10-1-2023

Attention-based human age estimation from face images to enhance public security

Md. Ashiqur Rahman

Shuhena S. Aonty

Kaushik Deb

Iqbal H. Sarker
Edith Cowan University

10.3390/data8100145
Rahman, M. A., Aonty, S. S., Deb, K., & Sarker, I. H. (2023). Attention-based human age estimation from face images to enhance public security. Data, 8(10), article 145. https://doi.org/10.3390/data8100145

data

MDPI

Article

Attention-Based Human Age Estimation from Face Images to Enhance Public Security

Md. Ashiqur Rahman [1], Shuhena Salam Aonty [1], Kaushik Deb [1,*] and Iqbal H. Sarker [1,2,*]

[1] Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh; ashiqurr343@gmail.com (M.A.R.); shuhena@cuet.ac.bd (S.S.A.)
[2] School of Science, Edith Cowan University, Perth, WA 6027, Australia
* Correspondence: debkaushik99@cuet.ac.bd (K.D.); m.sarker@ecu.edu.au (I.H.S.)

Abstract: Age estimation from facial images has gained significant attention due to its practical applications such as public security. However, one of the major challenges faced in this field is the limited availability of comprehensive training data. Moreover, due to the gradual nature of aging, similar-aged faces tend to share similarities despite their race, gender, or location. Recent studies on age estimation utilize convolutional neural networks (CNN), treating every facial region equally and disregarding potentially informative patches that contain age-specific details. Therefore, an attention module can be used to focus extra attention on important patches in the image. In this study, tests are conducted on different attention modules, namely CBAM, SENet, and Self-attention, implemented with a convolutional neural network. The focus is on developing a lightweight model that requires a low number of parameters. A merged dataset and other cutting-edge datasets are used to test the proposed model's performance. In addition, transfer learning is used alongside the scratch CNN model to achieve optimal performance more efficiently. Experimental results on different aging face databases show the remarkable advantages of the proposed attention-based CNN model over the conventional CNN model by attaining the lowest mean absolute error and the lowest number of parameters with a better cumulative score.

Keywords: age classification; attention module; convolution neural network; deep learning; transfer learning; public security

Citation: Rahman, M.A.; Aonty, S.S.; Deb, K.; Sarker, I.H. Attention-Based Human Age Estimation from Face Images to Enhance Public Security. *Data* **2023**, *8*, 145. https://doi.org/10.3390/data8100145

Academic Editor: Joaquin Torres Sospedra

Received: 4 August 2023
Revised: 16 September 2023
Accepted: 18 September 2023
Published: 25 September 2023

1. Introduction

The emergence of various patterns in facial appearance is thought to be a primary cause of human age, which is a significant personal characteristic. Age is a notion that describes a person's age at a specific period and is a crucial biological component. The aging process is ongoing and cannot be undone. The difficulties in predicting facial age are similar to those in other facial image identification tasks. Firstly, they must identify the facial region, and secondly, pinpoint the vital features for aging that are relevant to the assignment before formulating a vector of features and ultimately classifying the image [1]. Age estimation is the procedure of labeling a face image automatically with the age (year) or group of ages (year range) of each face using the facial image [2,3]. This age may be actual, perceived, apparent, or estimated.

Facial age estimation is a crucial aspect of public security and law enforcement. This technology has proven useful in identifying and verifying individuals, particularly in cases where proper identification documents are unavailable [4]. Additionally, it can aid in predicting the age of missing persons or abducted children, thereby facilitating faster recovery. The technology also serves a vital role in forensic investigations, age-restricted access enforcement, surveillance, and the monitoring of potential threats. By using age estimation for crime analysis, law enforcement can better understand criminal patterns and design more effective crime prevention strategies. Moreover, facial age estimation is

an essential tool in border control and immigration processes, as it helps to verify the age of individuals seeking entry into a country or applying for asylum. A previous paper [5] used a multi-scale convolution module and a width-channel attention module to assess the anxiety and depression levels related to public healthcare and security integration. However, it is crucial to address ethical considerations and privacy concerns to ensure the responsible and lawful implementation of this technology.

Globally, the aging process and growth rate vary. In contrast to South Asia, Europe has a different aging process. In addition to geographic effects, other variables include race, gender, environment, weather, habits, food, and living situation [6]. These factors globally affect the aging process. Under the same general circumstances, the aging process might also differ from person to person. Moreover, human aging processes and growth rates are caused by internal biological changes such as hormonal shifts and catastrophic diseases like AIDS and cancer. The human aging process is a natural phenomenon that cannot be stopped via artificial means. Because of this, human age plays a positive role in various fields, including human–computer interactions, identification, precision advertising, and biosecurity [6,7].

According to biological theory, a person's lifespan can be roughly divided into two phases, the formative or early years of life and the mature or aging years. The first span of human life can be examined from newborn to adult. The geometry and shape of the face are different in the early stages (craniofacial growth). According to craniofacial investigations, the shape of the human face changes from circular to oval [8]. The locations of the fiducial landmarks on the human face will shift due to these form modifications. The most obvious change in the second span of human life, from maturity to old age, is skin aging (texture change); however, a minor change in shape also occurs. In [9], the aging process is represented by eleven facial features closely related to aging. Figure 1 lists some of these characteristics, including the corners of the mouth, nose, and eyes. According to biology, as people age, their skin gets rougher, darker, thinner, and less elastic due to gravity's influence and collagen insufficiency [10]. Spots and wrinkles will start to form on the skin over time. To utilize the human age estimation system, these changes in texture and shape resulting from aging should be extracted [8].

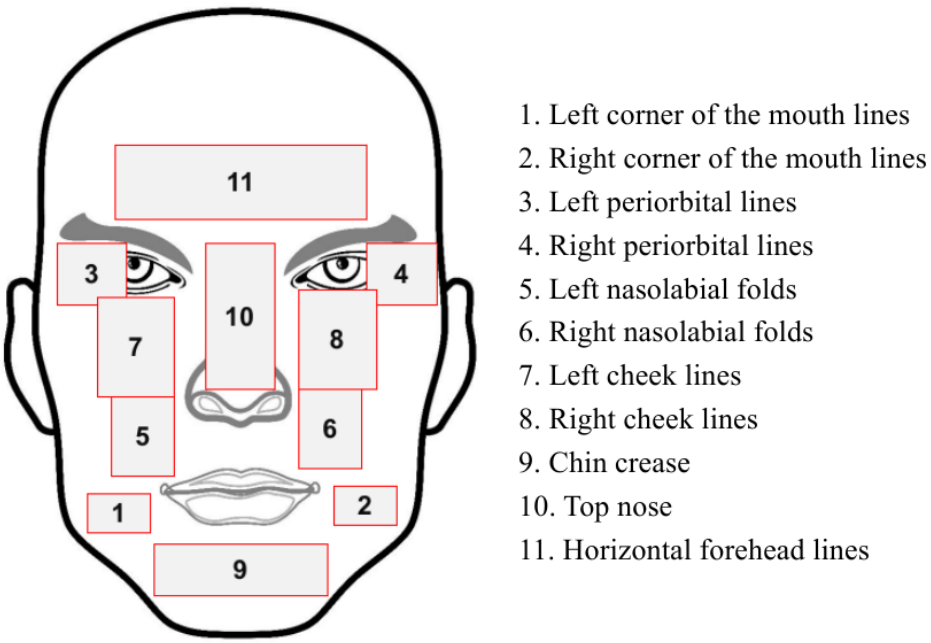

Figure 1. Eleven skin areas closely affected by the aging process.

For more than 20 years, human age estimation from facial images has been explored. Nevertheless, it is an incredibly challenging field for scholars. Numerous researchers have

added new approaches to the current methods to address the inadequate datasets and the similarity in photos for close ages. They established new methods, such as the conditional probability neural network (CPNN), IIS-Learning from Label Distribution (IIS-LLD) [3], and attention modules like the convolutional block attention module (CBAM) and squeeze-and-excitation networks (SENet) [11]. Several state-of-the-art deep learning algorithms like DenseNet [12] and MobileNetV2 [13] have been applied to image classification in recent years. Despite intensive research on the issue of age estimation, the accuracy and dependability of the current solutions still need to be revised in everyday life. Extrinsic appearance elements, such as pose, illumination, and expression factors [14], and intrinsic human aspects, such as gender, race, and health condition factors [15], contribute to the difficulty of this challenge. Due to the problem of resolving intrinsic factors, many earlier investigations have concentrated on extrinsic issues. This challenge results from inhomogeneous features brought on by two factors: (1) an extensive range of facial appearances among people of the same age, as shown in Figure 2, and (2) different age-related changes to the human face. For instance, rapid human aging in early life will lead to fewer changes in the face in adulthood, as shown in Figure 3.

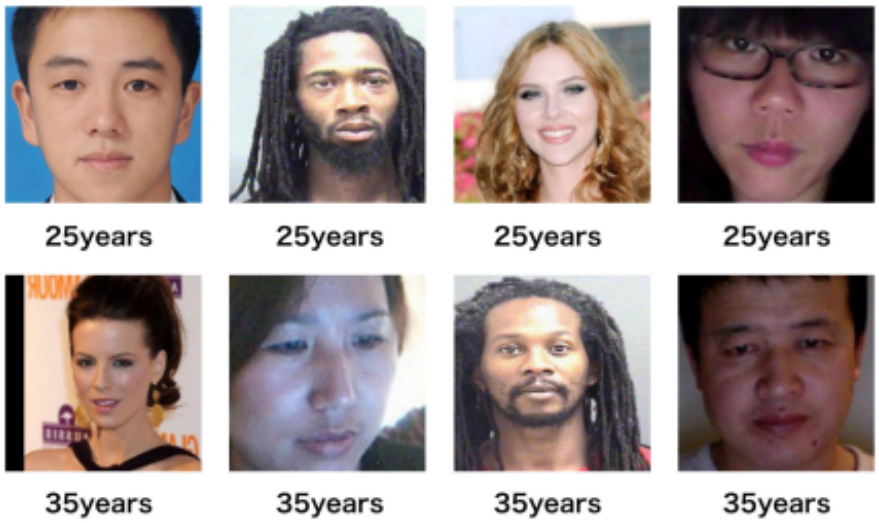

Figure 2. Variability in aging: exploring differences among individuals of the same age.

Figure 3. Metamorphosis: the journey of facial transformation from childhood to adulthood.

Due to the wide variations in aging patterns and populations, it is difficult to create an age estimator that can precisely estimate human age from various populations for various age ranges. Nowadays, deep learning techniques are often used for data analysis and feature extraction [16]. In this work, we attempted to classify human age from faces using a variety of existing CNN architectures using an attention module, which helps to extract more significant features from the human face.

The major contributions of this work are as follows:

1. By merging diverse datasets, we have crafted an extensive combined dataset. Through rigorous augmentation techniques, we have substantially increased image counts for each age group, enhancing our model's performance in age estimation through broader and more representative training data.
2. We have developed a CNN model integrating an attention mechanism to accurately estimate the human age from intricate facial expressions, harnessing the power of convolutional layers for feature extraction and a dedicated attention module to focus on salient facial regions, thus enabling precise age estimation through an end-to-end trainable architecture.
3. We have carried out thorough tests using a mix of combined datasets, separate datasets, and native human face pictures. This comprehensive approach helps us better evaluate the performance of the method in various scenarios and with different data types.

2. Related Work

The assessment of human facial age has been the subject of extensive research over the last few decades. The oldest research in this area is found in [17], based on an investigation of skin wrinkles and the theory of craniofacial growth. They divide faces into three groups of age: babies, teens, and adults. The Active Appearance Model (AAM) [18], a statistical method for modeling face images, is frequently used to depict facial appearance. The AAM approach, which outperforms anthropometric models, takes textural and geometric models into account to deal with any age.

For face and gender detection, bio-inspired features are frequently used. To estimate human age, in [15], the author presented biologically inspired elements first. They used the Gabor filter to extract many features in their work at various scales and orientations, which offered them an added advantage in handling minor transformations like rotations, translations, and scale changes in human faces.

In [3], the author proposed label distribution for age estimation and considered each picture connected with the label distribution, namely the Gaussian and triangular distribution, to increase the number of label images for age estimation. Two algorithms proposed by them, CPNN and IIS-LLD, achieved a better outcome in estimating human age. Learning algorithms require a significant amount of data. However, numerous images are on social media sites and labeled weekly. In [19], the authors suggested an age-difference learning method to use these weekly labeled photos through deep convolutional neural networks using label distribution.

Estimating age automatically is a difficult task due to the variety of facial appearances and factors to consider. In [20], the authors compared different deep learning frameworks using different datasets. The results show that CNNs are superior to hand-crafted features. The best model was tested for its robustness against noise, facial expressions, ethnicity, and gender changes, and it outperformed previous methods. Additionally, the study evaluated layer-wise transfer learning and knowledge transfer from face recognition.

Moving Window Regression (MWR), a novel technique for ordinal regression, was published in [21]. The input image's initial rank (age) estimate is iteratively adjusted in the MWR methodology for the estimation of the human face by calculating the ρ-rank inside a search window. The ρ-rank quantifies how much older or younger the input is than the references and the window is defined by the known ranks of two reference images. Every window centers on the previous outcome. To guarantee rank-monotonicity and consistent confidence scores, in [22], the author proposed a new framework called Consistent Rank Logits (CORAL). Their study assessed the performance of CORAL for age estimates from facial images using the ResNet-34 architecture.

In [23], the Attention-based Dynamic Patch Fusion (ADPF) architecture uses two distinct CNNs—AttentionNet and FusionNet—to estimate age based on facial features. AttentionNet employs a new technique called Ranking-guided Multi-Head Hybrid Attention (RMHHA) to locate and rank age-specific patches dynamically. Using the identified patches and the

subject's facial image, FusionNet estimates the subject's age. To lessen patch overlap and direct the training of the AttentionNet, ADPF also adds a diversity loss. On a number of benchmark datasets for age estimation, the suggested framework performs better than cutting-edge approaches. In [24], they also used fusion to train their model. They manually took five important patches from the face image and then concatenated them with the full image.

Recently, deep-learning-based algorithms have gained a lot of attention from researchers as they are very powerful in extracting more accurate features automatically. Still, a huge amount of data are needed to feed deep learning-based algorithms. In [7], CDCNN was proposed to utilize low-quality images using a cross-dataset learning method with high-quality images. They used VGG-16 architectures pre-trained on ImageNet for training images. In this paper, we utilize this method to overcome insufficient data. In [25], the attention mechanism played an important role in NLP, showing a prominent result in computer vision, a field in which it is widely used. Different attention modules are already used in computer vision for image classification, recognition, and segmentation [11]. We utilize different attention mechanisms in our research through a deep convolution neural network.

3. Research Methodology

Estimating a person's age through facial images involves analyzing specific features of their face, such as wrinkles, sagging skin, eye bags and dark circles, age spots, and facial contours. To accomplish this, the proposed methodology utilizes a CNN architecture for feature extraction, which allows for the extraction of various scale features through the use of different kernel sizes within the layer. A different attention module is also implemented between the feature extractor and classification module to identify the image's important regions and overcome insufficient datasets.

3.1. Overview of Framework

A computer-vision-based system is proposed to estimate human age from facial images. Figure 4 provides an overview of the basic steps that the methodology contains. The total architecture is divided into four significant steps:

1. Dataset preprocessing;
2. Feature extraction;
3. Attention module;
4. Classification.

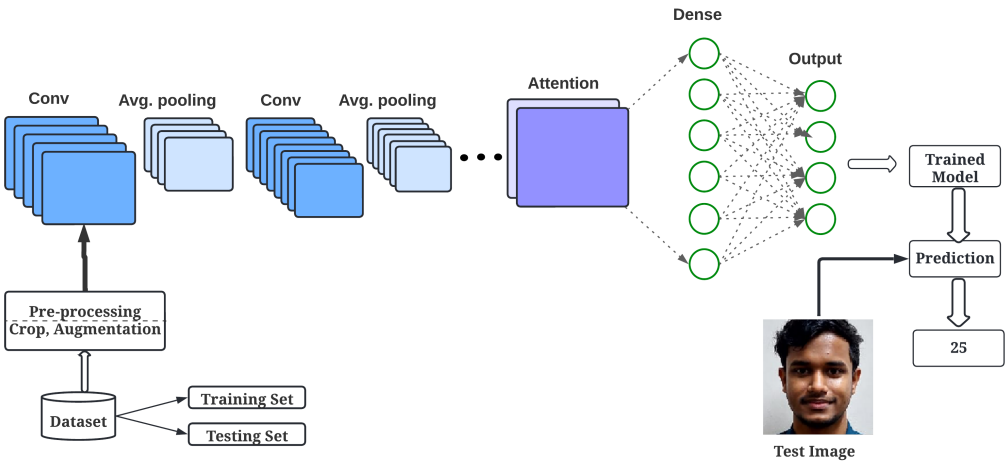

Figure 4. Block diagram of proposed age estimation methodology.

Initially, some preprocessing steps like scaling and resizing are performed on the input image. Then, the input image is passed through the different feature extractors (CNN). Some are transfer-learning-based, and some are from scratch. An attention module is added at the end of the feature extraction. An attention module is added in the proposed module

to improve the power of the identification of important areas in facial images. Finally, the produced feature of the attention module is flattened and passed through the classification module to estimate age.

3.2. Dataset and Preprocessing

Determining the best way to prepare visual input for convolutional neural network training is challenging. Data preprocessing includes scaling pixel values and utilizing techniques for improving image data throughout the model's training and evaluation. Considering the data preparation techniques, state-of-the-art models significantly achieve considerable performance on a challenging computer vision dataset as a practical shortcut, rather than trying a wide variety of options.

1. **Resizing:** One of the first tasks in data preprocessing is to create images of the same size. Images are resized and scaled arbitrarily within a specific range of dimensions. The annotations are changed appropriately. The deep learning model architecture needs images similar in size to function. Because convolution neural networks (CNN) only accept inputs of the same size, all images must be resized to a fixed size. On small images, deep learning algorithms frequently train more quickly. The network requires a long time to learn from larger input images, which increases the training time for the architecture. The ideal image size is therefore required. Depending on our GPU, all of the images are resized to ($228 \times 228 \times 3$) pixels to ensure all of the dataset images are the same size.
2. **Cropping:** Cropping is a popular image preprocessing method that entails eliminating a section of an image in order to accomplish various objectives. It crops images, resizes them, enhances their composition, or eliminates backgrounds. We used the Haar Cascades algorithm as it showed good results in face detection and cropping. Cropping assists in decreasing data noise and enhances the performance of models that use the images by deleting pointless portions of an image or shrinking it to an appropriate size. Cropping also enhances the image's composition and separates the topic from the background, which is beneficial for focusing on particular parts of the image or producing transparent backgrounds. All of the images are randomly cropped to $200 \times 200 \times 3$ pixels from the center of the image to make the model more vigorous, even if the image is partially seen.
3. **Data Augmentation:** Data augmentation is a powerful method used in deep learning to expand the dataset by generating new data from old data. Deep learning models need a lot of data to understand complicated patterns and generalize well to new data. However, large-scale data collection and labeling is time- and money-consuming. Data augmentation assists in developing new varieties of data similar to the original but with different properties by applying changes to the current images, such as rotation, flipping, scaling, or adding noise. Data augmentation helps the model be less overfitted, more accurate, and more resilient to changes in the input data. Overall, data augmentation is a crucial step in preprocessing image data for deep learning models, as it helps to improve model performance and reduces the need for large amounts of labeled data. We use random geometric transformations of the images in the dataset to increase the amount of training data in order to prevent the overfitting issue. Table 1 lists the used geometric transformations and their parameter ranges.

Table 1. We use different augmentation techniques to change the images and create diverse training examples, improving our model's ability to handle various situations.

Augmentation Technique	Parameter
horizontal_flip	True
rotation_clockwise_20degree	True
rotation_anti-clockwise_20degree	True
rotation_clockwise_40degree	True
rotation_anti-clockwise_40degree	True

3.3. Feature Extraction

Convolutional neural networks (CNNs) are used as the feature architecture in machine learning tasks, where the goal is to extract features from images or other types of data. In this approach, the pre-trained and from-scratch CNN model is used as a feature extractor rather than a classifier. CNNs are effective as feature architectures because they are designed to learn relevant features from image data automatically. In contrast to the deeper layers of a CNN, which learn more sophisticated features like shapes and textures, the early layers of a CNN learn basic elements like edges and corners.

3.3.1. Transfer Learning

For greater performance, deep learning algorithms require enormous amounts of data. Transfer learning is very beneficial when the dataset is relatively small because deep learning algorithms cannot generalize a pattern from a small dataset. Transfer learning is a method that uses a model that has already been developed on huge benchmark datasets like ImageNet [26] and COCO [27]. Various transfer learning models are used in our experiments, including DenseNet201 and MobileNetV2, trained on ImageNet.

DenseNet: In 2017, Gao Huang et al. [12] introduced a reliable convolutional neural network (CNN) architecture called DenseNet-201. This architecture features dense connections between layers, allowing for the more effective reuse of features and addressing the disappearing gradient issue. DenseNet-201 has 201 densely linked convolutional layers in each dense block, followed by a transition layer that reduces the output's spatial dimensions. These dense connections increase the flow of information through the network, enabling the model to capture more intricate details. The architecture of DenseNet-201 is divided into three parts: the initial convolutional layer, the dense blocks, and the classification layer. The network starts by extracting low-level features from an image through the initial convolutional layer. The output is then passed through multiple layers in each dense block, with each layer taking the concatenated feature maps of all preceding layers as input. This enables the network to reuse features more efficiently, resulting in fewer parameters required for training.

MobileNetV2: MobileNetV2 [13] is the updated version of MobileNetV1. In contrast to conventional residual models, which use symbolic representations in the input, the MobileNetV2 architecture is built on an inverted residual structure, where the input and output of the residual block are narrow bottleneck layers. When device capacity is an issue, this lightweight convolutional neural network design is highly helpful because it trains with fewer parameters. Lightweight depth-wise convolutions are used by MobileNetV2 to filter features in the intermediate expansion layer. This model has 53 convolution layers and 1 average pooling layer. It has two main components. These are inverted residual blocks and residual bottleneck blocks. There are two types of convolution layers in MobileNetV2 architecture. These are 1×1 convolution and 3×3 depth-wise convolution. By reducing the number of parameters (30% less than MobileNetV1) and necessitating fewer processes (2% fewer than MobileNetV1), these features make MobileNetV2 more efficient.

3.3.2. Scratch Model

CNN from Scratch: After analyzing various transfer-learning-based models, a convolution neural network (CNN) from scratch is also introduced to determine the human age

from face images to obtain a lower mean absolute error. The from-scratch model is mainly built using four convolutional layers (32, 64, 128, and 256 filters), two dense layers, and a dense output layer. All four convolution layers utilize smaller filters (3×3).

CNN from Scratch with Attention: An attention module is added at the end of the global average pooling of the from-scratch CNN model to further reduce the mean absolute error.

3.4. Attention Module

Instead of concentrating on the entire image, the attention mechanism aids in concentrating on a key area of the picture. In this study, we used several attention modules.

3.4.1. Self-Attention

Self-attention is trainable by backpropagation and generates attention weight by considering the entire image. The scaled dot-product is chosen as the attention-scoring function to perform a query and key vector element-wise dot production. The softmax function is used to normalize the output, which also removes any negative weight. It is a relation in which the pixel value or patch is the most important compared to other pixels or patches. The procedure described above is expressed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where $\sqrt{d_k}$ represents the dimension of the query vector Q and key vector K, and V represents the value vector. Figure 5 shows the architecture of the self-attention module.

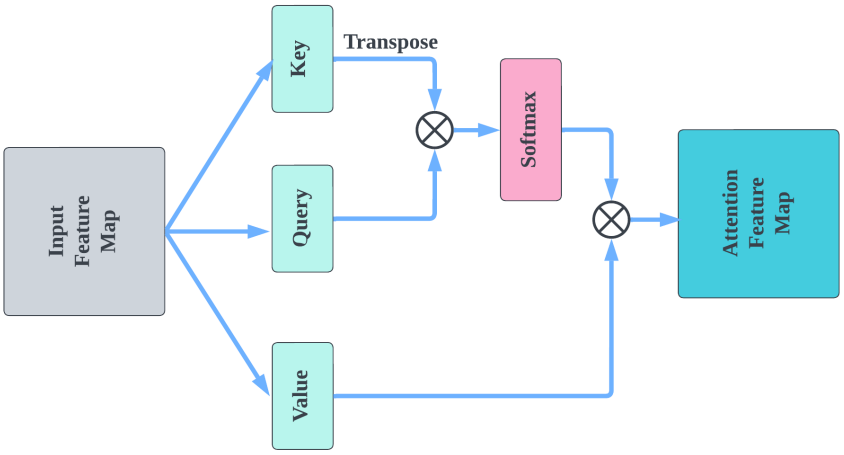

Figure 5. Self-attention module where $\otimes$ denotes the matrix multiplication.

3.4.2. Convolution Block Attention Module (CBAM)

CBAM is an attention module that makes use of channel attention (CA) and spatial attention (SA) in a sequential model to transfer the feature map to a refined feature map. These refined feature maps boost the network performance for networks to which this module is added. An overview of CBAM is given in Figure 6.

The formula for this attention module is represented by Equations (2) and (3):

$$F_c \quad : \quad CA(FeatureMap) \otimes FeatureMap \quad (2)$$

$$F_sc \quad : \quad SA(F_c) \otimes F_c \quad (3)$$

where the elementwise product is denoted by $\otimes$.

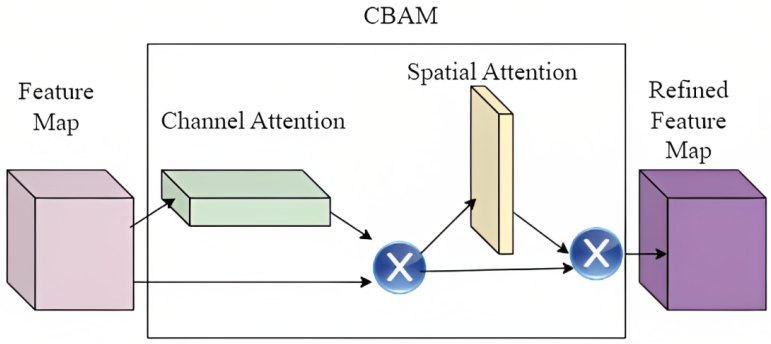

Figure 6. An overview of the convolution block attention module's structure.

3.4.3. Squeeze and Excitation Module

The squeeze-and-excitation module is used as an attention module by reconstructing the channel-wise feature in a dynamic manner, thus empowering the model to represent the extracted features more efficiently.

A detailed block diagram of the squeeze-and-excitation module is shown in Figure 7. As the figure depicts, the module has three sections: squeeze (SQ), excitation (EX), and scaling. Squeeze is carried out by applying pooling on a feature map that is global in nature. Next, excitation is carried out by applying relu and sigmoid activation functions in the dense layer, respectively. Finally, in the scaling section, each channel feature map is multiplied by the corresponding channel attention of the side network. The formula for the squeeze-and-excitation (SE) module is presented in Equations (4)–(6):

$$SQ \quad : \quad Avgpool(FeatureMap) \tag{4}$$

$$EX \quad : \quad \sigma(\delta(SQ)) \tag{5}$$

$$SE \quad : \quad FeatureMap \times EX \tag{6}$$

where the sigmoid and relu functions are denoted by σ and δ, respectively.

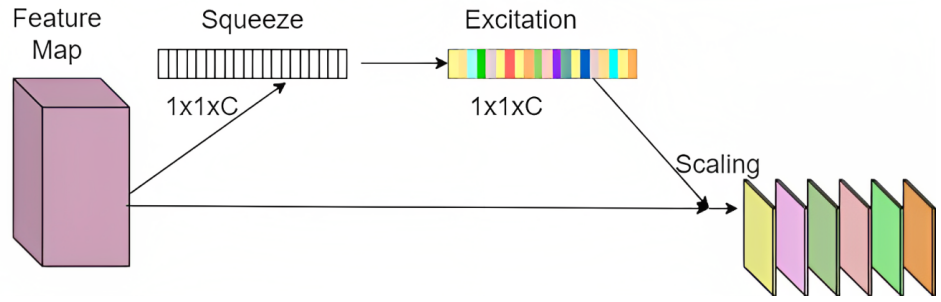

Figure 7. A squeeze-and-excitation channel attention module's architecture.

3.5. Classification Module

The classification module is added at the end of the attention module to estimate the human age from facial expressions. Two dense layers with several hidden units, 256 and 256, respectively, are used. The softmax activation in the output layer is used to create output probability distributions for human age classes.

3.6. Model Designing and Tuning

CNN from scratch with attention is selected as our foundation model. An input of a 2D convolution layer is paired with a 2D average pooling layer where the number of filters is 32 and the input shape of the image is 200 × 200. Three pairs of 2D convolutional layers paired with 2D average pooling layers, where the number of filters is 64, 128, and 256, are used as feature extractors. A 3 × 3 kernel is used in all four convolution layers, with Relu as the activation function, and the pool size is 2 × 2 for 2D average pooling. A

2D global average pooling layer is placed right after the final 2D average pooling layer. Figure 8 shows the architecture of the from-scratch CNN model.

In order to adjust the average weight of the feature before passing it to the dense layer based on the most significant feature located in the input image for a given problem domain, a self-attention module is added at the end of the 2D global average pooling, and before the dense layer. Two dense layers with several hidden units, 256 and 256, respectively, are used. The softmax activation in the output layer is used to create output probability distributions for human age classes.

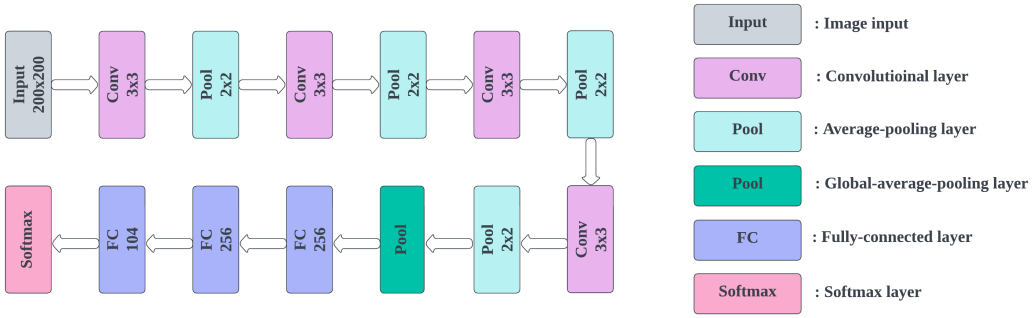

Figure 8. From-scratch CNN model architecture.

3.7. Training Details

For this work, we test a combined dataset that is made from UTKFace [28] and Facial-Age [29] to find the best hyperparameters for improved performance and we also utilize FG-NET [30] and CACD [31] datasets. We consistently used learning rates of 0.001 and 0.0001 throughout all experiments. Additionally, we utilized batch normalization for the convolution layer in our proposed model.

4. Result Analysis

In this section, we endeavor to present all of our experiments. Through the execution of these tests, we aimed to conduct a comprehensive evaluation of our proposed model.

To implement the suggested model, we conducted experiments on a Kaggle environment utilizing a computer equipped with an Intel(R) Xeon(R) CPU @ 2.00 GHz, 13 GB of RAM, and a high-performance 16 GB NVIDIA Tesla P100 GPU graphics card. The software stack employed is built upon Python version 3.7 and Keras version 2.11.0, which facilitated the development and execution of our deep learning models. Our training data consisted of 200 $\times$ 200 pixel images, and we utilized data augmentation to enhance performance. We divided our data into training and testing sets with a 70:30 ratio for this study.

4.1. Dataset Description

FG-NET: A total of 1002 photos from 82 participants make up the FG-NET [30] database, which was created to research actual ages. Both the color and grayscale resolutions of these photos are present. The ages of the participants range from 0 to 69. There are typically 12 photos for each person. The database also includes annotations for several human racial groups. There are several head postures, some facial expressions, and lighting effects in the pictures.

UTKFace: The UTKFace [28] dataset is a sizable face dataset with a broad age range (from 1 to 116 years old). The collection consists of over 23,707 face pictures with annotations for age, ethnicity, and gender. The images feature a wide variety of facial expressions, positions, lighting, resolution, occlusion, etc. A number of tasks, including age estimation, face detection, landmark localization, etc., could be performed using this dataset.

CACD: A sizable database called CACD [31] was created for face recognition and age retrieval tasks and assembled from the Internet. This database has 163,446 pictures of

2000 famous people. The age range is from 14 to 68 years old. A comparison of these (three) datasets is shown in Figure 9.

Combined Dataset: We developed a combined dataset by merging two existing datasets, namely UTKFace [28] and Facial-age [29]. Sample images of the combined dataset are shown in Figure 10. A total of 33,486 labeled facial images are combined to classify human age from age 1 to 116, as shown in Figure 11. We have the maximum number of images for the age of 26. We divided our dataset into 70:30 for training (23,440) and testing (10,046).

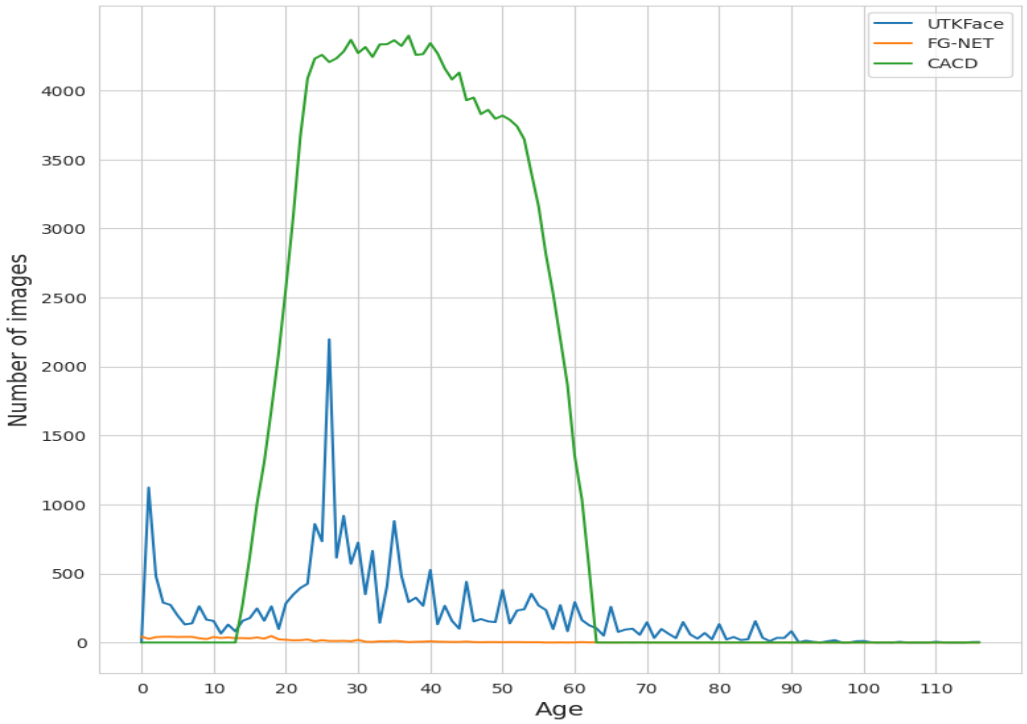

Figure 9. Distribution of the dataset used in our methodology.

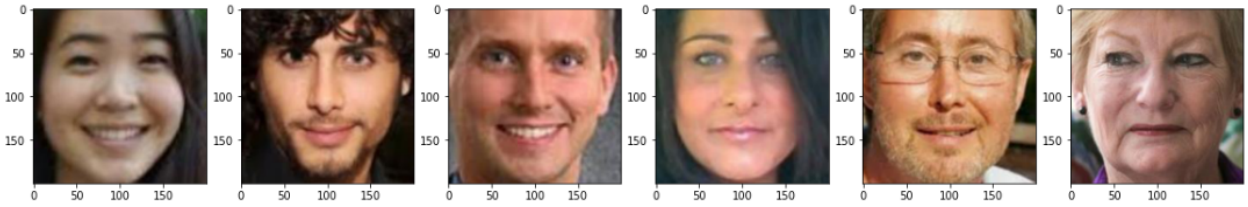

Figure 10. Sample of our merged dataset of UTKFace and Facial-age.

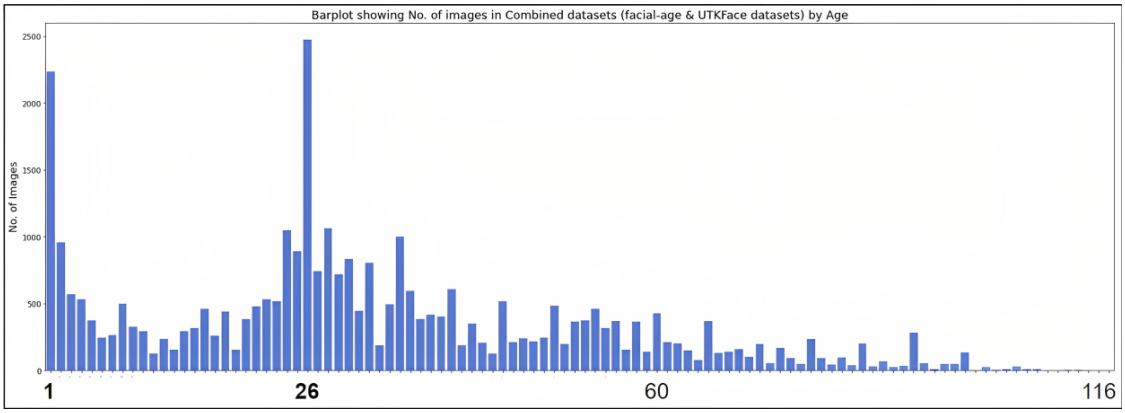

Figure 11. Distribution of our merged dataset.

4.2. Evaluation of Metric

Existing studies treat human age estimation from facial images as a regression or classification task, penalizing the discrepancies between the model-generated ages and the ages given by the dataset using MAE. We use MAE to evaluate human age estimation as a classification task. Mean absolute error, sometimes referred to as L1 loss, is a straightforward evaluation metric using simple loss functions. It is measured by averaging the absolute difference between the true and anticipated values over the given dataset. The average difference between the predicted age and the actual age is known as the MAE, and it is defined as follows:

$$\mathrm{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N} \tag{7}$$

where N is the total number of data samples, y_i represents the ground-truth age, and $\hat{y}_i$ represent the predicted age of the i-th sample.

To assess age estimation, another statistic is the cumulative curve (CS). The CS calculates the proportion of images that are accurately categorized within a particular age range, as follows:

$$\mathrm{CS(i)}(\%) = \frac{N_{e \leq i}}{N} 100\% \tag{8}$$

where $N_{e \leq i}$ is the number of test images for which an absolute inaccuracy cannot be more than i years.

4.3. Evaluation of Performance

We initially train two popular CNN networks—DenseNet201, MobileNetV2, and a from-scratch network—to conduct human age estimation using our combined dataset to choose a superior baseline network for the human age estimation task. As both DenseNet201 and MobileNetV2 are pre-trained with ImageNet, the combined dataset's RGB images ($200 \times 200 \times 3$) are used to train the models. In the other two models from scratch, with and without the attention module, images are converted to grayscale ($200 \times 200 \times 1$) to reduce the effect of lighting in color RGB images. We compared our four models over 50 epochs to determine the best model.

We first tested different batch sizes (128, 64, and 32) on our six models to choose the correct batch size and correct models for further processing. We established the number of samples used for each forward and backward pass during training. Although a bigger batch size necessitates more memory and processing power, it results in a faster convergence rate and a more reliable gradient estimate. On the other hand, a smaller batch size uses less memory and processing power but may result in a slower convergence rate and a less reliable gradient estimate. The ideal batch size is determined by the particular dataset, model design, and hardware configuration. In order to discover the best batch size for a particular scenario, it is usual practice to experiment with various batch sizes during model training. From the experiment, we found that using a batch size of 128 resulted in better outcomes for the CNN model built from scratch with attention. See Table 2 for the results.

Adam and stochastic gradient descent (SGD) optimizers are tested with different learning rates on the best model and their mean absolute errors are compared. When training machine learning models, optimizers and learning rates are essential hyperparameters. Distinct optimizers have distinct strengths and limitations, and they control how the model parameters are updated during backpropagation in order to minimize the loss function. The results shown in Table 3 show that the Adam optimizer with a learning rate of 0.0001 has better performance.

Table 2. Testing different batch sizes on the combined dataset helped us to find the optimal size for effective learning. In this table, we show a comparison of the effect of different batch sizes on the combined dataset.

Model	Batch Size	MAE (Year)
DenseNet210	128	5.557
DenseNet210	64	5.559
MobileNetV2	128	4.581
MobileNetV2	64	4.584
CNN from scratch	128	3.982
CNN from scratch	64	3.980
CNN from scratch	32	3.984
CNN from scratch with SE	128	3.93
CNN from scratch with SE	64	3.96
CNN from scratch with CBAM	128	3.6
CNN from scratch with CBAM	64	3.67
CNN from scratch with Self-Attention	128	**3.515**
CNN from scratch with Self-Attention	64	3.518
CNN from scratch with Self-Attention	32	3.527

Table 3. Comparison of the effect of different optimizers on the combined dataset.

Model	Optimizer	Learning Rate	MAE (Year)
CNN from scratch with Self-Attention	Adam	0.001	3.642
CNN from scratch with Self-Attention	Adam	0.0001	**3.515**
CNN from scratch with Self-Attention	SGD	0.001	3.593
CNN from scratch with Self-Attention	SGD	0.0001	3.611

4.3.1. Comparison with Existing Works

Comparison with existing works gives us an idea of what different researchers are thinking, and to what extent this work can be implemented. Many approaches are applied in solving this scenario, and we show a comparison with some.

Evaluation on Combined Dataset: We compared our model's performance with an existing model, which is shown in Table 4. Here, we used our combined dataset made with UTKFace and Facial-age on both our model and the existing model.

Table 4. Comparison of our proposed model on the combined dataset with existing methodology.

Papers	Methods	MAE (Year)
Cao et al. [22]	CORAL	4.0
Ours	CNN from scratch with Self-Attention	**3.5**

Evaluation on UTKFace: The evaluation findings of the suggested methodology for estimating age on the UTK-Face dataset using the MAE measure are shown in Table 5. To the best of our knowledge, this methodology performs better than other face-based age estimation methods already in use. This implies that our methodology is a viable option for age estimation in practical settings.

Table 5. Comparison of our proposed model on UTKFace with existing methodology.

Papers	Methods	MAE (Year)
Cao et al. [22]	CORAL	5.39
Berg et al. [32]	Randomized Bins	4.55
Shin et al. [21]	MWR	4.37
Ours	CNN from scratch with Self-Attention	**3.85**

Evaluation on CACD: The CACD dataset is publicly available to use for training and testing the face-based age estimation methodology. Many research works have used this dataset for their methodology and evaluated it to compare it with other methodologies. Our proposed methodology obtains the best mean absolute error (MAE) in comparison with the state-of-the-art methodology, which is demonstrated in Table 6.

Evaluation on FG-NET: For the FG-NET dataset, the MAE values are listed in Table 7 and the CS curve is shown in Figure 12. Again, not all methods for the FG-NET dataset present the findings under the CS metric. Table 7 indicates that our proposed methodology achieves an MAE value below 3.00, demonstrating its ability to perform effectively even with tiny datasets. From Figure 12, it is shown that the total number of correctly identified test images is the highest under an absolute error of 7.

Table 6. Comparison of our proposed model on CACD with existing methodology.

Papers	Methods	MAE (Year)
Cao et al. [22]	CORAL	5.35
Wang et al. [23]	ADPF	5.39
Zhang et al. [7]	CDCNN	4.58
Shin et al. [21]	MWR	4.41
Ours	CNN from scratch with Self-Attention	**4.24**

Table 7. Evaluation of our proposed model on FG-NET with existing methodology.

Papers	Parameters	MAE (Year)
IIS-LLD [3]	-	5.77
Feng et al. [33]	-	4.35
Mean-Variance Loss [34]	20M	4.10
DLDLF [35]	44M	3.71
DRF [35]	14M	3.47
ADPF [23]	14M	2.86
MWR [21]	-	2.23
Ours	**0.54M**	2.54

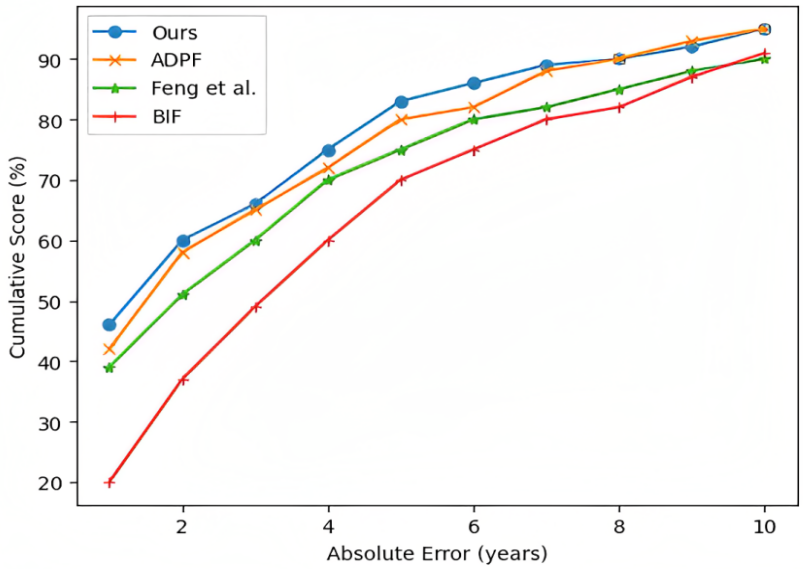

Figure 12. Cumulative score comparison with the existing model on FG-NET.

4.3.2. Analysis Through Generated Attention Map

From the preceding section, it is evident that our model (CNN from scratch with attention) has the lowest MAE on the UTKFace and CACD datasets. Figure 13 is presented

to demonstrate how well the model distinguishes between the many vital areas of the hand image. A heatmap is a visual representation that displays specific matrix values as colors. It is useful to use a heatmap to visualize the concentration of values between two dimensions of a matrix. This makes it easier to spot patterns and develop a sense of depth. Each image in the offered figures also includes the resulting attention maps.

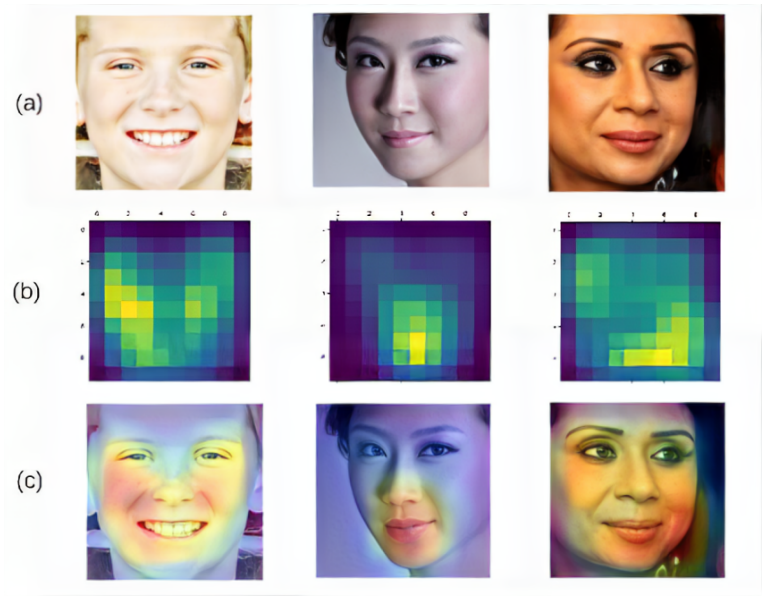

Figure 13. Heatmap and attention map for the combined dataset generated by the model (CNN from scratch with attention). (**a**) Original images from a combined dataset. (**b**) Heatmap generated for samples. (**c**) Resultant attention map for sample images.

4.3.3. Experiments with Real-World Images

We collect different face images with their actual ages (through National Identity Cards) and predict their ages using our proposed model. Some of them are shown in Figure 14. Table 8 shows the actual and predicted ages of the images shown in Figure 14 where MAE is 1.35. We also show the difference between them.

Figure 14. Real-world age estimation: predicting age from facial images.

Table 8. We put our model to the test with actual photos of people's faces. This table shows us how our model guesses their ages correctly in real situations.

Figure	Birth Date	Real	Predicted	AE (Year)
(a)	30 May 1999	24	25	1
(b)	16 October 1999	23.58	25	1.42
(c)	4 March 2004	19.25	17	2.25
(d)	25 July 1997	25.84	26	0.16
(e)	18 January 1999	24.34	26	1.66
(f)	14 December 1999	23.41	25	1.59

5. Discussion

Maintaining public safety and order is a critical aspect of any functioning society. The government and the authorities responsible for enforcing the law have a crucial role to play in ensuring the well-being and rights of both locals and visitors. The primary objective of public security is to identify and respond to potential threats and risks that could potentially undermine public safety, order, and peace. It is crucial to implement effective measures to tackle these challenges and create a secure environment for everyone. Deep learning and computer vision currently play important roles in ensuring cyber and public security [36]. By prioritizing public security, we can build a stronger, more resilient society that is well-equipped to face future challenges.

Facial age estimation is a powerful tool that can enhance public security and law enforcement efforts [4]. Its ability to identify and verify individuals is particularly important in cases where proper identification documents are not available. This technology can be invaluable in criminal investigations, especially in cases where missing persons or abducted children are involved. In addition, facial age estimation is useful in forensic investigations, as it can provide an estimate of the age of suspects or victims, aiding in criminal profiling [37]. It can also be used to enforce age restrictions and access control in various settings, ensuring compliance with age-related regulations. Additionally, facial age estimation can help in monitoring and surveillance, which can help identify and track potential threats to public safety. Age-based crime analysis can also benefit from facial age estimation, as it can help law enforcement to better understand criminal patterns and allocate resources for effective crime prevention strategies. Finally, accurate age verification can be an important tool in border control and immigration processes [38], adding an extra layer of security and compliance.

It is important to carefully consider ethical and privacy concerns when implementing facial age estimation technology. By finding a balance between public safety and individual privacy rights, we can ensure the responsible and lawful use of this technology. Collaboration between policymakers, technologists, and civil society can help us develop a framework that respects human rights while enhancing public security.

6. Conclusions

Age estimation from facial recognition is essential for security and law enforcement, but ethical and privacy concerns must be considered. In this work, experiments in the fields of human age estimation, using a CNN model with an attention mechanism, and leveraging transfer learning from two pre-trained CNN models—DenseNet201 and MobileNetV2—-are conducted. Images are resized and cropped as part of data preprocessing and an extensive augmentation to increase the number of images per class is carried out. A CNN model is chosen as the main feature extractor, and an attention module is used after the CNN model to adjust weight, focusing on the important patches of the images. Lastly, dense layers are used to predict human age. The experimental results show that the CNN model with attention performs better than the conventional CNN architecture for the same amount of data, with the lowest MAE for the UTKFace and CACD datasets (3.85 and 4.24, respectively). So, this proposed methodology has the ability to predict human age more precisely by increasing the size of the dataset. However, the model is susceptible to class imbalance and requires more images of older individuals for a robust system. This experiment has created numerous opportunities to improve human age estimation from facial expressions, which is still a complex and popular topic for researchers. In our future endeavors, we aim to refine the algorithm by incorporating a more advanced attention mechanism. This enhancement will help us achieve a higher degree of accuracy in human age classification across diverse real-world scenarios, which will ultimately lead to improved real-world performance. Additionally, we plan to explore novel experimentation strategies, such as utilizing features from multiple images to counteract noise and bias while augmenting the feature representation power. The collective aim of these strategies is to elevate the precision and robustness of our approach to age estimation.

Author Contributions: Conceptualization, K.D.; methodology, M.A.R.and K.D.; software, M.A.R.; validation, M.A.R.; formal analysis, M.A.R.; investigation, M.A.R.; resources, M.A.R., S.S.A. and K.D.; data curation, M.A.R.; writing—original draft preparation, M.A.R; writing—review and editing, M.A.R., S.S.A., K.D. and I.H.S. visualization, M.A.R.; supervision, S.S.A., K.D. and I.H.S.; project administration, S.S.A., K.D. and I.H.S.; funding acquisition, I.H.S. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The research data supporting this study are available on GitHub and Kaggle. FG-NET data are available at https://yanweifu.github.io/FG_NET_data/, UTKFace at https://susanqq.github.io/UTKFace/, CACD at https://bcsiriuschen.github.io/CARC/ and Facial-age at https://www.kaggle.com/datasets/frabbisw/facial-age .

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Scholarpedia. Facial Age Estimation—Scholarpedia.org. Available online: http://www.scholarpedia.org/article/Facial_Age_Estimation (accessed on 2 May 2023).
2. Fu, Y.; Guo, G.; Huang, T.S. Age synthesis and estimation via faces: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1955–1976. [PubMed]
3. Geng, X.; Yin, C.; Zhou, Z.H. Facial age estimation by learning from label distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2401–2412. [CrossRef] [PubMed]
4. Han, H.; Otto, C.; Jain, A.K. Age estimation from face images: Human vs. machine performance. In Proceedings of the 2013 International Conference on Biometrics (ICB), Madrid, Spain, 4–7 June 2013; pp. 1–8. [CrossRef]
5. Li, M.; Zhang, W.; Hu, B.; Kang, J.; Wang, Y.; Lu, S. Automatic assessment of depression and anxiety through encoding pupil-wave from HCI in VR scenes. *ACM Trans. Multimid. Comput. Commun. Appl.* **2022**. [CrossRef]
6. Al-Shannaq, A.S.; Elrefaei, L.A. Comprehensive analysis of the literature for age estimation from facial images. *IEEE Access* **2019**, *7*, 93229–93249. [CrossRef]
7. Zhang, B.; Bao, Y. Cross-dataset learning for age estimation. *IEEE Access* **2022**, *10*, 24048–24055. [CrossRef]
8. Angulu, R.; Tapamo, J.R.; Adewumi, A.O. Age estimation via face images: A survey. *EURASIP J. Image Video Process.* **2018**, *2018*, 42. [CrossRef]
9. Lemperle, G.; Holmes, R.E.; Lemperle, S.S.M. A classification of facial wri. *Plast. Reconstr. Surg.* **2001**, *108*, 1735–1750. [CrossRef]
10. Grazer, F.; Goldwyn, R. Abdominoplasty assessed by survey, with emphasis on complications. *Plast. Reconstr. Surg.* **1977**, *59*, 513–517. [CrossRef]
11. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
12. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 4700–4708.
13. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
14. Han, H.; Otto, C.; Liu, X.; Jain, A.K. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1148–1161. [CrossRef]
15. Guo, G.; Mu, G.; Fu, Y.; Huang, T.S. Human age estimation using bio-inspired features. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 112–119. [CrossRef]
16. Chen, H.; Wang, T.; Chen, T.; Deng, W. Hyperspectral image classification based on fusing S3-PCA, 2D-SSA and random patch network. *Remote. Sens.* **2023**, *15*, 3402. [CrossRef]
17. Kwon, Y.H.; da Vitoria Lobo, N. Age classification from facial images. *Comput. Vis. Image Underst.* **1999**, *74*, 1–21. [CrossRef]
18. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [CrossRef]
19. Hu, Z.; Wen, Y.; Wang, J.; Wang, M.; Hong, R.; Yan, S. Facial age estimation with age difference. *IEEE Trans. Image Process.* **2016**, *26*, 3087–3097. [CrossRef] [PubMed]
20. Othmani, A.; Taleb, A.R.; Abdelkawy, H.; Hadid, A. Age estimation from faces using deep learning: A comparative analysis. *Comput. Vis. Image Underst.* **2020**, *196*, 102961. [CrossRef]
21. Shin, N.H.; Lee, S.H.; Kim, C.S. Moving window regression: A novel approach to ordinal regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18760–18769.
22. Cao, W.; Mirjalili, V.; Raschka, S. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognit. Lett.* **2020**, *140*, 325–331. [CrossRef]

23. Wang, H.; Sanchez, V.; Li, C.T. Improving face-based age estimation with attention-based dynamic patch fusion. *IEEE Trans. Image Process.* **2022**, *31*, 1084–1096. [CrossRef]
24. Wang, H.; Wei, X.; Sanchez, V.; Li, C.T. Fusion network for face-based age estimation. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, Athens, Greece, 7–10 October 2018; pp. 2675–2679.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
26. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
28. UTKFace—Susanqq.github.io. Available online: https://susanqq.github.io/UTKFace/ (accessed on 4 May 2023).
29. Facial Age—Kaggle.com. Available online: https://www.kaggle.com/datasets/frabbisw/facial-age (accessed on 4 May 2023).
30. FG-NET Data by Yanwei Fu—Yanweifu.github.io. Available online: https://yanweifu.github.io/FG_NET_data/ (accessed on 13 May 2023).
31. Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval—Bcsiriuschen.github.io. Available online: https://bcsiriuschen.github.io/CARC/ (accessed on 13 May 2023).
32. Berg, A.; Oskarsson, M.; O'Connor, M. Deep ordinal regression with label diversity. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2740–2747.
33. Feng, S.; Lang, C.; Feng, J.; Wang, T.; Luo, J. Human Facial Age Estimation by Cost-Sensitive Label Ranking and Trace Norm Regularization. *IEEE Trans. Multimed.* **2017**, *19*, 136–148. [CrossRef]
34. Pan, H.; Han, H.; Shan, S.; Chen, X. Mean-variance loss for deep age estimation from a face. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-22 June 2018; pp. 5285–5294.
35. Shen, W.; Guo, Y.; Wang, Y.; Zhao, K.; Wang, B.; Yuille, A. Deep differentiable random forests for age estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 404–419. [CrossRef]
36. Sarker, I.H. Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. *Secur. Priv.* **2023**, *6*, e295. [CrossRef]
37. Aynsley-Green, A.; Cole, T.; Crawley, H.; Lessof, N.; Boag, L.; Wallace, R. Medical, statistical, ethical and human rights considerations in the assessment of age in children and young people subject to immigration control. *Br. Med. Bull.* **2012**, *102*, 17–42. [CrossRef] [PubMed]
38. Smith, E.; Marmo, M. Examining the Body through Technology: Age disputes and the UK border control system. *Anti-Traffick. Rev.* **2013**, *2*, 67–80. [CrossRef]