symmetry

MDPI

Article

Face Image Age Estimation Based on Data Augmentation and Lightweight Convolutional Neural Network

Xinhua Liu [1,2,*], Yao Zou [1,2,*], Hailan Kuang [1,2] and Xiaolin Ma [1,2]

[1] School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China; kuanghailan@whut.edu.cn (H.K.); maxiaolin0615@whut.edu.cn (X.M.)

[2] Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Wuhan University of Technology, Ministry of Education, Wuhan 430070, China

* Correspondence: liuxinhua@whut.edu.cn (X.L.); zouyao@whut.edu.cn (Y.Z.)

Received: 14 December 2019; Accepted: 9 January 2020; Published: 10 January 2020

Abstract: Face images contain many important biological characteristics. The research directions of face images mainly include face age estimation, gender judgment, and facial expression recognition. Taking face age estimation as an example, the estimation of face age images through algorithms can be widely used in the fields of biometrics, intelligent monitoring, human-computer interaction, and personalized services. With the rapid development of computer technology, the processing speed of electronic devices has greatly increased, and the storage capacity has been greatly increased, allowing deep learning to dominate the field of artificial intelligence. Traditional age estimation methods first design features manually, then extract features, and perform age estimation. Convolutional neural networks (CNN) in deep learning have incomparable advantages in processing image features. Practice has proven that the accuracy of using convolutional neural networks to estimate the age of face images is far superior to traditional methods. However, as neural networks are designed to be deeper, and networks are becoming larger and more complex, this makes it difficult to deploy models on mobile terminals. Based on a lightweight convolutional neural network, an improved ShuffleNetV2 network based on the mixed attention mechanism (MA-SFV2: Mixed Attention-ShuffleNetV2) is proposed in this paper by transforming the output layer, merging classification and regression age estimation methods, and highlighting important features by preprocessing images and data augmentation methods. The influence of noise vectors such as the environmental information unrelated to faces in the image is reduced, so that the final age estimation accuracy can be comparable to the state-of-the-art.

Keywords: CNN; age estimation; data augmentation; classification; regression

1. Introduction

Among the various biometric recognition technologies, face recognition is a biometric recognition technology with great development potential, and has broad application prospects in information security, public safety, and other fields. In academia, research topics related to classical face analysis usually include face detection [1], face recognition [2], face verification [3], face tracking [4], 3D facial expression recognition [5,6], etc. Among them, the analysis of face attributes such as age estimation, gender recognition, and ethnic recognition has attracted the interest of many researchers [7].

The face aging process generally follows some common aging modes. During the growth stage of children, the biggest change is the shape change caused by the growth of the skull. The aging process in adulthood is mainly reflected in changes in facial skin texture such as the appearance and deepening of wrinkles, loose skin, increased spots [8–10], etc. However, due to the complex facial features and

slow aging process, the degree of aging depends not only on the increase in age, but also due to various factors such as gender, race, genes, living habits, and health status [11]. In addition, the collection of face age images is very burdensome. The existing public face age datasets have many problems such as an imbalance in age, gender, and ethnicity, which makes it difficult to meet the requirements of most research work. The above reasons mean that the research on face age estimation still faces great challenges. Although facing huge challenges, face age estimation technology has a wide range of potential applications in the fields of surveillance and investigation, information management systems, intelligent human–computer interaction, social entertainment, and other fields.

The face age estimation process roughly includes image preprocessing, feature extraction, and age estimation. Image pre-processing methods include human face detection, face correction, and image cropping. Before neural network training in deep learning methods, image augmentation methods can be used to augment the dataset to alleviate the overfitting of the network. Image augmentation methods include filtering, sharpening, histogram enhancement, flipping, rotation, and scale transformation. In the feature extraction stage, traditional methods mostly use explicit feature extraction to obtain age features based on manual design [12,13]. Due to the limitations of hand-designed features, the extracted age features are not necessarily optimal. The modern feature extraction method based on convolutional neural networks can well capture the face-related feature information in the image, and has a strong robust adaptability to the noise in the image, which means the final estimation in the age estimation stage is more accurate. In the age estimation stage, there are roughly classification, regression, blending, and distributed learning methods. If age is regarded as a separate label, then age estimation is a classification problem. In addition, the age of the human face has a certain order, so that the age estimation can also be regarded as a regression problem. Some people have suggested that the facial features of adjacent ages have similarities. If the age correlation between adjacent ages is fully considered, the age estimation can be regarded as a distribution learning problem, so that multi-label learning or distribution learning can be used in the training of neural networks. Then, neighboring age tag information contributes to real age tags during the network training phase.

However, current convolutional neural networks (CNNs) are becoming larger and more complex, exposing many shortcomings such as too many model parameters, large footprints, large training dataset, long training time, and inconvenient deployment on mobile terminals. Therefore, a more lightweight network is needed, but at the same time, the accuracy of age estimation must be guaranteed.

Aiming at the above problems, this paper proposes an age estimation model based on ShuffleNetV2. Experiments show that the network proposed in this paper converges quickly during the training phase, has high accuracy during the age estimation phase, and has a small footprint of the network model. The contributions of this article are summarized as follows:

(1) A lightweight convolutional neural network age estimation model based on the mixed attention mechanism is constructed.
(2) The age estimation method combining classification and regression is easy to implement and the final age estimation accuracy is very high.
(3) Perform face detection and correction on the input face image, and perform image augmentation, so that the feature information related to the face age is amplified, which is helpful for network learning.

2. Related Work

2.1. Feature Extraction

At the feature extraction stage, Kwon et al. [12] first used anthropometric models to extract facial age features, and based on craniofacial development theory and skin wrinkle features, roughly divided ages into three categories: infants, youth, and elderly. Lanitis et al. [13] first applied Active Appearance Model (AAM) to the study of face age estimation. Based on AAM features, a quadratic regression function was used for age estimation. Geng et al. [11] proposed a concept of aging patterns

subspace (AGES). Guo et al. [14] introduced manifold learning into the age estimation of face images, and mapped high-dimensional face datasets to low-dimensional manifolds, that is, transforming face images into a low-dimensional age feature. The method based on manifold learning greatly relaxes the requirements for training data. Gunay et al. [15] applied local binary pattern (LBP) to age estimation, and achieved relatively good results. Since then, many improved face age estimation methods based on LBP have appeared. Guo et al. [16] also proposed a bio-inspired feature (BIF), which has attracted the attention of many domestic and foreign scholars on age estimation due to better experimental results. In recent years, deep learning technologies such as CNN have been gradually applied to age estimation, and have achieved better results than manually designed features. Dong et al. [17] used CNN in age estimation for the first time and designed a fully learned end-to-end age estimation system. This method introduces the multi-scale analysis strategy of traditional methods into CNN, which significantly improves the performance, but the CNN structure is very shallow. Wang et al. [18] applied CNN to the extraction of facial age features. Unlike the explicit feature extraction of traditional methods, the implicit features learned by CNN avoid the limitations brought by hand-designed features. However, in the research work of [18], CNN is only used for feature extraction, and then it is input into a separate classification or regression model for age estimation. Niu et al. [19] implemented an end-to-end learning method that uses deep convolutional neural networks to perform feature learning and regression modeling simultaneously. The absolute mean age error of the model on the MORPH2 and AFAD datasets was 3.27 ± 0.02 and 3.34 ± 0.08, respectively. Chen et al. [20] proposed a CNN-based age estimation network framework (ranking-CNN), which includes a series of basic CNN networks, each of which trains a label. Liu et al. [21] proposed a new face age estimation feature learning method-ordered deep feature learning. Gao et al. [22] regarded the age label as a multi-label problem and proposed deep label distribution learning. Kang et al. [23] used deep residual CNN to solve the problem of robustness of age estimation, but the accuracy of age estimation was not high. Jeong et al. [24] used a multi-task Siamese network to improve the accuracy of age estimation, but the network has a large structure and a bloated model.

The CNN networks above-mentioned in terms of age feature extraction are basically heavyweight networks such as VGG-16 and ResNet. With the development of lightweight networks in recent years, Mobilenetv2 [25], ShuffleNetV2 [26], SSR-NET [27] and other networks have greatly accelerated the deployment process of neural network models in mobile terminals. Although these lightweight networks have much fewer parameters than heavyweight networks, their feature extraction capabilities are very strong, and the network training time is not long.

In addition, in order to better extract the image feature information, the necessary data preprocessing is required before feature extraction. This paper used dlib [28] to first perform face detection, then performed rotation correction based on the relative positions of the eyes based on the feature points, and then cropped to remove unwanted information such as the environmental information, leaving only important information related to the face. In order to expand the training dataset and alleviate the overfitting phenomenon of the neural network during the training process, image augmentation on the training dataset can achieve good results. The specific method is to rotate, flip, and add noise to the images according to a certain probability during the training process.

2.2. Age Estimation

Age estimation methods include classification, regression, ranking, and combination. These age estimation methods can be used in conjunction with various feature extraction methods to complete the age estimation task. Many excellent methods have appeared in recent years.

Hu et al. [29] used the method of Uniform Local Binary Patterns (ULBP) + Principal Component Analysis (PCA) + Support Vector Machine (SVM) for age estimation. Guo et al. [30] proposed a locally adjusted robust regression (LARR) algorithm, which combines SVM and Support Vector Regression (SVR) when estimating age by first using SVR to estimate a global age range, and then using SVM to perform exact age estimation. Chao et al. [31] better understood the relationship between facial

features and age through metric learning and dimensionality reduction processing, and proposed an age-oriented local regression method for the age estimation of complex facial aging processes. Rothe et al. [32] used the deep expectation method to estimate the age. First, the Softmax operation was used to calculate the probability of each age category at the output layer of the network, and then the expected value was calculated as the estimated age. This method was the champion of the apparent age estimation competition in 2015. Its main contribution is to publish IMDB-WIKI dataset (images of celebrities from IMDb and Wikipedia), the largest database of face age and gender estimates to date (although there are many errors in it, such as not being human images). Chang et al. [33] proposed an ordinal hyperplane sorting algorithm OHRank, which estimates the age of humans through face images. The design of the algorithm is based on the relative order information between the age tags in the database. Human age is inferred by aggregating a set of preferences and their cost sensitivity from an ordered hyperplane.

The age estimation method used in this paper is a hybrid method of classification and regression. Considering age as independent multiple categories, then age estimation is a classification problem, a fully-connected layer of the neural network was constructed with a classification layer containing 101 neurons (representing the face 0–100 years old). In this way, the estimated age value can be calculated according to the method in [32], and the classification loss is used as part of the joint loss function. If you consider that age change is a continuous process, you can also consider the age estimation problem as a regression problem. This paper added a regression layer with only one neuron behind the classification layer, and calculated the absolute deviation of the estimated age from the real age as the regression loss and the previous classification loss constitute a loss function in the neural network training process of this paper.

3. Proposed Method

3.1. Architecture

The overall block diagram of our face age estimation system is shown in Figure 1. As can be seen from the figure, the original picture may be subjected to image augmentation operations before entering the neural network. There will be multiple sets of comparative experiments to compare and analyze the effectiveness of this image augmentation operation. The architecture of the neural network is based on the ShuffleNetV2 network, but the last output layer of the original ShuffleNetV2 has 1000 neurons. We first transformed it into 101 neurons, and then added an output with only one neuron behind this layer.

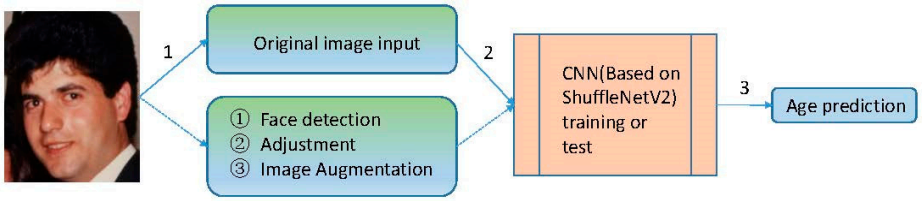

Figure 1. The framework of our method.

The focus of attention in computer vision is to let the system learn to focus on the places of interest. A neural network with an attention mechanism can, on one hand, learn the attention mechanism autonomously; on the other hand, the attention mechanism can in turn help us understand the world seen by the neural network.

In order to improve the feature extraction capability of the ShuffleNetV2 network, we also added a hybrid attention mechanism and a residual module. The attention mechanism originates from the study of human vision. In cognitive science, due to the bottleneck of information processing, humans selectively focus on a part of all information while ignoring other visible information. Later experiments proved that this operation can effectively improve the accuracy of the final age estimation.

3.2. Mixed Attention-ShuffleNetV2

ShuffleNetV2 is a lightweight convolutional neural network proposed by Ma et al. in 2018. It has advantages over networks such as Mobilenetv2 in terms of training time and feature extraction capabilities. The layers of the ShuffleNetV2 network are shown in Table 1. As the last layer of the original sfv2 network is a fully connected layer containing 1000 neurons, we made corresponding modifications based on the age estimation task. Specifically, first, the number of neurons in the last layer was changed to 101, and then a regression layer with only one neuron was added behind it.

Table 1. The overall architecture of ShuffleNetV2.

Layer	Output Size	KSize	Stride	Repeat	Output Channels
Image	224 × 224				3
Conv1 MaxPool	112 × 112	3 × 3	2	1	24
Stage2	28 × 28		1	3	244
Stage3	14 × 14		1	7	488
Stage4	7 × 7		1	3	976
Conv5	7 × 7	1 × 1			2048
GlobalPool	1 × 1	7 × 7			
FC					101
LastFC					1

The number of output channels of each Shuffle layer can be scaled. The last column of the table is the number of output channels of the Shuffle layer. Originally, there are various combinations, for example, 0.5 times, 1 time, 1.5 times, and 2 times. The corresponding output channels are [24, 48, 96, 192, 1024], [24, 116, 232, 464, 1024], [24, 176, 352, 704, 1024], and [24, 244, 488, 976, 2048]. [24, 244, 488, 976, 2048] is the 2x mode used in this paper to obtain stronger feature extraction capabilities. Very, very deep neural networks are difficult to train because there are gradient disappearance and gradient explosion problems, but then adding the residual network can alleviate this problem. The structure of this paper's convolutional neural network based on ShuffleNetV2 is shown in Figure 2.

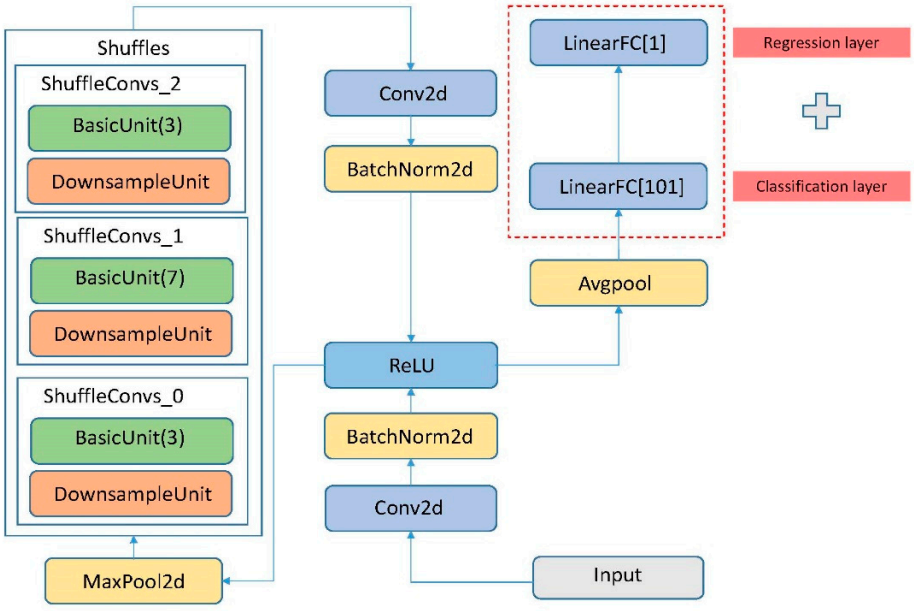

Figure 2. The composition of our Convolutional Neural Networks (CNN) based on ShuffleNetV2.

The basic unit of ShuffleNetV2 is shown in Figure 3. It can be seen that [14] designed an operation called channel splitting.

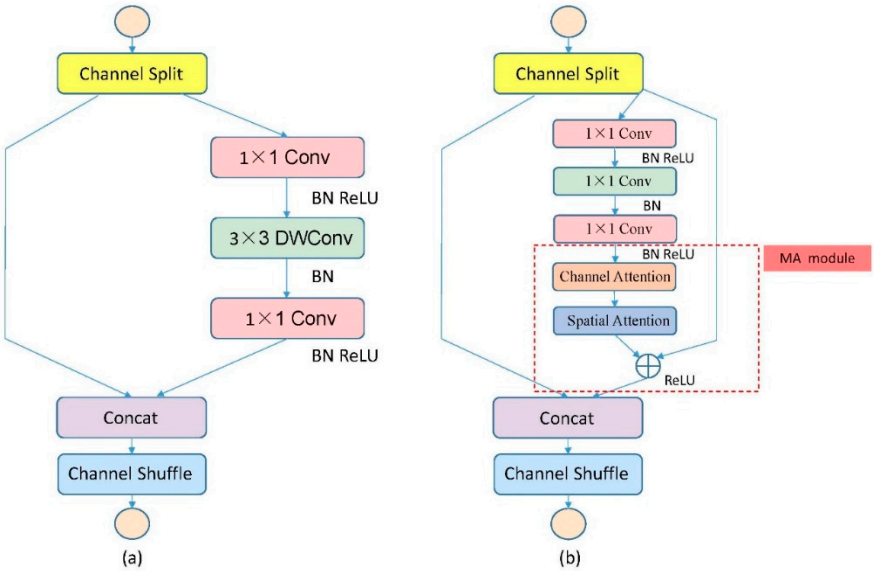

Figure 3. The basic unit of ShuffleNetV2. (**a**) The basic unit of original ShuffleNetV2. (**b**) The basic unit of our ShuffleNetV2 with mixed attention mechanism added.

At the beginning of each unit, the input channels are split into two branches, where one branch is used for identity mapping and the other branch passes three convolutional layers while ensuring that the number of output channels is the same as the number of input channels. After convolution, the two branches are spliced with the same number of channels. The subsequent channel cleaning operation enhances the information exchange between the two branches.

In recent years, the topic about attention mechanism has been very hot. Introducing the attention mechanism into the neural network can enlarge the area of interest in the image or assign different weights to multiple channels of the feature map. These can enhance the convolutional neural network. The feature extraction ability improves the accuracy of the final age estimation. Hu et al. [34] proposed Squeeze-and-Excitation networks (SENet), which is a channel attention mechanism. Woo et al. [35] proposed a convolution block attention module (CBAM). The convolution module attention mechanism is mainly composed of the channel attention module and the spatial attention module. The channel attention mechanism in CBAM is more complicated than SENet and the calculation operations are greater. For the spatial attention mechanism, we used CBAM's spatial attention module. We will introduce related attention modules below, and then introduce this mixed attention mechanism into the basic unit of ShuffleNetV2. Specifically, the output of the 1×1 convolution layer is passed through a channel attention module, and then goes through a spatial attention module and a residual module before the concatenation operation.

The channel attention module is shown in Figure 4. This module is mainly composed of the Squeeze operation and Excitation operation. Given an input x, the number of feature channels is c_1. After a series of general transformations such as convolution, a feature with the number of feature channels c_2 is obtained. Unlike the traditional CNN, the next step is to re-calibrate the features obtained by three operations. The first is the Squeeze operation. We performed feature compression along the spatial dimension, turning each two-dimensional feature channel into a real number. This real number has a global receptive field to some extent, and the dimensions of the output match the number of feature channels of the input. The second is the Excitation operation, which is a mechanism similar to a gate in a recurrent neural network. Weights are generated for each feature channel by a parameter w, where the parameter w is learned to explicitly model the correlation between feature channels. The last

is the Reweight operation. The weight of the output of Excitation is regarded as the importance of each feature channel after feature selection, and then it is weighted to the previous features channel by channel by multiplication to complete the recalibration of the original features in the channel dimension. The diagram of channel attention module is shown in Figure 4.

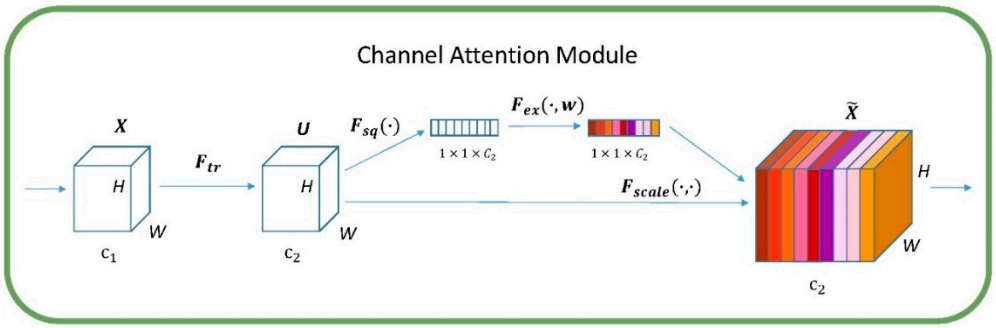

Figure 4. Diagram of the channel attention module.

The F_{tr} operation is a standard convolution operation. Strictly speaking, it does not belong to SENet. The definition of its input and output is expressed as follows: $F_{\mathrm{tr}} : X \rightarrow U, X \in \mathbb{R}^{H' \times W' \times C'}$, $U \in \mathbb{R}^{H \times W \times C}$ Then this F_{tr} convolution operation is shown in Equation (1) where v_c represents the c-th convolution kernel and x_s represents the s-th input.

$$u_{\mathrm{c}} = v_{\mathrm{c}} * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{1}$$

The U obtained by F_{tr} is the second three-dimensional matrix on the left in Figure 4, also called C, feature maps with the size $H \times W$. u_c represents the c-th two-dimensional matrix in U, and the subscript c represents the channel.

Squeeze operation is a global average pooling where the formula is as follows.

$$z_{\mathrm{c}} = F_{sq}(u_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i, j) \tag{2}$$

Therefore, Equation (2) converts the input of $H \times W \times C$ into the output of $1 \times 1 \times C$, corresponding to the F_{sq} operation in Figure 4.

The next step is the Excitation operation, as shown in Equation (3). The previous squeeze result is z. Here, we multiply z by W_1, which is a fully connected layer operation. Then, going through a ReLU layer, the output dimension is unchanged. Then, multiply with W_2, and this operation is also a fully connected layer process. Finally, go through the sigmoid function to get s.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{3}$$

After getting s, we can operate on the original tensor U, which is shown in Equation (4). u_c is a two-dimensional matrix, and s_c is a number, which is the weight, so it is equivalent to multiplying each value in the u_c matrix by s_c.

$$\widetilde{X}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{4}$$

The spatial attention module is shown in Figure 5. First, the input feature map is still compressed using average pooling and max pooling, but the compression here has become channel-level compression, and mean and max operations were performed on the input features in the channel dimension. Finally, two two-dimensional features were obtained, and were stitched together according to the channel dimensions to obtain a feature map with two channels. Then, a convolution operation

was performed using a hidden layer containing a single convolution kernel. The feature is consistent with the input feature map in the spatial dimension.

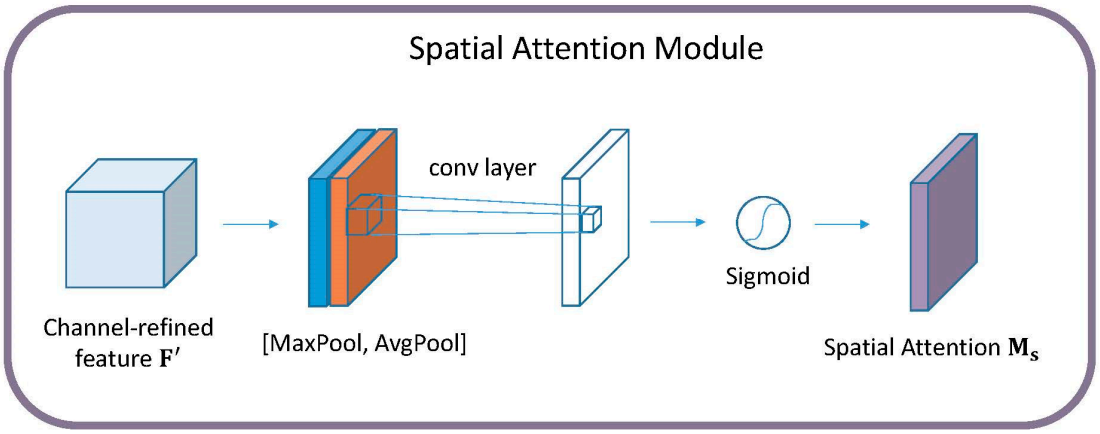

Figure 5. Diagram of the spatial attention module.

Define the feature map after max pooling and average pooling operations as $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F^s_{max} \in \mathbb{R}^{1\times H\times W}$. The mathematical processing of this part can be expressed by the following equation:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times 7}([F^s_{avg}; F^s_{\max})) \end{aligned} \tag{5}$$

Applying the attention mechanism to both the channel and spatial dimensions simultaneously can improve the feature extraction capability of the network model without significantly increasing the amount of calculation and parameters.

3.3. Data Augmentation

This section introduces data augmentation processing, which is a series of image preprocessing operations, and operations to alleviate overfitting of the network training.

As shown in Figure 6, the face in the picture is first detected using the dlib face image processing library, and the facial features are positioned, then the image is rotated according to the relative positions and angles of the eyes, so that the connection between the eyes is at a horizontal line. The image is then cropped according to the approximate range of the detected human face. Then, during the neural network training process, the output image is randomly subjected to image augmentation operations such as flipping left and right, rotating, scaling, and adding noise according to a certain probability. This can alleviate the over-fitting of the network and enhance the robustness of the model. Experiments show that after these operations, the training effect of the network will be better, and the age estimation accuracy of the final model will be higher.

Figure 6. Face feature point detection and pose correction.

3.4. Loss Function

As mentioned previously, the convolutional neural network in this article has two fully connected layers at the end. The penultimate layer contains 101 neurons, representing 101 age categories, that is, 0–100 years old. Calculating cross-entropy as a classification loss, the last layer contains a neuron, and the absolute deviation of the output of the neuron from the true age label is calculated as the regression loss. The two losses together constitute the objective function of the method in this paper. The gradient descent method is used to back-propagate the error and update the weight information of the neural network until the final loss curve converges.

1. Classification loss

Softmax is used in the multi-classification process. It maps the output of multiple neurons into the (0, 1) interval, which can be understood as a probability to perform multi-classification. For the penultimate layer of the neural network in this article, there were a total of 101 classifications, and the output was recorded as the array V, and V_{i} represents the i-th element in V, then the Softmax value of this element is shown in Equation (6).

$$V_{\mathrm{i}} = \frac{e^{i}}{\sum_{j} e^{j}} \tag{6}$$

The predicted age is based on the practice of the literature [20] to calculate the expectation after the Softmax operation, as shown in Equation (7).

$$E = \sum_{\mathrm{i}=0}^{|K-1|} y_i \cdot V_i \tag{7}$$

where K stands for 101 categories and $y_i \in [0, 101), 0 \leq i \leq 100$. In multi-classification problems, the loss function is a cross-entropy loss function. For the sample points (x, y), y is the real label. In a multi-classification problem, its value can only be a set of labels. As shown in Equation (8), $y_{i,k}$ is a one-hot vector; K represents 101 label values and the probability that the i-th sample is predicted to be the k-th label value is $p_{i,k}$. There are N samples in total.

$$L_{\mathrm{c}} = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{\mathrm{k}=0}^{K-1} y_{i,k} \log p_{i,k} \tag{8}$$

2. Regression loss

Regression loss is the absolute deviation of the output of the last layer of neurons from the real age label. As shown in Equation (9), y_1 represents the predicted age and y represents the true age.

$$L_{\mathrm{r}} = \left| y_1 - y \right| \tag{9}$$

Therefore, the joint loss function of classification loss and regression loss is as follows. Lambda is a hyperparameter used to control the proportion of regression loss in the joint loss function.

$$L = L_{\mathrm{c}} + \lambda \cdot L_r \tag{10}$$

4. Experiments

4.1. Datasets

The face age dataset is the data basis for age estimation. Due to the privacy of age, it is very difficult to collect face image datasets with real age. Currently, there are many face datasets with age labels, but most of them are undisclosed.

The MORPH (Craniofacial Longitudinal Morphological Face Database) [36] dataset is the largest face age dataset that has been published. It mainly includes two parts, namely the Album1 and Album 2 datasets. At present, only the whole Album1 dataset and some of the data of the Album2 dataset are disclosed. The public part of Album2 has a total of more than 55,000 pictures of more than 13,000 people, and each picture is marked with information such as the age and gender of the character. This experiment uses this part of the data, called the MORPH2 dataset. Some examples of the MORPH2 dataset are shown in Figure 7.

Figure 7. Some examples in the MORPH2 dataset.

FG-NET (Face and Gesture Recognition Research Network Aging Database) [37] is a popular public dataset currently used in face age estimation and aging synthesis research. It contains a total of 1002 color or grayscale images of 82 people. Since some images are obtained by scanning old photos, there are certain differences in the size, tone, background, and lighting conditions of the images.

Some examples of the FG-NET dataset are shown in Figure 8. There are people of different races and colors worldwide. At the same time, there are many studies on facial biometric recognition. Some studies have made good progress on age, race, and gender feature recognition. It should be pointed out that the objects studied in this paper were limited to the above two datasets. About 77% of the images in the MORPH2 dataset are black, about 19% are white, and about 4% are Asian, Indian, etc. The small number of the FG-NET dataset and the various features above-mentioned make it more suitable for simple testing and comparison with the methods of similar papers.

Figure 8. Some examples in the FG-NET (Face and Gesture Recognition Research Network Aging Database) dataset.

4.2. Evaluation Metrics

At present, the mean absolute error (MAE) and cumulative score (CS) are two objective evaluation indexes in the research of age estimation.

1. Mean absolute error (MAE)

   MAE is defined as the average of the absolute deviations of all age estimates and true values.

2. Cumulative Score (CS)

CS represents the ratio of the number of samples whose absolute error value between the estimated value and the true value is less than the specified threshold to the total number of test samples. The specific expression is:

$$CS(e) = \frac{1}{N}\sum_{i=1}^{N} 1\left\{\left|y_i - y_i'\right| \le e\right\} \tag{11}$$

Among them, $1\left\{\left|y_i - y_i'\right| \le e\right\}$ means that when the absolute error between the estimated value y_i' and the real value y_i is less than or equal to the e years, the value is 1, otherwise it is 0, which is the total number of test samples. The size of the CS value is related to the selection of the threshold value e. Under the condition of a fixed threshold value, the larger the CS value, the more samples in a specific error range, the better the age estimation effect.

4.3. Parameter Setting

Our experimental platform was as follows. Operating System (OS): Ubuntu 16.06; Graphics Processing Unit (GPU): Quadro P6000; Memory: 24 GB. The deep learning framework uses PyTorch. PyTorch was developed by the Torch7 team. As its name indicates, PyTorch differs from Torch in that PyTorch uses Python as the development language. It is a Python-first deep learning framework that can not only achieve powerful GPU acceleration, but also support dynamic neural networks. This is not supported by many mainstream frameworks such as Tensorflow. PyTorch can be seen as a numpy with GPU support, but also as a powerful deep neural network with automatic differentiation.

The input of the network is a 224×224 pixels RGB (Red, Green, Blue) face image. The network training optimizer uses the Adam optimizer ($\beta 1 = 0.9$, $\beta 2 = 0.999$), and its learning rate is 0.001. Step attenuation is used, and every 20 epochs, the learning rate is attenuated by a 0.2 attenuation rate. The batch size is 128, and a total of 80 epoch networks can be trained to converge. In the later experiments, the weight of the regression loss is 0.1. In subsequent experiments, the learning rate, attenuation coefficient, batch size, and number of training rounds are all the same. Whether to add the regression loss and whether to perform the image augmentation operation depends on the specific experimental settings.

4.4. Multiple Sets of Experiments

4.4.1. Experiment on MORPH2

The ratio of the MORPH2 training dataset, validation dataset, and test dataset is 7:2:1. A total of four groups of experiments were performed: the training dataset was divided into two types of MORPH2 and MORPH2_ALIGN, according to whether or not face alignment was performed. At the end of the network, the classification layer and the regression layer are used. Therefore, the network structure is divided into two types, according to whether the mixed attention mechanism is used: SFV2_l1 and MA_SFV2_l1. Then, the loss function curves and MAE curves of the four sets of experimental results are shown in Figure 9. Each training is performed for 80 epoch, and if face alignment is not performed, it takes about 116 min, otherwise it takes about 138 min.

In addition, we conducted experiments to explain the important role of image augmentation in mitigating overfitting. As shown in Figure 10, if there is no image augmentation, the network is severely overfitting.

It can be seen that after image augmentation, with the use of the mixed attention mechanism and the combined use of classification and regression loss, the final MAE value was relatively low, indicating that the method proposed in this paper contributes to the improvement of the accuracy of age estimation. As the LOSS curves and MAE curves of the experiments in each group are tested on the validation set that accounts for 20% of the total dataset, the final MAE needs to be tested on the test set that accounts for 10% of the total dataset using the trained model to represent the performance of the current model. We finally collected the results of a total of 33 experiments and plotted them using box and whisker plots. The results are shown in Figure 11.

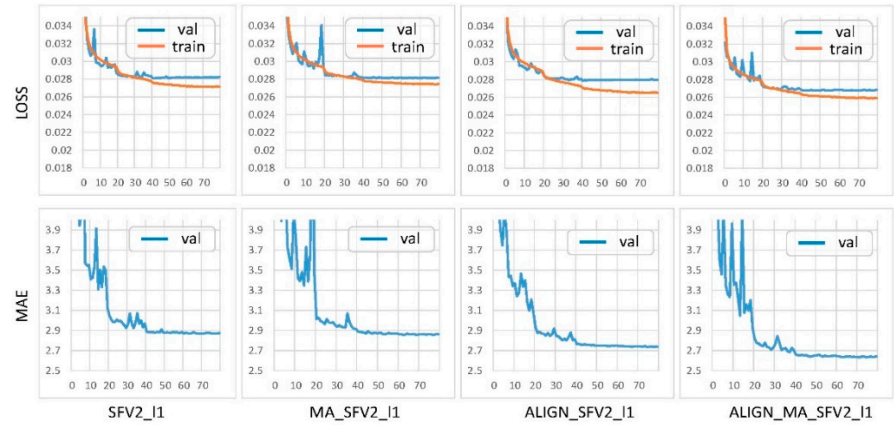

Figure 9. LOSS curves and MAE (Mean Absolute Error) curves of the four groups of experiments.

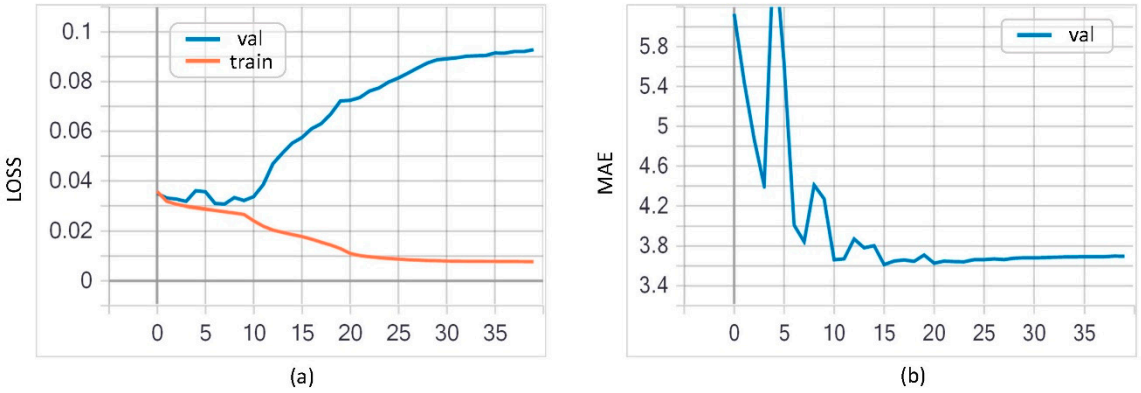

Figure 10. Final LOSS (**a**) and MAE (**b**) (Mean Absolute Error) curves without pre-processing data using image augmentation.

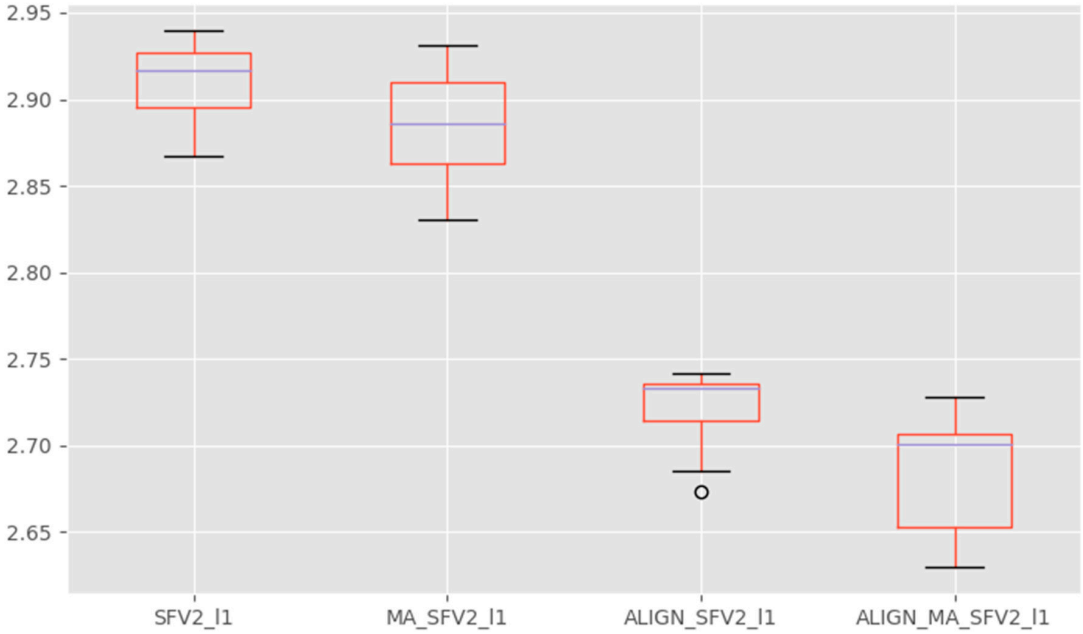

Figure 11. MAE results plotted by box and whisker plots of four group experiments.

There were 33 experiments in total, and each experiment had four groups of experiments, namely SFV2_l1, MA_SFV2_l1, ALIGN_FV2_l1, and ALIGN_MA_SFV2_l1. The final mean and variance of MAE were 2.91 (0.020), 2.88 (0.027), 2.72 (0.018), and 2.68 (0.030) respectively. We can see that the

variance of the MAE value was small, and that the change will not be very large, which fully shows that the improved method proposed in this paper improves the accuracy of age estimation.

4.4.2. Experiment on FG-NET (Face and Gesture Recognition Research Network Aging Database)

Since multiple sets of comparative experiments were performed on the MORPH2 dataset, the network model we used on the FG-NET dataset was MA-SFV2_l1. In addition, the FG-NET dataset was divided into two sets of comparative experiments based on whether or not image augmentation operations were performed. However, the capacity of the FG-NET dataset is not large, so according to previous practice, the method of leaving one person was adopted, that is, each specific experiment in each group of experiments uses all age pictures of one person as the test set, and all other pictures as the training set. Since FG-NET has a total of 82 individual pictures, a total of 82 experiments were performed, and the average value of each experiment was finally used as the final performance index.

We undertook 33 experiments, and the final mean and variance of MAE were 4.12 (0.025) and 3.81 (0.031) respectively.

4.5. Quantitative Comparison with Other Methods

4.5.1. Comparison on MORPH2

Table 2 shows the MAE values of the classic or latest age estimation methods such as DEX (Deep EXpectation), OHRank (Ordinal Hyperplanes Ranker), AD (Age Difference), AGEn (AGE group-n encoding), and C3AE (Compact yet efficient Cascade Context-based Age Estimation model) on the MORPH2 dataset. As can be seen from the comparison in the table, the age estimation method proposed in this paper had a good effect on the MORPH2 dataset, and achieved better results than the traditional age estimation methods, and excellent results based on deep learning models such as DEX, GA-DFL, and C3AE. In particular, even though methods such as DEX (IMDB-WIKI) and C3AE (IMDB-WIKI) use a large amount of additional training data during the training process, that is, pre-training on the IMDB-WIKI dataset first, these methods are not as effective as our method, which trains the network from scratch directly, which also fully proves that our system has strong competitiveness. In addition, the network model of methods such as AGEn is very large, training is time-consuming and takes up a lot of storage space, and in comparison, our lightweight network model has a short training time, takes up a small amount of space, and is convenient for deployment on mobile terminals.

Table 2. Comparison of MAE (Mean Absolute Error) on MORPH2 (Craniofacial Longitudinal Morphological Face Database).

Methods	MAE
AGES [11]	8.83
OR-CNN [19]	3.27
DEX [32]	3.25
DEX (IMDB-WIKI)	2.68
OHRank [33]	6.07
CS-LR [38]	4.59
Sparsity [39]	3.45
GA-DFL [40]	3.37
AD [41]	2.78
AGEn [42]	2.93
AGEn (IMDB-WIKI)	2.52
ARN [43]	3.00
ODFL-ODL [44]	3.12
C3AE [45]	2.78
C3AE (IMDB-WIKI)	2.75
MA-SFV2	2.68

According to the calculation method of the cumulative score (CS), the CS values under different error thresholds are calculated on the MORPH2 dataset, and the corresponding CS curves are plotted in Figure 12.

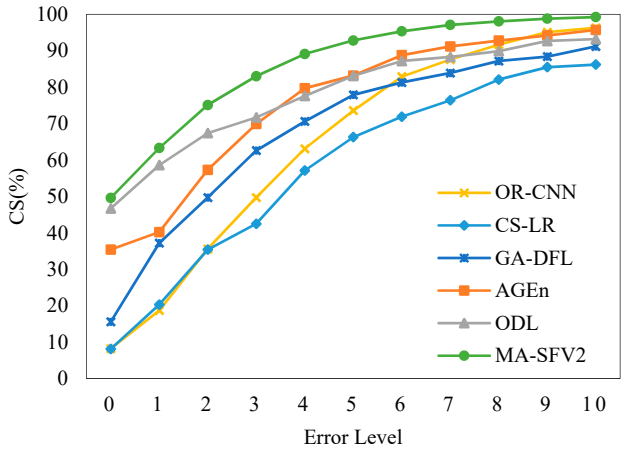

Figure 12. CS (Cumulative Score) curve on MORPH2.

As can be seen from the figure, with the increase of the allowable error, the CS curve in this paper showed a steady growth, and the value of the CS curve in this paper was higher from the beginning, which is obviously better than other methods.

4.5.2. Comparison on FG-NET

Table 3 shows the MAE values of the classic or latest age estimation methods such as DEX, OHRank, and C3AE on the FG-NET dataset.

Table 3. Comparison of MAE on FG-NET.

Methods	MAE
AGES [11]	6.77
Ranking-CNN [20]	4.13
LARR [30]	5.07
DEX [32]	4.63
DEX (IMDB-WIKI)	3.09
OHRank [33]	4.48
Sparsity [39]	4.25
GA-DFL [40]	4.16
C3AE [45]	4.09
C3AE (IMDB-WIKI)	2.95
CS-LBFL [46]	4.36
DRFs [47]	3.85
CA-SVR [48]	4.67
MA-SFV2	3.81

It can be seen from the comparison that the MAE value of the method (MA-SFV2) on the FG-NET dataset was 3.79, and the results were significantly better than the age estimation methods based on hand-designed features such as OHRank and CS-LBFL in the table, and also better than the deep learning-based age estimation models such as DEX, Ranking-CNN, and C3AE. However, the age estimation results of DEX (IMDB-WIKI) and C3AE (IMDB-WIKI) were better than MA-SFV2. The DEX (IMDB-WIKI) model and C3AE (IMDB-WIKI) were pre-trained on the IMDB-WIKI dataset during the construction process, and then fine-tuned on the FG-NET dataset, using a large number of extra training data.

5. Conclusions

This paper was based on a lightweight convolutional neural network, combining image augmentation, mixed attention mechanism, and age estimation algorithms combined with classification and regression. Under the premise of ensuring that the training time is reasonable and the network parameters are appropriate, there is no need to pass IMDB-WIKI pre-training of large face databases that can achieve higher age estimation accuracy than the current state-of-the-art. However, it is still not lightweight enough to deploy the model in this article to a mobile terminal, so the research direction in the future is to continue to explore how to use a more compact model to achieve the desired age estimation effect.

Author Contributions: In this work, X.L. conceived the face image age estimation method based on lightweight CNN and designed the experiments; Y.Z. proposed the MA-SFV2 network and optimized the experimental details. Y.Z. and H.K. performed the experiments and analyzed the data; Y.Z. drafted the manuscript; X.L., H.K., and X.M. edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Natural Science Foundation of China (Nos. 61772088 and 61502361).

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

References

1. Klare, B.F.; Klein, B.; Taborsky, E.; Blanton, A.; Cheney, J.; Allen, K.; Grother, P.; Mah, A.; Jain, A.K. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1931–1939.
2. Zhao, W.; Chellappa, R.; Phillips, P.J.; Rosenfeld, A. Face recognition: A literature survey. *ACM Comput. Surv. CSUR* **2003**, *35*, 399–458. [CrossRef]
3. Chen, J.-C.; Patel, V.M.; Chellappa, R. Unconstrained face verification using deep cnn features. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
4. Cao, C.; Hou, Q.; Zhou, K. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph. TOG* **2014**, *33*, 43. [CrossRef]
5. Vezzetti, E.; Tornincasa, S.; Moos, S.; Marcolin, F. 3D Human Face Analysis: Automatic Expression Recognition. *Biomed. Eng.* **2016**. [CrossRef]
6. Vezzetti, E.; Marcolin, F. 3D Landmarking in multiexpression face analysis: A preliminary study on eyebrows and mouth. *Aesthet. Plast. Surg.* **2014**, *38*, 796–811. [CrossRef] [PubMed]
7. Guo, G.; Mu, G. A framework for joint estimation of age, gender and ethnicity on a large database. *Image Vis. Comput.* **2014**, *32*, 761–770. [CrossRef]
8. Ramanathan, N.; Chellappa, R. Modeling shape and textural variations in aging faces. In Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, The Netherlands, 17–19 September 2008; pp. 1–8.
9. Yun, F.; Guodong, G.; Huang, T.S. Age synthesis and estimation via faces: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1955–1976. [CrossRef]
10. Suo, J.; Zhu, S.-C.; Shan, S.; Chen, X. A compositional and dynamic model for face aging. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 385–401.
11. Geng, X.; Zhou, Z.-H.; Smith-Miles, K. Automatic Age Estimation Based on Facial Aging Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 368. [CrossRef]
12. Kwon, Y.H.; da Vitoria Lobo, N. Age Classification from Facial Images. *Comput. Vis. Image Underst.* **1999**, *74*, 1–21. [CrossRef]
13. 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. [CrossRef]
14. Guo, G.; Fu, Y.; Dyer, C.R.; Huang, T.S. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Process.* **2008**, *17*, 1178–1188.

15. Gunay, A.; Nabiyev, V.V. Automatic age classification with LBP. In Proceedings of the 2008 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, 27–29 October 2008.
16. 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.
17. Dong, Y.; Zhen, L.; Li, S.Z. Age Estimation by Multi-scale Convolutional Network. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.
18. Wang, X.; Guo, R.; Kambhamettu, C. Deeply-Learned Feature for Age Estimation. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 6–9 January 2015; pp. 534–541.
19. Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; Hua, G. Ordinal Regression with Multiple Output CNN for Age Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4920–4928.
20. Chen, S.; Zhang, C.; Dong, M. Deep Age Estimation: From Classification to Ranking. *IEEE Trans. Multimed.* **2017**, *20*, 2209–2222. [CrossRef]
21. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Ordinal deep feature learning for facial age estimation. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 157–164.
22. Gao, B.B.; Xing, C.; Xie, C.W.; Wu, J.; Geng, X. Deep Label Distribution Learning with Label Ambiguity. *IEEE Trans. Image Process.* **2017**, *26*, 2825–2838. [CrossRef] [PubMed]
23. Kang, J.; Kim, C.; Lee, Y.; Cho, S.; Park, K. Age Estimation Robust to Optical and Motion Blurring by Deep Residual CNN. *Symmetry* **2018**, *10*, 108. [CrossRef]
24. Jeong, Y.; Lee, S.; Park, D.; Park, K. Accurate Age Estimation Using Multi-Task Siamese Network-Based Deep Metric Learning for Front Face Images. *Symmetry* **2018**, *10*, 385. [CrossRef]
25. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
26. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
27. Yang, T.-Y.; Huang, Y.-H.; Lin, Y.-Y.; Hsiu, P.-C.; Chuang, Y.-Y. SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation. *IJCAI* **2018**, *5*, 7.
28. King, D.E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
29. Hu, L.; Li, Z.; Liu, H. Age Group Estimation on Single Face Image Using Blocking ULBP and SVM. In *Proceedings of the 2015 Chinese Intelligent Automation Conference*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 431–438.
30. Guo, G.; Yun, F.; Huang, T.S.; Dyer, C.R. Locally Adjusted Robust Regression for Human Age Estimation. In Proceedings of the 2008 IEEE Workshop on Applications of Computer Vision, Copper Mountain, CO, USA, 7–9 January 2008.
31. Chao, W.-L.; Liu, J.-Z.; Ding, J.-J. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognit.* **2013**, *46*, 628–641. [CrossRef]
32. Rothe, R.; Timofte, R.; Van Gool, L. Deep Expectation of Real and Apparent Age from a Single Image without Facial Landmarks. *Int. J. Comput. Vis.* **2016**, *126*, 144–157. [CrossRef]
33. Chang, K.Y.; Chen, C.S.; Hung, Y.P. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
35. Woo, S.; Park, J.; Lee, J.-Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
36. Ricanek, K.; Tesafaye, T. Morph: A longitudinal image database of normal adult age-progression. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, New York, NY, USA, 10–12 April 2006; pp. 341–345.

37. Panis, G.; Lanitis, A. An Overview of Research Activities in Facial Age Estimation Using the FG-NET Aging Database. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
38. 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.* **2016**, *19*, 136–148. [CrossRef]
39. Dong, Y.; Lang, C.; Feng, S. General structured sparse learning for human facial age estimation. *Multimed. Syst.* **2017**, *25*, 49–57. [CrossRef]
40. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Group-aware deep feature learning for facial age estimation. *Pattern Recognit.* **2016**, *66*, 82–94. [CrossRef]
41. Hu, Z.; Wen, Y.; Wang, J.; Wang, M.; Hong, R.; Yan, S. Facial Age Estimation with Age Difference. *IEEE Trans. Image Process.* **2017**, *26*, 3087–3097. [CrossRef] [PubMed]
42. Tan, Z.; Wan, J.; Lei, Z.; Zhi, R.; Guo, G.; Li, S.Z. Efficient Group-n Encoding and Decoding for Facial Age Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2610–2623. [CrossRef] [PubMed]
43. Agustsson, E.; Timofte, R.; Van Gool, L. Anchored regression networks applied to age estimation and super resolution. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1643–1652.
44. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Ordinal Deep Learning for Facial Age Estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 486–501. [CrossRef]
45. Zhang, C.; Liu, S.; Xu, X.; Zhu, C. C3AE: Exploring the Limits of Compact Model for Age Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12587–12596.
46. Lu, J.; Liong, V.E.; Zhou, J. Cost-Sensitive Local Binary Feature Learning for Facial Age Estimation. *IEEE Trans. Image Process.* **2015**, *24*, 5356–5368. [CrossRef]
47. Shen, W.; Guo, Y.; Wang, Y.; Zhao, K.; Wang, B.; Yuille, A.L. Deep regression forests for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2304–2313.
48. Chen, K.; Gong, S.; Xiang, T.; Loy, C.C. Cumulative Attribute Space for Age and Crowd Density Estimation. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, UT, USA, 23–28 June 2013; pp. 2467–2474.