information

Article

Masked Face Recognition System Based on Attention Mechanism

Yuming Wang [1,†], Yu Li [1,†] and Hua Zou [2,*]

[1] School of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China
[2] School of Computer Sciences, Wuhan University, Wuhan 430010, China
* Correspondence: zouhua@whu.edu.cn
† These authors contributed equally to this work.

Abstract: With the continuous development of deep learning, the face recognition field has also developed rapidly. However, with the massive popularity of COVID-19, face recognition with masks is a problem that is now about to be tackled in practice. In recognizing a face wearing a mask, the mask obscures most of the facial features of the face, resulting in the general face recognition model only capturing part of the facial information. Therefore, existing face recognition models are usually ineffective in recognizing faces wearing masks. This article addresses this problem in the existing face recognition model and proposes an improvement of Facenet. We use ConvNeXt-T as the backbone of the network model and add the ECA (Efficient Channel Attention) mechanism. This enhances the feature extraction of the unobscured part of the face to obtain more useful information, while avoiding dimensionality reduction and not increasing the model complexity. We design new face recognition models by investigating the effects of different attention mechanisms on face mask recognition models and the effects of different data set ratios on experimental results. In addition, we construct a large set of faces wearing masks so that we can efficiently and quickly train the model. Through experiments, our model proved to be 99.76% accurate for real faces wearing masks. A combined accuracy of 99.48% for extreme environments such as too high or lousy contrast and brightness.

Keywords: deep learning; face recognition; mask; attention mechanism

Citation: Wang, Y.; Li, Y.; Zou, H. Masked Face Recognition System Based on Attention Mechanism. *Information* **2023**, *14*, 87. https://doi.org/10.3390/info14020087

Academic Editors: Xin Ning, Yizhang Jiang and Weiwei Cai

Received: 21 December 2022
Revised: 16 January 2023
Accepted: 22 January 2023
Published: 2 February 2023

1. Introduction

In recent years, face recognition has been the subject of research in various countries worldwide. It has advantages that cannot be compared with other identification methods, such as non-intrusiveness, intuitiveness and convenience. In a COVID-19 epidemic, the virus can be transmitted through physical contact and airborne droplets. Wearing a mask can effectively stop the entry of harmful substances into the respiratory tract. It also prevents droplets from getting into the air when we sneeze, cough or speak. Therefore, wearing a mask can be very effective in preventing the spread of COVID-19, which is essential for preventing and controlling the outbreak [1,2].

Traditional identification includes fingerprint, iris, facial, and palm print recognition. The different identification methods also have some disadvantages. For example, iris recognition is highly accurate but not very cost-effective [3] or as easy to access as fingerprints and palm prints. However, there are better times to use both of them. As we know, too much exposure allows viruses and bacteria to be carried and spread unknowingly, which can further escalate the severity of an outbreak. The face has been widely accepted as a reliable biometric parameter compared to other forms of identification [4]. The acceptability of the face is also higher. A key reason for this is that everyone has a face, which is usually easy to show.

In the past few years, face recognition has developed rapidly due to improvements in loss functions [5–10], elaborate convolutional neural networks [11–15] and the availability of large data sets [16–18]. It is not just these normal face recognition developments that are making rapid progress—some directions for face recognition in extreme conditions are also

progressing well. For example, in [19], a single network masking face recognition method named FROM is introduced. It can be trained end-to-end, while learning feature masks and depth-masking features. In [20], they built a mask dictionary using an innovatively designed two-by-two differential concatenation network (PDSN). Its feature discard mask is then generated by combining relevant dictionary entries to eliminate corrupted feature elements from recognition. The focus of these projects is to improve the accuracy of identifying people when wearing sunglasses and scarves. The method in [19] has a 100% accuracy rate in identifying people wearing sunglasses and scarves. The downside is that they are not specifically designed to identify the person wearing the mask. By extending those excellent face recognition models directly to face recognition models with masks, we find that their performance degrades significantly and is not sufficient for the intended use.

To prevent the spread of COVID-19, the World Health Organization was very quick to tell people to wear masks in public and to make their use a mandatory biosecurity measure. Since people need to wear masks from time to time, this poses a problem for the previous field of face recognition. Although face recognition has been an essential area of research in various countries worldwide [21,22], there are still areas for improvement in recognizing faces wearing masks.

Due to the lack of a large data set of real faces wearing masks, most masked face data sets are synthesized by software using faces from the original data set with masks (e.g., masks, sunglasses). We believe that they are not well suited to real-world applications and that some errors have an impact when we train the network. The result is a lower accuracy rate when testing real faces. In order to demonstrate that our network can work well for recognizing faces wearing masks, we created a data set of 14 people with 418 faces, as specified. This data set was also augmented to 1538 sheets by data enhancement and used to test our network.

In this paper, we improve on the original based on FaceNet [23] so that it can be used as a model to recognize faces wearing masks, and the contributions of this paper are summarized below:

1. Inspired by FaceNet [23], we have improved the original model so that it can learn face features better. Moreover, there is a significant improvement in the accuracy of recognizing faces wearing masks, which is vital in the COVID-19 era for public places where masks are required.
2. We place great importance on the use of attention mechanisms. We believe that suitable attention mechanisms can effectively enable the network to learn more helpful information while paying less attention to other invalid information and even sifting out unrelated information. Moreover, we use ConvNeXt-T [24] as a new backbone of FaceNet, which has trouble with larger models. Using a suitable attention mechanism can solve the information overload problem well. Therefore, we tested the feasibility of most of the currently available attention mechanisms for recognizing faces wearing masks
3. We produced a data set containing 1538 images of real faces wearing masks. Our network achieved excellent results under extreme conditions (such as too bright, too dark, too high or too low contrast) and under normal conditions. It also has a good accuracy rate for normal faces, indicating that it has some robustness for normal face recognition as well.

2. Related Work

2.1. General Face Recognition and Face Recognition with a Mask

Over the past decade, scientists have worked on developing face recognition technology. Some of the latest methods have achieved almost 100% accuracy in recognizing normal faces and have high accuracy on some low-quality face data sets. Currently, the main effort in face recognition in general is to improve the loss function, with the core being to enhance the discrimination of the model by maximizing the inter-class variance and minimizing the intra-class variance. There are two main methods, the first of which is to optimize the

model by optimizing the comparison between the two distances and then obtaining a loss based on the relationship between the input samples [6,25,26]. An alternative approach is to define model training as classification task learning, which can make use of large data sets to strengthen the classification capabilities of the model [5,7,8,27]. Center loss [9] learns the features and feature centers of each identity, and then uses this obtained information to reduce the intra-class variance. CosFace [8] proposed a cosine interval term to amplify the inter-class variance. Large margin softmax [27] achieves an increase in the intergroup distance by adding an angle constraint, which in turn compresses the intragroup distance. Angular softmax [7] introduces the hypersphere space by normalizing the weights so that the points on the features are mapped to the unit hypersphere. ArcFace [5] optimizes geodesic distance margins by normalizing the exact correspondence between angles and arcs in the hypersphere. UniformFace [28] makes the class centers spread uniformly in the feature space thus maximizing the inter-class distance. Adaface [29] introduces image quality as a factor, proposing a generalized loss function by using the feature paradigm to approximate image quality, and can move arbitrarily between ArcFace [5] and CosFace [8], improving the recognition accuracy of low-quality images without losing the accuracy of high-quality images.

Due to the effects of COVID-19, people's daily activities are restricted on a large scale. At the height of the epidemic in China, almost all outdoor activities were forbidden, even shopping was not allowed, and supplies were delivered only by special staff. The effect of these things is a significant drop in people's productivity. People's quality of life declined accordingly. Despite this, we should still understand that human health and safety come first, which is why biosecurity measures are necessary to restrict the spreading of the virus. In addition, people's work and factory production have become much less efficient as a result of the epidemic, so many measures have been implemented, and many technologies have been developed. These measures ensure that people can move around as freely as possible and interact face-to-face during the epidemic while safeguarding their health. CNN is a convenient and effective tool during this epidemic. Some systems use CNNs and sensors for security and control, where CNNs are used as the basis for face recognition, and sensors can detect the temperature of the person being tested to determine if they have a fever, etc. Ref. [30] used CNN networks to detect whether people were wearing masks and to confirm the identity of faces. Ref. [31] used FaceNet as the base network framework, the MobileNetV2 architecture as the feature extractor and the OpenCV face detector to produce a facial recognition system for people wearing masks and without masks. They tested a data set of real faces wearing masks that they photographed themselves. The recognition rate for faces wearing masks reached 99.52%, and the accuracy rate for face recognition of people not wearing masks reached 99.96%. Inspired by [31], we improved FaceNet to produce better results for faces wearing masks.

2.2. FaceNet

The general face recognition process is divided into four steps: face detection, face localization, face feature extraction and face retrieval. In the face recognition process, several key points are generally located on the detected face. The multi-dimensional floating-point vector of face features and face confidence are then calculated from these key points, and finally, similar faces are retrieved from the face feature library based on the face features. The cosine angle or Euclidean distance often measures similarity.

As illustrated in Figure 1, the authors use a CNN model to extract face features directly. FaceNet [23] uses end-to-end learning of face images directly in order to map them into the same Euclidean space and to be measurable. After learning how to encode from image to Euclidean space, the network undergoes face recognition, face verification and face clustering based on this encoding.

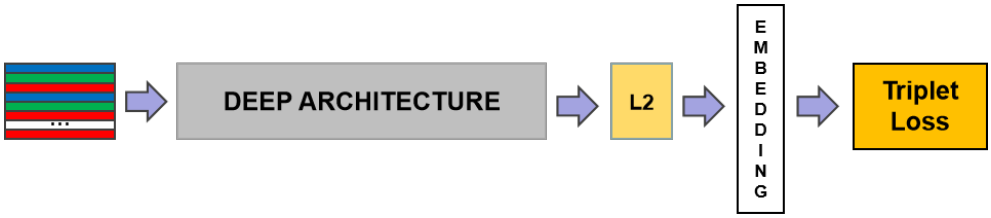

Figure 1. Architecture of FaceNet.

Triplet Loss is an essential feature of FaceNet. The mapped vector representation $f(x)$ can be made metrizable in Euclidean space by Triplet Loss. The intent of Triplet Loss is to enable closer Euclidean distances for vectors of the same face image in Euclidean space and further Euclidean distances for vectors of different face images in Euclidean space. The process can be expressed as follows: suppose the input face image is x_i^a, called the anchor. The image of the same person's face as x_i^p, is called the positive. Another image of a different person's face as x_i^n, is called the negative. We need to make the vectors between x_i^a and x_i^p closer together and between x_i^a and x_i^n further apart. It can be expressed as Equation (1).

$$\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 + \alpha < \| f(x_i^a) - f(x_i^n) \|_2^2, \forall f(x_i^a), f(x_i^p), f(x_i^n) \in \tau \tag{1}$$

$$L = \sum_i^N \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha \right] \tag{2}$$

where α is the threshold that sufficiently separates the positive and negative samples. τ is the set of all triples (x_i^a, x_i^p, x_i^n), the set size is N. The training process is shown in Figure 2. After learning, it makes the Euclidean distance between the anchor and positive closer and the Euclidean distance between the anchor and negative further and further. Equation (2) shows the calculation of the loss (L). In our approach, we use Triplet Loss because it allows for efficient training of the model in the absence of data.

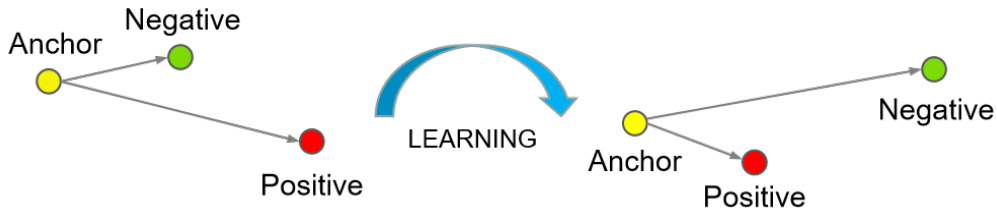

Figure 2. The Triplet Loss minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity.

3. Method

Our improved FaceNet is a clear and unambiguous end-to-end approach. It follows the architecture of Figure 3. First, it takes normal face images and masked faces as input. The feature extractor with ConvNeXt-T [24] as the backbone, thus extracting rough face features, and then further optimized by the ECA module. It calculates the metric to generate vectors of real characteristics. The feature vectors are compared with the face feature vectors in the face database, and the final result is obtained.

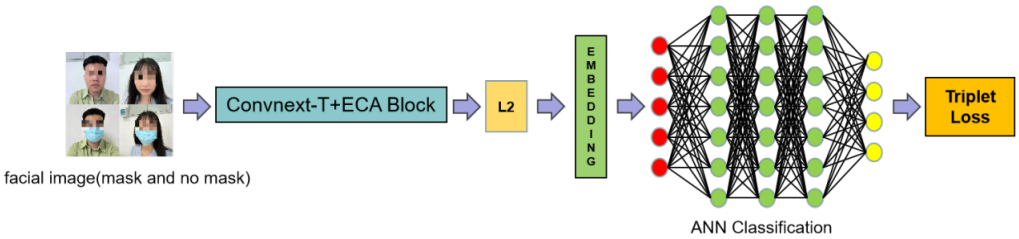

Figure 3. An overview of the proposed framework. First, a batch of masked and unmasked face images are used as the input to the network. Then, deep CNN feature extraction and L2 normalization are used to achieve face embedding.

3.1. Feature Extractor

We use ConvNeXt-T as our backbone network and obtain both spatially-aware features for exact mask learning and discriminant features for recognition. ConvNeXt [24] is based on resnest50, which uses the training strategy of VIT to train the original network model on top of resnet50, and uses it as a baseline to imitate the stage compute ratio of the swim transform. In addition, a series of improvements, such as using depthwise convolution, using the new Inverted Bottleneck, and changing the activation function. These improvements have led to a huge breakthrough, while maintaining the simplicity and efficiency of standard ConvNeXt. With the same flops, ConvNeXt has much better accuracy and inference speed than Swim Transform in Imgnet-1k, which is why we choose it. It follows the architecture of Figure 4.

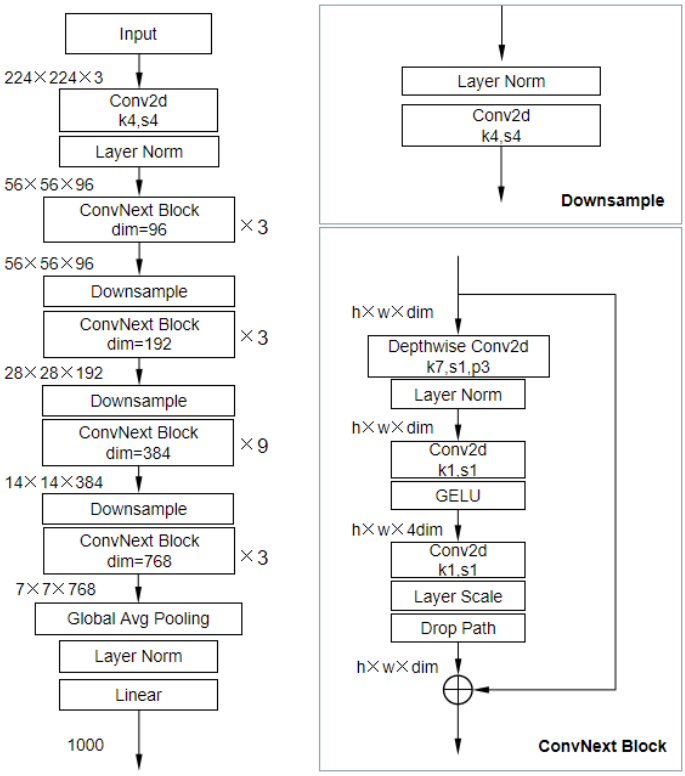

Figure 4. Architecture of ConvNeXt-T.

3.2. Attention

In recent years, attention mechanisms have been widely used in various areas of deep learning. Attention mechanism models can be found almost everywhere in projects such as image enhancement, semantic segmentation, and target detection. The attention mechanism in deep learning is similar to the human visual attention mechanism in that

it essentially prioritizes attention to more interesting targets or information that is more appropriate for the task at hand [32]. The human visual system cooperates with the brain to quickly focus on the area of interest when observing things. The brain then devotes more resources to this area that needs attention, thus obtaining more detailed information while suppressing other useless information.

More carefully, Attention can be divided into Channel attention, Spatial Attention, Channel and Spatial Attention and Self-attention mechanisms. Using the channel attention mechanism allows the importance of different channels to be learned based on a specific task by applying the weights of each feature channel to each of the original feature channels once they are obtained. In our model, we use an efficient channel attention (ECA) block [33], which uses a 1D convolution to avoid channel dimensionality reduction when learning channel attention information. The ECA block is very similar to the SE block [34], as shown in Figure 5. The SE block can be split into two parts, including a squeeze module for aggregating global information and an efficient excitation module for simulating cross-channel interactions. Global information is gathered in the squeeze module through global pooling.

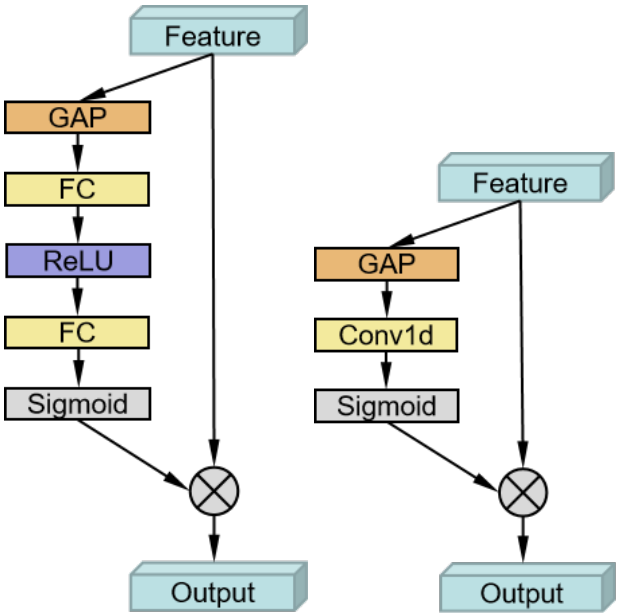

Figure 5. The left one is the SE block. The right one is the ECA block.

Then, the fully-connected layers in the excitation module are used to remove the redundant information and output the channel attention vector. Finally, the output is obtained by multiplying each channel information of the input feature by the channel attention vector. However, the SE module uses too many fully-connected layers resulting in too much complexity of the model. The ECA module uses a local cross-channel interaction strategy to generate channel weights by performing a one-dimensional convolution of convolution kernel size k, effectively reducing the model complexity while maintaining performance. In summary, the formulation of an ECA block is as follows:

$$s = F_{eca}(X, \theta) = \delta(Conv1D(GAP(X))) \tag{3}$$

$$Y = sX \tag{4}$$

where $Conv1D(\cdot)$ denotes $1D$ convolution with a kernel of shape k across the channel domain, to model local cross-channel interaction. The covering range of the interaction is

dependent on parameter k, where the size of k is determined adaptively, according to the channel dimension C, using cross-validation:

$$k = \psi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{5}$$

4. Experiment

4.1. Data Sets and Evaluation Metrics

Data sets: In the following, we present the face data sets used for training and testing the model, respectively.

CASIA-WebFace [35]: The CASIA-WebFace data set is currently the dominant data set in face recognition, containing 494,414 images of 10,575 individuals. According to [18], The CASIA-WebFace data set is applied as the training data set. We also produced a muzzled CASIA-WebFace data set (Substitute for WebFace-Mask).

Labeled faces in the wild (LFW) [17]: LFW is an unconstrained natural scene face recognition data set consisting of more than 13,000 face images of famous people worldwide in natural scenes with different poses, expressions and lighting environments. Each face image has a corresponding identity. Because of these factors, even images of the same person can vary greatly. In addition to the original LFW data set, we also produced a muzzled LFW data set (substitute for LFW-Mask).

Test data set: Since there is currently no database to publicly identify people wearing masks, we took a total of 417 images of faces wearing masks for a total of 14 people (9 males and 5 females) per the relevant regulations and with the permission of the people photographed. The data set was also augmented to 1538 using data enhancement, randomly decreasing or increasing the contrast and brightness, and was used to test the accuracy of our model under extreme conditions. Figure 6 shows some examples.

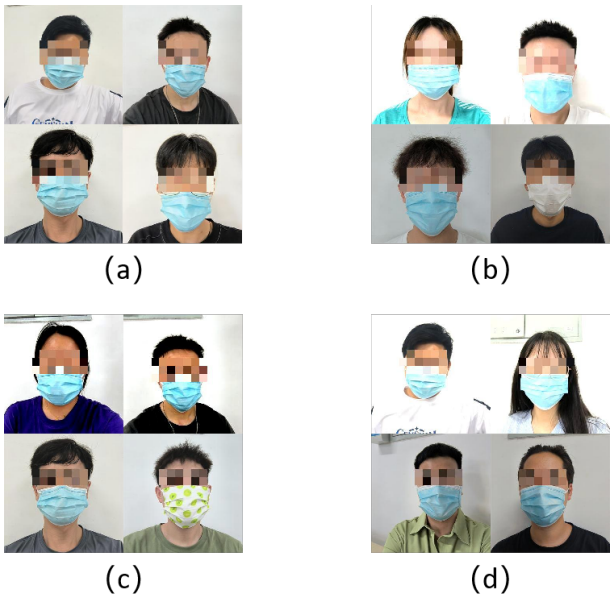

Figure 6. Test data set. (**a**) Unprocessed face images. (**b**) Face images with brightness adjustment. (**c**) Contrast-adjusted face images. (**d**) Face image with simultaneous contrast adjustment and brightness adjustment.

Evaluation Metrics: We employ four widely-used metrics to quantitatively evaluate the face recognition performance in our experiments, including *Accuracy*, *Precision*, *Recall* and *F1-score*.

4.2. Implementation Details

Pre-processing: We apply the corresponding similarity transformation to the captured face image wearing a mask, align and crop the face image appropriately, and obtain a face image of 300 × 300 size. Then, the original image is image enhanced. We divide the produced face data set into two categories. One is the face data set (Face-Nor) without data enhancement and the other face data set (Face-En) includes normal and data-enhanced face images.

Train: The part of training can be split into two phases. In the first stage, we learn backbone networks (i.e., Figure 4) with Triplet Loss for general face recognition on the CASIA-WebFace data set. Then, we use the trained model in the first stage as our pre-trained model and train the entire network for 30 epochs, including the feature extractor and ECA blocks, on the WebFace-Mix data set, which is a 1:1 ratio mix of WebFace and WebFace-Mask. By the way, we trained each epoch in about 40 min on an NVIDIA GeForce RTX 3090 graphics card.

4.3. Ablation Study

4.3.1. Backbone

In order to verify the effectiveness of ConvNeXt and ensure that the network's model is manageable, we first choose ConvNeXt-T as the pre-selected model to experiment with the current face recognition models. We use the LFW data set for training in the first stage, use the LFW-Mask data set training directly in the second stage without adding the attention mechanism, and train 100 epochs in both stages. As shown in Table 1, we see that after simple training, the model with a ConvNeXt-T backbone has an accuracy of 85.85% on the Face-Nor data set, 84.52% on the Face-En data set, and the others are less accurate. Therefore, ConvNeXt-T is selected as the backbone for our later experiments. Incidentally, the accuracy is higher when we use a larger data set as the training set.

Table 1. Face verification comparison (%) on Face-Nor and Face-En when using different backbones.

Method	Face-Nor	Face-En
inception_resnetv1 [36]	72.66	64.95
iresnet50 [37]	73.38	71.58
Mobilenetv1 [38]	83.93	82.51
ConvNeXt-T [24]	90.16	89.40

4.3.2. Data Set Settings

It is well known that training a face recognition model with a larger face data set will improve the recognition ability of the model. In the first phase, we train the model using Webface, and then in the second phase, we adjust the training data set. Because we need to satisfy the need to recognize both masked and normal faces, we made three data sets with different scales using the WebFace and WebFace-Mask. In this paper, three important baselines are considered.

- Web-NOR: The backbone network is only trained on the WebFace-Mask data set for 40 epochs.
- Web-AUG: 30 epochs of training in the second stage using a mixed data set consisting of the WebFace and WebFace-Mask in a 1:2 ratio.
- Web-MD: Mix the WebFace and WebFace-Mask in the ratio of 1:1 as the training set. This is also our final choice for the second stage training set.

As shown in Table 2, we have compared the three previously mentioned baselines in the same configuration in this experiment. The Web-NOR is significantly less accurate in recognizing faces wearing masks, which is consistent with our intuition. Next, Web-AUG with the training hybrid data set performs significantly better than Web-NOR, which may be because faces with masks and normal faces correlate even for the part of the face with the

mask on. However, when we compare Web-AUG and Web-MD, we can see that Web-MD has a clear advantage, which indicates that Web-MD is more effective. From a general face recognition perspective, Web-MD is also more generalizable. In the following experiments, we use Web-MD as our training method.

Table 2. Face verification comparison (%) on Face-Nor and Face-En when using different data set proportions.

Method	Face-Nor	Face-En
Web-NOR	87.76	84.65
Web-AUG	95.20	93.17
Web-MD	97.12	95.92

4.3.3. Discussion of Attention

After the initial extraction of face features by ConvNeXt-T, we believe that adding the attention mechanism can aggregate face features more effectively. Because wearing a mask will obscure a considerable part of the face, we would like to increase the weight of the effective features of the mask-wearing face by the attention mechanism, so we conducted a series of experiments on the attention mechanism. As shown in Table 3, we selected some representative attention mechanisms for testing. The SE block and ECA block are clearly superior to other attention mechanisms. Channel attention has a very positive effect on our model. We speculate that Polarized Self-Attention [39] may not have paid attention to the correlation between feature channels, so the effect is poor. Regarding the CBAM block [40], we believe that mixed attention may not fit our model and that pure channel attention may be better suited.

Table 3. Face verification comparison (%) on Face-Nor and Face-En when using different attention mechanisms.

Method	Face-Nor	Face-En
Polarized Self-Attention [39]	94.50	94.02
CBAM [40]	98.56	96.94
SE [34]	99.28	98.57
ECA [33]	99.76	99.48

4.3.4. Comparison with SOTA Methods

To compare with SOTA methods, we have selected some representative methods. We uniformly use WebFace and Web-MD as training sets when testing these methods, with other settings unchanged, and then test them on Face-Nor and Face-En. Test results are shown in Tables 4 and 5, which show that our approach can significantly improve the performance of the CNN model on the faces of people with realistic mask coverings.

Table 4. Comparison with other state-of-the-art methods on Face-Nor.

Method	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Facenet [23]	83.93	88.64	92.04	90.31
PDSN [2]	98.80	99.18	99.45	99.31
FROM [19]	99.52	99.46	99.73	99.59
Pre-Facenet [31]	99.04	99.18	99.46	99.32
Ours	99.76	99.73	100	99.86

Table 5. Comparison with other state-of-the-art methods on Face-EN.

Method	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Facenet [23]	82.51	82.71	93.27	87.67
PDSN [2]	98.57	98.89	99.48	99.18
FROM [19]	99.35	99.48	99.78	99.63
Pre-Facenet [31]	99.02	99.26	99.63	99.44
Ours	99.48	99.56	99.85	99.70

4.3.5. Result

With the above discussion, we arrive at the best training scheme and network model. Figure 7 shows the training graph of the face recognition model.

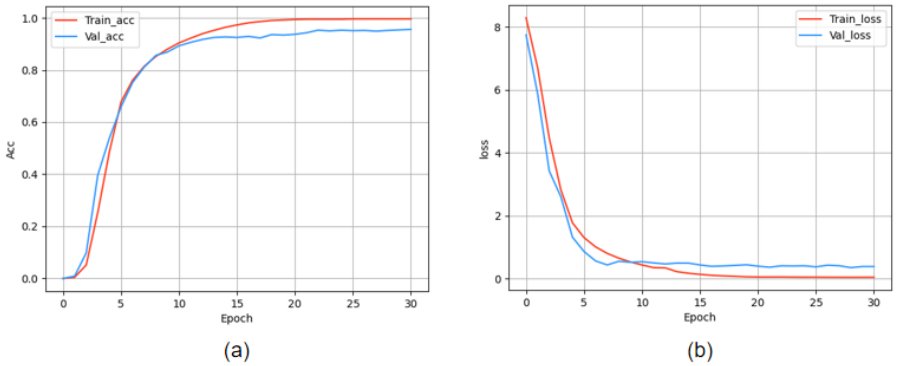

Figure 7. Training graph of accuracy and loss with the number of epochs. (**a**) Accuracy graph. Train_acc indicates the accuracy curve of the training set. Val_acc indicates the accuracy curve of the validation set. (**b**) Loss graph. Train_loss indicates the loss curve of the training set. Val_loss indicates the loss curve of the validation set.

Our model achieves an accuracy rate close to 100% during the training process. When the model was evaluated with test data with face masks, the unprocessed face recognition accuracy was 99.72%. The 99.48% accuracy of face recognition after data enhancement. This indicates that our model is also robust to faces after data enhancement.

In order to verify that our face recognition module is realistic, we tested the module with real face images and simulated similar outdoor and indoor access control environments, both of which yielded good accuracy and optimization results. As shown in Figure 8, The model successfully recognizes a person's face. Moreover, when a person's face is in the face database, their name and confidence rate are displayed on the image. Furthermore, if a person's face is not in the face database, the face will still be detected and the word "unknown" will be added to the label, representing a stranger.

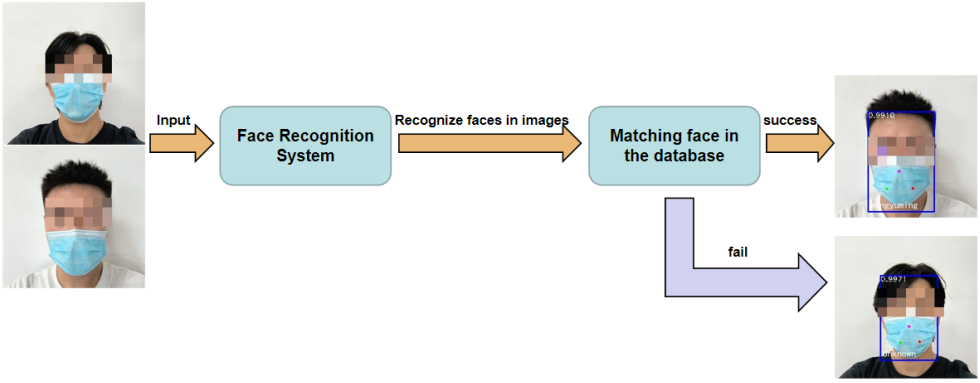

Figure 8. Results of identification between strangers and registered people.

5. Conclusions and Future Works

Our method is a very effective way to bring the COVID-19 pandemic under control. The method is used to identify people wearing masks, determine whether they are stored in the database, and identify people who are stored in the database. If applied correctly, our method could be used to ensure our safety and others. This model can realize faces while wearing a mask, which can be applied as a low-cost solution for controlling the movement of people in situations where health and safety are required. In this sense, our model yields satisfactory results in experiments. Our model also lays the foundation for further expansion of research in this area.

However, our approach still has many problems. When making the data set for testing, we chose to photograph real faces to make our experiments closer to real life. Since only 14 people sampled the data set, it resulted in a small sample of faces in the data set that did not have much variability. We randomly processed faces wearing masks with data enhancement in order to simulate environments with different illumination or contrast, but these data-enhanced images of faces did not particularly match the realistic environment. However, the model still shows potential for use in differentiated facial recognition applications. Therefore, as future work, a real face mask data set is indispensable for both training and testing, and we should consider improving the loss function to enhance the discriminative ability of the model by maximizing the inter-class variance and minimizing the intra-class variance, to enhance the recognition ability of the model for faces wearing masks.

Author Contributions: Conceptualization, H.Z. and Y.L.; methodology, Y.W. and H.Z.; software, H.Z. and Y.L.; validation, H.Z. and Y.L.; formal analysis, Y.W. and H.Z.; data curation, Y.W. and H.Z.; writing—original draft preparation, Y.W. and H.Z.; writing—review and editing, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Bingtuan Science and Technology Program (grants no. 2019BC008 and no. 2022DB005).

Institutional Review Board Statement: Not applicable.

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

Data Availability Statement: Publicly available data sets were used in this study. The CASIA-WebFace data set can be found here: https://pan.baidu.com/s/1cnnKrYQDheNfoEhcDoShyA, (accessed on 25 January 2023, code: vk36). The LFW data set can be found here: https://pan.baidu.com/s/1DR600XJFhm8lqfHZ6mOU_A, (accessed on 25 January 2023, code: akbi).

Acknowledgments: We would like to thank the authors of the methods compared, including Facenet, FROM, Convnext. Our deepest gratitude goes to the reviewers and editors for their careful work and thoughtful suggestions that have helped improve this paper substantially.

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

References

1. Coccia, M. The impact of first and second wave of the COVID-19 pandemic in society: Comparative analysis to support control measures to cope with negative effects of future infectious diseases. *Environ. Res.* **2021**, *197*, 111099. [CrossRef] [PubMed]
2. Cheng, V.C.C.; Wong, S.C.; Chuang, V.W.M.; So, S.Y.C.; Chen, J.H.K.; Sridhar, S.; To, K.K.W.; Chan, J.F.W.; Hung, I.F.N.; Ho, P.L.; et al. The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2. *J. Infect.* **2020**, *81*, 107–114. [CrossRef] [PubMed]
3. Daugman, J. How iris recognition works. In *The Essential Guide to Image Processing*; Elsevier: Amsterdam, The Netherlands, 2009; pp. 715–739.
4. Van Noorden, R. The ethical questions that haunt facial-recognition research. *Nature* **2020**, *587*, 354–359. [CrossRef] [PubMed]
5. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
6. Hoffer, E.; Ailon, N. Deep metric learning using triplet network. In Proceedings of the International Workshop on Similarity-based Pattern Recognition, Copenhagen, Denmark, 12–14 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 84–92.
7. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. Sphereface: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 212–220.

8. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
9. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 499–515.
10. Kaur, P.; Krishan, K.; Sharma, S.K.; Kanchan, T. Facial-recognition algorithms: A literature review. *Med. Sci. Law* **2020**, *60*, 131–139. [CrossRef] [PubMed]
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
12. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
15. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
16. 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 87–102.
17. 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, 12–18 October 2008.
18. 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/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4873–4882.
19. Qiu, H.; Gong, D.; Li, Z.; Liu, W.; Tao, D. End2End occluded face recognition by masking corrupted features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6939–6952. [CrossRef] [PubMed]
20. 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, Republic of Korea, 27 October–2 November 2019; pp. 773–782.
21. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics* **2020**, *9*, 1188. [CrossRef]
22. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
23. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
24. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
25. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 539–546.
26. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
27. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-margin softmax loss for convolutional neural networks. *arXiv* **2016**, arXiv:1612.02295.
28. Duan, Y.; Lu, J.; Zhou, J. Uniformface: Learning deep equidistributed representation for face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3415–3424.
29. Kim, M.; Jain, A.K.; Liu, X. AdaFace: Quality Adaptive Margin for Face Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18750–18759.
30. Kaur, G.; Sinha, R.; Tiwari, P.K.; Yadav, S.K.; Pandey, P.; Raj, R.; Vashisth, A.; Rakhra, M. Face mask recognition system using CNN model. *Neurosci. Inform.* **2021**, *2*, 100035. [CrossRef]
31. Talahua, J.S.; Buele, J.; Calvopiña, P.; Varela-Aldás, J. Facial recognition system for people with and without face mask in times of the covid-19 pandemic. *Sustainability* **2021**, *13*, 6900. [CrossRef]
32. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
33. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. Supplementary material for 'ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13–19.

34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv* **2014**, arXiv:1411.7923.
36. 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/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
37. Duta, I.C.; Liu, L.; Zhu, F.; Shao, L. Improved residual networks for image and video recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9415–9422.
38. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
39. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv* **2021**, arXiv:2107.00782.
40. 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.