{2048}$, subsequently projected to $F'_R \in R^{1024}$, via a learned linear layer, and a pre-trained Vision Transformer (ViT) that yields a 768-dimensional global feature vector $F_V \in R^{768}$, from its class token, similarly projected to $F'_V \in R^{1024}$. The two projected representations are then concatenated into a unified 2048-dimensional vector$F_{concat}[\,F'_R; F'_V] \in R^{2048}$ and passed into the Adaptive Feature Fusion Module (AFFM), which applies a channel-wise attention mechanism to dynamically weight and produce the final fused representation $F_{fused} \in R^{2048}$. This fused representation is subsequently fed into a three-layer MLP classification head that progressively reduces the dimensionality from 2048 to 512, then to 256, and finally to Nc output neurons whereN_c denotes the number of age classes specific to each dataset followed by a Softmax activation to produce the final age prediction. Mixup augmentation is incorporated during training to further improve generalization across age group boundaries. The specific details of each component are described individually in the following subsections.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

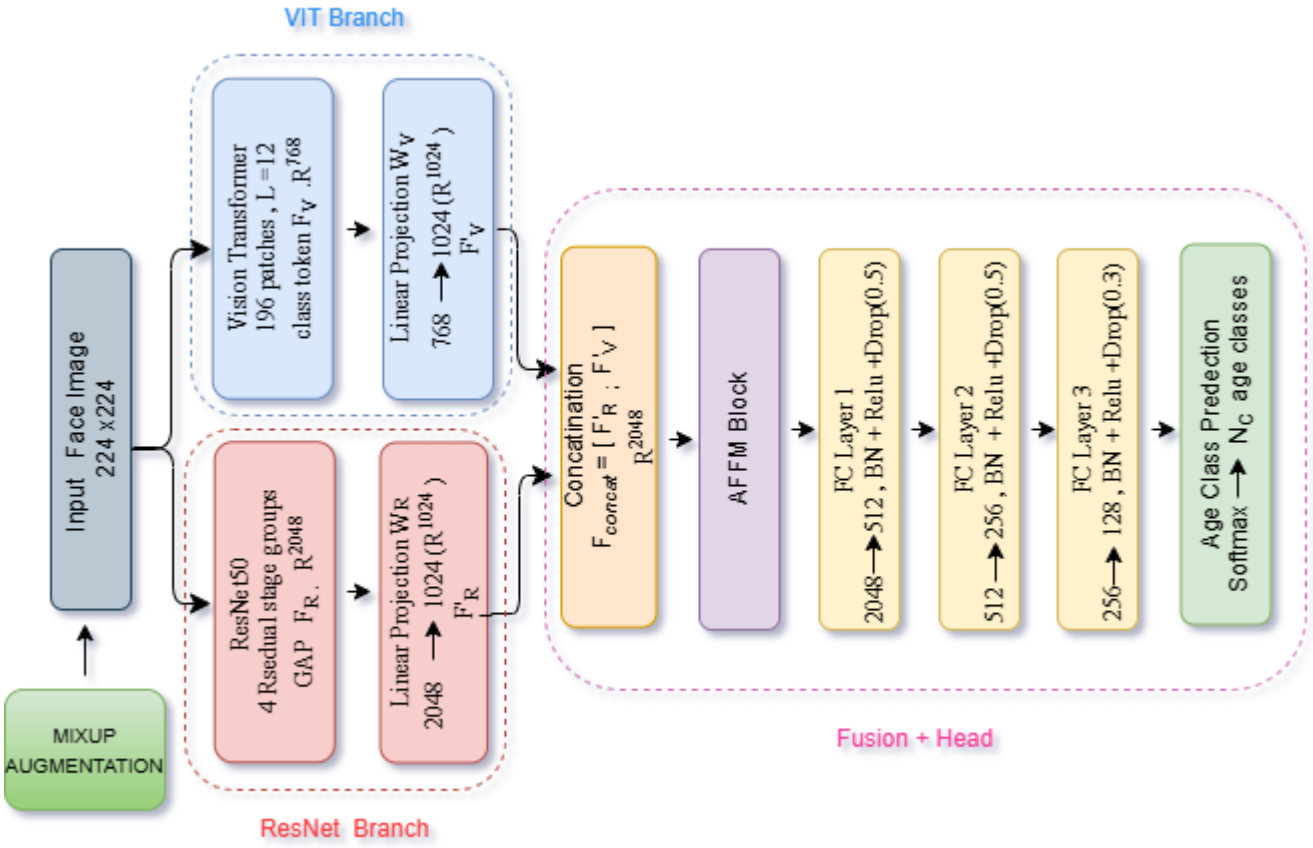

Figure 1 Overall architecture of the proposed dual-branch adaptive fusion network for facial age classification.

I.2. ResNet50 Branch for Local Feature Extraction

ResNet50 is a 50-layer deep CNN introduced by He et al.[16] that excels in extracting fine-grained local features from facial images through its residual learning framework, which enables the training of very deep networks via skip connections that alleviate the vanishing gradient problem. As illustrated in Figure 2, the architecture is organized as a series of stacked residual blocks, each implementing a shortcut connection that bypasses one or more layers, allowing gradients to flow directly through the network during backpropagation. In our implementation, we utilize ResNet50 pre-trained on ImageNet as the backbone for local feature extraction, taking facial images of size 224×224×3 pixels as input and processing them through several key stages: an initial 7×7 convolutional layer with 64 filters and stride 2, followed by batch normalization and ReLU activation, and then four groups of residual blocks (conv2_x, conv3_x, conv4_x, conv5_x) with increasing channel dimensions of 64, 128, 256, and 512 filters respectively, each implementing the identity mapping $y = F(x, \{W_i\}) + x$, where x is the input, $F(x, \{W_i\})$ represents the residual mapping, and y is the output. For feature extraction, we remove the final fully connected layer from the pre-trained ResNet50 and extract features from the global average pooling layer, yielding a 2048-dimensional feature vectordenoted as $F_R \in R^{2048}$ that captures local facial characteristics such as wrinkles, skin texture, facial contours, and age-related morphological changes, encoding rich local spatial information critical for identifying fine-grained age-related patterns in facial images.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

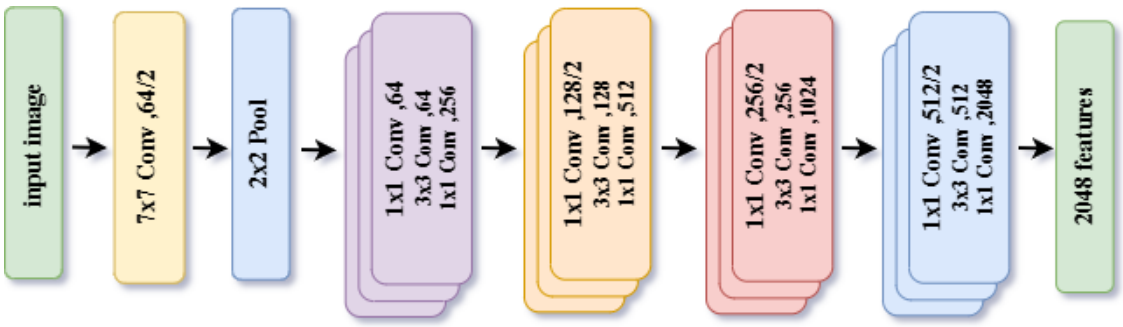

Figure 2 Architecture of the ResNet50 network showing the stacked residual blocks with skip connections. Each residual block implements the identity mapping y=F(x,{Wi})+x, enabling effective training of very deep networks by alleviating the vanishing gradient problem. Figure reproduced from He et al[16].

I.3. Vision Transformer Branch for Global Context Modeling

The Vision Transformer (ViT), first proposed by Dosovitsky et al. [9] in 2020, offers a novel approach to image understanding by breaking down visual inputs into discrete token sequences, analogous to how transformers handle words in natural language processing.As illustrated in Figure 3 (reproduced from [9]), Rather than relying on convolutional operations, this architecture captures spatial relationships across the entire image through a structured pipeline. In our work, the input facial image of size 224×224×3 is partitioned into a grid of N=196 fixed-size, non-overlapping patches, each of dimensions P×P×3, with P=16 and the value 3 accounting for the three RGB color channels. These patches undergo a linear projection that transforms the raw input $X^{(0)} \in \mathbb{R}\hat{}(N \times P^2 \times 3)$ into a compact embedding $X^{(1)} \in \mathbb{R}\hat{}(N \times D)$, where D=768 denotes the embedding dimension. Since this projection discards positional information, learnable positional embeddings are incorporated and added to the patch embeddings to retain the spatial layout of the image as formulated in Equation 1.

$$Z_0 = [X_{class}; X_1^E; X_2^E; \dots\dots; X_N^E] + E_{pos} \quad \textbf{Equation 1}$$

Where X_{class} denotes a learnable classification token prepended to the sequence, and $E_{pos} \in R^{(N+1)\times D}$ encodes the positional information of each token. The resulting sequence is subsequently fed into a stack of L=12 transformer encoder layers, each composed of two core components: A Multi-Head Self-Attention (MSA) mechanism (Equation 2)that enables every patch to attend to all others, learning complex inter-region dependencies:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad \textbf{Equation 2}$$

where Q, K, and V refer to the query, key, and value projections respectively, and d_k is the scaling factor tied to the key dimension, and a Feed-Forward Network (FFN) applying a two-layer transformation with GELU non-linearity(Equation 3).

$$FFN(X) = gelu(X_{w1} + b_1)w_2 + b_2 \quad \textbf{Equation 3}$$

Upon passing through all encoder layers, the final representation associated with the class token, denoted $\mathbf{F_V \in R^{768}}$, serves as a holistic descriptor of the facial image, encapsulating global structural cues and long-range dependencies that are essential for robust facial age estimation. The output of the final transformer layer corresponding to the class token, denoted as $\mathbf{F_V \in R^{768}}$, encapsulates global contextual information about the facial image, capturing holistic age-related patterns and inter-patch relationships.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

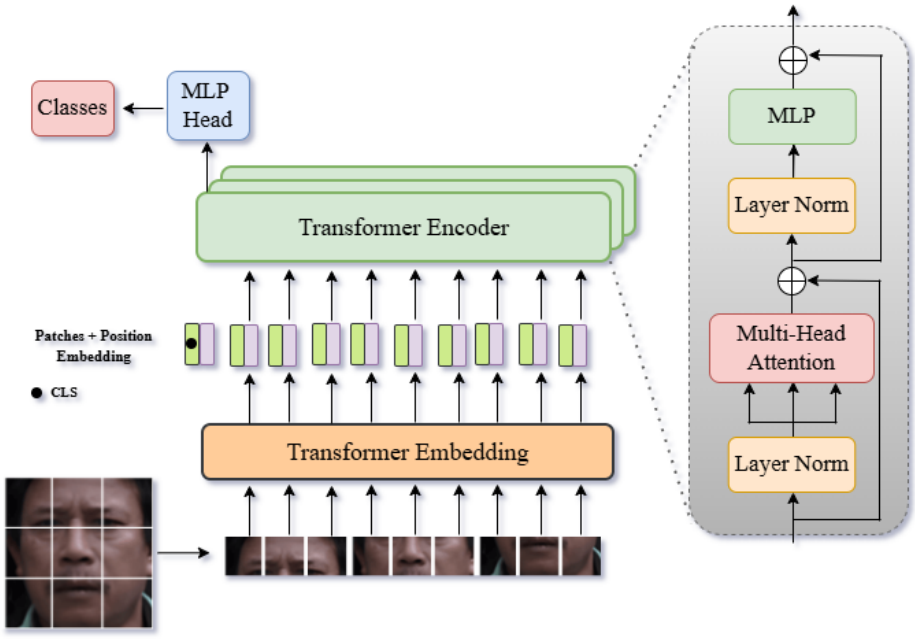

Figure 3 Overview of the Vision Transformer (ViT) architecture. The input image is split into fixed-size patches, linearly embedded, combined with positional encodings, and processed through standard transformer encoder layers. Figure reproduced from Dosovitskiy et al. [9].

I.4. Adaptive Feature Fusion Mechanism

Unlike static fusion methods that rely on fixed weights, our adaptive feature fusion mechanism employs an attention-based strategy to dynamically integrate the local structural features extracted by ResNet50 with the global contextual representations produced by ViT, yielding a richer and more discriminative descriptor for facial age classification[17][18]. Since ResNet50 and ViT produce feature vectors of different dimensionalities 2048 and 768 respectively the first step consists of applying learned projection layers to map both into a common 1024-dimensional space:

$$F'_R = W_R . F_R + b_R \quad ; \quad F'_V = W_V . F_V + b_V \textbf{Equation 4}$$

where $W_R \in R^{1024\times2048}$ and $W_V \in R^{1024\times768}$ are learnable projection matrices. The projected features are then concatenated to form $F_{concat} = [F'_R, F'_V] \in R^{2048}$. We apply a channel attention mechanism to dynamically weight the importance of different feature components:

$$\alpha = \sigma(W_a . F_{concat} + b_a) \quad ; F_{fused} = \alpha \odot F_{concat} \textbf{Equation 5}$$

where σ denotes the sigmoid activation function, $W_a \in R^{2048\times2048}$ is a learnable weight matrix, and ⊙ represents element-wise multiplication.

I.5. Age Classification Head:

The fused representation F_{fused} is passed through a multi-layer perceptron (MLP) composed of four fully connected layers for age classification. Following established best practices for deep classification networks, we apply batch normalization and dropout regularization after each layer, along with ReLU activation functions to introduce non-linearity and stabilize training. Specifically, the MLP progressively reduces the feature dimensionality from 2048 through three hidden layers of sizes 512, 256, and 128 respectively, before mapping to the final Nc-dimensional output, where Nc denotes the number of age group classes defined for each dataset. Dropout is applied with a probability of p=0.5

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

for the first two layers and p=0.3 for the third layer to prevent overfitting. The final predicted age value is computed as the expected value over all age group classes, defined as the weighted sum of each class value multiplied by its corresponding predicted probability:

$$\hat{y} = \sum_{k=0}^{N_C} K.P(K) \quad \textbf{Equation 6}$$

where *P(K)* denotes the predicted probability of the input belonging to age group class k, obtained via a Sotmax activation over the Nc output logits. This formulation combines the benefits of classification and regression by producing a continuous and interpretable age estimate that naturally accounts for the uncertainty distributed across neighboring age group classes, rather than committing to a single hard class decision, which has been shown to yield more accurate and well-calibrated predictions in the presence of inherent label ambiguity at age boundaries[19].

I.6. Mixup Data Augmentationand Training Strategy

Deep learning models for facial age classification require sufficient training data to learn discriminative age-related representations, and the inherent class imbalance and visual similarity between neighboring age groups make data augmentation a necessary step in our training pipeline. Common augmentation methods such as random cropping, horizontal flipping, rotation, and brightness adjustment were applied, but we additionally incorporated the Mixup data augmentation technique, introduced by Zhang et al.[12], to further improve the generalization of our network. Mixup builds a new artificial training sample $\tilde{x}$by mixing the pixels of two original facial images $x_i$and x_j, belonging to age group classes $y_i$and y_j. In formula:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \quad \textbf{Equation 7}$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \quad \textbf{Equation 8}$$

where the mixing coefficient λ∈ [0,1] is sampled from a Beta distribution λ~Beta (α, α) with α=0.2 to favor values close to 0 or 1 while allowing moderate mixing. Inspired by the successful application of Mixup in various classification tasks, we adopted this technique to improve the robustness of our model against the inherent ambiguity at age group boundaries, where the gradual and continuous nature of facial aging makes neighboring classes visually indistinguishable. Experimental results confirmed that the incorporation of Mixup yielded notable improvements in classification performance compared to conventional augmentation approaches alone, particularly in reducing overconfident predictions at age group boundaries across all three benchmark datasets.

Our complete training pipeline leverages transfer learning by initializing both ResNet50 and Vision Transformer with ImageNet-1K pre-trained weights to exploit learned visual representations. We employ a two-stage fine-tuning strategy: Stage 1 freezes the backbone networks (ResNet50 and ViT) and trains only the fusion layers and prediction head for 10 epochs, followed by Stage 2 which unfreezes all layers for end-to-end fine-tuning with a reduced learning rate. The optimization process utilizes the Adam optimizer with an initial learning rate of 1×10^{-3} (Stage 1) and 1×10^{-4} (Stage 2), β1=0.9, β2=0.999, and weight decay of 1×10^{-4}, with learning rate reduction by a factor of 0.5 when validation loss plateaus.Input images are resized to 224×224 pixels, normalized using ImageNet statistics (μ= [0.485,0.456,0.406], σ= [0.229,0.224,0.225]), and augmented using random horizontal flipping (probability=0.5) and rotation (±10 degrees) in addition to Mixup augmentation with α=0.2. Training is conducted with a batch size of 32 for 100 epochs with early stopping (patience=15 epochs) based on validation MAE. This methodology combines the strengths of CNNs for local feature extraction ($F_R \in R^{2048}$), transformers for global context modeling ($F_V \in R^{768}$), adaptive fusion for optimal feature integration (F_{fused}), and Mixup (mixing coefficient λ~Beta(α,α)) for improved regularization, enabling end-to-end training that yields robust age predictions ($\hat{y}$) from mixed target ages ($\tilde{y}$) .

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Experimental Setup

I.7. Dataset Description

To evaluate the performance of our proposed dual-branch architecture, we conducted experiments on three publicly available benchmark datasets that are widely adopted in the facial age group classification literature: MORPH II, UTKFace, and UAGD.

We selected the MORPH II dataset, introduced by Ricanek and Tesafaye[13], as our first benchmark owing to its status as one of the largest and most extensively used longitudinal face databases in age-related research. Its second album, which we used in our experiments, contains over 55,000 facial images of more than 13,000 subjects collected between 2003 and 2007, covering ages ranging from 16 to 77 years. The controlled acquisition conditions of this dataset allowed us to focus our evaluation on fine-grained age-related facial variations without interference from environmental noise. For our classification setup, we followed the standard random protocol, assigning 80% of the dataset to training and reserving the remaining 20% for testing, and we mapped the continuous age labels into discrete age class to align with our classification objective.

As our second benchmark, we employed the UAGD (Uniform Age and Gender Dataset)[14], which we chose specifically because it addresses the age distribution imbalance problem that affects most existing facial datasets. UAGD contains 11,852 facial images with age labels spanning from 1 to 80 years, where the number of images per age class is kept nearly equal, yielding a uniform age distribution. Additionally, the number of female and male images is roughly balanced within each age class, a property we found particularly valuable for training a classifier that generalizes fairly across demographic groups without being skewed toward over-represented categories. We partitioned this dataset using a stratified 80/20 train-test split to preserve class balance across both subsets throughout our experiments.

As our third and most challenging benchmark, we used the UTKFace dataset[15], which we selected for its large scale and highly diverse in-the-wild imaging conditions. The dataset contains over 20,000 facial images with age labels spanning from 1 to 116 years, annotated with age, gender, and ethnicity attributes, and collected under significant variations in pose, facial expression, illumination, occlusion, and image resolution, reflecting the complexity of real-world scenarios. For our classification setup, we followed a standard 80/20 stratified train-test split and mapped the continuous age labels into discrete age group categories to align with our classification objective. The wide age range and highly unconstrained nature of UTKFace, combined with its diversity in demographic attributes including gender, race, and ethnicity, made it the most demanding testbed in our experimental setup, and we used it as the primary measure of our model's robustness and generalization ability under real-world conditions that are far removed from the controlled environments of MORPH II and the balanced distribution of UAGD.

I.8. Implementation Details

I.8.1. Preprocessing and Data Augmentation

To ensure consistent and high-quality input to our proposed dual-branch architecture, we applied a unified preprocessing pipeline across all three datasets. We began with face detection using the Multi-Task Cascaded Convolutional Network (MTCNN)[20], which we selected for its robustness in detecting faces under varying poses, scales, and lighting conditions, as it simultaneously performs face detection and facial landmark localization in a single cascaded framework. Following detection, we performed facial alignment using the five facial landmarks predicted by MTCNN specifically the two eye centers, the nose tip, and the two mouth corners to geometrically normalize each face by correcting for in-plane rotation and ensuring that key facial structures are consistently positioned across all images. This alignment step was carried out using affine transformations implemented via

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

OpenCV. We then applied Dlib's shape predictor to refine facial landmark localization on the aligned faces, allowing us to achieve more precise structural alignment before feeding the images into the network. All detected and aligned face regions were subsequently resized to 224×224 pixels to match the expected input dimensions of both ResNet50 and the Vision Transformer backbone. Finally, we normalized each image by subtracting the ImageNet channel-wise mean values and dividing by the corresponding standard deviations, following the convention of models pre-trained on ImageNet, so as to preserve the statistical properties learned during pre-training and accelerate fine-tuning convergence.

To improve the generalization ability of our model and mitigate overfitting, particularly given the class imbalance present in MORPH II and UTKFace, we applied a set of data augmentation strategies during training. Standard geometric and photometric transformations were employed, including random horizontal flipping, random rotation within a range of ±15 degrees, random cropping with resizing, and random adjustments to brightness and contrast. In addition to these conventional augmentations, we incorporated the Mixup strategy,which generates synthetic training samples by computing convex combinations of pairs of training images and their corresponding labels.All preprocessing and augmentation operations were implemented using a combination of OpenCV, Dlib, and standard PyTorch data transformation pipelines, and were applied consistently across training folds for each dataset.

I.8.2. Training Configuration

We trained our proposed model using the Adam optimizer with an initial learning rate of 1×10^{-3} and momentum parameters $\beta1=0.9$, $\beta2=0.999$, and $\varepsilon=1\times10^{-8}$, combined with a weight decay of 1×10^{-4} to regularize the network and prevent overfitting. To adaptively adjust the learning rate during training, we employed a ReduceLROnPlateau scheduler that reduces the learning rate by a factor of 0.5 whenever the validation loss fails to improve for 5 consecutive epochs. Since our problem is framed as a classification task, we adopted the cross-entropy loss function, and we incorporated Mixup augmentation with $\alpha=0.2$ to further regularize the decision boundaries between adjacent age classes. All models were trained for a maximum of 100 epochs with a batch size of 32, and we applied early stopping with a patience of 15 epochs based on validation accuracy to avoid unnecessary computation and overfitting. We adopted a two-stage fine-tuning strategy to stabilize training: during the first stage (epochs 1-10), we froze both the ResNet50 and ViT backbones and trained only the fusion layers and classification head at a learning rate of 1×10^{-3}, allowing the newly initialized layers to converge before affecting the pre-trained weights; in the second stage (epochs 11-100), we unfroze all layers and performed end-to-end fine-tuning at a reduced learning rate of 1×10^{-4} to gently adapt the backbone representations to our target task. All experiments were implemented in PyTorch 2.0 and conducted on an NVIDIA RTX 3090 GPU with 24 GB of memory.

I.9. Evaluation Metrics

To thoroughly evaluate the performance of our proposed model across all three benchmark datasets, we adopted three complementary metrics widely used in the facial age estimation literature. The first is the Mean Absolute Error (MAE), which quantifies the average absolute deviation between predicted and ground-truth age group labels over all N test samples:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y_i} - y_i| \textbf{Equation 9}$$

where lower values indicate more accurate predictions.

and serves as our primary evaluation criterion given the classification nature of our problem formulation. The The second metric is the Cumulative Score at threshold 5 (CS-5), which measures the percentage of predictions falling within 5 years of the ground-truth label:

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$CS(5) = \frac{1}{N}\sum_{i=1}^{N} 1(|\hat{y_i} - y_i| \leq 5) \textbf{Equation 10}$$

This metric is particularly relevant for age group classification, as it accommodates the natural ambiguity at the boundaries between adjacent age groups. Together, these three metrics provide a comprehensive and balanced assessment of our model's accuracy, robustness, and practical effectiveness.

Experimental Results

Table 1 summarizes the MAE and CS-5 results obtained by our proposed model across all three benchmark datasets. Our dual-branch adaptive fusion architecture achieves consistent performance under diverse evaluation conditions, recording an MAE of 3.42 and a CS-5 of 79.64% on MORPH II, an MAE of 4.42 and a CS-5 of 61.85% on UTKFace, and an MAE of 5.63 and a CS-5 of 59.27% on UAGD. The best results are obtained on MORPH II, which can be attributed to its controlled and well-structured imaging conditions, while the relatively higher MAE values on UTKFace and UAGD reflect the greater difficulty introduced by unconstrained in-the-wild conditions and uniform age distribution respectively. Overall, the results in Table 1 confirm that our proposed architecture generalizes effectively across three complementary and challenging benchmarks, demonstrating the robustness of the adaptive feature fusion mechanism in capturing discriminative age-related representations under varying imaging conditions and age distributions.

Table 1: Performance of the proposed model on the three benchmark datasets.

Dataset	MAE	CS-5
MORPH II	3.42	79.64%
UTKFace	4.42	61.85%
UAGD	5.63	59.27%

I.10. Results on the MORPH II Dataset

In this section, we compare our proposed dual-branch adaptive fusion architecture with the most classic and representative models on the MORPH II dataset. In our experiments, the MORPH II dataset was randomly partitioned into two distinct subsets, with 80% allocated for training and 20% reserved for testing, following the standard random protocol denoted as 80–20. The specific results are shown in Table 2.

Among the comparison methods, ranking-based algorithms such as OR-CNN [21]and Ranking-CNN [6]are classical CNN-based approaches that exploit the ordinal relationship between age labels to improve estimation accuracy, achieving MAE values of 3.34 and 2.96 respectively. Regression-based methods such as SGD [22]and CDCNN [23]treat age estimation as a direct mapping problem, with SGD recording a relatively high MAE of 5.69 due to its sensitivity to dataset imbalance, while CDCNN achieves a more competitive MAE of 2.76. Among the more recent methods, Multi-Stage DNN [24]achieves an MAE of 2.59 with a CS-5 of 86.66% by progressively refining feature representations across multiple stages, and Relative Age [25] further reduces the error to 2.47 with a CS-5 of 86.33% by leveraging relative age relationships between facial images to improve discriminability. Our proposed method achieves the best overall performance among all compared methods, with an MAE of 2.42 and a CS-5 of 88.64%, outperforming the strongest baseline, Relative Age, by 0.05 years in MAE and by 2.31 percentage points in CS-5. These results demonstrate that our dual-branch

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

architecture, which integrates fine-grained local features from ResNet50 with global contextual representations from ViT through an adaptive feature fusion mechanism, is highly competitive and achieves superior performance even when compared against specialized and complex architectures designed specifically for the MORPH II benchmark.

Table 2 The comparisons of our module on MORPH

METHODS	MAE	CS-5
OR-CNN[21]	3.34	81.5%
RANKING-CNN[6]	2.96	85.2%
SGD[22]	5.69	-
CDCNN [23]	2.76	-
Multi-Stage DNN[24]	2.59	86.66%
RELATIVE AGE[25]	2.47	86.33%
OUR	2.42	88.64%

I.11. Results on the UTKFace Dataset

The UTKFace dataset constitutes the most challenging benchmark in our experimental setup, owing to its large scale and highly diverse in-the-wild imaging conditions covering ages from 1 to 116 years with significant variations in pose, illumination, occlusion, and ethnicity. Following the same experimental protocol as MORPH II, the dataset was partitioned using a stratified 80–20 train-test split, and the specific results are shown in Table 3.

Among the comparison methods, MobileNet [29] and OR-CNN [21] are CNN-based approaches achieving MAE values of 5.44 and 5.74 respectively, reflecting the limitations of standard convolutional architectures when confronted with the highly diverse and unconstrained imaging conditions of UTKFace. CORAL [26], a classical ordinal regression approach that models the inherent ordering of age labels through rank-consistent binary classification constraints, achieves a slightly better MAE of 5.47, yet still falls short of more recent methods. MobileAgeNet[27] and Axel Berg et al. [28] adopt lightweight and deep ordinal regression frameworks respectively, achieving more competitive MAE values of 4.65 and 4.55, demonstrating the clear benefit of explicitly modeling age uncertainty and ordinal structure over standard regression and classification approaches. Among hybrid and transformer-based methods, CSCS-Swin[29] achieves an MAE of 4.87, which, while competitive against classical CNN baselines, remains higher than probabilistic and ordinal regression approaches, suggesting that transformer-based architectures alone may not sufficiently capture the fine-grained local texture cues that are critical for age estimation under the highly diverse conditions present in UTKFace. Our proposed method achieves the best overall performance among all compared methods, recording an MAE of 4.42 and a CS-5 of 61.85%, outperforming the strongest baseline, Axel Berg et al. [28], by 0.13 years in MAE and surpassing CSCS-Swin[29] by 0.45 years. These results confirm that the complementary integration of local structural features from ResNet50 and global contextual representations from ViT through our adaptive feature fusion mechanism provides a significant advantage on large-scale in-the-wild datasets, where both fine-grained local details and holistic facial context are essential for robust facial age classification.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 3 The comparisons of our module on UTKFACE

METHODS	MAE
MobileNet[30]	5.44
MobileAgeNet[27]	4.65
OR-CNN[21]	5.74
CORAL[26]	5.47
Axel Berg et al.[28]	4.55
CSCS-Swin[29]	4.87
OUR	**4.42**

I.12. Results on the UAGD Dataset

The UAGD dataset presents a unique evaluation challenge due to its uniform and balanced age distribution, which prevents models from exploiting demographic bias toward over-represented age groups. Given its relatively recent introduction in 2021, the number of available comparison methods remains very limited compared to more established benchmarks such as MORPH II and UTKFace. Following the standard 80–20 stratified train-test split, the specific results are shown in Table 4.

Among the comparison methods, Kong et al.[14] , who introduced the UAGD dataset, established the first baseline using the DEX CNN model, achieving an MAE of 6.84, which reflects the inherent difficulty of this benchmark. D2MO[31], a more recent deep multi-input multi-stream ordinal model that incorporates spatial attention learning to capture multi-context age representations, achieves a lower MAE of 5.34 and a CS-5 of 60.15%, demonstrating the benefit of explicitly modeling ordinal relationships and spatial context on uniformly distributed datasets. Our proposed method achieves an MAE of 5.63 and a CS-5 of 59.27%, outperforming the baseline of Chang et al. by a considerable margin of 1.21 years in MAE while remaining competitive with D2MO, confirming the robustness and generalization ability of our adaptive feature fusion mechanism across uniformly distributed age classes.

Table 4 The comparisons of our module on UAGD

METHODS	MAE	CS-5
CHANG ET AL[14]	6.84	-
D2MO[31]	5.34	60.15
OUR	**5.63**	**59.27%**

I.13. Ablation Study

To validate the contribution of each component in our proposed architecture, we conducted a systematic ablation study on the UAGD dataset, progressively adding one component at a time and evaluating the impact on MAE and CS-5. The specific results are shown in Table 5.

We first evaluated each backbone branch independently. The ResNet50-only variant, which relies solely on local convolutional feature extraction without any global contextual modeling, yields the weakest performance with an MAE of 7.12 and a CS-5 of 48.35%, confirming that local features alone are insufficient to capture the full complexity of age-related facial appearance across a uniformly distributed age range. The ViT-only variant achieves a slightly better MAE of 6.89 and a CS-5 of

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

50.14%, suggesting that global self-attention representations provide marginally richer age-discriminative cues than purely local convolutional features, yet remain limited when used in isolation due to the absence of fine-grained texture information critical for distinguishing neighboring age groups.

When we combine both branches through simple feature concatenation without any adaptive weighting, the MAE drops significantly to 6.21 and the CS-5 improves to 54.73%, demonstrating that the complementary nature of local and global representations provides a substantial gain over either branch alone. This result confirms that the dual-branch design is a key driver of our model's discriminative power. The further addition of our Adaptive Feature Fusion Module (AFFM), which replaces naive concatenation with a channel-wise attention mechanism that dynamically weights the most informative feature dimensions, reduces the MAE to 5.94 and raises the CS-5 to 57.82%, proving that adaptive fusion is more effective than static integration and that not all feature dimensions contribute equally to age group classification. Finally, incorporating Mixup augmentation into the full model achieves the best overall performance with an MAE of 5.63 and a CS-5 of 59.27%, confirming that smoothing the decision boundaries between adjacent age group classes through synthetic interpolated training samples further enhances generalization, particularly on the UAGD dataset where the uniform age distribution makes inter-class boundaries inherently ambiguous. Overall, the ablation results demonstrate that each component of our proposed architecture contributes meaningfully and cumulatively to the final performance, and that the combination of dual-branch feature extraction, adaptive fusion, and Mixup regularization is essential for achieving robust facial age group classification.

Table 5 Ablation study results on the UAGD dataset.

Model Variant	MAE	CS-5
ResNet50 only (no ViT)	7.12	48.35%
ViT only	6.89	50.14%
ResNet50 + ViT (simple concat)	6.21	54.73%
ResNet50 + ViT + AFFM (no Mixup)	5.94	57.82%
ResNet50 + ViT + AFFM + Mixup (full model)	5.63	59.27%

Discussion

The results obtained across all three benchmark datasets demonstrate that our proposed dual-branch adaptive fusion architecture consistently outperforms or remains highly competitive with existing state-of-the-art methods, and we attribute this to three key design choices. First, the complementary integration of ResNet50 local features and ViT global representations allows our model to simultaneously capture fine-grained age-sensitive cues such as skin texture and wrinkles alongside holistic facial structure and inter-region relationships, which neither branch achieves alone as confirmed by our ablation study. Second, the AFFM dynamically assigns higher importance to the most discriminative feature dimensions rather than treating all features equally as static fusion methods do, making the fused representation more robust to irrelevant or noisy facial attributes. Third, Mixup augmentation effectively regularizes the model against the inherent label ambiguity at age group boundaries, where the gradual and continuous nature of facial aging makes neighboring classes visually indistinguishable, resulting in smoother and more calibrated decision boundaries. Despite these strengths, our approach presents certain limitations that are worth acknowledging. The

Research Article

performance gap between MORPH II on one hand and UTKFace and UAGD on the other reveals that our model, like most existing methods, benefits considerably from controlled imaging conditions and struggles when confronted with significant variations in pose, occlusion, illumination, and image resolution. In particular, we observe that our model tends to underperform on elderly subjects, where age-related facial changes become subtler and individual variation becomes more pronounced, making it harder to assign confident class predictions. Similarly, heavily occluded faces, low-resolution images, and extreme head poses present persistent challenges for both the ResNet50 and ViT branches, as the former relies on local texture patterns that are disrupted by occlusion while the latter depends on complete patch sequences that are compromised by pose variation. The variation in results across the three datasets is consistent with their respective characteristics: the controlled conditions of MORPH II yield the lowest MAE of 3.42, the diverse in-the-wild conditions of UTKFace result in a moderate MAE of 4.42, and the uniquely challenging uniform distribution of UAGD produces the highest MAE of 5.63, since our model, like all existing methods on this benchmark, cannot exploit any demographic bias toward over-represented age groups and must generalize evenly across the entire age spectrum from 1 to 80 years. Addressing these limitations through more targeted data augmentation strategies, lightweight occlusion-handling modules, or age-group-specific loss weighting represents a promising direction for future work.

Conclusion

In this paper, we proposed a novel dual-branch deep learning architecture for facial age group classification that effectively combines the complementary strengths of a pre-trained ResNet50 backbone and a Vision Transformer through an Adaptive Feature Fusion Module. The ResNet50 branch captures fine-grained local facial features such as skin texture, wrinkles, and age-related morphological changes, while the ViT branch models long-range global dependencies and holistic contextual relationships across the entire facial image through multi-head self-attention mechanisms. The two feature streams are projected into a common 1024-dimensional embedding space, concatenated, and dynamically weighted through a channel-wise attention mechanism within the AFFM, producing a rich and discriminative fused representation that is subsequently classified by a four-layer MLP head into discrete age group categories. Mixup data augmentation was further incorporated during training to smooth decision boundaries between inherently ambiguous neighboring age group classes and improve generalization across diverse demographic distributions.

We conducted comprehensive experiments on three benchmark datasets of markedly different characteristics MORPH II, UAGD, and UTKFace and demonstrated that our proposed model achieves competitive and consistent performance across all three settings, recording an MAE of 3.42 and CS-5 of 79.64% on MORPH II, an MAE of 4.42 and CS-5 of 61.85% on UTKFace, and an MAE of 5.63 and CS-5 of 59.27% on UAGD. Our systematic ablation study further confirmed that each architectural component contributes meaningfully to the final performance, with the full model consistently outperforming all single-branch and static fusion baselines across all evaluated metrics.

The main contributions of our work are fourfold. We introduced a dual-branch feature extraction pipeline that jointly leverages convolutional local descriptors and transformer-based global representations for facial age classification. We designed an attention-based adaptive fusion mechanism that dynamically integrates the two feature streams in a content-sensitive manner, outperforming naive concatenation approaches. We validated our architecture on three diverse benchmarks including the relatively underexplored UAGD dataset, for which our work constitutes one of the few independent evaluations reported in the literature. Finally, we demonstrated the effectiveness of Mixup augmentation as a regularization strategy specifically tailored to the ordinal and ambiguous nature of age group classification tasks.

Research Article

Despite these contributions, several avenues remain open for future investigation. Incorporating lightweight occlusion-handling and pose-normalization modules could further improve robustness under unconstrained in-the-wild conditions. Exploring age-class-specific loss weighting strategies or ordinal-aware classification objectives may help reduce errors at the boundaries between adjacent age classes, particularly for elderly subjects where inter-class visual differences are subtle. Additionally, extending the proposed framework to multi-task learning settings that jointly predict age class, gender, and ethnicity could provide complementary supervisory signals that further improve the quality of learned facial representations, opening promising directions for the development of more robust and generalizable facial age classification systems.

References

[1] O. Agbo-Ajala and S. Viriri, "Deep learning approach for facial age classification: a survey of the state-of-the-art," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 179–213, Jan. 2021, doi: 10.1007/s10462-020-09855-0.

[2] I. Huerta, C. Fernández, C. Segura, J. Hernando, and A. Prati, "A deep analysis on age estimation ☆," vol. 68, pp. 239–249, 2015, doi: 10.1016/j.patrec.2015.06.006.

[3] G. Guo, "Human Age Estimation : What is the Influence Across Race and Gender ?," *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work.*, pp. 71–78, 2010, doi: 10.1109/CVPRW.2010.5543609.

[4] O. Abhulimen, "Facial Age Estimation Using Deep Learning : A Review," vol. 8, no. 5, pp. 13927–13946, 2021.

[5] X. Geng, Z. Zhou, S. Member, K. Smith-miles, and S. Member, "Automatic Age Estimation Based on Facial Aging Patterns," vol. 29, no. 12, pp. 2234–2240, 2007.

[6] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using Ranking-CNN for Age Estimation," 2017.

[7] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age Estimation Using Expectation of Label Distribution Learning *," 2018. doi: https://doi.org/10.24963/ijcai.2018/99.

[8] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "BridgeNet: A Continuity-Aware Probabilistic Network for Age Estimation," Apr. 2019.

[9] A. Dosovitskiy*et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[10] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.14030

[11] C. Shi, S. Zhao, K. Zhang, Y. Wang, and L. Liang, "Face-based age estimation using improved Swin Transformer with attention-based convolution," *Front. Neurosci.*, vol. 17, 2023, doi: 10.3389/fnins.2023.1136934.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–13, 2018.

[13] K. Ricanek and T. Tesafaye, "MORPH: A Longitudinal Image Database of Normal Adult Age-Progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, IEEE, pp. 341–345. doi: 10.1109/FGR.2006.78.

[14] C. Kong, Q. Luo, and G. Chen, "A comparison study: The impact of age and gender distribution on age estimation," *ACM Int. Conf. Proceeding Ser.*, no. 1, 2021, doi: 10.1145/3469877.3490576.

[15] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 4352–4360,

2017, doi: 10.1109/CVPR.2017.463.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[17] A. Singh and V. K. Singh, "A hybrid transformer–sequencer approach for age and gender classification from in-wild facial images," *Neural Comput. Appl.*, vol. 36, no. 3, pp. 1149–1165, Jan. 2024, doi: 10.1007/s00521-023-09087-7.

[18] G. Maroun, S. E. Bekhouche, J. Charafeddine, and F. Dornaika, "Integrating ConvNeXt and vision transformers for enhancing facial age," *Comput. Vis. Image Underst.*, vol. 262, no. October, p. 104542, 2025, doi: 10.1016/j.cviu.2025.104542.

[19] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-Variance Loss for Deep Age Estimation from a Face." [Online]. Available: http://www.

[20] K. Zhang, Z. Zhang, Z. Li, S. Member, Y. Qiao, and S. Member, "Joint Face Detection and Alignment using Multi - task Cascaded Convolutional Networks," no. 1, pp. 1–5.

[21] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 4920–4928. doi: 10.1109/CVPR.2016.532.

[22] A. Akbari, M. Awais, S. Fatemifar, S. S. Khalid, and J. Kittler, "A Novel Ground Metric for Optimal Transport-Based Chronological Age Estimation," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 9986–9999, 2022, doi: 10.1109/TCYB.2021.3083245.

[23] B. Zhang and Y. Bao, "Cross-Dataset Learning for Age Estimation," *IEEE Access*, vol. 10, pp. 24048–24055, 2022, doi: 10.1109/ACCESS.2022.3154403.

[24] S. E. Bekhouche, A. Benlamoudi, F. Dornaika, H. Telli, and Y. Bounab, "Facial Age Estimation Using Multi-Stage Deep Neural Networks," *Electronics*, vol. 13, no. 16, p. 3259, Aug. 2024, doi: 10.3390/electronics13163259.

[25] R. Sandhaus and Y. Keller, "Relative Age Estimation Using Face Images," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.04852

[26] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec. 2020, doi: 10.1016/j.patrec.2020.11.008.

[27] A. Kumar, "MobileAgeNet: Lightweight Facial Age Estimation for Mobile Deployment".

[28] A. Berg, M. Oskarsson, and M. O'Connor, "Deep ordinal regression with label diversity," *Proc. - Int. Conf. Pattern Recognit.*, vol. 2, pp. 2740–2747, 2020, doi: 10.1109/ICPR48806.2021.9412608.

[29] L. Xu, C. Hu, X. Shu, and H. Yu, "Cross spatial and Cross-Scale Swin Transformer for fine-grained age estimation," *Comput. Electr. Eng.*, vol. 123, no. PD, p. 110264, 2025, doi: 10.1016/j.compeleceng.2025.110264.

[30] Andrey V. Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), pages 119–124, 2021.

[31] C. Kong, H. Wang, Q. Luo, R. Mao, and G. Chen, "Deep Multi-Input Multi-Stream Ordinal Model for age estimation: Based on spatial attention learning," *Futur. Gener. Comput. Syst.*, vol. 140, pp. 173–184, Mar. 2023, doi: 10.1016/j.future.2022.10.009.