Research on the Performance of Multiple Deep Learning Models in Facial Age Recognition

Huiying Zhang *, Haoyi Xie, Yule Sun, Zelang Chen, Chaoyong Rong, Wenshun Sheng

Posted Date: 23 September 2025

doi: 10.20944/preprints202509.1947.v1

Keywords: age recognition; convolutional neural network; mean absolute error; cumulative score

Article

Research on the Performance of Multiple Deep Learning Models in Facial Age Recognition

Huiying Zhang *, Haoyi Xie, Yule Sun, Zelang Chen, Chaoyong Rong and Wenshun Sheng

School of Computer and Communication Engineering, Pujiang Institute, Nanjing Tech University, Nanjing 211200, China

* Correspondence: 111403@njpji.edu.cn; Tel.: +86-186-6272-3490

Abstract

Facial age recognition,as an important research in the field of computer vision, has broad application prospects in areas such as security monitoring and human-computer interaction. This study aims to compare the performance of classic (VGGNet, GoogleNet, DenseNet) and lightweight (MobileNet, EfficientNet) deep learning frameworks in fa-cial age recognition, providing a scientific basis for model selection. The research is based on the public datasets MORPH (large-scale) and FG-NET (small-scale), compre-hensively evaluating from dimensions of accuracy, model efficiency (parameter scale, training time), and resource consumption. Experimental results show that there is no significant linear relationship between model parameters, accuracy, and training time. DenseNet121 achieves the best accuracy with MAE is 2.50 on MORPH, suitable for high-precision requirements; GoogleNet performs best on FG-NET with MAE is 4.02; MobileNet and other lightweight models are suitable for mobile devices but have slightly lower accuracy than large models. This study quantifies the adaptation char-acteristics of different models, providing core evidence for the selection of scenar-io-based face age recognition models, facilitating the practical application of this tech-nology.

Keywords: age recognition; convolutional neural network; mean absolute error; cumulative score

1. Introduction

Age, as one of the core features conveyed by the face, plays a crucial role in facilitating and hindering interper-sonal communication [1]. Age not only shapes an individual's expression of their own intentions,such as young people tending to use emerging vocabulary and the elderly relying on traditional expression patterns, but also in-fluences the interpretation of others' intentions, such as understanding biases that are formed in the inherent cog-nition of behaviors of specific age groups. As an important factor influencing interpersonal interaction, age can not only promote resonance but also may cause communication barriers due to cognitive differences [2].

Aging is an irreversible and continuous process. The morphological changes caused by aging are not only reflected in the face but also involve other body parts such as voice and gait [3]. The impact of aging on the face is more significant than that on other body parts. Since facial structure undergoes sequential changes with age, a correlation model between age and facial structure changes can be constructed to achieve age prediction based on facial images. From the existing researchs, facial images have more significant advantages over other modalities such as voice in age recognition tasks [4]. Age recognition tasks in academic and applied research can be divided into two major categories: physiological age estimation (focusing on the prediction of biological age) and apparent age estimation (based on visual perception) [5].

The face undergoes significant changes and shows a high degree of individual variation with aging [6,7]. As people aging, the increase of melanin in the skin leads to changes in color (pigmentation, dullness), while water loss and the reduction of collagen cause changes in texture (roughness and fine lines deepen into wrinkles). As muscle tension decreases, it gradually moves

 doi:10.20944/preprints202509.1947.v1

towards the direction of gravity, the nasolabial fold deepens, the facial tissue structure migrates and the contour is reshaped [8]. The individual differences in the aging process are essentially the result of the combined effects of genetics (endogenous) and the environment (exogenous), with genetics determining the basic rhythm of aging. Environmental factors accelerate or delay the aging process, such as ultraviolet rays, smoking, air pollution and lifestyle habits, etc [9].

The complexity of the aging process increases the difficulty and generalization pressure in the design of auto-matic age recognition models, posing challenges to age recognition technology [10]. Traditional facial age recognition methods, due to their reliance on manual feature extraction, struggle to cope with the complex patterns of facial aging. However, deep learning has achieved breakthroughs through its ability to automatically learn features, with con-volutional neural models (CNNs) capable of automatically learning multi-level aging features such as skin texture, color and facial contour structure [11].

In recent years, early deep models such as VGGNet and GoogleNet have laid the foundation for facial age recognition technology through hierarchical feature extraction [12–14]. ResNet effectively solves the problem of van-ishing gradients through residual connection, significantly improving the accuracy of facial age recognition [15]. DenseNet enhances feature reusability through dense connections, providing strong support for facial age recogni-tion [16]. ShuffleNet、MobileNet and EfficientNet have achieved innovations in balancing recognition accuracy and efficiency, and have also promoted the development of facial age recognition technology [17,18].

Furthermore, the Transformer demonstrates potential in capturing the global aging correlations of human faces through its global attention mechanism, and has emerged as a new direction in the field of age estimation. For in-stance, Aakash and Kumar [23] utilized the multi-head attention of the Transformer to focus on cross-region aging features and achieved an improvement in accuracy. However, such methods have the drawbacks of large parameter quantities and high inference latency, making them difficult to adapt to resource-constrained scenarios. Current research mostly independently validates the performance of the Transformer or a single CNN, lacking studies on the adaptation rules of classic CNNs and lightweight CNNs on different datasets. The age estimation is affected by the imbalance of the dataset and the ambiguity of the labels. Relevant studies have proposed solutions. Gao et al. [13] proposed label distribution learning, which models the label ambiguity using a Gaussian distribution. Zhang et al. [9] further optimized the label distribution range to alleviate the data shortage.

Nevertheless, the adaptability of different models in the facial age recognition task varies significantly, and their performance is influenced by feature learning capabilities and parameter scales. Previous studies have briefly ex-plored the application of different CNN architectures. Liu et al. [17] designed a hybrid attention model based on the lightweight ShuffleNetV2, verifying the potential of lightweight models in age estimation. Aruleba and Viriri [18] compared the performance of the EfficientNet series on the UTKface and Adience datasets, emphasized the efficiency advantages of the compound scaling strategy, and further clarified the differentiated value of this study.

So far, there has been no study that systematically compares and analyzes the performance of mainstream deep learning models with the existing state-of-the-art CNN architectures in the task of facial age recognition. The absence of such a comparative study makes it difficult to clearly define the advantages and limitations of different deep learning methods in the age estimation scenario, and also hinders the summary of model design rules and the optimization of subsequent technical routes. Therefore, this paper focuses on the classic models, configures them on public datasets, and conducts multiple sets of comparative experiments to evaluate each framework from multiple dimensions, revealing the performance differences in age recognition, and providing a basis for model selection.

It should be specially noted that this study does not introduce label distribution learning, not because it ignores the value of this technology, but based on the core goal of focusing on the comparison of model architectures. Label distribution learning belongs to an optimization strategy at the data level, which will introduce additional interfering variables, making it impossible to

 doi:10.20944/preprints202509.1947.v1

accurately distinguish the sources of performance differences, thereby affecting the objective evaluation of the adaptability of different models.

This research has achieved the following innovations in the field of facial age recognition:

- The mainstream CNN architectures were systematically compared, filling the gap in horizontal evaluation across multiple models. The experimental results show that the performance of the models is closely related to the dataset size and structural design. The number of parameters is not linearly correlated with accuracy. Feature reuse mechanisms (such as dense connections) and multi-scale feature fusion (such as Inception modules) have greater advantages in specific scenarios.
- Combining model efficiency metrics (parameter size, training time) with accuracy metrics (MAE, CS) for the first time, a comprehensive evaluation system was established. This provides clear selection criteria for re-source-constrained scenarios such as mobile device deployment.
- Revealed the pattern of model structure adaptation to dataset characteristics, quantified the impact of sample size on model performance, and provided a new direction for cross-dataset generalization research.
- This study systematically verified the practicality of lightweight models in age recognition, providing theo-retical support for the design of real-time age recognition systems.

2. Related Work

The human visual system is inherently equipped with the ability to quickly infer age based on facial features. People judge the age of a face by comprehensively analyzing multiple dimensions of characteristics such as facial contours, skin texture, and the degree of soft tissue relaxation [19]. After inferring the age of the other person through facial features, quickly identifying their social status, people will subconsciously adjust their communica-tion methods to improve communication efficiency. However, age estimation based on intuition has obvious limita-tions. Artificial modifications such as tanning, makeup, and plastic surgery can significantly alter facial visual fea-tures, thus affecting the accuracy of the judgment [20].

The age attribute in facial images is not only the focus of technical research but also a link connecting biometric recognition, social governance and commercial applications. The value of age identification continues to increase with the development of artificial intelligence technology [21]. In traditional facial age recognition, the manually extracted features mainly focus on the visual aspects related to facial aging, covering multiple dimensions such as geometric structure, texture changes, and skin color status. Geometric structure features focus on the morphology, position, and proportion relationship of facial organs, reflecting the changes in contour caused by the aging process of bones and soft tissues [22]. Facial contour features include the curvature and clarity of the mandibular line (be-coming blurred due to fat accumulation or skin relaxation as one ages), the prominence of the cheekbones, the ratio of the length and width of the face, etc. The positioning and proportion features of facial key points, such as the distance between the eyes, the depth of the nasolabial folds, the angle of the corners of the mouth, and the position of the eyebrows (which may lower with aging), are quantified by calculating the relative distance or angle based on the coordinates of key points [23]. The geometric features of wrinkles, such as the distribution range and main direc-tion (horizontal or radial) of crow's feet and forehead wrinkles. Texture features mainly include microstructural changes on the skin surface caused by aging, relying on local variations in image grayscale or color. Wavelet trans-form can extract facial texture and spots, while Gabor filtering can capture facial features at different directions and scales [24]. These feature extraction rules rely on manual tuning, and the effectiveness of feature extraction heavily depends on algorithm parameters. They cannot automatically adapt to image translations, rotations, and scaling, and their accuracy drops sharply when faced with complex scenarios.

As age increases, skin color changes, pigmentation deposits, and the number of age spots increase, leading to a shift in mean color and an increase in variance. The skin tone of younger people is usually more uniform (with smaller variance). Combine the distance features of geometric contours

 doi:10.20944/preprints202509.1947.v1

with the texture features to form a more comprehensive feature vector, such as the shape parameters (reflecting geometric changes) and texture parameters (reflecting skin condition), decomposed by the Active Appearance Model (AAM) [25]. The handcrafted features rely on manually designed rules and are difficult to cover complex aging patterns, such as the differences in wrinkle distribution among different individuals, or the interference of lighting or expressions on texture. They are also sensitive to the parameters of feature extraction algorithms, such as the scale of Gabor filters. Automatically capturing image translation, rotation, and scale invariance becomes difficult, such as with SIFT and HOG [26,27]. Additionally, traditional computer vision algorithms have high computational complexity and limited generalization ability. Its practicality is limited in natural settings.

CNNs are important branches of deep learning, automatically extracting abstract features from images through multiple layers of nonlinear transformations [28]. Shallow convolutional layers extract low-level features of images such as edges, colors, textures, and changes in brightness. Deeper convolutional layers, based on the features ex-tracted by shallow layers, form high-level features by combining and abstracting them (such as corners, texture combinations, parts, etc.), ultimately obtaining semantic features used for classification (such as 'eyes', 'faces', 'wheels',etc.) [29]. CNNs use local connections and weight sharing, reducing parameter redundancy and solving the problem of parameter explosion when fully connected models process images. Pooling layers usually follow convolutional layers, used for dimensionality reduction, reducing redundancy, and feature enhancement, while also enhancing model robustness and providing more efficient input for deep feature learning.

The powerful automatic hierarchical feature extraction capability of CNNs, their adaptability to space, ro-bustness to translations and other transformations enable CNNs to break through the limitations of traditional handcrafted features and automatically learn discriminative features suitable for tasks. This has led to significant success in computer vision, medical imaging, autonomous driving and natural language processing. By combining large-scale datasets (such as ImageNet) and GPU parallel computing, CNNs continuously improve their perfor-mance through the expansion of depth and width, breaking through the limitations of traditional methods [30].

The automatic feature learning capability of deep learning models enables them to capture subtle patterns of facial aging, extract effective facial features, and improve the accuracy and reliability of age recognition [31,32]. The strong dependence of deep learning models on large-scale high-quality labeled training data is a core bottleneck in their application to face age recognition tasks. Constrained by factors such as privacy protection, high collection costs, and unbalanced age distribution, the available actual data is difficult to meet the training requirements of the models, leading to overfitting and decreased generalization ability of the models. In recent years, the transfer learning technology has provided an effective solution to the problem of insufficient data by reusing the general features of pre-trained models on general visual datasets (such as ImageNet) [23]. However, this technology has obvious limitations in the facial age recognition scenario. The semantic differences between general image classification and age recognition are significant, and model training often exhibits saturation in accuracy. To broaden the coverage dimension of feature representation, researchers have further explored fusion strategies of multiple CNN models to comprehensively capture multi-dimensional facial aging features and effectively compensate for the lim-itations of single models in feature representation [34,35]. Additionally, the combination of attention mechanisms with age recognition models enables dynamic adjustment of attention weight allocation to focus on the optimiza-tion of the output layer [36]. The optimization effect of deep learning is highly dependent on the characteristics of the model architecture, and there is no universal adaptation scheme applicable to all architectures yet.

Although CNNs can capture key age-related information such as wrinkles and skin tone from facial images, significantly improving estimation accuracy, practical application still faces significant bottlenecks due to the com-plexity of the aging process, including significant individual differences in aging speed, non-age factors interfering with feature extraction,such as lighting, expressions,

 doi:10.20944/preprints202509.1947.v1

makeup, etc., and limitations of data and models, such as un-even distribution of training data and insufficient generalization ability of classic models for extreme age samples.

3. Introduction to Classic Models

CNNs such as VGGNet, GoogLeNet, ResNet, MobileNet, ShuffleNet, DenseNet, and EfficientNet are milestone models in the field of computer vision. They are widely used in tasks like image classification, object detection, se-mantic segmentation, and more. Each model addresses key issues in the development of CNNs, such as depth bot-tlenecks, computational efficiency, and feature utilization, through unique design concepts. The following provides an introduction to these models focusing on their core designs and advantages.

1) VGG [37] was proposed by the University of Oxford and performed exceptionally well in the ImageNet competition in 2014. Its core design involves uniformly using 3x3 convolution stacks instead of large-sized convolu-tions, combined with 2x2 pooling layers, to achieve deepening through a modular structure. Although it has a large number of parameters, its structure is simple and can stably capture multi-scale features, making it a classic benchmark model in the field of deep learning.
2) GoogLeNet [38] was the champion in the ImageNet classification task in 2014. Its core innovation was the Inception module, which used parallel multi-scale convolutions (1x1, 3x3, 5x5) and pooling, and concatenated the outputs to fuse multi-scale information; by using 1x1 convolutions to reduce dimensions and introduce auxiliary classifiers to alleviate gradient vanishing, it abandoned fully connected layers and replaced them with global aver-age pooling, significantly reducing parameters and overfitting risks.
3) ResNet [39] is a milestone model proposed in 2015, with the core being residual connections: through re-sidual blocks, allowing gradients to propagate directly, solving the problem of gradient vanishing and performance decline in deep models. The modular design (BasicBlock, Bottleneck) increases depth while controlling the number of parameters, has extremely strong generalization ability, and has become the standard backbone model for com-puter vision tasks.
4) MobileNet [40] is a lightweight model designed by Google for mobile devices, with the core being depth-wise separable convolutions (depthwise convolutions capture spatial features + 1x1 pointwise convolutions for channel fusion), combined with the width multiplier α to flexibly scale the number of channels, significantly re-ducing parameters and computational costs while maintaining accuracy, achieving a balance between efficiency and performance.
5) ShuffleNet [41] is a lightweight model proposed by Megvii, with the core being group convolutions fol-lowed by channel shuffling to break the information barriers between groups; by dynamically adjusting the model size through scaling factors, it enhances cross-channel fusion while retaining the advantages of lightweight design, adapting to different scenarios with limited computing resources.
6) DenseNet [42] innovates in dense connections, where each layer is directly connected to all previous layers, achieving maximum feature reuse through channel concatenation, promoting information and gradient flow, effec-tively alleviating gradient vanishing, and significantly reducing parameters while improving performance, suitable for training deep models.
7) EfficientNet [43] achieves breakthroughs in accuracy and efficiency through systematic design: adjusting the model depth, width, and input resolution in a fixed proportion, combining the MBConv module (inverted bottle-neck structure + depthwise convolutions) and the SE attention mechanism, enhancing feature expression without significantly increasing computational costs, becoming a benchmark for efficient model design.

 doi:10.20944/preprints202509.1947.v1

4. Experiments

4.1. Datasets

The selected datasets are MORPH [44] (large-scale, multi-ethnic, and accurately labeled) and FGNET [45] (small-scale, longitudinal tracking, and complete age sequence). These two datasets complement each other in terms of scale and features, allowing for a comprehensive evaluation of the overall performance of different deep learning frameworks in facial age estimation tasks. The MORPH dataset was collected under natural conditions and better reflects real-world scenarios compared to datasets collected in controlled environments. As a large-scale longitudinal facial dataset, MORPH includes 55,134 images of nearly 13,618 subjects, covering an age range from 16 to 77 years old, with an average of 4 images per person, effectively simulating changes in facial appearance over time in real-life situations. It contains rich data such as race and gender, as well as diverse acquisition conditions like variations in lighting and pose, making it effective for testing model robustness in scenarios with large volumes of data and high complexity. For large-scale models like VGGNet, MORPH can fully leverage its fitting capabilities, while its performance on lightweight models like EfficientNet reflects its efficiency advantages in large-scale data.

FG-NET is a small dataset, includes 1,002 images from 82 subjects, with an age range of 0-69 years. It features longitudinal tracking, with each person having an average of 10 images, with time intervals up to 18 years, focusing more on the gradual aging process of individuals. FG-NET primarily tests the model's generalization ability in scenarios with small sample sizes, such as whether DenseNet's dense connection mechanism can more efficiently utilize limited data, or if GoogleNet's inception module can reduce overfitting when data is insufficient. Additionally, the complete age sequence from children to the elderly in FG-NET compensates for the lack of samples over 50 years old, testing the model's adaptability across all age groups.

4.2. Image Preprocessing

In automatic age recognition tasks, the input facial images often contain numerous interfering factors, which directly affect the model's extraction of age-related features. Face preprocessing is usually the first step, which directly affects the accuracy of subsequent feature extraction and age identification. The goal of face preprocessing is to reduce non-age-related interference factors (such as background and posture) from the original image and convert it into standardized facial data that meets the model's input requirements, enabling the model to more efficiently learn age-related features.

We adopt a unified preprocessing procedure, which includes face detection, key point detection, and alignment. In the face detection stage, we use a DPM model with simple parameter settings [13] for face detection and localization to extract facial regions, excluding backgrounds and other irrelevant information. Detect five key facial points, namely the centers of the eyes, the tip of the nose, and the corners of both sides of the mouth. Based on these facial key points, we adjust the face to an upright posture to ensure that the relative spatial relationships of key features remain consistent, reducing the impact of pose variations on the model. Using the center of the aligned face as a reference, the aligned face is cropped to a fixed size of 224×224. Preprocessed standard raw data is used as input to ensure that performance differences arise from the model itself, enhancing the intuitiveness and persuasiveness of the conclusions. Based on the need to preserve the integrity of age-related features, this study did not employ data augmentation techniques such as rotation or color jittering.

4.3. Experimental Settings

The experimental environment uses Microsoft Windows 11 operating system and Python 3.12 programming language, implementing model training and inference based on the PyTorch framework. The hardware configuration includes a single NVIDIA GeForce RTX 3060 GPU (12GB)

 doi:10.20944/preprints202509.1947.v1

with CUDA acceleration for computation. The core objective of this study is to conduct a horizontal evaluation of the performance differences of various deep learning models in the task of facial age estimation. Therefore, all models were trained from scratch to eliminate the interference of pre-trained features on the performance comparison of different architectures. Although age is a continuous value, human judgment of age is essentially interval-based cognition (such as around 25 years old). Additionally, classification tasks can more stably optimize models using cross-entropy loss. Therefore, this study defines the age recognition task as a multi-classification task rather than a regression task, with age labels discretized. For the MORPH dataset (age range 16-77 years), it is divided into 62 categories, and for the FGNET dataset (age range 0-69 years), it is divided into 70 categories. The batchsize was set to 32 for the MORPH and 1 for the FG-NET due to its smaller sample size.

4.4. Evaluation Criteria

Mean Absolute Error (MAE) and Cumulative Score (CS) are used to evaluate the performance of each model. MAE is defined as the average of all absolute errors between predicted ages and physiological ages. The formula for calculating MAE is:

$$y_{\mathrm{MAE}} = \frac{1}{N}\sum_{n=1}^{N} | y - y^{*} | \tag{1}$$

In formula (1), N represents the number of test face samples; y represents the physiological age; y^{*} represents the predicted age. The smaller the y_{MAE}, the closer the model's predicted age is to the physiological age, indicating better performance of the model.y_{CS} represents the accuracy rate for different error values measured in years, with error levels ranging from 0 to 10, defined as follows:

$$y_{\mathrm{CS}} = \frac{N_m}{N} \times 100\% \tag{2}$$

In the formula, N_m represents the number of test facial images that satisfy $|y - y^{*}| \le m$, N represents the number of test faces.

4.5. Models Training and Experimental results

The MORPH and FG-NET datasets were randomly divided into two parts, with 90% used for training and the remaining 10% for testing. Fixing the 90/10 split ensures that all models use the same training set and test set, and the performance differences can be directly attributed to the model design (such as dense connections, multi-scale convolutions), rather than the validation strategy. This is crucial for the research goal of comparing multiple architectures horizontally.

Several models, including VGGNet, GoogLeNet, ResNet, MobileNet, ShuffleNet, DenseNet, and EfficientNet were used. Due to the large number of parameters in VGGNet, the learning rate was set to 0.0001 on MORPH and 0.00001 on FG-NET. For the other models, the learning rate was set to 0.001 on MORPH and 0.0001 on FG-NET. The number of epochs for both datasets was set to 20. Learning rate scheduling, weight decay, and early stopping strategies were not used during the training process. After training for 20 epochs, the model with the smallest MAE on the validation set during the training process was selected as the final model.

All models use uniform preprocessing, training strategies, loss functions, etc. By comparing the performance differences of the same model on two datasets, the impact of data volume and the model on performance can be separated, providing a more objective evaluation of the model's generalization ability. The parameters of each model and the training time (in seconds) per epoch on MORPH and FG-NET are shown in Figure 1. The MAE of each model on MORPH and FG-NET is shown in Figure 2. Figures 3 and 4 show the CS under different models. On the MORPH dataset during training, the result is MAE ± 0.08, as shown Figure 1. Taking DenseNet121 as an example, σ = 0.08 indicates that the results fluctuate between 2.42 and 2.58 after 5 rounds. On the FG-NET dataset, the result is MAE

 doi:10.20944/preprints202509.1947.v1

± 0.15 . The fluctuation in the small sample scenario is slightly larger, which is in line with expectations.

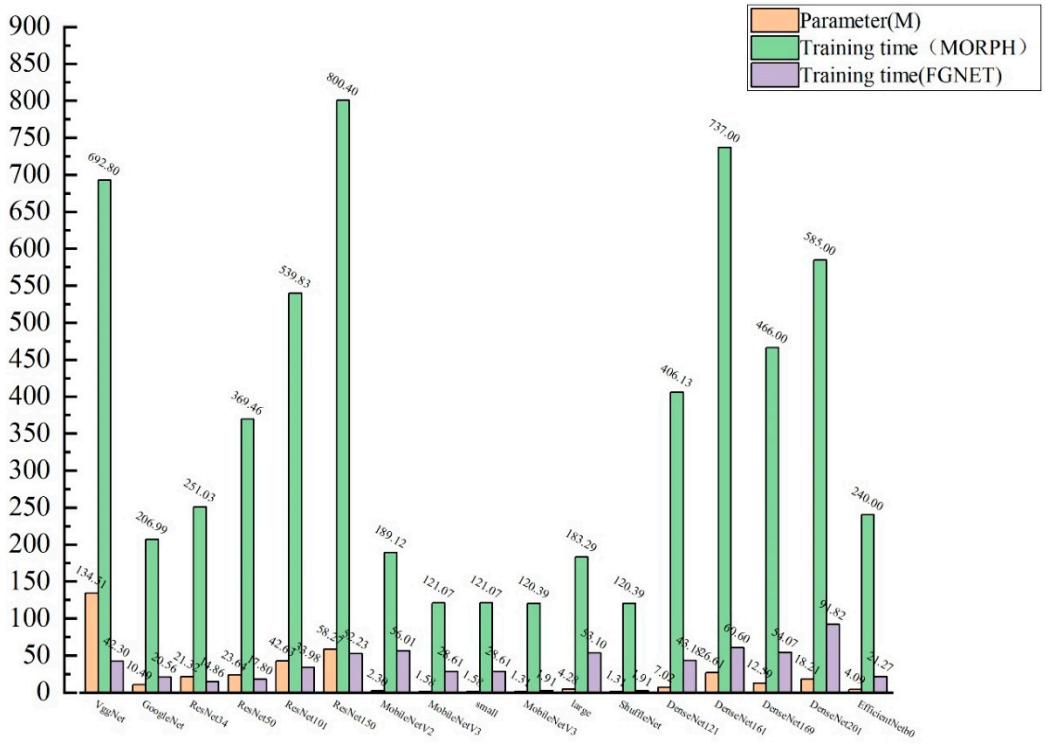

Figure 1. The parameters of each model and the training time on MORPH and FG-NET.

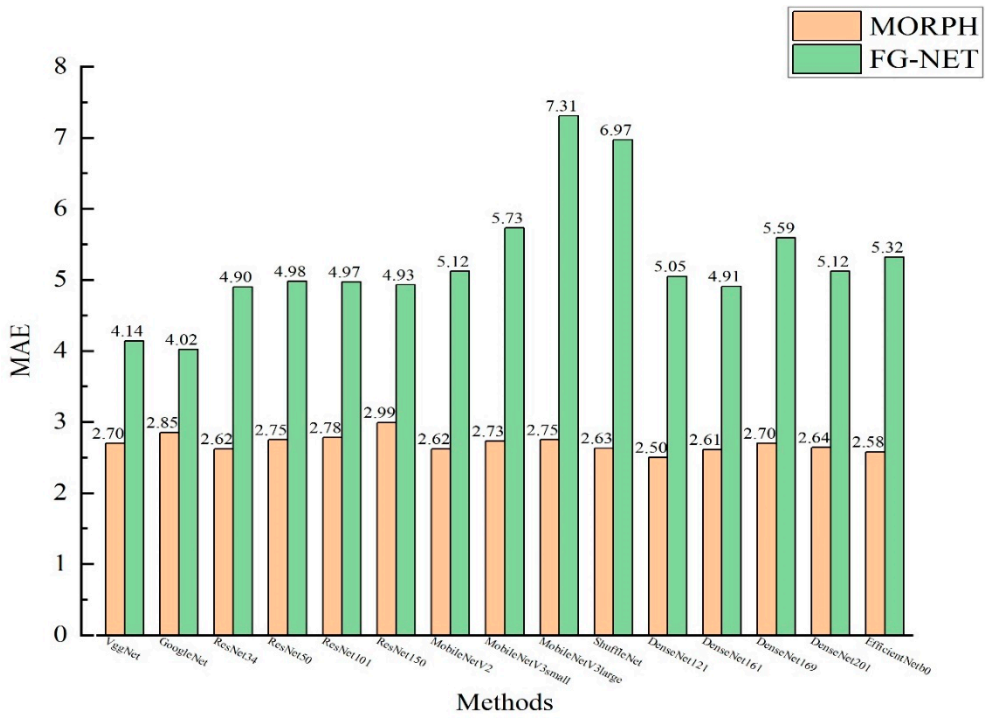

Figure 2. MAE based on MORPH and FG-NET under different models.

 doi:10.20944/preprints202509.1947.v1

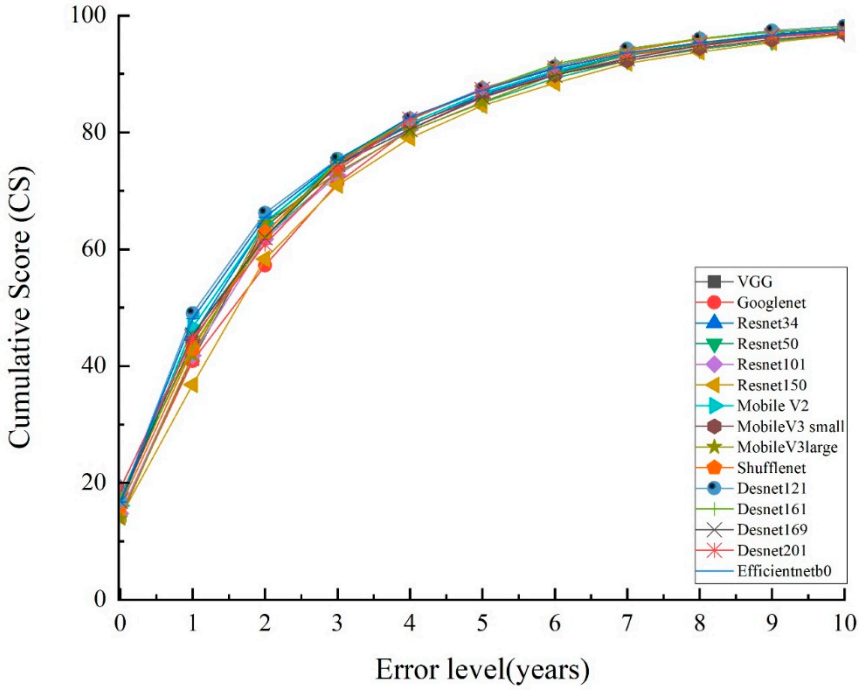

Figure 3. Comparisons of CS on MORPH.

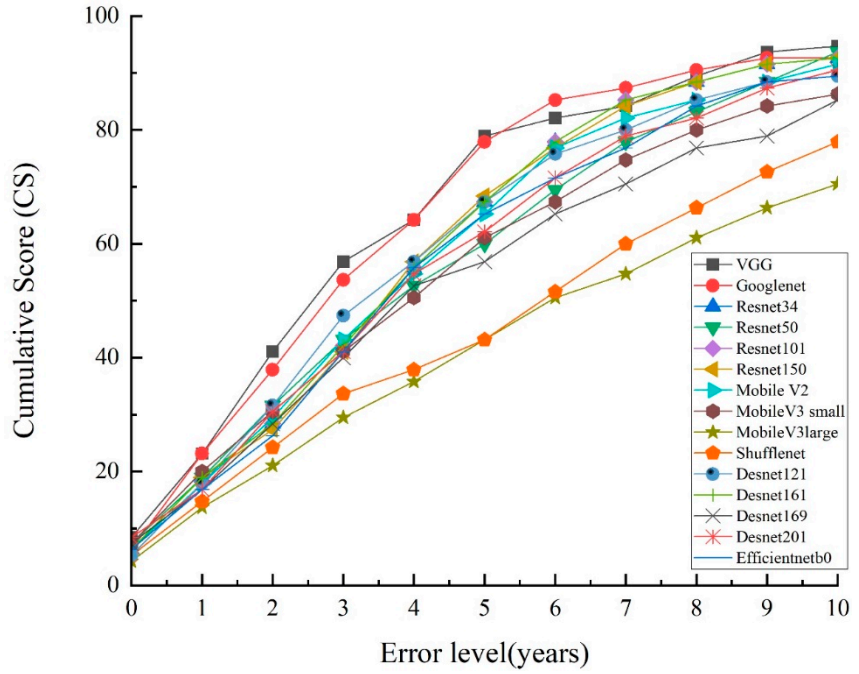

Figure 4. Comparisons of CS on FG-NET.

Compared with two representative age recognition methods (MA-SFV2 [1] and MSDNN [17]), MA-SFV2 is an improved ShuffleNetV2 based on a hybrid attention mechanism, used for facial age recognition. MSDNN adopts a multi-stage deep neural model and utilizes the feature maps at different levels of the CNN's backbone to generate distinctive features. Among them, on MORPH, the average absolute error (MAE) of MA-SFV2 is 2.68, while that of MSDNN is 2.59. On FGNET, the MAE of MA-SFV2 is 3.81.

From Figure 1,2,3, and 4, it can be seen that parameter scale affects model storage and computational cost. VGGNet has large parameters (134.51M) and each training epoch takes the

 doi:10.20944/preprints202509.1947.v1

longest time of 692.8 seconds, but on MORPH, the MAE is 2.70, because stacking small convolutional kernels can extract finer texture features. On FG-NET, MAE is 4.14. Due to the small sample size of FG-NET, VGGNet is prone to overfitting, and its complex structure amplifies the noise in the small dataset. On MORPH, GoogleNet has an MAE of 2.85. With large samples, the multi-scale advantage of GoogleNet is overshadowed by the deeper features of other models. On FG-NET the MAE of GoogleNet is 4.02. It performs relatively well among the comparison models. The multi-scale convolution in the Inception module can capture the fine-grained age features of different areas of the face. As the depth of the ResNet series increases (ResNet34 - 150), the MAE on MORPH first decreases and then increases by 2.62 -2.99. The gradient propagation in the shallow ResNet (34 layers) is smooth and suitable for the large sample size of MORPH; The deep ResNet (150 layers) has an MAE of 4.93 on FG-NET. However, a small dataset is insufficient to support the feature abstraction of the deep model. On the MORPH, the MAE of DenseNet121 reached 2.50, achieving the highest accuracy. The dense connection can fully utilize multi-scale features and is suitable for the complex age characteristics of MORPH. However, on the FG-NET, the MAE was 5.05, as the small dataset can not support the dense connected feature redundancy learning, which instead leads to overfitting and a decrease in accuracy.

MobileNet, ShuffleNet and EfficientNet belong to lightweight models, and their lightweight features are very prominent. MobileNet can flexibly adapt to various computing resource constraints under different configurations. The MobileNetV3 small has only 1.58M parameters. On MORPH, the MAE is 2.73. The depthwise separable convolution simplifies the computation, but it loses some of the complex feature extraction capabilities. On FG-NET, the MAE is 5.73. The lightweight structure is unable to capture the fine-grained age differences of FG-NET, and its generalization ability is even weaker under small sample conditions. ShuffleNet has 1.31M parameters. On MORPH, the MAE is 2.63, but on FG-NET, the MAE is 6.79. Channel shuffling and grouped convolution overly simplify the channel interaction, and in the small sample and multi-noise scenarios of FG-NET, the feature expression ability is seriously insufficient. The training time of EfficientNetb0 is 240 seconds on MORPH and 21.27 seconds on FG-NET. The MAE is good on both datasets (2.58/5.32). Because the composite scaling strategy balances depth, width and resolution, it can converge stably under different data volumes.

The performance differences of models in facial age recognition tasks are essentially the result of the interaction between dataset properties (potential biases) and model architecture characteristics. Public datasets show significant differences in age distribution, sample diversity, acquisition conditions, and labeling reliability, which form potential biases. These biases need to be matched with the model's parameter scale, feature extraction methods, etc., to maximize performance. This rule is particularly evident in the FG-NET (small-scale) and MORPH (large-scale) datasets. Specifically, the small FG-NET easily leads to overfitting risks, while GoogleNet's Inception multi-scale module captures facial features of different scales in parallel, adapting to the age differences in the dataset, ultimately achieving the best MAE on this dataset. In contrast, for the large-scale MORPH, GoogleNet's limitations of insufficient depth and low parameter efficiency become apparent, and its performance is surpassed by models better at deep feature mining, among which DenseNet121, with its dense connection feature reuse architecture, can fully utilize the scene diversity and fine-grained aging features in the vast samples of MORPH, achieving the best accuracy on this dataset.

Based on the aforementioned dataset and model adaptation rules, model selection in actual scenarios should closely align with demand orientation. If high precision is pursued and computational resources are sufficient, GoogleNet can be prioritized for small-scale datasets like FG-NET, while DenseNet121 is recommended for large-scale datasets like MORPH. If the focus is on model lightweight and training efficiency, lightweight models such as ShuffleNet and EfficientNetb0, which have the advantages of small parameter scale and short training time, can reduce computational resource consumption while meeting certain precision requirements. Their specific adaptability can be further verified and adjusted based on the scale and diversity of the target dataset.

 doi:10.20944/preprints202509.1947.v1

The application scenarios of facial age recognition have different priorities for accuracy and efficiency. For example, in medical aging analysis scenarios, accuracy takes the highest priority, followed by efficiency, and DenseNet121 can be chosen to improve diagnostic accuracy. In mobile real-time recommendation scenarios, efficiency takes the highest priority, followed by accuracy, and ShuffleNet or MobileNet can be selected. Subsequently, the model improvement strategies can be further explored to enhance the generalization ability across datasets and overall performance.

The performance on the MORPH dataset is generally better than that on FG-NET. The core reasons are closely related to the sample size and task adaptability of the two datasets. 1.Large sample sizes make it easier for various models, especially deep ones like ResNet and EfficientNet, to learn age-related generalization features such as wrinkles, skin texture, and hairstyle changes; whereas the limited number of samples in FG-NET makes it difficult for the model to cover diverse patterns of aging, resulting in weaker generalization ability.2.MORPH has a wide age coverage and a relatively balanced distribution, which can provide the model with a complete age feature gradient from youth to old age. The uneven distribution of FG-NETs may lead to an increase in the prediction error of the model for a few age groups, such as middle-aged and elderly people.3. The images from MORPH are mostly standardized captures with uniform lighting and posture, making it easier for models to extract stable age features. In contrast, the images from FG-NET are often taken in natural settings with lower resolution, increasing the difficulty of feature extraction.4. MORPH's age labels are based on official records, with relatively small label errors. However, the age labels of FG-NET partially rely on manual intervention, which may lead to errors. The label noise can interfere with the model's learning of age features. Therefore, MORPH's advantages in sample size, diversity, quality, and annotation accuracy better meet the feature learning needs of CNN models, allowing each model to more stably extract age-related features and outperform FG-NET.

Through multi-dimensional evaluation, two types of models suitable for real-time deployment and offline analysis have been identified. Lightweight models (such as MobileNet series, ShuffleNet, EfficientNetb0) support real-time deployment. The core features of lightweight models match the core requirements of real-time deployment, such as resource constraints and low latency. For example, the parameter size of MobileNetV3 small (1.58M), ShuffleNet (1.31M), and EfficientNetb0 (4.09M) is only 1/5 to 1/30 of classic models (such as VGGNet 134.51M, DenseNet121 7.02M), making them easily adaptable to the limited storage resources of mobile and embedded devices without the need for high-performance GPUs. In training on the MORPH, the time taken for a single epoch by ShuffleNet is 120 seconds, and by EfficientNetb0 is 240 seconds, significantly lower than that of classic models such as DenseNet121 (406 seconds) and VGGNet (692.8 seconds), indicating their smaller computational load. Classic models (such as VGGNet, DenseNet121, and ResNet series) are more suitable for offline analysis. The core advantage of classic models is high accuracy, but their large parameter size and high computational load make them unsuitable for real-time deployment due to resource and latency constraints. They are better suited for offline analysis scenarios where there is no real-time pressure and high accuracy is pursued.

Lightweight models (such as MobileNet, EfficientNetb0, etc.) are adapted for real-time deployment (on mobile devices, embedded devices), supporting real-time interaction scenarios such as personalized services and real-time monitoring. Classic models (such as DenseNet121, etc.) are suitable for offline analysis (on servers, cloud), and are used in high-precision demand scenarios such as medical retrospection and security batch statistics.

5. Conclusions

This study systematically compared the performance of seven classic deep learning models, including VGGNet, GoogleNet, and ResNet, among others, on facial age recognition tasks using the MORPH (large-scale) and FG-NET (small-scale) datasets. The evaluation was conducted from multiple dimensions, such as accuracy (MAE), parameter scale, and training efficiency, revealing the adaptation characteristics and application value of different models. The experimental results show

 doi:10.20944/preprints202509.1947.v1

that model performance is closely related to the scale and structural design of the dataset. The number of parameters is not linearly and positively correlated with the accuracy.

On the large-scale MORPH, DenseNet121 leverages the advantage of dense connectivity for feature reuse, making it particularly suitable for mining complex age features. EfficientNet demonstrates outstanding compre-hensive performance by balancing accuracy and efficiency through a composite scaling strategy. On the small-scale FG-NET, GoogleNet achieves the optimal MAE through multi-scale feature fusion and moderate parameter scale. Its Inception module effectively reduces the risk of overfitting in small samples. The accuracy of VGGNet is inferior to that of lightweight models like EfficientNet; ResNet shows an increase in MAE on MORPH as depth increases, indicating that over-deepening may lead to feature redundancy. Lightweight models such as ShuffleNet have out-standing training efficiency, but they exhibit the worst MAE performance on FG-NET, revealing their limitations in complex tasks with small sample sizes.

This study provides quantitative criteria for model selection in different scenarios. For high precision requirements, DenseNet121 should be chosen for large datasets, while GoogleNet is preferred for small datasets. In re-source-constrained scenarios, EfficientNet can be selected. Additionally, it reveals the matching patterns between model structures and dataset characteristics, offering directions for future optimization of age recognition models by integrating attention mechanisms and multimodal fusion.

Acknowledgments: Special thanks to the reviewers of this paper for their valuable feedback and constructive suggestions, which greatly contributed to the refinement of this research.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Huiying Zhang; data collection: Yule Sun; analysis and interpretation of results: Huiying Zhang, Haoyi Xie; draft manuscript preparation: Huiying Zhang, Zelang Chen, Chaoyong Rong. All authors reviewed the results and approved the final version of the manuscript.

Funding: This research was funded by (1) 2024 Key project of natural science, NanjingTech University Pujiang Institute, grant number NJPJ2024-1-01. 2025 Association of Fundamental Computing Education in Chinese Universities, grant number 2025-AFCEC-113.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Bekhouche, S.E.; Benlamoudi, A.; Dornaika, F.; Telli, H.; Bounab, Y. Facial age estimation using multi-stage deep mod-els. Electronics 2024, 13, 3259. https://doi.org/10.3390/electronics13163259
2. Jumbadkar, R.; Kamble, V.; Parate, M. Development of facial age estimation using modified distance-based regressed CNN model. Traitement du Signal 2025, 42(2),1041. DOI:10.18280/ts.420237.
3. Abbas, Z. H.; Shaker, S.H.Prediction of human age based on face image using deep convolutional neural network.AIP Con-ference Proceedings, 2024, 3009(1):11.DOI:10.1063/5.0190537.
4. Zhang, Y.; Shou, Y.; Ai, W.; et al. GroupFace: imbalanced age estimation based on multi-hop attention graph convolutional network and group-aware margin optimization. IEEE Trans. Inf. Forensics Secur. 2025, 20, 605–619. DOI:10.1109/TIFS.2024.3520020.
5. Liu, X.; Qiu, M.; Zhang, Z.; et al. Enhancing facial age estimation with local and global multi-attention mechanisms. Pattern Recognit. Lett. 2025, 189, 71–77. DOI:10.1016/j.patrec.2025.01.005.
6. Abbes, A.; Ouarda, W.; Ayed, Y.B. Age-API: are landmarks-based features still distinctive for invariant facial age recognition? Multimed. Tools Appl. 2024, 83, 67599–67625. DOI:10.1007/s11042-024-18227-7.
7. Zhang,H.Y.;Zhang,Y.;Geng,X.Recurrent age estimation.Pattern Recognition Letter.2019,125, 271–277.DOI：10.1016/j.patrec.2019. 05.002.

8. Geng, X.; Zhou, Z.H.; Smith-Miles, K. Automatic age estimation based on facial aging patterns. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 2234–2240. DOI:10.1109/TPAMI.2007.70733.
9. Zhang, H.Y.; Zhang, Y.; Geng, X.; Chen, F.Y. Practical age estimation using deep label distribution learning. Frontiers of Computer Science. 2021, 15(3), 73–78. DOI：10.1007/s11704-020-8272-4.
10. Babu, A.A.; Sudhavani, G.; Madhav, P.V.; Sadharmasasta, P.; Chowdary, K.U.; Balaji, T.; Jaya, N. Deep learning centered methodology for age guesstimate of facial images. Journal of Theoretical and Applied Information Technology. 2024, 102, 2568–2572.
11. Zhang, Z.; Yin, S.; Cao, L. Age-invariant face recognition based on identity-age shared features.The Visual Computer, 2024, 40(8):5465-5474.DOI:10.1007/s00371-023-03116-1.
12. Zhu, Y.; Li, Y.; Mu, G.; Guo, G. A study on apparent age estimation. In Proceedings of the IEEE Int. Conf. Comput. Vis. Workshop, Santiago, Chile, 07 December 2015, 267–273.
13. Gao, B.B.; Xing, C.; Xie, C.W.; Wu, J.; Geng, X. Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing, 2017, 26, 2825–2838.DOI:10.1109/TIP.2017.2689998.
14. Zhang, H.Y.; Lin, J.; Zhou, L.; et al. Facial age recognition based on deep manifold learning. Mathematical Biosciences & Engineering. 2024, 21, 4485–4500. DOI:10.3934/mbe.2024198.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, 27 June 2016,770–778.
16. Zhu, B,; Li, L.; Hu,X.,et al.DEFOG: Deep Learning with attention mechanism enabled cross-age face recognition.Tsinghua Science and Technology, 2025, 30(3):1342-1358.DOI:10.26599/TST.2024.9010107.
17. Liu, X.; Zou, Y.; Kuang, H.; Ma, X. Face Image Age estimation based on data augmentation and lightweight convolutional neural network. Symmetry 2020, 12, 146. DOI:10.3390/sym12010146.
18. Aruleba, I.; Viriri, S. Deep learning for age estimation using efficientnet. In Advances in Computational Intelligence, 1st ed.; Rojas, I., Joya, G., Català, A., Eds.; Springer: Cham, Switzerland, 2021,12861, 407–419. DOI:10.1007/978-3-030-85030-2_34.
19. Zhang, H.Y.; Sheng, W. An optimized algorithm for facial age recognition based on label adaptation. Computer Engineering. 2024, 9, 1–10. DOI:10.19678/j.issn.1000-3428.00EC0068561.
20. Smith-Miles, K.; Geng, X. Revisiting facial age estimation with new insights from instance space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44, 2689–2697. DOI:10.1109/TPAMI.2020.3038760.
21. Gupta, S.K.; Nain, N. Review: Single attribute and multi attribute facial gender and age estimation. Multimed. Tools Appl. 2023, 82, 1289–1311. DOI:10.1007/s11042-022-12678-6.
22. Kwon, Y.H.; Vitoria Lobo, N. Age classification from facial images. Comput. Vis. Image Underst. 1999, 74, 1–21. DOI:10.1006/cviu.1999.0742.
23. Aakash, S.; Kumar, S.V. A hybrid transformer–sequencer approach for age and gender classification from in-wild facial images. Neural Comput. Appl. 2024, 36, 1149–1165. DOI:10.1007/s00521-023-09087-7.
24. Guo, G.; Mu, G.; Fu, Y.; Huang, T.S. Human age estimation using bio-inspired features. In Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL, USA, 20 June 2009,112–119.
25. Lanitis, A.; Taylor, C.J.; Cootes, T.F. Toward automatic simulation of aging effects on face images. IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 442–455. DOI:10.1109/TPAMI.2002.1007097.
26. Lu, D.; Wang, D.; Zhang, K.; et al. Age estimation from facial images based on Gabor feature fusion and the CIASO-SA al-gorithm. CAAI Transactions on Intelligence Technology. 2023, 8, 518–531.
27. Guo ,G.; Mu,G.; Fu,Y. ; Huang,T. S. Human age estimation using bio-inspired features,In 2009 IEEE Conference on Computer Vision and Pattern Recognition, Florida, Miami , 20 -25 June,2009, 112–119.https://doi.org/10.1109/CVPR.2009.5206681.
28. Zhang, H. Y. ; Sheng, W. S. ; Zeng Y. Z. Face age recognition algorithm based on label distribution learning, Journal of Jiangsu University. 2023,44(02), 180–185. doi: 10.3969 /j. issn.1671 -7775.2023. 02.008.
29. An, W.X.; Wu, G.S. Hybrid spatial-chanel attention mechanism for cross-age face recognition. Electronics 2024, 13(7), 1257.
30. Muliawan, N.H.; Angky, E.V.; Prasetyo, S.Y. Age estimation through facial images using deep CNN pretrained model and particle swarm optimization. E3S Web Conf. 2023, 426, 8. DOI:10.1051/e3sconf/202342601041.

31. Bao, Z.H.; Luo, Y.T.; Tan, Z.C.; et al. Deep domain-invariant learning for facial age estimation. Neurocomputing 2023, 534, 86–93. DOI：10.1016/j.neucom.2023.02.037.
32. Abbas, Z.H.; Shaker, S.H. Prediction of human age based on face image using deep convolutional neural network. AIP Conference Proceedings 2024, 3009(1), 11. DOI:10.1063/5.0190537.
33. Naaz, S.; Pandey, H.; Lakshmi, C. Deep Learning based age and gender detection using facial images. In: 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). 0 [2025-09-11]. DOI:10.1109/ACCAI61061.2024.10601975.
34. Zhao, Q.; Liu, J.; Wei, W. Mixture of deep models for facial age estimation. Information Sciences,2024,679. DOI:10.1016/j.ins.2024.121086.
35. Jiang, S.; Ji, Q.; Shi, H.; et al. Spatial correlation guided cross scale feature fusion for age and gender estimation. Scientific Reports 2025, 15(1). DOI:10.1038/s41598-025-03081-w.
36. Liu ,X.; Qiu, M.; Zhang, Z.;et al.Enhancing facial age estimation with local and global multi-attention mecha-nisms.Pattern Recognition Letters, 2025, 189(000),71-77. DOI:10.1016/j.patrec.2025.01.005.
37. Simonyan, K.; Zisserman, A. Very deep convolutional models for large-scale image recognition. In Proceedings of the Int. Conf. Learn. Representations, Banff, Canada, 14 April 2014. DOI:10.48550/arXiv.1409.1556.
38. Szegedy, C.; Liu, W.; Jia, Y.; et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition, Boston, MA, USA, 07 June 2015,1–9.
39. He, K.; Zhang, X.; Ren, S.; et al. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 26-30 June 2016, 770–778. DOI:10.1109/CVPR.2016.90.
40. Howard, A.G.; Zhu, M.; Chen, B.; et al. MobileNets: Efficient convolutional neural models for mobile vision applica-tions. Proc. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ,Hawaii Convention Center,USA, 21 - 26 July 2017. DOI:10.48550/arXiv.1704.04861.
41. Zhang, X.; Zhou, X.; Lin, M.; et al. ShuffleNet: An extremely efficient convolutional neural network for mobile de-vices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii Convention Cen-ter,USA, 21 - 26 July 2017. DOI:10.48550/arXiv.1707.01083.
42. Huang, G.; Liu, Z.; Laurens, V.D.M.; et al. Densely connected convolutional models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Las Vegas, USA, 26-30 June 2016. DOI:10.1109/CVPR.2017.243.
43. Tan, M.; Le, Q.V. EfficientNet: rethinking model scaling for convolutional neural models.Proceedings of Machine Learning Research, California, USA , 9-15 June, 2019, 6105-6114. DOI:10.48550/arXiv.1905.11946.
44. FG-NET Database. Available online: https://yanweifu.github.io/FG_NET_data/index.html (accessed on 04 January 2020).
45. MORPH Database. Available online: https://ebill.uncw.edu/C20231_ustores/web/store_main.jsp?STOREID=4 (accessed on 17 May 2019).