electronics

MDPI

Article

Attention-Mechanism-Based Models for Unconstrained Face Recognition with Mask Occlusion

Mengya Zhang [1,2,3], Yuan Zhang [1,2,3] and Qinghui Zhang [1,2,3,*]

[1] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China; monicazhang@haut.edu.cn (M.Z.); 2020930677@stu.haut.edu.cn (Y.Z.)
[2] Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education, Zhengzhou 450001, China
[3] Henan Grain Big Data Analysis and Application Engineering Research Center, Henan University of Technology, Zhengzhou 450001, China
* Correspondence: zqh131@haut.edu.cn

Abstract: Masks cover most areas of the face, resulting in a serious loss of facial identity information; thus, how to alleviate or eliminate the negative impact of occlusion is a significant problem in the field of unconstrained face recognition. Inspired by the successful application of attention mechanisms and capsule networks in computer vision, we propose ECA-Inception-Resnet-Caps, which is a novel framework based on Inception-Resnet-v1 for learning discriminative face features in unconstrained mask-wearing conditions. Firstly, Squeeze-and-Excitation (SE) modules and Efficient Channel Attention (ECA) modules are applied to Inception-Resnet-v1 to increase the attention on unoccluded face areas, which is used to eliminate the negative impact of occlusion during feature extraction. Secondly, the effects of the two attention mechanisms on the different modules in Inception-Resnet-v1 are compared and analyzed, which is the foundation for further constructing the ECA-Inception-Resnet-Caps framework. Finally, ECA-Inception-Resnet-Caps is obtained by improving Inception-Resnet-v1 with capsule modules, which is explored to increase the interpretability and generalization of the model after reducing the negative impact of occlusion. The experimental results demonstrate that both attention mechanisms and the capsule network can effectively enhance the performance of Inception-Resnet-v1 for face recognition in occlusion tasks, with the ECA-Inception-Resnet-Caps model being the most effective, achieving an accuracy of 94.32%, which is 1.42% better than the baseline model.

Keywords: face recognition; mask occlusion; SE attention mechanism; ECA attention mechanism; capsule network

Citation: Zhang, M.; Zhang, Y.; Zhang, Q. Attention-Mechanism-Based Models for Unconstrained Face Recognition with Mask Occlusion. *Electronics* **2023**, *12*, 3916. https://doi.org/10.3390/electronics12183916

Academic Editor: Silvia Liberata Ullo

Received: 24 July 2023
Revised: 13 September 2023
Accepted: 14 September 2023
Published: 17 September 2023

1. Introduction

Face recognition is an important research field in biological information security. The commonly used steps of face recognition are face detection, face alignment, feature extraction and classification. Among these, good feature extraction is essential for obtaining an effective face recognition model [1]. In the real environment, face images are obscured by occlusion factors in the recognition process, and the loss of key facial features caused by occlusion is a significant problem to solve in the field of face recognition [2]. The sudden outbreak and spread of COVID-19 have posed great challenges to face recognition systems in China and other countries. To mitigate the spread of COVID-19, wearing masks in public has become common. However, mask occlusion has a significant negative impact when feature extraction is performed on specific facial regions (e.g., the mouth, nose, and chin), prompting face recognition applications to consider the problem of mask occlusion [3].

Face recognition with mask occlusion can be applied to community security systems to identify individuals in public when there is a need to wear masks to prevent the spread of contagious diseases. It can also serve tasks such as tracking suspects who wear masks

intentionally in security scenarios. However, the success of face recognition with mask occlusion in real applications is attributed to the continuous optimization of the application terminal and human active cooperation. It cannot be adapted to unconstrained scenarios wherein factors such as lighting, posture, expression, age, clarity, and resolution cannot be controlled. Mask occlusion in unconstrained environments is still a big challenge for face recognition. How to alleviate or eliminate the negative impact of occlusion is one of the important problems in the field of unconstrained face recognition [4]. From the theoretical perspective, studying the problem of unconstrained face recognition with mask occlusion not only helps solve the general occlusion problem encountered in face recognition but also gives certain inspiration for the fields of image synthesis and image restoration.

At present, there are two mainstream methods for occluded face recognition in unconstrained scenes. One is robust feature extraction, which focuses on obtaining robust features of non-occluded face regions using the corresponding facial landmarks of a given face image [5]. Robust feature extraction mainly involves dimensionality reduction decomposition of high-dimensional features such as expression and pose in images [6]. Hariri et al. [7] aimed at locating and removing the occlusion area of the face, extracted deep features of the eyes and forehead of an occluded face, and further obtained a lightweight representation and more generalized occluded face recognition model with a Bag-of-Features paradigm. Mandal et al. [8] aimed to obtain the identity of individuals with a mask from images of the same individuals without a mask, achieved by finetuning ResNet-50 and simulating masks on an unmasked dataset. Song et al. [9] aimed to learn the correspondence between occluded facial areas and corrupted feature elements with the Pairwise Differential Siamese Network (PDSN), and finally eliminate corrupted feature elements from recognition by establishing a mask dictionary with the Feature Discarding Mask (FDM). Li et al. [10] proposed an attention-based approach that combines cropping-based and attention modules to focus on the regions around the eyes. The cropping helps the model gain more attention when extracting robust features, and the attention is embedded in the convolution module of ResNet to refine the face features. Deng et al. [11] proposed a method called MFCosface; they designed an Att-inception module to pay more attention to the area that is not covered by the mask and enlarged the unoccluded area's contribution to the identification process by minimizing a large margin cosine loss.

The other method is occluded region recovery, which recovers a whole face from an occluded face such that the new face has no occlusion by using reconstruction-based techniques or inpainting techniques [12]. This method can obtain the semantic information of face images by means of face image restoration, face information reconstruction of occluded parts and face feature transfer; it can then reduce the distance between facial features with occlusion and facial features without occlusion corresponding to the same subject, and can finally alleviate the lack of face structure information and incomplete face information caused by face self-occlusion and external occlusion. Li et al. [13] proposed an end-to-end de-occlusion distillation framework to eliminate image blurring and recover facial features, which uses a Generative Adversarial Network (GAN) to construct de-occlusion modules for image completion. Din et al. [14] proposed a GAN-based network using two discriminators to learn the global facial structure and focus the learning on the missing regions.

Above all, the robust feature extraction methods all have limited attention performance and unsatisfactory accuracy on mask-wearing occlusion tasks. On the other hand, most of the occluded region recovery methods require additional information of occluded regions, which becomes challenging when dealing with large occlusion in large-scale masked face recognition benchmarks [15,16]. Mask-wearing occlusion is still a big challenge in face recognition, especially in real-world uncooperative situations. How to alleviate or eliminate the negative impact of occlusion is a significant issue to address in the field of unconstrained face recognition.

In this paper, we use Inception-ResNet-v1 [17], which is a Convolutional Neural Network (CNN) proposed in recent years and has good convergence and performance, as the benchmark network to learn deep face features in unconstrained mask-wearing conditions. To eliminate the negative impact of occlusion during feature extraction of Inception-ResNet-v1, the effects of adding a Squeeze-and-Excitation (SE) module [18] and Efficient Channel Attention (ECA) module [19] are compared and analyzed. To increase the interpretability and generalization of Inception-ResNet-v1, the effect of adding capsule modules [20] is analyzed. Finally, a novel mask-occluded face recognition framework, ECA-Inception-Resnet-Caps, is proposed to increase the attention of unoccluded face areas for obtaining more discriminative face features. The proposed approach enhances Inception-ResNet-v1 with attention mechanisms and capsule modules to perform face recognition, addressing the limitation of the insufficient utilization of unoccluded information. Unlike other methods, the framework uses both masked and unmasked datasets for model training, which ultimately enables both masked and unmasked face recognition. Compared to the baseline evaluated using the Labeled Faces in the Wild (LFW) database [21], the performance of adding embedded attention mechanisms and a capsule module is improved by 1.25% and 1.42%, respectively.

The main contributions of this paper are as follows:

1. We propose incorporating the SE attention module into the Inception-Resnet-v1 network structure to obtain more discriminative deep face features. This is achieved by changing the network structure, increasing the weights of non-occluded regions and decreasing the weights of occluded regions.
2. In order to increase the attention of non-occluded face areas, we propose the embedding of the ECA module into Inception-Resnet-v1. The ECA module helps capture key information from the input and improves the generalization capability of the model.
3. By taking advantage of both the CNN and capsule network, we propose the ECA-Inception-Resnet-Caps framework to further enhance the performance and generalization of mask-occluded face recognition. Experimental results on different loss settings show the effectiveness of the proposed ECA-Inception-Resnet-Caps.

2. Related Work

2.1. Inception-Resnet-v1 Network

In recent years, deep convolutional neural networks have made great progress in image recognition and analysis performance and have become a hot research topic. Currently, the commonly used convolutional neural networks include Inception, AlexNet [22], ResNet, VGG, GoogLeNet [23], etc. After that, the Inception network has been optimized based on GoogLeNet, and several improved versions have been introduced to enhance the classification efficiency. The latest versions primarily include Inception-ResNet-v1 and Inception-ResNet-v2 [17], with Inception-ResNet-v2 exhibiting superior recognition performance but requiring more parameters. Considering the comparable effectiveness of Inception-ResNet-v1 and its better convergence and performance, this paper adopts Inception-ResNet-v1 as the benchmark network. This choice helps avoid issues such as gradient explosion and vanishing gradient while enhancing the network's recognition performance.

Inception-ResNet-v1 is a deep convolutional neural network comprising three main modules: Stem, Inception-ResNet, and Reduction modules. Among them, the Stem module provides input to the initial convolutional network, the Inception-Resnet module extracts image features without changing the grid size, and the Reduction module can not only extract and abstract features but can also compress the grid size. Figure 1 shows Inception-Resnet-v1 as the baseline framework.

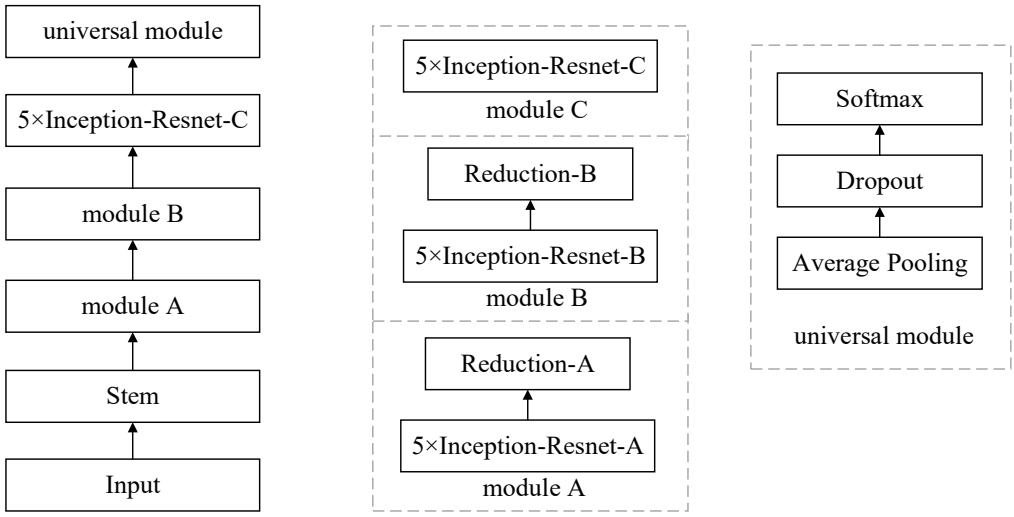

Figure 1. Inception-Resnet-v1 framework.

2.2. Attention Mechanism

In recent years, attention mechanisms have been widely applied in various fields such as image segmentation, natural language processing, and speech recognition. Many researchers have started integrating attention modules into deep convolutional neural networks, especially channel attention modules, which have gained a lot of attention from researchers. The attention mechanism for improving the efficiency and accuracy of perceptual information processing is mainly characterized by learning suitable weight feature maps. Some fields use attention mechanisms combined with the Inception-Resnet network. However, there is limited research on applying this approach to face recognition in occluded scenes. Additionally, attention mechanisms can improve the network efficiency without significantly increasing the computational complexity. In this paper, two different attention mechanisms are utilized, namely Squeeze-and-Excitation (SE) and Efficient Channel Attention (ECA), combined with Inception-ResNet, to increase the attention on non-occluded face areas and to alleviate the negative impact of occlusion during feature extraction in unrestricted scenes. Figures 2 and 3 show the structure of the SE module and ECA module, respectively.

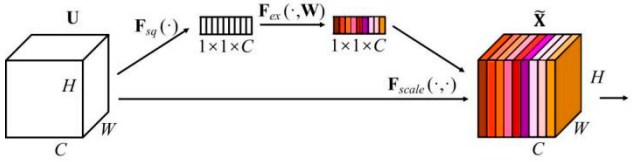

Figure 2. The structure of SENet.

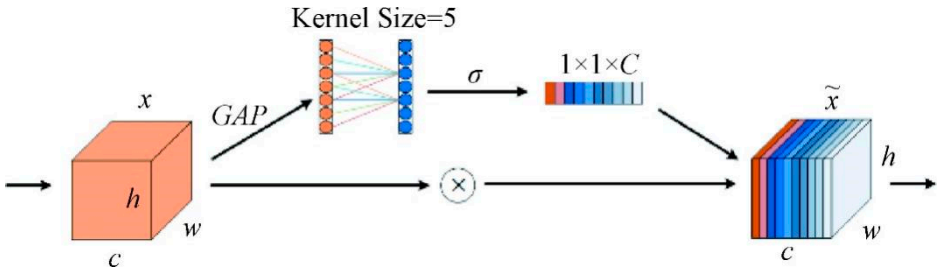

Figure 3. ECANet structure diagram.

2.2.1. SE Attention Mechanism

The Squeeze-and-Excitation Network (SENet) [18], which focuses on the relationship between feature channels, is an attention mechanism that adds attention to the channel dimension and mainly consists of squeeze and excitation. SENet learns different degrees of channels in a feature graph in a neural network and gives weight values that boost the channels of feature maps that are effective for the current task and suppress the channels of features that are less effective so that the network focuses more on the important channels. SENet is a simple channel attention mechanism with the advantages of a carefully designed structure, easy deployment, and low cost. The basic structure of SENet is shown in Figure 2.

Each channel of the feature map has the same degree of importance. However, after applying the SE operation (right color map C), different colors represent different weights. This means that each feature channel has different degrees of importance, enabling the network model to focus more on the channels with higher weight values. In Figure 2, U represents C feature maps of size H × W, where C, H, and W represent the number of channels, height, and width of the image, respectively. F_{sq} denotes the Squeeze operation, which is equivalent to a global average pooling operation.

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

Among them, z_c is the output after the F_{sq} operation, c denotes the channel, and u_c denotes the cth two-dimensional matrix in U. In Figure 2, F_{ex} is the Excitation operation, which is equivalent to two fully connected operations:

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

where s is the output after the F_{ex} operation; δ and σ denote ReLU and Sigmoid activation functions, respectively. After obtaining s, the F_{scale} operation is performed, which involves channel-wise multiplication between the feature map u_c and the scalar s_c. The output feature results are represented as:

$$X_c = F_{scale}(u_c, s_c) = s_c u_c \tag{3}$$

2.2.2. ECA Attention Mechanism

Although the fully connected dimensionality reduction operation in SENet reduces the complexity of the network model, it is still not good enough to describe the correspondence between weights and channels, which will affect the correlation prediction. At the same time, the correlation analysis of all channels will also reduce the computational efficiency of the network. To achieve a deeper analysis, this paper also incorporates the ECA mechanism and evaluates the influence of both attention mechanisms on the Inception-Resnet-v1, independently, for further comparison.

In 2020, Wang et al. [19] proposed the Efficient Channel Attention Network (ECANet) based on SENet, which uses one-dimensional convolutions instead of fully connected layers to avoid dimensionality reduction and thus significantly reduces the network parameters. The ECANet attention mechanism is an extension and improvement of SENet to achieve cross-channel communication in an efficient way, introducing very small parameters that provide excellent results without significantly increasing the computational complexity while avoiding dimensionality reduction. ECANet achieves weight analysis of the importance of different channel feature maps by cross-channel interaction while not reducing dimensionality, allowing the network to extract more discriminative features for classification. The structure of ECANet is shown in Figure 3.

ECANet considers only the interaction between each channel and its k-nearest neighbors and generates weights for each channel by a one-dimensional convolution of size k, i.e.,

$$w = \sigma(C1D_k(y)) \tag{4}$$

where $C1D_k$ denotes a one-dimensional convolution with kernel size k, y is the result of applying GAP to all channels and σ is a sigmoid activation function. The value of k is related to the channel dimension. A larger number of channels will result in a larger range of local cross-channel interactions. The value of k is determined adaptively by a function related to the channel dimension, i.e.,

$$C = 2^{(\gamma * k - b)} \tag{5}$$

Thus, we can obtain the following equation:

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{6}$$

In the equation, $|t|odd$ is the nearest odd number to t, and the values of γ and b are set to 2 and 1, respectively.

2.2.3. Capsule Network

Sabour et al. [20] proposed the Capsule Network for the first time. Since then, it has caught the attention of many authors aiming to employ this novel network for image classification [24]. For example, Zhang et al. [25] proposed an enhanced capsule network for image classification, introducing feature decomposition modules and multi-scale feature extraction modules into the basic capsule network. This network can extract richer features, reduce computational complexity, and reach a state of convergence in a short time. Other researchers [26] proposed a single model Capsule network with focal loss for Kaggle Toxic Comment Classification, which beats other architectures with a total Area Under the Curve (AUC) of 98.46%. They have also shown that the problem that occurs during extensive preprocessing and augmentation of data can be tackled using Capsule networks.

The core concept of a Capsule Network is a "capsule", which is a collection of neurons that represents a specific type of feature or pattern; this property allows the capsule to learn the features of an image in addition to its deformations and viewing conditions. In traditional CNNs, features are extracted effectively through source and pooling layers and then classified through a fully connected layer. Traditional CNNs process images by focusing on the localization and spatial location of features while ignoring the relationships between targets. The Capsule network aims to overcome the limitations of traditional neural networks in dealing with pose variations, system features and spatial relationships.

The most important component of the capsule network is the capsule [20]. A capsule is a vector composed of neurons that characterizes the instantiation parameters and existence of multidimensional entities of the detection type.

Figure 4 shows the relationship between two layers of capsules in a capsule network; the transformation is as follows:

$$\hat{u}_{j|i} = W_{ij} u_i \tag{7}$$

where capsule i and capsule j are both vectors, u_i is the activation value of capsule i, and W_{ij} and C_{ij} are the transformation matrix and weights between the two capsules. $\hat{u}_{j|i}$ is called the prediction vector of low-level capsule i to high-level capsule j.

For all but the first layer of capsules, the total input to a capsule s_j is a weighted sum over all "prediction vectors" $\hat{u}_{j|i}$ from the capsules in the layer below and is produced by multiplying the output u_i of a capsule in the layer below by a weight matrix W_{ij}:

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \tag{8}$$

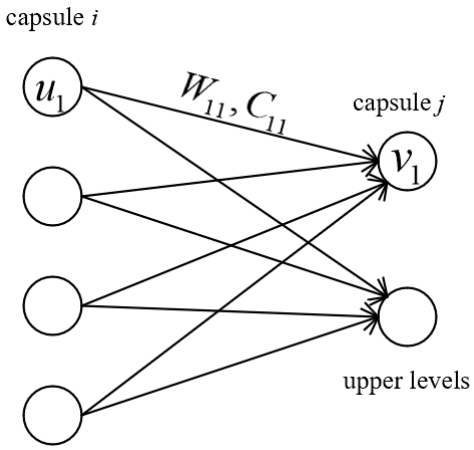

Figure 4. Capsule network structure.

We use a non-linear "squashing" function to obtain the activation value v_j for capsule j:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \tag{9}$$

where v_j is the vector output of capsule j and s_j is its total input. c_{ij} is defined as the coupling coefficient, which is calculated by

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \tag{10}$$

and b_{ij} is the update weight. The prediction vector $\hat{u}_{j|i}$ of the next layer is computed with the capsule output v_j of the previous layer to update b_{ij}. The formula is represented as:

$$b_{ij} = b_{ij} + \hat{u}_{j|i} \times v_j \tag{11}$$

3. Proposed Models and Methods

To overcome the limitation of inadequate utilization of unoccluded information and improve the effect of occluded face recognition, we propose a novel framework by incorporating attention mechanisms and capsule networks into the Inception-ResNet architecture. This method allows for making changes in the internal structure to classify images, without relying on occluded region information. The entire flowchart of mask-occluded face recognition is shown in Figure 5.

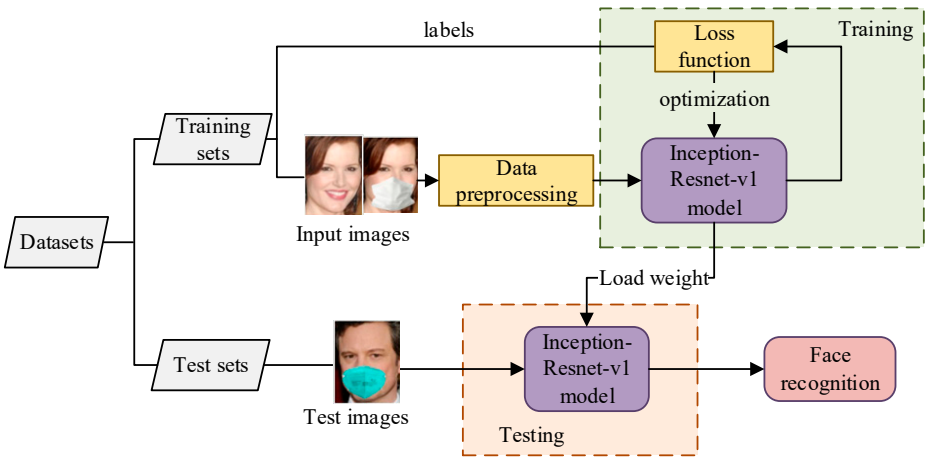

Figure 5. Flowchart of mask-occluded face recognition based on the Inception Resnet-v1 model.

3.1. SE-Based Attention Network

Compared with the residual structure of the ResNet network, this paper replaces the backbone network with the Inception-Resnet network to enhance the network's characterization ability. The Inception-Resnet network combines the advantages of Inception and ResNet networks, similar to model fusion, to further enhance the overall performance. As shown in Figure 6, the left figure shows the Inception-ResNet module without adding the SE attention mechanism, the middle figure shows the SE-Inception-ResNet module with adding the SE attention mechanism, and the right figure shows the ECA-Inception-ResNet module with the addition of the ECA attention mechanism.

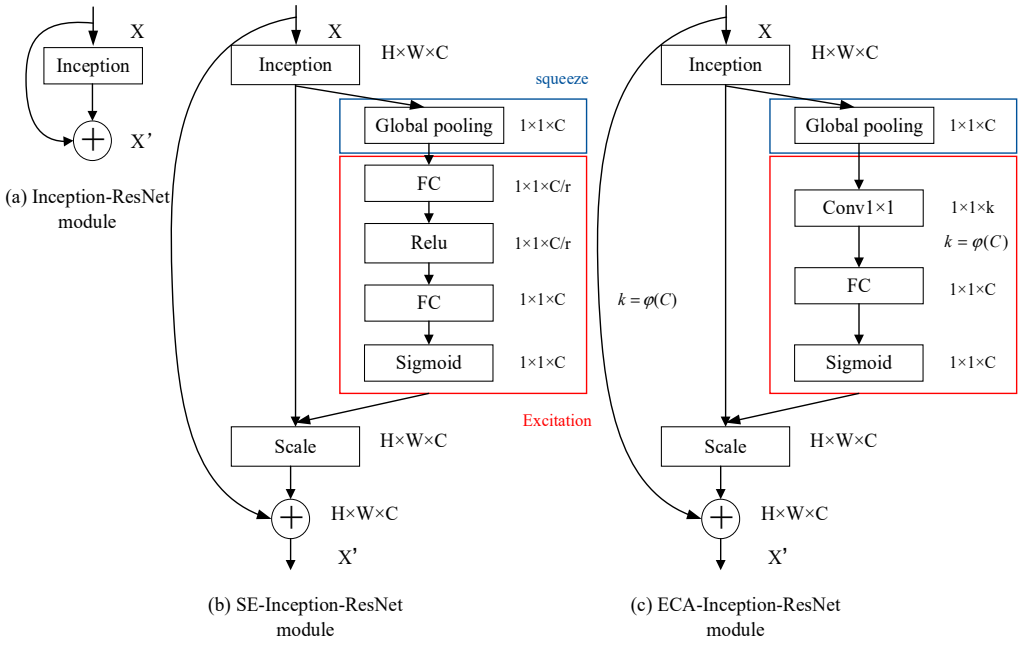

Figure 6. (**a**) Inception-Resnet module; (**b**) SE-Inception-Resnet module; (**c**) ECA-Inception-Resnet module.

According to Figure 1, Inception-Resnet-v1 consists of three important modules: Inception-Resnet-A, Inception-Resnet-B and Inception-Resnet-C. Among them, Inception-Resnet-C is a deeper one, which may extract higher-level face features than the other two modules. Based on the assumption, SE attention is first embedded into Inception-Resnet-C to evaluate the performance.

The Inception-ResNet-C module consists of three branches. The first branch is directly output without any processing. The second branch undergoes a convolutional kernel size of 1×1, and then it is run through a SENet module. The third branch undergoes three consecutive convolutions with kernel sizes of 1×1, 1×3 and 3×1, with output channels of 192, 192, and 192, respectively. It then goes through a SENet module. Finally, the obtained output is added to the first branch to obtain the new module, SE-Inception-ResNet-C. The structure of the module is shown in Figure 7.

To give a more comprehensive comparison, the SE module is also added to module A and model B of Inception-ResNet-v1, simultaneously. The three modules are combined to form the SE-Inception-ResNet (A+B+C) model. The SE-Inception-ResNet-A and SE-Inception-ResNet-B modules are shown in Figures 8 and 9, respectively.

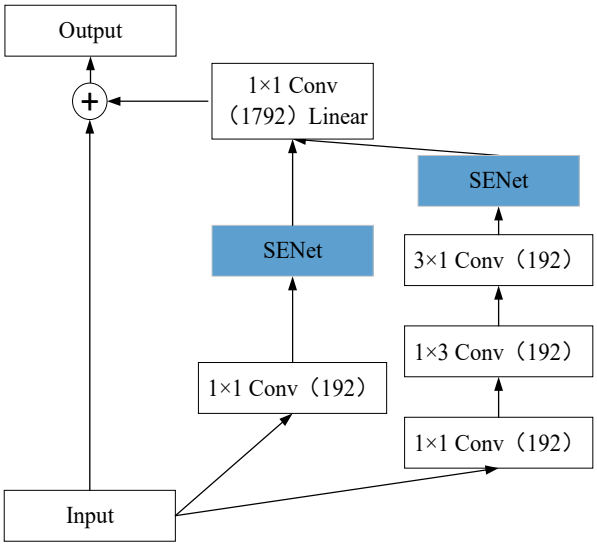

Figure 7. SE-Inception ResNet-C module.

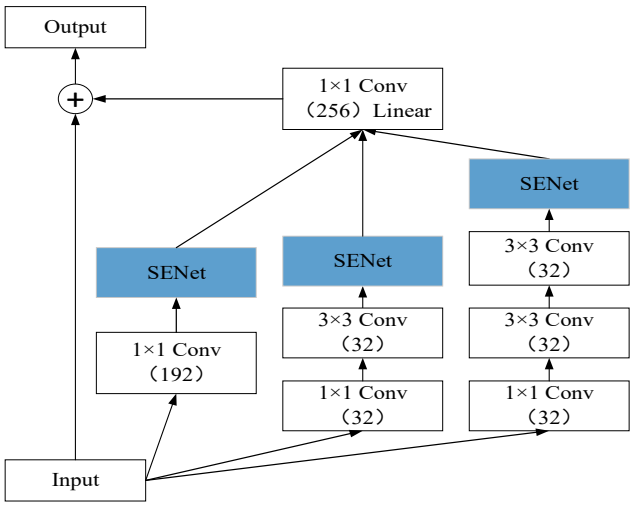

Figure 8. SE-Inception-ResNet-A module.

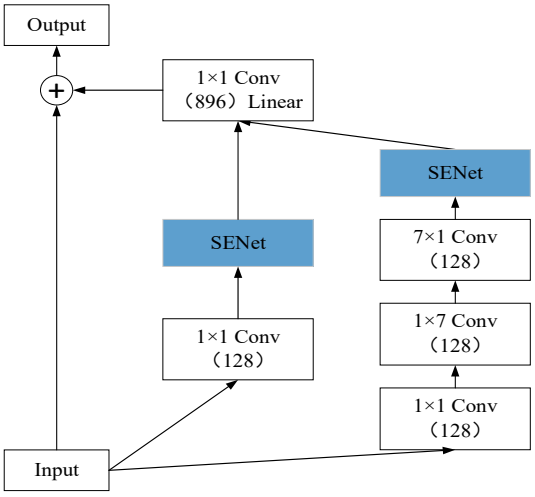

Figure 9. SE-Inception-ResNet-B module.

3.2. ECA-Based Attention Network

In order to extract effective facial features efficiently, this paper embeds the ECANet module into both the C module and the A+B+C full module of Inception-ResNet-v1, resulting in two new modules: ECA-Inception-ResNet-C and ECA-Inception-ResNet (A+B+C). The ECA module is embedded separately into the A, B, and C models of Inception-ResNet, as shown in Figure 10, Figure 11, and Figure 12, respectively.

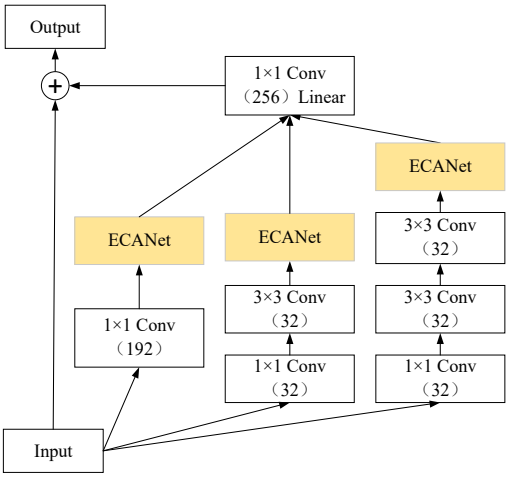

Figure 10. ECA-Inception-ResNet-A module.

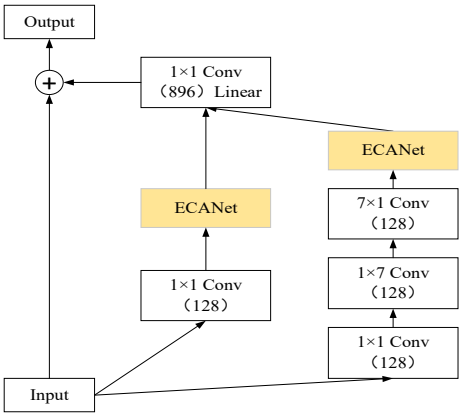

Figure 11. ECA-Inception-ResNet-B module.

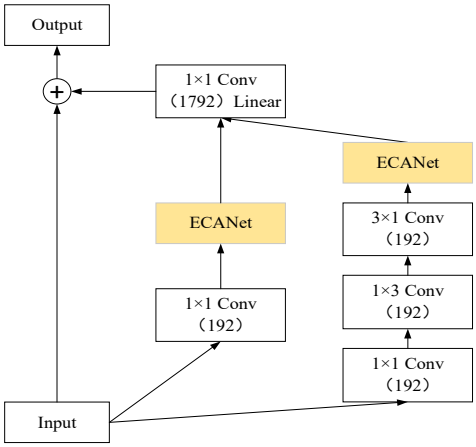

Figure 12. ECA-Inception-ResNet-C module.

3.3. EIRC Network

In unconstrained scenes, the existing methods tend to rely excessively on CNN features, and it is difficult to overcome the CNN defects using the aforementioned method. Based on the combination of the attention mechanism and Inception-Resnet, we propose a mask-occluded face recognition method based on ECA-Inception-ResNet-Caps (EIRC). This method integrates deep convolutional neural networks with capsule network modules, allowing for better fusion of features at different scales. This fusion enhances the accuracy and robustness of the model to a certain extent.

The EIRC model mainly combines the Inception model, Resnet ResNet architecture, ECA attention module, and capsule network. It aims to extract face features to the maximum extent and improve the accuracy of occluded face recognition in unrestricted scenes. The network structure of EIRC is shown in Figure 13. The proposed method has the following main improvements:

(1) The ECA attention module and capsule network module are introduced in module C of Inception-Resnet-v1.

(2) The network structure has been improved by using the convolutional capsule module to compute the embedding.

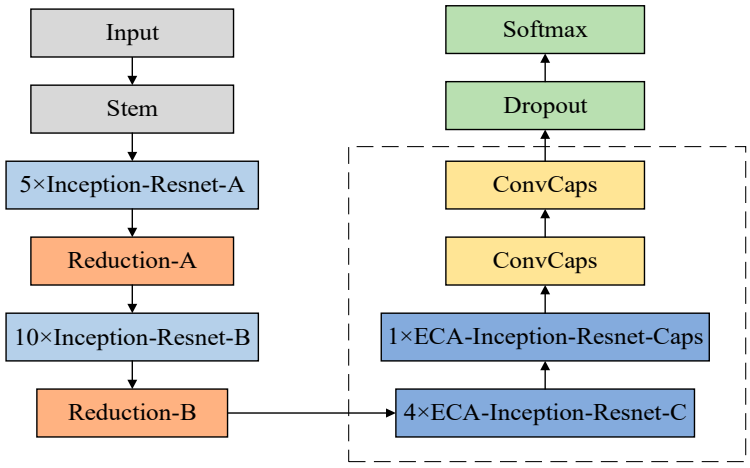

Figure 13. EIRC network structure.

Compared with traditional CNNs, the capsule network can learn the relationship between objects and extract deep face features with fewer parameters to better handle complex scenes and tasks. Therefore, many models embed capsule networks at the end of the network. By embedding the capsule network within the Inception-ResNet architecture, the model's robustness and generalization can be further enhanced. The ECA-Inception-Resnet-Caps module, which is based on the Inception-Resnet network embedded in the ECA module, embeds the capsule network into the Inception module in which the model performs feature fusion.

The EIRC model is similar in structure to the ECA-Inception-Resnet-C model proposed in the previous section, embedding the capsule network into the higher layers of this network. The EIRC divides the input into three parts: the first part is directly output and the second part is operated by a 1×1 convolution kernel and then passed through an ECANet module. The third part is processed by 1×1, 1×3, and 3×1 three convolution kernels, and the output channels are processed in three convolutional capsule layers of 192, 192, and 192, respectively, and then passed through an ECANet module. Finally, the resulting output is summed with the first branch to obtain the main ECA modules in EIRC, which is shown in Figure 14.

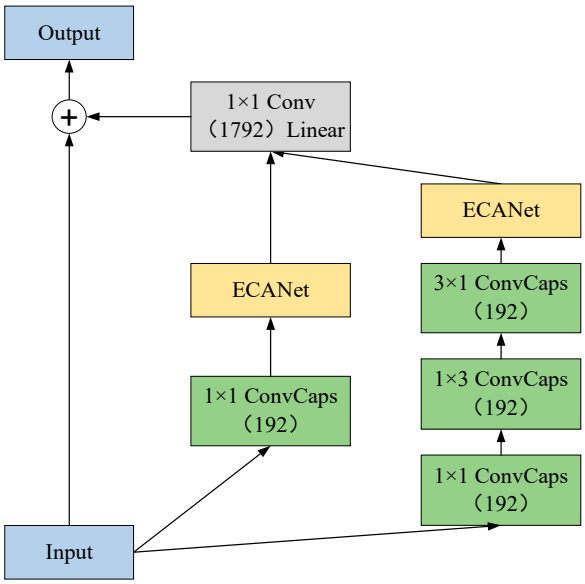

Figure 14. ECA modules embedded in EIRC.

4. Experiments

4.1. Experimental Setup

All experiments were implemented using an Ubuntu 20.04 PC, the TensorFlow library, and an NVIDIA Geforce GTX 1080 graphics card. We used Adam optimization for training and set the learning rate to 5×10^{-4}. The batch size was set to 128, and a total of 60 epochs were trained. The network hyper-parameters are shown in Table 1.

Table 1. Hyper-parameter setting for network training.

Hyper-Parameters	Setting
Optimizer	Adam
Learning rate	0.0005
Batch size	128
Epoch	60
Logits_margin	0.5
Logits_scale	64
Loss_method	Cross-entropy
weight decay	0.0005
Embed_length	128
Shuffle	10,000/20,000
Input shape	(112,112,3)

4.2. Employed Datasets

The experimental datasets are two well-known, publicly available face datasets, CASIA-WebFace [27] and LFW [21], with CASIA-WebFace as the training set and LFW as the test set. CASIA-WebFace is a commonly used public face dataset, which contains 10,575 identities of a total of 494,144 face images with rich variations in pose, light and occlusion. We chose it as the training database since it is almost independent of the LFW and can dispel the chaos of evaluation.

In this paper, the CASIA-WebFace dataset is pre-processed and mainly divided into three steps: face alignment, data cleaning and mask application. Face alignment is a crucial step in dataset processing, which is mainly performed using a Multi-task Convolutional Neural Network (MTCNN) [28]. Data cleaning is carried out to remove duplicate, unclear and redundant data to improve the quality of the dataset (https://codered.eccouncil.org/course/deep-learning-masked-face-detection-recognition, accessed on 13 September 2023). Mask application uses the open-source tool MaskTheFace [29], which utilizes the dlib

facial landmark detector to identify facial tilt and the six key features of the face needed to apply the mask, which is used to convert the existing face dataset into a masked face dataset. In this paper, we add multiple types of masks to the face images in the CASIA-WebFace dataset to generate the mask-occluded dataset Z-CASIA, which effectively creates a large mask-occluded dataset. In this paper, the model training uses a hybrid dataset of unmasked CASIA-WebFace and masked Z-CASIA, including 10,575 identities with 177,004 face images. Among these images, 90,883 were unmasked and 86,121 were masked. The Z-CASIA with an occlusion dataset is shown in Figure 15.

Figure 15. Example images for CASIA, Z-CASIA and LFW.

LFW is a public base test set for face verification collected in unconstrained scenes. The LFW dataset contains 13,233 images of 5749 identities. These images are collected from the internet, and the identities in these two datasets are entirely independent. Similarly, the LFW dataset undergoes face alignment using the MTCNN algorithm, as well as data cleaning and data augmentation techniques. The processed LFW dataset is shown in Figure 15.

4.3. Evaluation Criteria and Parameters

The evaluation metrics used in this paper for experiments include the macro average accuracy, macro average recall, and macro average F1 value. The final ranking is based on the accuracy of the test set, which represents the proportion of correctly classified samples out of the total samples. Assuming that there are n categories $C_1, \ldots, C_n$, P_i is the accuracy for categories (i.e., subjects) C_i and R_i is the recall for each category C_i. The formula for macro average accuracy, macro average recall, macro average F1 value and accuracy are as follows:

$$\text{macro average accuracy} = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{12}$$

$$\text{macro average recall} = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{13}$$

$$\text{macro average F1} = \frac{1}{n}\sum_{i=1}^{n} \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{14}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

where *TP* denotes the number of correctly classified positive samples, *TN* denotes the number of correctly classified negative samples, *FP* denotes the number of misclassified positive samples, *TN* denotes the number of misclassified negative samples, and *TP* + *FP* + *FN* + *TN* denotes the total number of test samples.

4.4. Experimental Results

To validate the effectiveness of the proposed algorithm, the experiments on face recognition were grouped as follows. First, the baseline model was employed. Second, we added the SE attention module to the baseline model. Third, the ECA module is employed instead of the SE module. The classification results are shown in Table 2. In this paper, Inception-Resnet-v1 serves as the baseline model, and two attention mechanisms, SE and ECA are embedded into this network independently.

Table 2. Experimental results on LFW with the proposed models.

Method	Precision	Recall	F1
Inception-Resnet-v1	92.09%	90.03%	91.02%
SE-Inception-Resnet-C	93.49%	92.00%	92.79%
SE-Inception-Resnet-ABC	93.07%	91.01%	92.01%
ECA-Inception-Resnet-C	**94.00%**	**92.11%**	**92.97%**
ECA-Inception-Resnet-ABC	93.28%	91.88%	92.55%

SE-Inception-Resnet-C represents the SE attention mechanism embedded into the C module, and SE-Inception-Resnet-ABC represents the SE attention mechanism embedded into the A+B+C module, they are used to compare the effect of embedding the SE module in different stages of the baseline model. Similarly, ECA-Inception-Resnet-C represents the ECA attention mechanism embedded into the C module, and ECA-Inception-Resnet-ABC represents the ECA attention mechanism embedded into the A+B+C module, they are used to compare the effect of embedding the ECA module in different stages of the baseline model. The experimental results on LFW for the proposed models are shown in Tables 2 and 3, respectively.

Table 3. Testing accuracy on LFW with the proposed models.

Method	Test Accuracy
Inception-Resnet-v1	92.9%
SE-Inception-Resnet-C	93.82%
SE-Inception-Resnet-ABC	93.63%
ECA-Inception-Resnet-C	**94.15%**
ECA-Inception-Resnet-ABC	94.03%

From Tables 2 and 3, it can be observed that the four proposed network models outperform the baseline model, Inception-Resnet-v1, indicating the effectiveness of the attention mechanisms. From Table 2, the ECA attention module achieves a higher F1 value compared with the SE attention module, and the ECA-Inception-Resnet-C model exhibits the best recognition performance. Therefore, this model is selected for occluded face recognition tasks. From Table 3, the recognition accuracy of networks embedded with attention mechanisms is higher compared to the baseline network, while the inclusion of ECANet has the highest accuracy. The ECA-Inception-Resnet-C model exhibits the best recognition performance, with a 1.25% improvement over the baseline Inception-Resnet-v1.

This paper conducted a series of comparative experiments to demonstrate the performance advantages of the proposed model improvements. Figures 16 and 17 show the accuracy and loss curves of Inception-Resnet-v1, SE-Inception-Resnet-C, SE-Inception-Resnet-ABC, ECA-Inception-Resnet-ABC and the proposed model in this paper on the

training module, which are represented by the black, purple, blue, green, and orange curves, respectively.

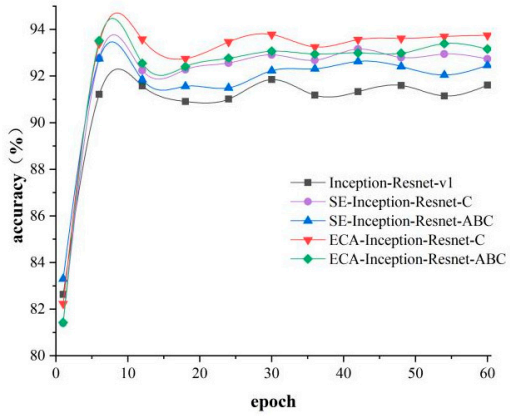

Figure 16. Comparison of accuracy curves.

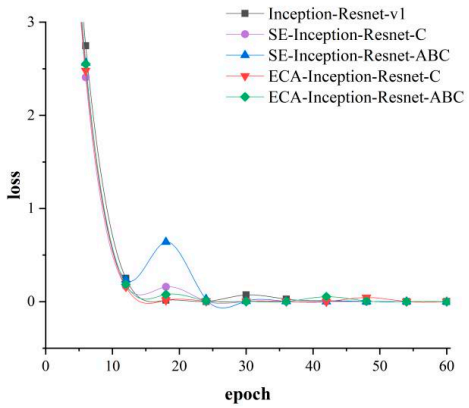

Figure 17. Training loss curves comparison for the proposed models.

As shown in Figure 16, all curves eventually converge. The black curve is the baseline model, Inception-Resnet-v1. All four proposed methods have higher accuracy than the baseline model, and the two ECA embedding methods outperform the two SE embedding methods. Among them, ECA-Inception-Resnet-C achieves the highest accuracy, followed by ECA-Inception-Resnet-ABC, SE-Inception-Resnet-C, and SE-Inception-Resnet-ABC. The experimental results demonstrate that ECA attention can improve the recognition performance of the model.

From Figure 17, all curves eventually converge. At epoch 20, SE-Inception-Resnet-ABC achieves the highest value, followed by SE-Inception-Resnet-C, while ECA-Inception-Resnet-ABC and ECA-Inception-Resnet-C have lower values. This indicates that embedding either SE or ECA attention mechanisms into the ABC module of the network is unstable. Therefore, embedding too many attention mechanisms increases the complexity of the network model and might impose a certain burden on the effect of recognition.

The comparison of ROC (Receiver Operating Characteristic) curves is shown in Figure 18. From the figure, it can be observed that the proposed four attention-based improvement algorithms outperform the baseline model in terms of recognition rate. When the SE module is added to the C module and the A+B+C module of Inception-Resnet, the AUC values reach 96.0% and 95.6%, respectively, which are 5.4% and 5.0% higher than the baseline model. When the ECA module is added to the C module and the A+B+C module of Inception-Resnet, the AUC values reach 98.5% and 97.0%, respectively, which are 7.9% and 6.4% higher than the baseline model.

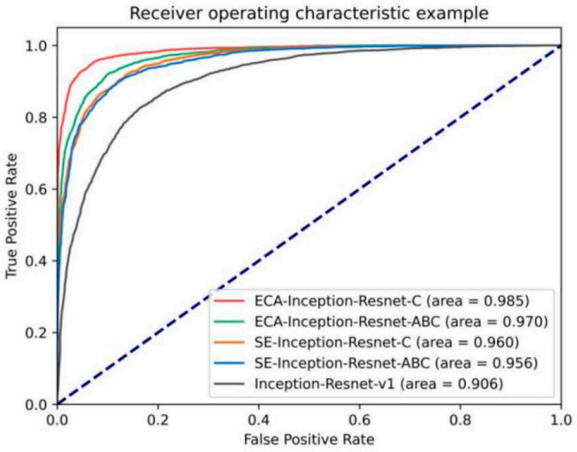

Figure 18. Comparison of ROC curves.

Before training the proposed model EIRC, it is first necessary to determine the value of the hyper-parameters for the Dropout layer of the network Inception-Resnet, namely dropout_keep_prob. For this experiment, four different values of dropout_keep_prob are selected for model training, and the cross-entropy loss function is used to verify the recognition accuracy under mask occlusion. As can be seen from Figure 19, the highest accuracy of 94.15% is obtained when keep_prob takes the value of 0.8.

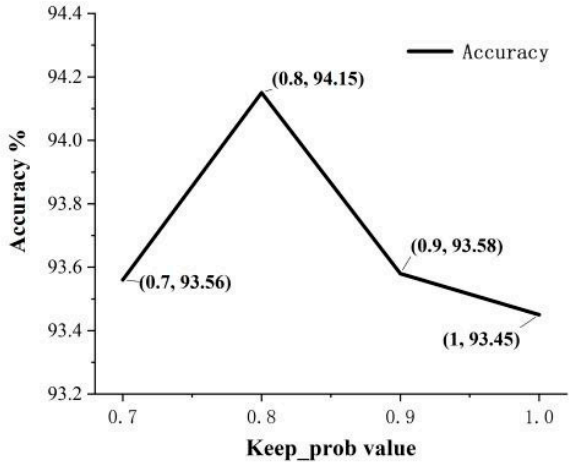

Figure 19. Recognition accuracy on LFW for different keep_prob values.

After determining the keep_prob value, this paper continues to train different network models. We compare the performance of Inception-Resnet-v1, ECA-Inception-Resnet-C, and EIRC when trained under different losses, the results are shown in Figure 20. The proposed model ECA-Inception-Resnet-Caps (EIRC) has higher recognition rates of 1.42% and 0.17% than the benchmark model and ECA-Inception-Resnet-C, respectively. This demonstrates that embedding the capsule network can improve the recognition rate under mask occlusion, verifying the effectiveness of the improved network. In addition, the experimental results show that using the cross-entropy loss function is better than the triplet loss function. The proposed model in this paper can achieve a recognition rate of 94.32% by using the cross-entropy loss function. The experimental results are shown in Figure 20 below.

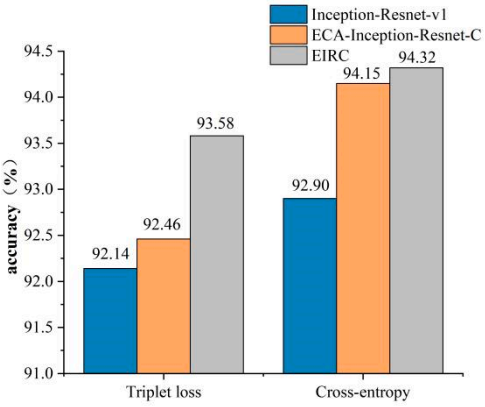

Figure 20. Statistical chart of algorithms for various models.

5. Conclusions

Compared to traditional learning-based methods that are only suitable for small-scale datasets, deep-learning-based methods have powerful learning capabilities and can capture more effective facial features when dealing with large-scale recognition tasks with mask occlusion. However, mask occlusion in unconstrained environments is still a big challenge since it has a significant negative impact on feature extraction on key face regions.

To further enhance the accuracy of face recognition under mask occlusion and address the limitation that the network model fails to fully utilize the information of the occluded region, this paper proposes an improved residual network model based on deep learning. The model is trained using a large-scale public dataset, CASIA-WebFace, to achieve accurate face recognition under mask occlusion. By introducing the SE and ECA attention mechanisms into the network modules, the network is enhanced to learn channel-wise information and capture richer facial features. In addition, the capsule network is introduced, and through comparative classification experiments, it is found that the EIRC hybrid model achieves a higher test accuracy of 94.32% compared to other models and accurately classifies faces with masks.

Author Contributions: Writing—original draft preparation, resources, Y.Z.; conceptualization, methodology, M.Z.; supervision, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Research Project of Henan Province, grant number 222102320039, and Major Public Welfare Project of Henan Province, grant number 201300311200.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors choose not to disclose the data.

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

References

1. Damer, N.; Grebe, J.H.; Chen, C.; Boutros, F.; Kirchbuchner, F.; Kuijper, A. The effect of wearing a mask on face recognition performance: An exploratory study. In Proceedings of the 2020 International Conference of the Biometrics Special Interest Group, Darmstadt, Germany, 16–18 September 2020; pp. 1–6.
2. Chen, C.; Sixiang, W.; Wenfeng, W. Research on face recognition system and standards for epidemic prevention and control. *Inf. Technol. Stand.* **2020**, *6*, 11–13.
3. Yaoling, X.; Tao, L.; Guohui, T.; Wenjuan, Y.; Dajun, X.; Shapeng, L. Overview of face recognition methods in occlusive environments. *Comput. Eng. Appl.* **2021**, *57*, 46–60.
4. Hemathilaka, S.; Aponso, A. A comprehensive study on occlusion invariant face recognition under face mask occlusion. *arXiv* **2022**, arXiv:2201.09089. [CrossRef]

5. Zhang, J.; Yan, X.; Cheng, Z.; Shen, X. A face recognition algorithm based on feature fusion. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e5748. [CrossRef]
6. Chen, Y.; Liu, S. Deep partial occlusion facial expression recognition via improved cnn. In Proceedings of the International Symposium on Visual Computing, San Diego, CA, USA, 5–7 October 2020; Springer: Cham, Switzerland, 2020; pp. 451–462.
7. Hariri, W. Efficient masked face recognition method during the COVID-19 pandemic. *Signal Image Video Process.* **2022**, *16*, 605–612. [CrossRef] [PubMed]
8. Mandal, B.; Okeukwu, A.; Theis, Y. Masked face recognition using resnet-50. *arXiv* **2021**, arXiv:2104.08997.
9. Song, L.; Gong, D.; Li, Z.; Liu, C.; Liu, W. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Repbulic of Korea, 27 October–2 November 2019; pp. 773–782.
10. Li, Y.; Guo, K.; Lu, Y.; Liu, L. Cropping and attention based approach for masked face recognition. *Appl. Intell.* **2021**, *51*, 3012–3025. [CrossRef] [PubMed]
11. Deng, H.; Feng, Z.; Qina, G.; Lv, X.; Li, H.; Li, G. MFCosface: A masked-face recognition algorithm based on large margin cosine loss. *Appl. Sci.* **2021**, *11*, 7310. [CrossRef]
12. Zeng, D.; Veldhuis, R.; Spreeuwers, L. A survey of face recognition techniques under occlusion. *IET Biom.* **2021**, *10*, 581–606. [CrossRef]
13. Li, C.; Ge, S.; Zhang, D.; Li, J. Look through masks: Towards masked face recognition with de-occlusion distillation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3016–3024.
14. Din, N.U.; Javed, K.; Bae, S.; Yi, J. A novel gan-based network for unmasking of masked face. *IEEE Access* **2020**, *8*, 44276–44287. [CrossRef]
15. Deng, J.; Guo, J.; An, X.; Zhu, Z.; Zafeiriou, S. Masked face recognition challenge: The insightface track report. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1437–1444.
16. Qi, D.; Hu, K.; Tan, W.; Yao, Q.; Liu, J. Balanced masked and standard face recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1497–1502.
17. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; p. 31.
18. 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.
19. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECANet: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
20. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
21. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Proceedings of the Workshop on faces in'Real-Life'Images: Detection, Alignment, and Recognition, Marseille, France, 16 September 2008.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
23. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
24. Patrick, M.K.; Adekoya, A.F.; Mighty, A.; Edward, B.Y. Capsule networks–A survey. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1295–1310.
25. Zhang, Z.; Ye, S.; Liao, P.; Liu, Y.; Su, G.; Sun, Y. Enhanced capsule network for medical image classification. In Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, Montreal, QC, Canada, 20–24 July 2020; pp. 1544–1547.
26. Srivastava, S.; Khurana, P.; Tewari, V. Identifying aggression and toxicity in comments using capsule network. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, Santa Fe, NM, USA, 25 August 2018; pp. 98–105.
27. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv* **2014**, arXiv:1411.7923.
28. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
29. Anwar, A.; Raychowdhury, A. Masked face recognition for secure authentication. *arXiv* **2020**, arXiv:2008.11104.