electronics

MDPI

Article

Hybrid Spatial-Channel Attention Mechanism for Cross-Age Face Recognition

Wenxin An and Gengshen Wu *

Faculty of Data Science, City University of Macau, Macao SAR 999078, China; d22091100308@cityu.edu.mo
* Correspondence: gswu@cityu.edu.mo; Tel.: +853-8590-2289

Abstract: Face recognition techniques have been widely employed in real-world biomimetics applications. However, traditional approaches have limitations in recognizing faces correctly with large age differences because of significant changes over age in the same person, leading to unsatisfactory recognition performance. To address this, previous studies propose to decompose and identify age and identity features independently in facial images across diverse age groups when optimizing the discriminative model so as to improve the age-invariant face recognition accuracy. Nevertheless, the interrelationships between these features make it difficult for the decomposition to disentangle them properly, thus compromising the recognition accuracy due to the interactive impacts on both features. To this end, this paper proposes a novel deep framework that incorporates a novel Hybrid Spatial-Channel Attention Module to facilitate the cross-age face recognition task. Particularly, the proposed module enables better decomposition of the facial features in both spatial and channel dimensions with attention mechanisms simultaneously while mitigating the impact of age variation on the recognition performance. Beyond this, diverse pooling strategies are also combined when applying those spatial and channel attention mechanisms, which allows the module to generate discriminative face representations while preserving complete information within the original features, further yielding sounder recognition accuracy. The proposed model is extensively validated through experiments on public face datasets such as CACD-VS, AgeDB-30, and FGNET, where the results show significant performance improvements compared to competitive baselines.

Keywords: age-invariant face recognition; attention mechanism; generative model; deep network optimization

Citation: An, W.; Wu, G. Hybrid Spatial-Channel Attention Mechanism for Cross-Age Face Recognition. *Electronics* **2024**, *13*, 1257. https://doi.org/10.3390/electronics13071257

Academic Editor: Byung-Gyu Kim

Received: 24 February 2024
Revised: 21 March 2024
Accepted: 26 March 2024
Published: 28 March 2024

1. Introduction

Face recognition technology is a method of identifying and verifying individuals by analyzing their facial features using computer vision and pattern recognition techniques [1–5]. This technology has been widely adopted in various fields such as security, identity verification, and social media [1,2]. However, traditional face recognition approaches have limitations when it comes to recognizing faces with large age differences. This is because the human face changes significantly over time due to factors such as skin loosening and subtle bone changes, making it difficult for ordinary face recognition systems to identify the same person at different ages [6]. To deal with this, Age-Invariant Face Recognition (AIFR) has been proposed as a promising solution [4]. Nevertheless, significant differences in facial features and poor image quality still pose challenges in recognizing faces across diverse age groups, which can compromise the overall recognition accuracy [5,7]. Figure 1 shows examples of face images of different people over time.

Previous studies have explored ways to differentiate between faces of various ages by analyzing age-related features such as skin texture, wrinkles, and eye bags [8]. To do this, they have proposed various deep neural network-based techniques that learn age-specific feature representations. Additionally, these researchers have developed face image datasets that contain different age groups to train and evaluate cross-age face recognition

algorithms [9]. For example, many researchers have used generative models like Generative Adversarial Networks (GANs) [10] (e.g., Conditional Generative Adversarial Networks (cGANs) [11]) to improve the quality of synthetic face images. A face-aging model is then constructed to extract features from the input image and generate a face image with a specific age, where subsequently face recognition is performed using a target image of the same age generated by this model. However, it is important to note that generative models often have drawbacks such as high computational complexity and even model instability.

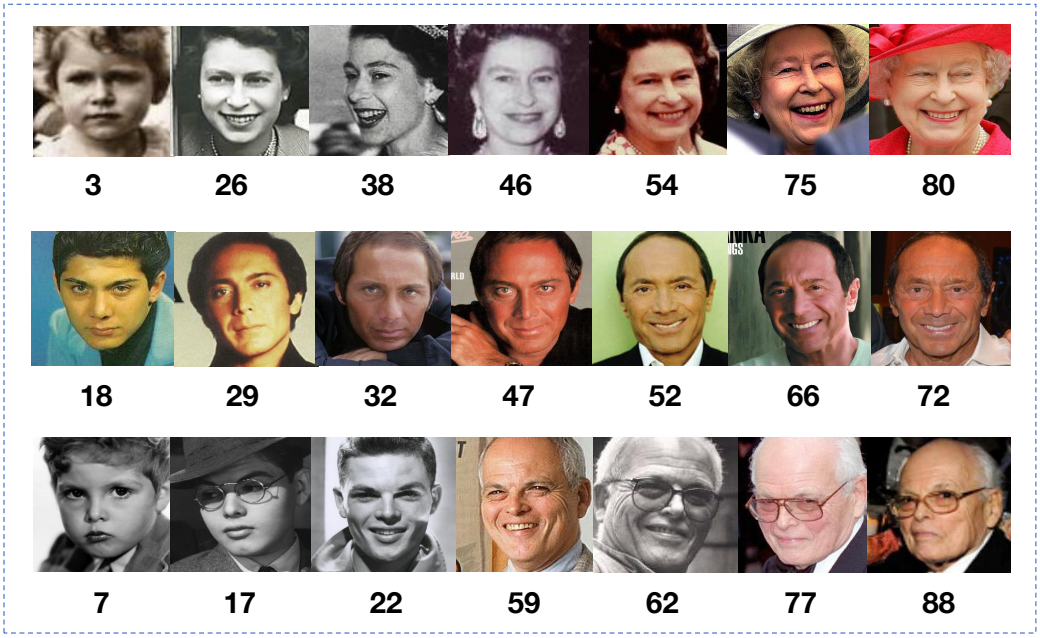

Figure 1. Example of the CACD-VS dataset [12] showing images of each subject at different ages, using three of them as examples.

On the contrary, the approach based on discriminative models is considered a more reliable option for face recognition as it is less prone to model instability. It is supposed to directly extract age-independent identity feature information from hybrid features to mitigate the age impact to facilitate the subsequent cross-age recognition. For instance, the hidden factor analysis is used to eliminate age factors from hybrid face features, resulting in high robustness and improved cross-age face recognition accuracy [5,9]. However, it is worth noting that there can still be strong interrelationships between the decomposed identity components, which may contain age information to some extent and could in turn affect the recognition results.

To this end, this paper proposes a deep-face recognition framework that adopts a novel hybrid attention mechanism to address the issue of poor identity discrimination in current discriminative models, where the latter struggle to extract complete and undisturbed identity features from face attributes. Particularly, the proposed Hybrid Spatial-Channel Attention Module (HSCAM) can decompose the face features completely by using spatial and channel attention simultaneously while preserving the rich information within images to improve the accuracy of cross-age face recognition. The main contributions of this paper are summarized as follows:

- In this work, we propose a new deep-face recognition framework that incorporates a novel Hybrid Spatial-Channel Attention Module (HSCAM) for better facial feature representation to enhance cross-age face recognition performance.
- The proposed HSCAM significantly benefits cross-age face recognition by decomposing the hybrid face features with both spatial and channel attention mechanisms to eliminate the age variations on the recognition performance. Moreover, it integrates various pooling strategies when performing those attention mechanisms, which

enables the preservation of complete information within original features, further boosting recognition accuracy with more distinctive face representations.
- Extensive experiments on three benchmark datasets demonstrate the superiority of the proposed method in improving the accuracy of cross-age face recognition in the testing phase. These results further validate the effectiveness of the HSCAM.

The remaining part of this paper is arranged as follows: Section 2 discusses some related works. In Section 3, we elaborate on the proposed deep framework, including the overall network structure, learning modules, objective function, and optimization. Next, we evaluate and analyze the proposed method on public datasets and compare it with recent baselines in the experiments. We report the results in Section 4. Finally, we summarize the conclusion of this work in Section 5.

2. Related Works

Many deep-based approaches have been proposed in the field of AIFR. This paper aims to introduce relevant works by categorizing them into two groups: general AIFR methods and multi-task learning-based AIFR methods. In this case, general AIFR methods focus on building and optimizing discriminative models to better distinguish facial features of different age groups. These methods extract specific age and identity information from images to enhance the accuracy and robustness of cross-age face recognition. On the other hand, multi-task learning-based AIFR methods combine multiple tasks such as face synthesis and recognition into a unified framework. These approaches improve the overall performance of the system in the field of cross-age facial recognition by involving diverse learning strategies.

2.1. General AIFR Methods

Discriminative models extract age-independent features of individuals directly and use them for matching in face recognition. Among them, the method termed hidden factor analysis (HFA) that was proposed in [13] captures two factors affecting face features—an age-invariant identity factor and an age factor affected by aging. Observed facial appearance features containing age and identity factors are then modeled. In 2016, a deep-face recognition framework with two parallel networks (LF-CNN) [14] was launched to learn age-invariant deep-face features. Subsequently, the researchers proposed Orthogonal Embedding CNN (OE-CNN) [15] to learn age-invariant identity features. OE-CNN decomposes facial features into two orthogonal components, trains identity-related features as a multi-class classification task with A-Softmax loss supervised by identity labels, and trains age-related features as a regression task supervised by age labels [16].

Then, the Decorrelated Adversarial Learning (DAL) algorithm proposed in [17] aims to find the maximum correlation between pairs of features generated by a backbone mesh by introducing a canonical mapping module (CMM) while training the backbone mesh and decomposition module to generate features with reduced correlation. This results in the removal of age-related components from the hybrid features. The algorithm decomposes the hybrid face features into two uncorrelated components: the identity component and the age component. The identity component provides useful information for face recognition. Therefore, the model decomposes age and identity features to significantly reduce their correlation. To ensure that both contain the correct information, the identity and age features are supervised by ID and age-preserving signals, respectively.

Consequently, a recent study has proposed a method for age-invariant face recognition called Implicit and Explicit Feature Purification (IEFP) [18]. This method helps to remove age information from facial features and obtain a purer representation of the same. Another method, called Low-Complexity Attention Module (LCAM) [19], uses three parallel branches of the attention mechanism with only one convolutional operation in each branch. It is a lightweight method and shows better performance in face recognition tasks. Another approach, called Lightweight Attentive Angular Distillation (LIAAD) [20], has been introduced to extract age-invariant attentional and angular knowledge into a

lightweight student network. This approach uses two high-performance heavy networks as teachers with different expertise, making it more robust with higher FR accuracy and robustness to age factors.

2.2. Multi-Task Learning-Based AIFR Methods

Researchers have been working on methods to improve cross-age face recognition by combining multiple models [21]. One approach is to use multi-task learning to simultaneously learn feature representations of multiple attributes such as age and identity. This improves the robustness and generalization ability of cross-age face recognition algorithms by transforming and expanding training data.

For example, a model called Parallel Multi-path Age Distinguish Network (PMADN) has been proposed in a recent study [22]. This model comprises two networks: Age Distinguish Mapping Network (ADMN) and Cross-Age Feature Recombination Network (CFRN). These networks can effectively extract age-robust features for identity features and identity classification, especially for younger ages. In 2021, a unified multitasking framework called MTLFace [23] was proposed, which jointly handles both tasks. The approach decomposes features into two uncorrelated components like identity-related and age-related features, and de-correlates them through multi-task training and continuous domain adaptation. A new identity conditional module is also proposed to improve the age smoothing of synthesized faces.

In addition, a method called MT-MIM (i.e., mutual information minimization) [24] has been proposed to train a de-entanglement network using a multi-task learning framework that minimizes the mutual information between the identity and age components. This is achieved by using the learning of de-entangled representations as an information-constrained objective. The aim is to unravel age and identity embeddings of facial images, which is achieved by using a multi-task learning and Wasserstein distance discriminator to minimize the mutual information between these embeddings [25]. Another semi-supervised learning method, termed Cross-Age Contrastive Learning (CACon) [26], has been proposed to address the problem of limited supervised data in age-invariant face recognition. CACon introduces a new contrastive learning method that uses additional synthetic samples from the input image and proposes a new loss function associated with contrastive learning on ternary samples.

Although the discriminant model can improve the accuracy of cross-age face recognition, there are still potential interrelationships between the decomposed components, and the decomposed identity component still contains age information [27]. This largely affects the recognition effect. The face synthesis method in the joint model still has the problem of generating faces that are too smooth or with poor model interpretability. Therefore, there is much room for improvement [28]. Our research focuses on using a discriminative model to improve the decomposition of face image features. We proposed to use HSCAM to decompose face feature images into age- and identity-related features to reduce the relevance of age information in identity features and vice versa. This improves the accuracy of cross-age face recognition by improving the age smoothing degree [15,29].

3. Proposed Method

To start this section, the proposed framework is summarized first. Then, the involved modules and optimization goals are elaborated part by part.

3.1. Overall Framework

This paper proposes a novel face recognition framework that incorporates a Hybrid Spatial-Channel Attention Module (HSCAM) to accurately extract identity features from facial images while disregarding age information [30]. The overall framework consists of three modules: the face feature extraction module, the separated age identity feature module, and the cross-age recognition module, as illustrated in Figure 2. Initially, hybrid face features are extracted from the input face images. Next, the feature decomposition module

separates the face image features into age-related and identity-related features [31]. Finally, the cross-age recognition module trains the identity-related features and age features through dedicated modeling for better performance in cross-age face recognition tasks.

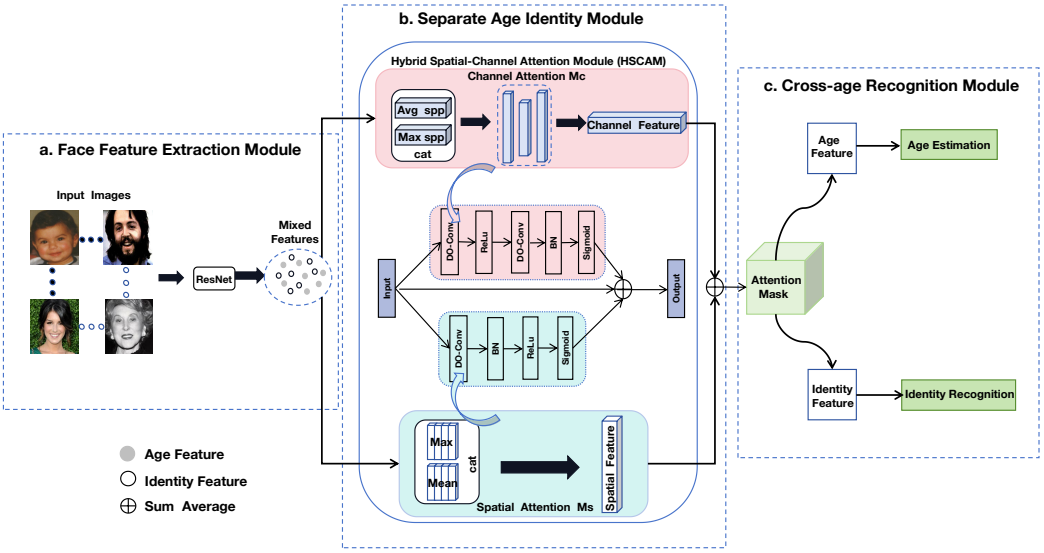

Figure 2. The overall structure of the model consists of three modules: (**a**) Face Feature Extraction Module, (**b**) Separation Age Identity Module, and (**c**) Cross-age Recognition Module. In the first stage, we extract the facial representations from the input face images through the improved ResNet network to obtain the face image features. In the second stage, we decompose the extracted face image features into age-related information and identity-related information using the HSCAM to reduce the degree of rubrication of age and identity information. Finally, we perform the age estimation task and identity recognition task with the age and identity-related information obtained from decomposition.

3.2. Face Feature Extraction

The face feature extraction module in this paper uses an improved ResNet network [32] as its backbone. This network addresses the gradient vanishing and gradient explosion problems in deep neural networks by using residual modules, allowing for the construction of very deep neural networks [33]. The ReLU activation function is replaced with the PReLU activation function, and the BN layer and PReLU activation layer are placed in front of the CONV layer [34]. This enhances the training efficiency, stability, and generalization performance of the network. Mathematically, the PReLU function can be represented as $f(x)$, where x is the input data and $f(x)$ is the output value after the convolutional layer operation.

$$f(x) = \begin{cases} x, & \text{if} \quad x > 0; \\ ax, & \text{if} \quad x \leq 0. \end{cases} \tag{1}$$

The PReLU function outputs x when the input x is greater than 0. Otherwise, it outputs ax, where a is a learnable parameter with a value less than 1.

To aid in explaining the modules, this paper defines some basic symbols beforehand. The initial input image is represented by $\mathbf{I}$, the mixed facial feature map resulting from feature extraction as $\mathbf{X}$, the feature map obtained through the hybrid channel attention mechanism in the separated age identity module as $\mathbf{M}_c$, and the feature map obtained through the hybrid spatial attention mechanism as $\mathbf{M}_s$. The HSCAM is referred to as $\mathbf{M}$. The face feature extraction module $\mathbf{H}$ extracts the initial hybrid face feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ from the input image $\mathbf{I}$, indicated as $\mathbf{X} = \mathbf{H}(\mathbf{I})$. This hybrid face feature map contains both age and identity-related information. The formula for the composition of the face feature map is as follows:

$$\mathbf{X} = \mathbf{X}_{id} + \mathbf{X}_{age}. \tag{2}$$

Among them, $\mathbf{X}_{age}$ and $\mathbf{X}_{id}$ represent age-related and identity-related information, respectively.

3.3. Age Identity Separation

While decomposing the extracted face feature images, we introduce the HSCAM to segregate age features and identity features from their hybrid source more effectively. This is for pointing out that the hybrid learning mechanism has two meanings: (1) we use both channel and spatial attention mechanisms to better fully decompose age and task information; (2) within the two attention mechanisms, we combine diverse pooling strategies (e.g., max pooling and average/mean pooling) to preserve facial image information as much as possible from different dimensions. This can result in better facial representation and achieve improved cross-age recognition performance.

3.3.1. Hybrid Channel Attention

The channel attention mechanism is a technique that dynamically adjusts the importance of the channels in a neural network. This allows the network to focus more on the channels that are most relevant for solving a particular task. This technique has been introduced in recent works like [35,36]. On the other hand, Global Average Pooling (GAP) calculates the average activation of each channel, producing a weight vector. This weight vector is then used to adjust the feature map of each channel to obtain the attention-adjusted feature representation.

Firstly, we have improved the traditional channel attention mechanism by optimizing the use of GAP. This method helps in capturing the weight of each channel and compressing the spatial dimension of the feature graph, thereby improving efficiency. GAP captures the average activity value of each channel in the feature map, which better represents the contribution of each channel to the overall features. In Convolutional Neural Networks (CNNs) [37], max pooling and average pooling [38] are typically used to reduce the spatial dimensions of the feature map, extract important features, and reduce computational complexity. The choice between them in channel attention mechanisms typically depends on the nature of the task and the data characteristics. Max pooling emphasizes the salient features in the channel, while average pooling is more beneficial for the smoothing process and overall feature capture.

To this end, we propose a hybrid channel attention mechanism that combines average pooling and max pooling along the channel dimensions to work simultaneously. The model diagram of the hybrid channel attention mechanism is shown in Figure 3. These two pooling operations are used in conjunction with a Multi-Layer Perceptron (MLP) [39] to model the complex relationships between channels and generate weights for tuning the attention. Ultimately, the channel attention mechanism is implemented by applying these weights to the feature graph so that the activation values of each channel are weighted with different weights.

Here, we denote the attention weight obtained through the traditional channel attention mechanism as $\mathbf{A}_i$, which will be used to weight the feature maps of the corresponding channels. The formula for traditional channel attention is as follows:

$$\mathbf{A}_i = \sigma(MLP(AvgPool(x_i))), \tag{3}$$

where x_i is the c-th channel of the input feature map, *AvgPool* denotes the global average pooling operation, and σ is the activation function (usually Sigmoid or Softmax). In the overall framework, we define $\mathbf{M}_c$ as the hybrid channel feature map $\mathbf{X}$ obtained through the hybrid channel attention mechanism. Therefore, the formula for the hybrid channel attention mechanism proposed in this paper is as follows:

$$\mathbf{M}_c(\mathbf{X}) = \sigma(MLP(AvgPool(\mathbf{X})) \oplus MLP(MaxPool(\mathbf{X}))). \quad (4)$$

The equation mentioned above uses the feature map denoted as $\mathbf{X}$, which is generated by the modified ResNet network backbone. In the equation, the operation *MaxPool* represents max pooling, whereas $AvgPool(\mathbf{X})$ and $MaxPool(\mathbf{X})$ represent average pooled features and maximum pooled features, respectively. The symbol $\oplus$ denotes the splicing operation in the channel attention mechanism. The resulting feature then goes through an MLP network to generate the final channel attention feature map, which is denoted as $\mathbf{M}_c(\mathbf{X})$.

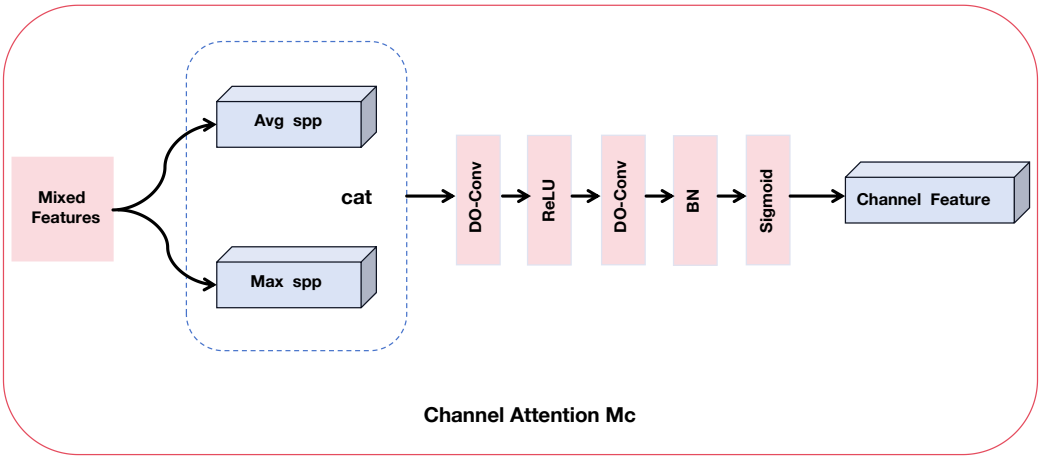

Figure 3. The hybrid channel attention mechanism combines average pooling and max pooling operations along the channel dimension to work simultaneously. These two pooling operations are combined with an MLP using the DO-Conv layer [40], which replaces the conventional convolutional layer. The output is then passed through the ReLU and batch normalization (BN) layer, followed by another DO-Conv layer. Finally, the channel features are obtained by mapping the output through the Sigmoid activation function.

3.3.2. Hybrid Spatial Attention

Spatial attention mechanisms are a commonly used technique in image processing that help to highlight the importance of different areas in an image. Figure 4 shows the model diagram of the hybrid spatial attention mechanism, which allows the model to focus on the areas of the image that contain essential information [41]. To process the input data more effectively, the spatial locations are weighted. The feature map consists of representations of features at different locations in the image, and we use global pooling for each channel to obtain global information for each channel. We use mean and max pooling operations to better capture the global information. By combining the results of these two pooling methods, the model can obtain weights for the spatial information, which helps it focus on the features in different areas of the input image.

In the spatial attention mechanism, the global mean pooling step calculates the average of the feature values within each channel of the input feature map, which outputs a single value. To illustrate this concept, we consider a simplified example of the mean pooling operation. Suppose we have an input feature map, denoted by $\mathbf{X}$, where $\mathbf{X}_{i,j,c}$ represents the eigenvalue at position (i, j) in channel c. To perform the mean pooling operation, we take the average of all the feature values within each channel of the input feature map:

$$MeanPool(\mathbf{X}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_{i,j,c}, \quad (5)$$

The equation demonstrates how to calculate the *MeanPool* value by taking the average eigenvalue of all positions in the c-th channel, where H and W represent the height and width of the feature map, respectively.

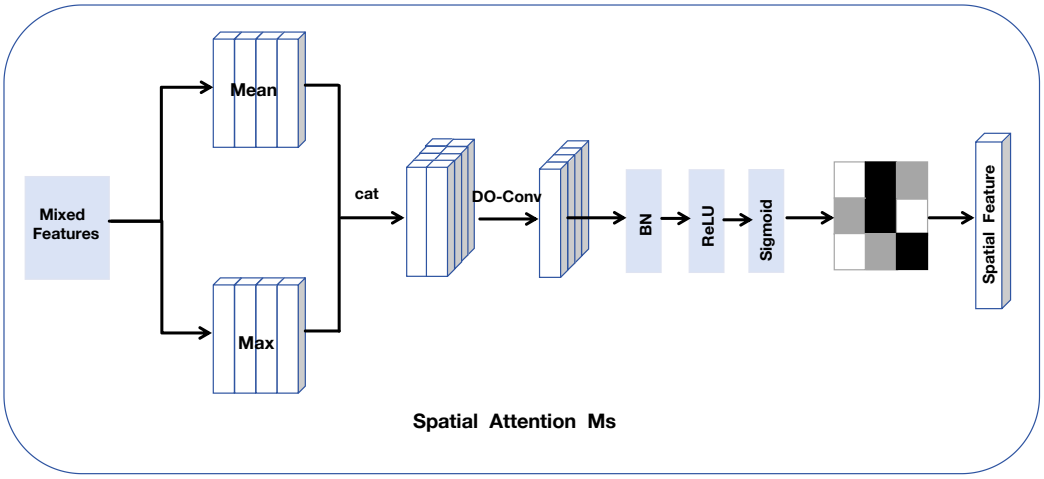

Figure 4. The hybrid spatial attention module uses max pooling and mean pooling operations to extract important features from each spatial position. It replaces the traditional convolutional layer with a DO-Conv layer, which then goes through the BN layer, as well as ReLU and Sigmoid activation functions to learn the weights for each position. The weights learned are then applied to every spatial position of the feature map, enhancing its spatial importance, and generating more informative feature representations.

To obtain the output for each channel, the maximum eigenvalue of all positions in that channel is calculated over the entire feature map. This maximum value is then used as the output for that specific channel. This operation is known as global max pooling and can be represented as follows:

$$MaxPool(\mathbf{X}_c) = Max_{i,j}(\mathbf{X}_{i,j,c}). \tag{6}$$

In the process of calculating spatial attention weights, a linear transformation and activation function is performed on the concatenated feature maps to generate the final spatial attention weights. Firstly, the feature map is combined with max pooling and average pooling, and then it passes through the DO-Conv layer, BN layer, ReLU layer, and sigmoid layer in sequence.The resulting hybrid spatial feature map obtained through the hybrid spatial attention mechanism is defined as $\mathbf{M}_s$. Additionally, the formula for the hybrid channel attention mechanism in this process can be expressed by the following equation:

$$\mathbf{C}(\mathbf{X}) = MeanPool(\mathbf{X}) \oplus MaxPool(\mathbf{X}), \tag{7}$$

$$\mathbf{M}_s(\mathbf{X}) = Sigmoid(ReLU(BN(DO - Conv(\mathbf{C}(\mathbf{X}))))). \tag{8}$$

The notation used in this formulation is as follows: $\oplus$ denotes the element-level splicing operation, while $\mathbf{C}(\mathbf{X})$ combines the maximum and average pooling results along the channel dimensions. *DO-Conv* refers to the convolution operation, which can be a learnable convolution kernel that transforms features in the spliced results. *BN* denotes the batch normalization operation which normalizes the output of the convolution layer. *ReLU* is the modified linear unit operation that introduces non-linearities. *Sigmoid* is the operation that maps the output to the interval $[0, 1]$ to obtain the final spatial attention weights.

The ReLU layer can help alleviate the problem of vanishing gradient, while the derivatives of the Sigmoid layer have better gradients both near zero and near one, which helps with backpropagation. In this paper, we have added the ReLU layer between the BN layer and the Sigmoid layer to enhance the non-linear capability of the model and make it

more adaptable to complex data distributions. It also enables the model to better learn the differentiation of features and improves the sensitivity of the model to different features. Such an operation can be used to adjust the importance of channels or spaces in the feature map to better focus on the information of interest in subsequent computations.

3.3.3. Hybrid Spatial-Channel Attention Module

The proposed HSCAM in the research combines channel attention and spatial attention mechanisms to the feature map. This means that the feature maps are essentially given weights using both channel attention and spatial attention weights. The process is described by the following equations:

$$\mathbf{X}' = \mathbf{M}_c(\mathbf{X}) \odot \mathbf{X}, \tag{9}$$

$$\mathbf{X}'' = \mathbf{M}_s(\mathbf{X}) \odot \mathbf{X}, \tag{10}$$

$$\overline{\mathbf{X}} = \mathbf{X}' + \mathbf{X}'' = (\mathbf{M}_c(\mathbf{X}) \oplus \mathbf{M}_s(\mathbf{X})) \odot \mathbf{X}. \tag{11}$$

To summarize, we involve the input feature map $\mathbf{X}$, the channel attention feature map $\mathbf{M}_c(\mathbf{X})$, and the spatial attention feature map $\mathbf{M}_s(\mathbf{X})$ to complete a hybrid spatial-channel attention mechanism termed $\mathcal{M}$. The channel attention mechanism operation is denoted as $\mathbf{X}'$, and the spatial attention mechanism operation is denoted as $\mathbf{X}''$. The weighted feature map by the HSCAM is represented as $\overline{\mathbf{X}}$.

To enhance the performance, computational efficiency, and generalization ability of the model, we substitute the conventional convolutional layer with a DO-Conv layer [40]. The DO-Conv layer comprises an additional depthwise convolution, which establishes an over-parameterized convolutional layer. First, we apply a depthwise convolution operation to the input features, followed by a regular convolution operation to the output intermediate results. The use of the DO-Conv layer optimizes the deep learning model's parameters, effectively integrates the attention mechanism, and leads to better performance and faster training.

3.4. Cross-Age Recognition

In the above-mentioned HSCAM, we separate the features related to age and identity from their hybrid features. This allows us to perform age estimation and identity recognition tasks using the separated information [42]. The formulas used for separating the age and identity-related information are presented below.

$$\overline{\mathbf{X}} = \underbrace{\overline{\mathbf{X}} \odot \mathbf{M}(\mathbf{X})}_{\mathbf{X}_{age}} + \underbrace{\left(1 - \overline{\mathbf{X}} \odot \mathbf{M}(\mathbf{X})\right)}_{\mathbf{X}_{id}}, \tag{12}$$

where age-related information is separated from the feature map using $\mathbf{M}$, namely HSCAM, leaving identity-related information intact. The symbol $\odot$ denotes element-by-element multiplication.

3.4.1. Age Estimation Task

The hybrid facial features are divided into two categories: age-related and identity-related information using the hybrid channel space attention module. Then, to estimate the age change, we use the L1 loss function [43]. This function calculates the absolute difference between the predicted age and the true age, adds up all the differences, and calculates the average. We use the L1 loss function because it is more resistant to outliers and improves the model's robustness for age estimation [42]. The loss function for age estimation can be defined as follows:

$$\mathbf{L}_{\text{age}} = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{X}_{age_i} - \mathbf{Y}_{age_i}|, \tag{13}$$

where the decomposed age feature is $\mathbf{X}_{age}$ and the true age label is $\mathbf{Y}_{age}$, while N denotes the number of samples. $\mathbf{X}_{age_i}$ denotes the age information obtained from the decomposition of the feature map for the i-th sample and $\mathbf{Y}_{age_i}$ is the corresponding true age label.

3.4.2. Identity Recognition Task

In the identity recognition task, each label for an identity is usually treated as a category. Therefore, the cross-entropy loss function is a suitable choice for this task, as noted in Zhang's work [44]. We have opted to use the cross-entropy loss function for the identity recognition task, as it aids the model in learning effective identity features and makes it more resilient in multi-category classification problems, as suggested by [45]. The loss function used for this task can be defined as:

$$\mathbf{L}_{\text{id}} = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_{id_i}\cdot\log(\mathbf{X}_{id_i}), \tag{14}$$

where the identity feature decomposed by the model is $\mathbf{X}_{id}$ and the real identity label is $\mathbf{Y}_{id}$, $\mathbf{Y}_{id_i}$ is the real identity label corresponding to the i-th sample, and $\mathbf{X}_{id_i}$ is the identity information obtained from the model decomposition [46]. From the above, we define the total loss function as $\mathbf{L}_{\text{total}}$, and the overall loss function can be derived as follows:

$$\mathbf{L}_{\text{total}} = \alpha\cdot\mathbf{L}_{\text{age}} + \beta\cdot\mathbf{L}_{\text{id}}, \tag{15}$$

where α and β are the weights of the loss function. The weights are dynamically adjusted according to the model performance during training.

4. Experiments and Results

We conducted extensive experiments on three of the most popular face image datasets, namely AgeDB-30 [47], CACD-VS [12], and FGNET [48]. The experiments are divided into two parts. Firstly, the methodology will be evaluated by comparing the experimental results with state-of-the-art methods in Section 4.2. Secondly, the effects of different modules as well as hyperparameters from the proposed framework will be evaluated in Section 4.3.

4.1. Experimental Details

4.1.1. Data Preparation

During the pretraining stage, we utilized two mega-datasets, namely MS1M-ArcFace [49] and CASIA-Webface [50], due to their vast and diverse facial image data and label information. We divided these datasets into eight different age groups with no overlap: 0–12, 13–18, 19–25, 26–35, 36–45, 46–55, 56–65 years, and 66+ years. This allowed us to perform the age estimation task based on the grouped age labels, following previous studies [23,25].

In the performance evaluation, we trained three benchmark datasets: AgeDB-30 [47], CACD-VS [12], and FGNET [48]. We used their test sets to assess the recognition results. Specifically, we divided these three datasets into training and testing sets in a 4:1 ratio during the implementation phase. Four copies were used for the training process, and one copy was used for the testing process. To ensure fair comparisons, we employed these settings for all baselines. We discuss more detailed information about the datasets in the corresponding experiment sections, where some examples are presented in Figure 5.

4.1.2. Training Settings

The ResNet network modified for improved performance, pre-trained on the MS-Celeb-1M [49] and CASIA WebFace [50] datasets, is chosen as the backbone network to extract identity and age recognition features. The model is trained iteratively 200 times, with the initial learning rate set to 0.1 for the classifier and 0.01 for the features part in the age estimation task. We use the SGD optimizer with a momentum value of 0.9. α and β are set to 0.01 and 0.01, respectively. For the identity recognition task, we use the Adam

optimizer with a learning rate set to 0.01 for the classifier and 0.0001 for the features part. We also employ a weight decay mechanism with an intensity of $1 \times e^{-4}$.

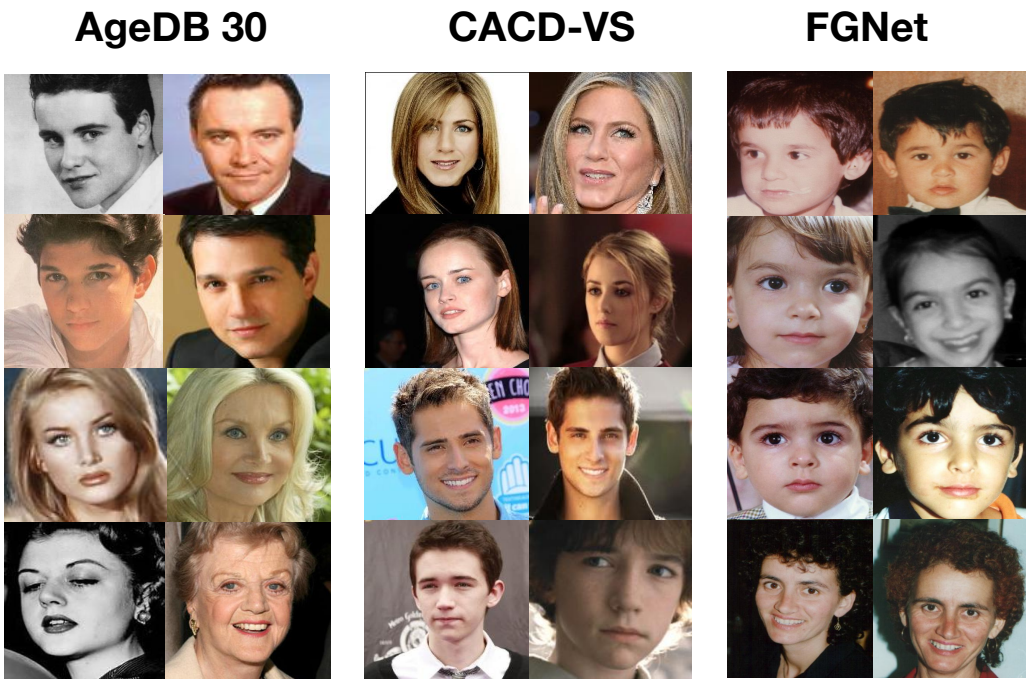

Figure 5. The figure displayed above showcases the facial images utilized in our cross-age facial recognition research. These images are taken from the AgeDB-30, FGNET, and CACD-VS datasets. We have incorporated these datasets to ensure that our study covers a wide range of age spans and face image qualities, thereby making it diverse and challenging. The sample images from each dataset are presented in a grid format, making it easier for readers to comprehend the context of our experiments.

4.2. Comparison Results

Our next step involves conducting a thorough evaluation of our proposed method on various widely used face recognition datasets, including the AgeDB-30 [47], CACD-VS [12], and FGNET [48] datasets. Our goal is to gain a comprehensive understanding of how effectively our method performs under different age spans and data distributions. In our experiments, we compared our approach with existing state-of-the-art methods and evaluated its accuracy and improvement. The improvement indicates the accuracy difference between our approach and existing methods.

4.2.1. Experiments on CACD-VS Dataset

Researchers in the field of cross-age face recognition often use public datasets to verify the accuracy of their methods. One such dataset is the cross-age celebrity dataset (CACD), which contains 163,446 images of 2000 celebrities. However, in this study, we utilized the CACD verification subset (CACD-VS) [12] to evaluate our method. This subset was manually sampled and annotated to minimize mislabelling and duplicate images, making it more reliable than the CACD dataset. Our results, as shown in Table 1, demonstrate that our method outperformed other state-of-the-art methods, achieving 99.77% accuracy. Our proposed method has achieved a 1.27% improvement in accuracy compared to LF-CNN [14] (i.e., 98.50% $\rightarrow$ 99.77%), and a 0.2% improvement compared to WMI-AI [25] (i.e., 99.57% $\rightarrow$ 99.77%). The results of our experiment indicate that our hybrid attention mechanism is effective in disentangling age and identity features in facial images. In fact, it outperformed the WMI-AI method that uses the Wasserstein distance discriminator for multitask learning. This suggests that our approach is successful in minimizing the mutual information between age and identity features, thus improving the performance

of cross-age face recognition tasks [51]. Our hybrid attention mechanism considers both spatial and channel information, which enables it to fully utilize local and global features in the images. By doing so, the mechanism comprehensively understands the features in the images, allowing it to more effectively distinguish age and identity features.

Table 1. Face verification results evaluated using the CACD-VS dataset, with the highest accuracy result highlighted in bold.

Method	Accuracy (%)	Improvement (%)
LF-CNN [14]	98.50	1.27
OE-CNN [15]	99.20	0.57
DAL [17]	99.40	0.37
WMI-AI [25]	99.57	0.20
MTLFace [23]	99.55	0.22
PMADN [22]	99.20	0.57
Ours	**99.77**	-

4.2.2. Experiments on AgeDB-30 Dataset

The AgeDB dataset is a publicly available dataset that consists of 16488 face images belonging to 568 different individuals, each of which is manually labeled with age information. This dataset divides all data into four age groups (age differences of 5 years, 10 years, 20 years, and 30 years). Each age group consists of 300 positive and 300 negative samples. A new dataset named AgeDB-30 has been created from the AgeDB dataset [47]. The AgeDB-30 dataset is a validation set using facial images of people with an age span of over 30 years. As a result, this dataset covers a wider age range and is considered more challenging than the original dataset. This enhances the ability to assess the performance of models across different age groups. The increased number of subsets in AgeDB-30 also makes it more difficult to generalize to different ages, thus requiring the model to be more robust. The performance of the proposed model has been evaluated using the AgeDB-30 dataset. Table 2 displays the superior performance of our model. Our proposed method improves accuracy by 1.95% over IEFP [18] (i.e., 95.82% $\rightarrow$ 97.77%) and 4.07% over LCAM [19] (i.e., 93.70% $\rightarrow$ 97.77%), demonstrating its effectiveness. Based on experimental results, it can be observed that while the Low-Complexity Attention Module (LCAM) approach has a slight advantage over past modules that employed channel and spatial attention, our proposed hybrid channel-spatial mechanism offers higher flexibility. Our model utilizes facial features in the image and recognition target, employing a hybrid attention mechanism that combines spatial and channel attention to dynamically allocate attention. This allows the model to better adapt to diverse image features, leading to improved recognition performance and increased accuracy in cross-age face recognition.

Table 2. Face verification results evaluated using the AgeDB-30 dataset, with the best accuracy result highlighted in bold.

Method	Accuracy (%)	Improvement (%)
MT-MIM [24]	96.10	1.67
MTLFace [23]	96.23	1.54
IEFP [18]	95.82	1.95
PocketNet [52]	96.78	0.99
LCAM [19]	93.70	4.07
Ours	**97.77**	-

4.2.3. Experiments on FGNET Dataset

The FGNET dataset, as described in [53], is a relatively small dataset that includes 82 different individuals of various ages, from infants to seniors. This dataset features multiple age images per individual and contains approximately one thousand images in

total. It provides face images of different races, genders, and backgrounds, covering a diverse range of appearance features, along with age annotations for each face image. As a result, it has been widely used in previous face recognition research. Our model will be experimented on the FGNET dataset, and we will also use the Megaface challenge 1 (MF1) protocol [54]. Table 3 displays the results. Our approach outperforms other models, improving by 5.22% compared to CACon [26] (i.e., 64.37% $\rightarrow$ 69.59%) and by 11.67% compared to DAL [17] (i.e., 57.92% $\rightarrow$ 69.59%). While the CACon method introduces a contrast learning method along with the identity preservation capability based on the face synthesis model, our method is improved based on the discriminative model and does not introduce a contrast learning method. These texture-based methods may have stronger expressive power and robustness; however, the experimental results indicate that our method still holds the dominant position, which also demonstrates the superiority of the HSCAM in this paper. Additionally, Figure 6 is presented to showcase the success and failure of the matching cases for the three datasets.

Table 3. State-of-the-art approaches are evaluated on the FGNET dataset in comparison to our method. The best accuracy result is highlighted in bold.

Method	Accuracy (%)	Improvement (%)
OE-CNN [15]	58.18	11.41
DAL [17]	57.92	11.67
MTLFace [23]	57.18	12.41
LIAAD [20]	60.11	9.48
CACon [26]	64.37	5.22
Ours	**69.59**	-

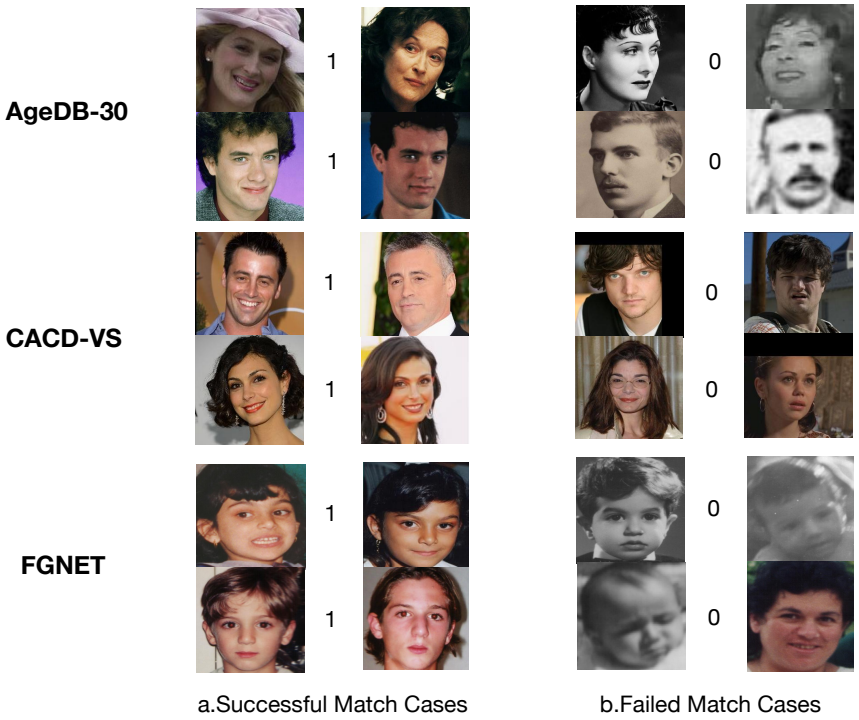

Figure 6. Some examples of matching and non-matching pairs in randomly selected images in three datasets, AgeDB, CACD-VS, and FGNET, using the method proposed in this paper. In this figure, (**a**) Successful Match Cases and (**b**) Failed Match Cases represent examples of successful and failed matches, where 1 indicates successful face detection and 0 indicates failure. Two sets of cases were extracted from the AgeDB-30, CACD-VS, and FGNET datasets in sequence for display.

4.3. Ablation Study

The ablation experiments were designed to further analyze the impact of the HSCAM on the cross-age face recognition task as well as the hyperparameter settings.

4.3.1. Analysis on Attention Module Settings

In the first experiment, the hybrid features were decomposed into age- and identity-related information, and then directly used for identity recognition. The other experiments were designed to validate the validity of the modules in the hybrid channel attention mechanism and to verify the effectiveness of the max-pooling and mean-pooling operations in the hybrid channel attention mechanism. Particularly, different settings were employed:

1. Baseline: The hybrid features were decomposed into age- and identity-related information, and then directly used for identity recognition without any attention mechanism components;
2. SA + Channel avg: The hybrid spatial attention (i.e., SA) mechanism was kept unchanged, and the hybrid channel attention mechanism was tested by eliminating the splicing of max pooling and average pooling in the channel dimension and only performing the max pooling operation;
3. SA + Channel max: Similar to the second setting, but only the global average pooling operation was performed;
4. CA + Spatial mean: The hybrid channel attention (i.e., CA) mechanism was kept unchanged, and the effectiveness of the hybrid max-pooling and mean-pooling modules was tested by canceling the splicing of max-pooling and mean-pooling in the channel dimension and only performing max-pooling operations;
5. CA + Spatial max: Similar to the fourth setting, but only max-pooling operations were carried out to verify the difference in the effect of splicing the max-pooling and the mean-pooling with the effect obtained from separate pooling operations;
6. Baseline + DO-Conv: A normal conv layer was replaced with a DO-Conv layer without using any attention mechanism model;
7. Baseline + CA'+ SA': Among them, CA': CA ($max \wedge avg$); SA': SA ($mean \wedge max$). $\wedge$ denotes the splicing role. The traditional conv layer was not replaced with a DO-Conv layer, and the attention of the hybrid channel space was kept unchanged for the experiment;
8. Ours: The traditional conv layer was replaced with a DO-Conv layer along the HSCAM that was used to evaluate the three face validation sets.

The results of the ablation experimental tests, as shown in Table 4, indicate that the performance improvement on these three datasets was achieved by our methods. It also verifies that HSCAM performs better compared to the normal channel attention mechanism and spatial attention mechanism. Additionally, the introduction of DO-Conv to change the traditional convolution also improves the performance of the experiment.

Table 4. Comparison of the performance of ablation experiment results in the form of accuracy (%) for different age prediction model components on the CACD-VS, AgeDB-30, and FGNET datasets. $\wedge$ denotes the splicing role, i.e., the splicing of both maximum and average pooling along the channel dimension. The bold values indicate the best results.

Network Settings	CACD-VS [12]	AgeDB-30 [47]	FGNET [48]
Baseline	88.38	92.56	50.20
SA + Channel avg	93.25	91.82	63.94
SA + Channel max	90.89	93.65	55.98
CA + Spatial mean	92.79	95.02	62.22
CA + Spatial max	94.68	94.38	60.52
Baseline + DO-Conv	95.02	95.19	62.35
Baseline + CA' + SA'	97.66	96.94	65.26
Ours	**99.77**	**97.77**	**69.59**

4.3.2. Analysis on Hyperparameter Settings

As mentioned in the previous section, we introduce hyperparameters α and β to balance the different loss terms in the whole loss function. In order to investigate the effects of α and β, we conduct a series of experiments on the CACD-VS dataset to judge their effectiveness by face verification accuracy. In this experiment, we set the hyperparameter values of α and β interchangeably to $[0.001, 0.01, 0.1, 1]$ while keeping one of them fixed. By thoroughly optimizing the model performance with diverse parameter settings, the best performance can be achieved when both α and β are set at 0.01. Figure 7 briefly presents the accuracy variations under different α and β values like 0.01 and 1, which further confirms the parameter settings are consolidated.

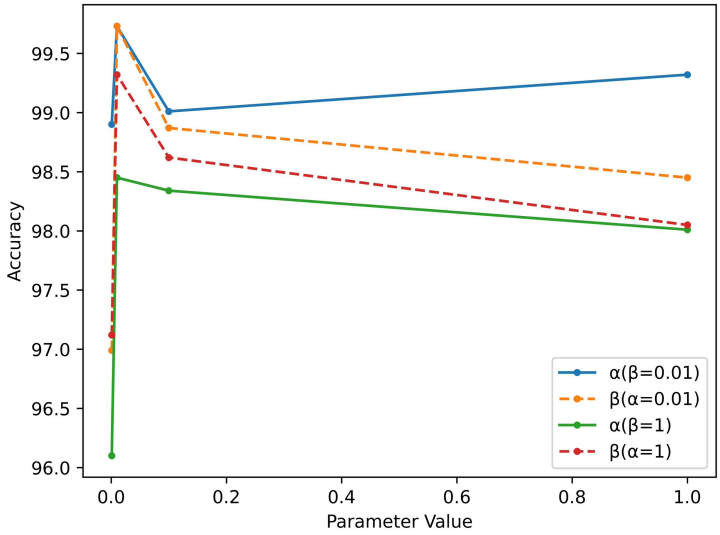

Figure 7. The face verification accuracy is evaluated on the CACD-VS dataset at different α and β values. The results show that the accuracy reaches the highest level when both α and β are 0.01.

5. Conclusions

In this work, we have presented a new deep framework for age-invariant face recognition. Our proposed HSCAM effectively distinguishes between age-related and identity-related features in the feature decomposition process. By incorporating spatial and channel attention simultaneously, we can fully capture the information in the face images, which leads to an improvement in the accuracy of cross-age face recognition. The proposed module also dynamically learns the weights, which helps to mitigate the influence of these two features and ultimately achieves better face representation. We have conducted extensive experiments on three public face datasets, including AgeDB-30, CACD-VS, and FGNET, which validate our approach. The proposed method has demonstrated significant improvements in recognition accuracy compared to previous methods. Notably, our model successfully addresses the challenge of face recognition across age groups of 30 and above in the AgeDB-30 dataset, indicating its excellent performance in dealing with wide age differences. While there is still room for improvement, we plan to incorporate more effective learning techniques into our model in the future.

Author Contributions: Conceptualization, W.A. and G.W.; methodology, W.A. and G.W.; software, W.A.; validation, W.A. and G.W.; formal analysis, W.A. and G.W.; investigation, W.A.; data curation, W.A. and G.W.; writing—original draft preparation, W.A.; writing—review and editing, W.A. and G.W.; visualization, W.A. and G.W.; supervision, G.W.; funding acquisition, G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Development Fund, Macao SAR. Grant number: 0004/2023/ITP1.

Data Availability Statement: The datasets used in this study can be downloaded from the urls in their official websites. The code is accessible from the corresponding author upon reasonable request.

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

References

1. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
2. Li, S.Z.; Lu, J. Face recognition using the nearest feature line method. *IEEE Trans. Neural Netw.* **1999**, *10*, 439–443. [CrossRef] [PubMed]
3. Wu, H.; Wu, G.; Hu, J.; Xu, S.; Zhang, S.; Liu, Y. Cityuplaces: A new dataset for efficient vision-based recognition. *J. Real Time Image Process.* **2023**, *20*, 109. [CrossRef]
4. Shakeel, M.S.; Lam, K.M. Deep-feature encoding-based discriminative model for age-invariant face recognition. *Pattern Recognit.* **2019**, *93*, 442–457. [CrossRef]
5. Kong, Y.; Zhang, K.; Zhang, L.; Wu, G. Deep facial attribute analysis. *Front. Neurosci.* **2023**, *17*, 1280831. [CrossRef]
6. Sun, Y.; Chen, Y.; Wang, X.; Tang, X. Deep learning face representation by joint identification-verification. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. [CrossRef]
7. Dhamija, A.; Dubey, R. An approach to enhance performance of age invariant face recognition. *J. Intell. Fuzzy Syst.* **2022**, *43*, 2347–2362. [CrossRef]
8. Zhao, J.; Yan, S.; Feng, J. Towards age-invariant face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 474–487. [CrossRef] [PubMed]
9. Zhao, J.; Cheng, Y.; Cheng, Y.; Yang, Y.; Zhao, F.; Li, J.; Liu, H.; Yan, S.; Feng, J. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9251–9258.
10. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9. [CrossRef]
11. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
12. Chen, B.C.; Chen, C.S.; Hsu, W.H. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimed.* **2015**, *17*, 804–815. [CrossRef]
13. Gong, D.; Li, Z.; Lin, D.; Liu, J.; Tang, X. Hidden factor analysis for age invariant face recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2872–2879.
14. Wen, Y.; Li, Z.; Qiao, Y. Latent factor guided convolutional neural networks for age-invariant face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4893–4901.
15. Wang, Y.; Gong, D.; Zhou, Z.; Ji, X.; Wang, H.; Li, Z.; Liu, W.; Zhang, T. Orthogonal deep features decomposition for age-invariant face recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2018; pp. 738–753.
16. Gibertoni, G.; Borghi, G.; Rovati, L. Vision-Based Eye Image Classification for Ophthalmic Measurement Systems. *Sensors* **2022**, *23*, 386. [CrossRef] [PubMed]
17. Wang, H.; Gong, D.; Li, Z.; Liu, W. Decorrelated adversarial learning for age-invariant face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3527–3536.
18. Xie, J.C.; Pun, C.M.; Lam, K.M. Implicit and explicit feature purification for age-invariant facial representation learning. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 399–412. [CrossRef]
19. Hoo, S.C.; Ibrahim, H.; Suandi, S.A.; Ng, T.F. LCAM: Low-Complexity Attention Module for Lightweight Face Recognition Networks. *Mathematics* **2023**, *11*, 1694. [CrossRef]
20. Truong, T.D.; Duong, C.N.; Quach, K.G.; Le, N.; Bui, T.D.; Luu, K. LIAAD: Lightweight attentive angular distillation for large-scale age-invariant face recognition. *Neurocomputing* **2023**, *543*, 126198. [CrossRef]
21. Wang, Z.; He, K.; Fu, Y.; Feng, R.; Jiang, Y.G.; Xue, X. Multi-task deep neural network for joint face recognition and facial attribute prediction. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, Bucharest, Romania, 6–9 June 2017; pp. 365–374.
22. Wu, Y.; Du, L.; Hu, H. Parallel multi-path age distinguish network for cross-age face recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 3482–3492. [CrossRef]
23. Huang, Z.; Zhang, J.; Shan, H. When age-invariant face recognition meets face age synthesis: A multi-task learning framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7282–7291.
24. Hou, X.; Li, Y.; Wang, S. Disentangled representation for age-invariant face recognition: A mutual information minimization perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 3692–3701.
25. Dahan, E.; Keller, Y. Age-Invariant Face Embedding using the Wasserstein Distance. *arXiv* **2023**, arXiv:2305.02745.

26. Wang, H.; Sanchez, V.; Li, C.T. Cross-Age Contrastive Learning for Age-Invariant Face Recognition. *arXiv* **2023**, arXiv:2312.11195.
27. Ermao, L.; Min, Z. Review of Cross-Age Face Recognition in Discriminative Models. In Proceedings of the 2023 8th International Conference on Image, Vision and Computing (ICIVC), Dalian, China, 27–29 July 2023; pp. 124–130. [CrossRef]
28. Deb, D.; Zhang, J.; Jain, A.K. Advfaces: Adversarial face synthesis. In Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September–1 October 2020; IEEE: New York, NY, USA, 2020; pp. 1–10.
29. Yan, C.; Meng, L.; Li, L.; Zhang, J.; Wang, Z.; Yin, J.; Zhang, J.; Sun, Y.; Zheng, B. Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2022**, *18*, 1–18. [CrossRef]
30. Ren, X.; Wang, J.; Li, S. MAM: Multiple Attention Mechanism Neural Networks for Cross-Age Face Recognition. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 8546029. [CrossRef]
31. Du, L.; Hu, H. Cross-age identity difference analysis model based on image pairs for age invariant face verification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2675–2685. [CrossRef]
32. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (JMLR Workshop and Conference Proceedings), Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
33. Babbar, S.; Dewan, N.; Shangle, K.; Kulshrestha, S.; Patel, S. Cross-age face recognition using deep residual networks. In Proceedings of the 2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India, 15–17 November 2019; IEEE: New York, NY, USA, 2019; pp. 257–262.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
35. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. 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. Fang, S.; Wu, G.; Liu, Y.; Feng, X.; Kong, Y. Dual enhanced semantic hashing for fast image retrieval. *Multimed. Tools Appl.* **2024**, 1–20. [CrossRef]
37. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]
38. Bieder, F.; Sandkühler, R.; Cattin, P.C. Comparison of methods generalizing max-and average-pooling. *arXiv* **2021**, arXiv:2103.01746.
39. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
40. Cao, J.; Li, Y.; Sun, M.; Chen, Y.; Lischinski, D.; Cohen-Or, D.; Chen, B.; Tu, C. Do-conv: Depthwise over-parameterized convolutional layer. *IEEE Trans. Image Process.* **2022**, *31*, 3726–3736. [CrossRef] [PubMed]
41. 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, 18–23 June 2018; pp. 7132–7141.
42. Zheng, T.; Deng, W.; Hu, J. Age estimation guided convolutional neural network for age-invariant face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1–9.
43. Barron, J.T. A general and adaptive robust loss function. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4331–4339.
44. Zhang, Z.; Song, Y.; Qi, H. Age progression/regression by conditional adversarial autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5810–5818.
45. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 448–456.
46. Wang, H.; Sanchez, V.; Li, C.T. Age-oriented face synthesis with conditional discriminator pool and adversarial triplet loss. *IEEE Trans. Image Process.* **2021**, *30*, 5413–5425. [CrossRef] [PubMed]
47. Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; Zafeiriou, S. Agedb: The first manually collected, in-the-wild age database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 51–59.
48. Nithyashri, J.; Kulanthaivel, G. Classification of human age based on Neural Network using FG-NET Aging database and Wavelets. In Proceedings of the 2012 Fourth International Conference on Advanced Computing (ICoAC), Chennai, India, 13–15 December 2012; IEEE: New York, NY, USA, 2012; pp. 1–5.
49. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 87–102.
50. Dong, Y.; Zhen, L.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv* **2014**, arXiv:1411.7923.
51. Chen, X.; Lau, H.Y. The identity-level angular triplet loss for cross-age face recognition. *Appl. Intell.* **2022**, *52*, 6330–6339. [CrossRef]
52. Boutros, F.; Siebke, P.; Klemt, M.; Damer, N.; Kirchbuchner, F.; Kuijper, A. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access* **2022**, *10*, 46823–46833. [CrossRef]

53. Fu, Y.; Hospedales, T.M.; Xiang, T.; Xiong, J.; Gong, S.; Wang, Y.; Yao, Y. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 563–577. [CrossRef] [PubMed]
54. Kemelmacher-Shlizerman, I.; Seitz, S.M.; Miller, D.; Brossard, E. The megaface benchmark: 1 million faces for recognition at scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4873–4882.