ResSaNet: A Hybrid Backbone of Residual Block and Self-Attention Module for Masked Face Recognition

Wei-Yi Chang, Ming-Ying Tsai, and Shih-Chieh Lo
Fujian KuKe3D Technology Co., Ltd.
{weiyi.chang, sam.tsai, roger.lo}@kuke3d.com

Abstract

*In recent years, the performances of face recognition have been improved significantly by using convolution neural networks (CNN) as the feature extractors. On the other hands, in order to avoid spreading COVID-19 virus, people would wear mask even when they want to pass the face recognition system. Thus, it is necessary to improve the performance of masked face recognition so that users could utilize face recognition methods more easily. In this paper, we propose a feature extraction backbone named ResSaNet that integrates CNN (especially **Res**idual block) and **S**elf-**a**ttention module into the same network. By capturing the local and global information of face area simultaneously, our proposed ResSaNet could achieve promising results on both masked and non-masked testing data.*

1. Introduction

With the rapid growth of deep learning techniques, and the increasing of large scale training data [11, 1, 51], the performance of face recognition has been improved a lot in recent years [33, 34, 39]. Thus, face recognition algorithms have been deployed for various applications, e.g., access control, border control, and payment systems. However, there are various factors that would affect the performance of face recognition, for example, large-pose variation, unsuitable illumination, and wearing mask. How to alleviate these influences has become important. For example, in order to prevent the spread of COVID-19 virus, almost everyone wears a facial mask in their daily life. Thus, it is necessary for face recognition model to overcome the effect of facial mask. If the masked face could not be recognized accurately, it would be inconvenient for users since they need to take off their mask, but this would increase the risk of infection. To deal with this problem, Deng *et al.* [4] organize the Masked Face Recognition (MFR) challenge to benchmark deep face recognition methods. In this paper, we will describe the details of our solution for the InsightFace Track in this challenge.

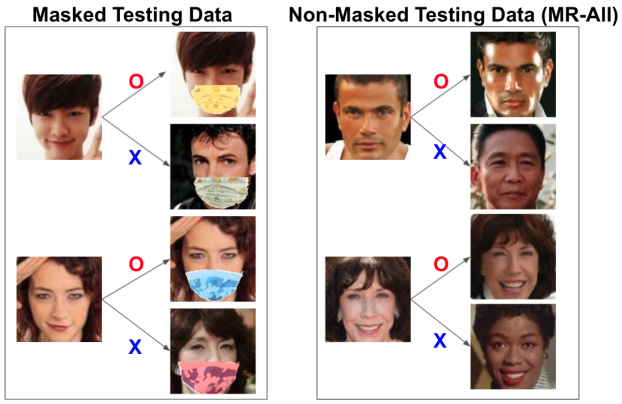

Figure 1. The example scenarios of testing data in MFR challenge.

To overcome the problem of occlusion or masked face recognition, researchers have proposed various solutions from different viewpoints [47, 8]. For example, from the viewpoint of recovering facial image, Li *et al.* [18] utilized a generative adversarial network (GAN) to perform face completion so that the content under the mask could be recovered, and then the distillation module was employed to obtain more realistic faces for face recognition. From the viewpoint of occlusion robust feature extraction, Song *et al.* [31] introduced a mask learning strategy to find and discard the feature elements that have been corrupted by occlusions. From the viewpoint of loss function, Montero *et al.* [22] proposed a loss function named MTArcFace that combined the ArcFace loss [5] and mask-usage classification loss so that it could force the network to learn when a face is wearing a mask. In this paper, since GAN-based recovering methods are hard to reproduce the results, and we could only submit a model for feature extraction. Thus, we prefer to design a robust and powerful backbone, and then train it with both masked and non-masked images to handle these two kinds of evaluation protocols in MFR challenge, so that it could achieve better performance in both testing sets. The example scenarios of testing data are illustrated

in Figure 1. In masked testing set, we need to verify that the non-mask and mask pairs are the same person or not. In non-masked testing set (MR-All), the testing pairs are facial images without masks.

In recent years, several CNN architectures were introduced to improve the performance on the ImageNet [26] challenge, e.g., VGGNet [30], ResNet [13], EfficientNet [35], and RegNet [25], and some of them also have been applied to the task of face recognition [39]. On the other hands, the Transformer architectures [37] have become more and more popular in both natural language processing and vision tasks, and they also have shown the competitive results with respect to CNN structures in various vision tasks [46, 20]. In this paper, inspired by non-local network [41] and BoTNet [32], we propose a backbone structure named **ResSaNet** that integrates convolution blocks (especially **Res**idual block) and **S**elf-**a**ttention module for face recognition. Moreover, we also add some useful blocks such as SE block [14] and FReLU [21] activation function, so that the recognition rate could be improved. As to the effect of facial mask, we also do data augmentation by synthesizing the masked facial images in original training data (MS1M-RetinaFace [6]). With aforementioned techniques, we will show that our proposed model could achieve better results in both masked and non-masked testing set with respect to the baseline model (ResNet100). Moreover, in the section of experiment, we will also show the effectiveness of each new block that we add, so that these results could be viewed as the reference for designing a stronger one.

2. Related Work

2.1. Loss function for face recognition

As to the backbone model for face recognition, it is slightly different from the traditional image classification task (i.e., face recognition model is served as a feature extractor). To train this model, we usually add a fully-connected (FC) layer (as a classifier, and the output of channel is the number of identity) after the backbone, and then calculate loss based on the output of FC. For the common loss function of image classification (e.g., softmax loss), it would be hard to train this model since the number of class is so large (equal to the number of identity). Thus, how to design a more suitable loss function has attracted many attentions for researchers in computer vision community. In order to increase the power of discrimination, marginal softmax loss functions (e.g., CosFace [38] and ArcFace [5]) has been proposed to minimize the intra-class variance and maximize the inter-class variance. By training with large-scale data, these methods could outperform the models trained by softmax loss.

In addition to margin-based loss function, Huang *et al.* [16] proposed an adaptive curriculum learning loss which would mainly address easy samples in the early training stage and hard ones in the later stage. By incorporating the idea of curriculum learning into face recognition, the CurricularFace loss could obtain competitive results with respect to margin-based loss function. In this paper, since the performance of margin-based loss functions are good enough for masked face recognition, we directly apply them to train our model.

2.2. Transformer structure

Self-attention models like Transformers [37] have shown the excellent performance in natural language processing. Recently, many works have been proposed that explored Transformers to solve various tasks in computer vision. For example, Dosovitskiy *et al.* [7] proposed the Vision Transformer (ViT) structure that could achieve reasonable performance on ImageNet. However, one drawback of ViT is that requires large-scale datasets such as ImageNet-21k and JFT-300M (which is a private dataset) to obtain the pre-trained model. In order to overcome this limitation, Yuan *et al.* [45] introduced a layer-wise Tokens-To-Token transformation to progressively structurize the image to tokens and model the local structure information. Moreover, they also designed a T2T-ViT backbone with a deep-narrow architecture. In addition to use Transformers for image classification, Liu *et al.* [20] presented Swin Transformer that could achieve state-of-the-art in various vision tasks such as image classification, object detection, and semantic segmentation.

Different from using pure attention models, some researchers proposed hybrid methods that utilize both convolutions and self-attention module in the same architecture. For example, Yuan *et al.* [44] proposed convolution-enhanced image Transformer (CeiT) that utilized CNN to extract low-level features, and then employed Transformers to establish long-range dependencies. Srinivas *et al.* [32] presented BoTNet that integrated self-attention module into ResNet, and this structure could achieve better results with respect to ResNet in the task of image classification and object detection. Similarly, Dai *et al.* [3] proposed a simple yet effective network structure named CoAtNet which is composed of MBConv block [27] and the Transformer block. Different from BoTNet, CoAtNet employed MBConv block as main component rather than residual block, and the position of Transformer blocks were placed on the last two stages rather than the last stage. By using this design, CoAtNet could enjoy both good generalization like ConvNets and superior model capacity like Transformers. Moreover, Guo *et al.* [10] introduced a novel CMT (CNNs meet Transformers) block, and Wu *et al.* [43] presented a new architecture named CvT (Convolutional vision Transformer) to integrate convolution layers and Transformers into a same block. Similar to ResNet, the CMT and CvT architecture have multiple stages to generate different sizes of feature

maps, while each stage is composed of CMT/CvT block. One difference between CMT and CvT is that the former has four stages while the latter only has three stages. In this paper, since the ResNet structures were popular for face recognition, and have shown the excellent performance, so that we choose it as baseline, and try to integrate different structures such as self-attention into it.

2.3. Occluded face recognition

Due to the information loss on the face area, if we do not utilize some techniques to alleviate this influence, there would be a degradation on the performance of face recognition. In general, there are three kinds of method proposed to overcome this problem [47]: occlusion robust feature extraction method, occlusion aware face recognition, and occlusion recovery based face recognition. As to the first one, for example, Triguerosa *et al.* [36] proposed a method to find out the important parts of face, and train face recognition model with the proposed batch triplet loss. On the other hands, occlusion aware based methods only utilized visible facial parts for recognition. For instance, Liao *et al.* [19] developed and alignment-free face representation based on Multi-Keypoint Descriptors (MKD) to describe the holistic or partial face, so that it could handle the problem of partial faces. Moreover, Weng *et al.* [42] introduced an approach based on feature set matching to solve the problem of partial face recognition, while the geometric features and textural features were considered for simultaneous matching. As to occlusion recovery based method, Zhao *et al.* [48] proposed a robust LSTM-Autoencoders model which consists of two LSTM components: occlusion-robust face encoding and recurrent occlusion removal, and these two networks collaborate with each other to localize and remove the facial occlusion. Moreover, Ge *et al.* [9] introduced the identity-diversity inpainting method by integrating GAN with a pre-trained CNN face recognizer so that the output image of generator would have the similar representation in identity space. In this paper, occlusion recovery based methods are not suitable for the MFR challenge since we could only submit one model to the evaluation system. Thus, we develop a structure that integrates convolution blocks and self-attention module into the same model, so that it could enjoy the benefit of occlusion robust based approach and occlusion aware based approach.

3. Our Proposed Method

In this section, we will introduce the detailed structure of our proposed ResSaNet which is based on the IResNet in InsightFace Project [1], Bottleneck Transformer (BoT) block [32], Squeeze-and-Excitation (SE) block [14], and FReLU activation [21]. The main differences between IResNet and ResSaNet are illustrated in Table 1, and we would introduce these structures in the following sections.

[1] https://github.com/deepinsight/insightface/

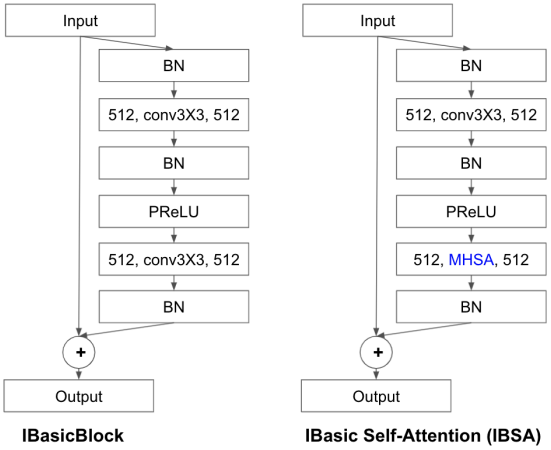

Figure 2. The example structure of the original IBasic Block in the IResNet of InsightFace project (left), and our proposed IBasic Self-Attention (right).

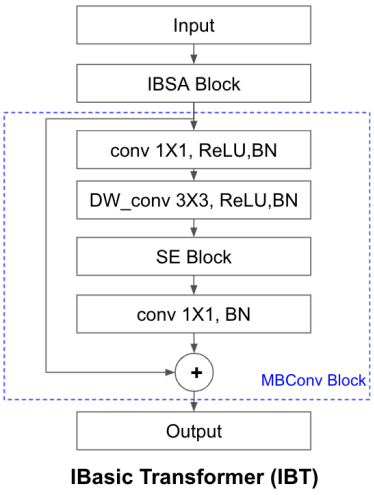

Figure 3. The structure of proposed IBasic Transformer (IBT).

3.1. IBasic Self-Attention (IBSA) and IBasic Transformer (IBT)

In recent years, Transformer models have shown their excellent performances [45, 20, 46] in vision tasks such as image classification, object detection, and image segmentation. However, if we directly apply these models as the backbone for face recognition, we could not ensure that the same structure would achieve suitable results with respect to CNNs due to the smaller input size of face recognition model. For example, the general input size for image classification is 224×224 or bigger, and some Transformers are designed based on this input size. As to the input size for face recognition models, the general size is 112×112 which is smaller. Thus, maybe we have to do some modifications so that we could directly train Transformer models

Table 1. The architectures of IResNet series (50 and 100), and our proposed ResSaNet series (ResSaNet-50 and ResSaNet-100) . In IResNet, the details of IBasic block are shown in Figure 2. For our ResSaNet, the structure of IBT block is depicted in Figure 3, and SE-IBasic block is demonstrated in Figure 4. As to SE-IBasic_F block, the PReLU activation function in SE-IBasic block is replaced by FReLU, and the structure of FReLU is depicted in Figure 5. As to the inference time, the results are evaluated on Tesla V100 GPU.

stage	output (size, #channel)	IResNet-50	ResSaNet-50	IResNet-100	ResSaNet-100
conv1	$112 \times 112, 64$	Conv 3×3	Conv 3×3	Conv 3×3	Conv 3×3
stage1	$56 \times 56, 64$	IBasic $\times 3$	IBasic $\times 3$	IBasic $\times 3$	IBasic $\times 3$
stage2	$28 \times 28, 128$	IBasic $\times 4$	SE-IBasic_F $\times 4$	IBasic $\times 13$	SE-IBasic_F $\times 13$
stage3	$14 \times 14, 256$	IBasic $\times 14$	SE-IBasic_F $\times 14$	IBasic $\times 30$	SE-IBasic_F $\times 30$
stage4	$7 \times 7, 512$	IBasic $\times 3$	IBT $\times 3$	IBasic $\times 3$	IBT $\times 3$
FC	$1 \times 1, 512$				
#Params		43.57×10^6	44.13×10^6	65.16×10^6	66.35×10^6
#FLOPs		6.31×10^9	6.29×10^9	12.12×10^9	12.11×10^9
Inference Time (ms)		4.51	5.18	7.03	9.69

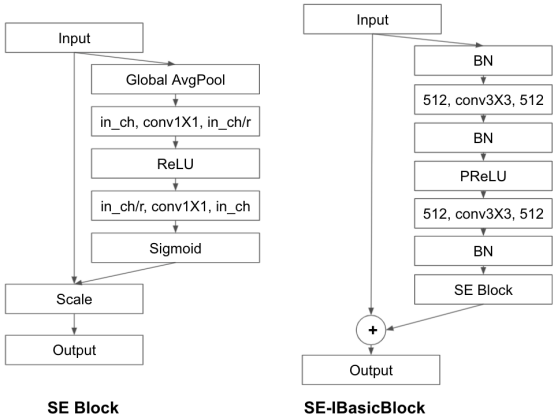

Figure 4. The example structure of SE block (left) and the integration of IBasic block and SE block: SE-IBasic block (right).

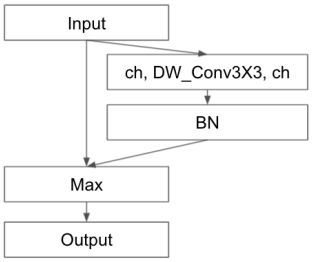

Figure 5. The structure of FReLU.

for face recognition. On the other hands, since CNN structures have demonstrated excellent results for face recognition, one thing we are curious to know : if the performance of face recognition could be improved by integrating the self-attention module into convolution blocks ? Since convolution blocks could extract the local information, and self-attention module could capture long-term dependency, integrating both structures into the same backbone could alleviate the weakness of each other. Moreover, the long-term dependency is also important for face recognition because we could not recognize one person accurately only by seeing the partial face (e.g., only using eyes or nose).

Inspired by non-local network [41] and BoTNet [32], we modify the IBasic block in IResNet by replacing one 3×3 convolution layer with Multi-Head Self-Attention (MHSA) layer, while batch normalization [17] layer and PReLU [12] layer are preserved. Figure 2 shows the difference between the original IBasic block and our IBasic Self-Attention (IBSA) block. In MHSA layer, the attention logits are $qk^\top + qr^\top$ where q, k, r represent query, key and position encodings respectively. Here, we follow the experiments in BoTNet [32], and utilize the relative position encodings [29] in order to achieve better performance.

Moreover, since the BoT block originally was designed for bottleneck block which is slightly different from the IBasic block in IResNet. The order of convolution layers in bottleneck block is 1×1, 3×3, and 1×1, but there are only two 3×3 convolution layers in IBasic block. If we follow the design of BoTNet to replace the second 3×3 convolution layer with MHSA, there would not exist feed-forward network (FFN) after MHSA. Thus, similar to Transformer models like ViT [7] and CMT [10], we employ a MBConv block [27] as FFN behind our proposed IBSA block, and the detailed structure of our proposed IBasic Transformer (IBT) is depicted in Figure 3.

3.2. Squeeze-and-Excitation (SE) Block and FReLU Activation

Since we select ResNet as the base model, there are several simple yet effective structures that could be integrated to the base model. For example, SENet [14] has demonstrated the better performance with respect to ResNet by just adding the SE blocks. Thus, we also employ this channel attention to perform feature recalibration so that the informative features could be emphasized, and the worse ones could be suppressed. The detailed structures of SE block and SE-IBasic block are illustrated in Figure 4, where the

Table 2. The experimental results (measured by TAR) on mask and non-masked (MR-All) testing data with different settings of data and backbone structures.

Base Structure	Add Masked Data	Use IBSA	Use FReLU	Add SE block	Mask Dataset	MR-All Dataset
R50					63.850	80.533
R50		√			66.700	82.401
R50		√	√		66.370	82.669
R50	√	√	√		70.699	83.095
R50	√	√	√	√	**70.994**	**83.488**
R100					69.091	84.312
R100	√	√	√	√	**77.649**	**88.093**

Figure 6. Examples of our synthetic mask data.

Table 3. The results of different positions (i.e., different stages in ResNet) for SE block, while the training data is with masked faces.

Base Structure	SE Position	Mask	MR-All
R50_FReLU_IBSA	None	70.699	83.095
R50_FReLU_IBSA	[2,3]	70.994	**83.488**
R50_FReLU_IBSA	[1,2,3]	**71.374**	83.428
R50_FReLU_IBSA	[2,3,4]	69.838	83.279

Table 4. The results of different positions for IBSA, while the training data is with masked faces.

Base Structure	IBSA Setting	Mask	MR-All
R50_FReLU	stage4 ×3	70.699	**83.095**
R50_FReLU	stage4 ×6	69.335	80.389
R50_FReLU	stage3 ×2 stage4 ×3	**70.915**	82.812

scale function in SE block is utilized to do channel-wise multiplication, and r is the reduction ratio.

In addition to SE block, we also notice that FReLU [21] could outperform the popular activation functions such as ReLU [24] and PReLU [12] in the task of classification, detection, and segmentation. Thus, we replace the PReLU with FReLU in SE-IBasic block or IBasic block. Moreover, according to the experiments in FReLU [21], we do not do this replacement for all stages. The structure of FReLU is shown in Figure 5, and the main idea is using an additional depth-wise convolution layer to increase the pixel-wise modeling capacity.

4. Experiments

4.1. Dataset and Settings

For the Masked Face Recognition Challenge [4], the training data and testing data are restricted by the organizers. In InsightFace track, the training data could be MS1M-RetinaFace [6] or Glint360k [1], and the testing data is a private dataset that was collected by the organizers. For training data, we utilize the tool of MaskTheFace[2] to wear masks for the images in MS1M-RetinaFace dataset, and then we merge these images into the original dataset to generate our training data. The ratio of synthetic masked face is about 5 % of total image. Figure 6 shows the examples of these masked images. Note that all of the models in following sections are trained by our synthetic data or the original one (no mask). For testing data, the detailed rules and statistics are shown in the website of this track of challenge[3]. In a short summary, the evaluation metric would be measured on two types of testing data: Mask and Multi-racial (MR-All), and number of identity in Mask data is 6,964, while there are 6,964 masked images and 13,928 non-masked images. In MR-All dataset, there are four racial sets: African, Caucasian, Indian, and Asian, while it contains about 0.24 million identities and 1.62 million images. For evaluation, the metric is TAR (True Accept Rate) on all-to-all 1:1 protocol, with FAR (False Accept Rate) less than different thresholds ($1e^{-4}$ for Mask dataset, and $1e^{-6}$ for MR-All dataset).

As to the detailed settings for model training, the configuration is based on the project of Arcface_torch[4]: the initial learning rate is 0.05, and the optimizer is SGD. Besides, we

[2] https://github.com/aqeelanwar/MaskTheFace

[3] https://github.com/deepinsight/insightface/tree/master/challenges/iccv21-mfr

[4] https://github.com/deepinsight/insightface/tree/master/recognition/arcface_torch/

Table 5. The results (measured by TAR) of baseline models (R50 and R100) and our proposed models: ResSaNet-IBSA series and ResSaNet-IBT series on multiple racial and masked face datasets in MFR challenge.

Model \ Data Set	Mask	Children	Affrican	Caucasian	South Asian	East Asian	MR-All
R50	63.850	60.457	75.48	86.115	84.305	57.352	80.533
ResSaNet-IBSA-50	70.994	64.785	79.025	87.499	86.141	62.552	83.488
ResSaNet-IBT-50	71.877	66.809	80.061	87.628	87.671	62.090	83.621
R100	69.091	66.864	81.083	89.040	88.082	62.193	84.321
ResSaNet-IBSA-100	77.649	71.408	85.462	91.388	90.953	68.839	88.093
ResSaNet-IBT-100	**78.123**	**72.833**	**85.942**	**92.099**	**91.151**	**69.273**	**88.333**

also utilize the method of partial_fc [1] to accelerate the time for model training, while the loss function is Cosface [38] ($s = 64$, $m = 0.4$). As to the settings for backbone, the reduction ratio for SE block is 0.25, and the number of head for MHSA is 4.

4.2. Performance of different types of blocks and the effectiveness of masked data

In this section, we show the performance of different types of blocks (e.g., IBSA, FReLU, and SE) in Table 2. Firstly, the results of baseline model of ResNet-50 (R50) and ResNet-100 (R100) come from the website of challenge, and our goal is to achieve better results w.r.t the baselines on the testing data of Mask and MR-All. Secondly, with the same training data of MS1M-RetinaFace, R50 structure would have an obvious improvement by using IBSA blocks, this shows that self-attention blocks are useful for face recognition. Then, based on this structure, the replacement of FReLU would have slight improvement in MR-All dataset. Although FReLU could not obtain better results on Mask dataset, we still adopt this structure since the improvement in MR-All dataset is also important.

Once the structure is almost fixed (IBSA and FReLU), we add the synthetic masked images for training, so that the results on Mask testing data could be improved. Moreover, we also add SE block to emphasize the more important feature maps, and we show that by adding this channel attention, the performance on face recognition could be enhanced. Finally, because training one ResNet-100 model would take too much time, in order to obtain the experimental results faster, we put these structures on ResNet-50 and do comparison with respect to ResNet50. Once we have confirmed that all structures are useful, we directly adopt these structures on ResNet-100, and as shown in Table 2, our structure could also outperform the ResNet-100 (R100) baseline on both Mask and MR-All testing data.

4.2.1 Performance of different settings for SE block

As to SE block, the positions in different stages are also important since they would influence not only the accuracy but also the inference time. As shown in Table 3, we notice that adding SE blocks into the stage two and three in ResNet would have an obvious improvement in both Mask and MR-All testing set. Meanwhile, the inference time would also be increased from 4.2 ms to 4.8 ms (evaluated on Tesla V100 GPU). Moreover, we also try to integrate SE blocks into stage one or four of ResSaNet, but both settings do not have significant improvement. Maybe this is due to that there are only three blocks in stage one and four so that the effectiveness of adding SE blocks is not so obvious. Based on the experimental results on ResSaNet-50, we only add SE blocks into stage two and three.

4.2.2 Performance of different settings for IBSA

Although IBSA block could improve the performance of face recognition by replacing the IBasic blocks in the last stage of ResNet, we also have done some experiments by changing the positions of IBSA blocks to investigate the effectiveness of different settings. As shown in Table 4, we add more IBSA blocks into stage four of the backbone, but it could not get better result. Besides, we also include IBSA blocks in stage three (by replacing the last two blocks), but we could not obtain better result in MR-All. In a short summary, the suitable position of IBSA is in stage four.

After deciding the position of SE block and IBSA block, we also investigate the different structures in IBSA block: without FFN and with FFN (denoted as IBT). Table 5 shows the TAR of ResNet, ResSaNet-IBSA, and ResSaNet-IBT on Mask dataset, MR-All dataset, and the detailed results for each racial dataset. Comparing ResSaNet-IBSA with ResNet, our proposed ResSaNet-IBSA could outperform ResNet in the structures of 50 layers and 100 layers. Moreover, we could also notice that by adding a FFN after IBSA block, the performances of ResSaNet could be improved no matter the number of layer is 50 or 100. Thus, the robustness of ResSaNet could be shown according to these results.

4.3. Results on popular face recognition benchmark

In addition to the testing data in MFR challenge, we also show the results of our proposed method on several bench-

Table 6. The results (measured by accuracy) of our proposed models ResSaNet-50 and ResSaNet-100 on different popular datasets. Note that both the models of R50 and R100 come from the InsightFace project, and IR-152 model comes from face.evoLVe project [40].

Data Set \ Model	R50	ResSaNet-50	R100	ResSaNet-100	IR-152
LFW	0.9980	0.9978	**0.9982**	**0.9982**	**0.9982**
CFP_FF	0.9970	0.9976	0.9980	0.9979	**0.9983**
CFP_FP	0.9780	0.9821	0.9854	**0.9857**	0.9837
AgeDB	0.9815	0.9777	**0.9827**	0.9807	0.9807
CALFW	0.9593	0.9602	0.9593	**0.9607**	0.9603
CPLFW	0.9158	0.9230	0.9283	0.9297	**0.9305**
VGG2_FP	0.9504	0.9516	0.9544	**0.9568**	0.9550

marks of face recognition, e.g., LFW [15], CFP_FF [28], CFP_FP [28], AgeDB [23], CALFW [50], CPLFW [49], and VGG2_FP [2], while the training data for baseline models (R50 and R100) is the MS1M-RetinaFace, and our proposed models (ResSaNet-50 and ResSa-Net-100) utilize the masked data for model training (the total number of identity is equal to the MS1M-RetinaFace dataset). Moreover, we also compare our model with the larger model: IR152 (which comes from the project of face.evoLVe [40] while it was trained by MS1M dataset [11]. As shown in Table 6, firstly, the trends of R50 and R100 are similar. In other words, the performance of R100 series models are better than the results of R50 series models. Secondly, our proposed ResSaNet models could achieve better performance with respect to the counterpart in most of the testing data. Thirdly, although the comparison between ResSaNet-100 and IR-152 is not so fair (the training data and the inference time are not very closed), our ResSaNet-100 still could obtain promising results in these testing data.

5. Conclusion and Future Work

In this paper, we propose a backbone named ResSaNet by integrating residual blocks and self-attention module for face recognition. By training with both masked facial images and non-masked facial images, our backbone has an obvious improvement with respect to the baseline model in both mask and non-masked testing data. In experiments, we also investigate the usage of SE blocks and self-attention module by changing the positions of them. Besides, we also use public benchmark to evaluate the performance of our proposed ResSaNet, and we demonstrate that ResSaNet could achieve promising results on several datasets.

For future work, we would like to train our ResSaNet with the larger dataset (e.g., Glint360k and Webface260M) to evaluate the performance. Besides, we would like to design new loss function for the masked data, and also investigate the effectiveness of different margin-based loss functions. Moreover, there are more and more great works related to Transformer, and they have shown the advantage for several vision tasks. Thus, we also plan to integrate these structures into CNNs to check the performances.

References

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Ying Fu. Partial FC: Training 10 million identities on a single machine. *arXiv: 2010.05222*, 2020. 1, 5, 6

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *IEEE FG*, 2018. 7

[3] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying convolution and attention for all data sizes. *arXiv: 2106.04803*, 2021. 2

[4] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The InsightFace track report. *IEEE ICCV Workshop*, 2021. 1, 5

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE CVPR*, 2019. 1, 2

[6] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv: 1905.00641*, 2019. 2, 5

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 4

[8] Fadi Boutros et al. Mfr 2021: Masked face recognition competition. *arXiv: 2106.15288*, 2021. 1

[9] Shiming Ge, Chenyu Li, Shengwei Zhao, and Dan Zeng. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE TCSVT*, 2020. 3

[10] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. CMT: Convolutional neural networks meet vision transformers. *arXiv: 2107.06263*, 2021. 2, 4

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-Celeb-1M: A dataset and benchmark for large-scale face recognition. *ECCV*, 2016. 1, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE ICCV*, 2015. 4, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE CVPR*, 2016. 2

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation networks. *IEEE CVPR*, 2018. 2, 3, 4

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Technical Report 07-49, University of Massachusetts, Amherst*, 2007. 7

[16] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. *IEEE CVPR*, 2020. 2

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015. 4

[18] Chenyu Li, Shiming Ge, Daichi Zhang, and Jia Li. Look through masks: Towards masked face recognition with de-occlusion distillation. *ACM Multimedia*, 2020. 1

[19] Shengcai Liao, Anil K. Jain, and Stan Z. Li. Partial face recognition: Alignment-free approach. *IEEE TPAMI*, 2013. 3

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv: 2103:14030*, 2021. 2, 3

[21] Ningning Ma, Xiangyu Zhang, and Jian Sun. Funnel activation for visual recognition. *ECCV*, 2020. 2, 3, 5

[22] David Montero, Marcos Nieto, Peter Leskovsky, and Naiara Aginako. Boosting masked face recognition with multi-task arcface. *arXiv: 2104.09874*, 2021. 1

[23] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. *IEEE CVPR Workshop*, 2017. 7

[24] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. *ICML*, 2010. 5

[25] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. *IEEE CVPR*, 2020. 2

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 2

[27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *IEEE CVPR*, 2018. 2, 4

[28] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. *WACV*, 2016. 7

[29] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv: 1803.02155*, 2018. 4

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 2

[31] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. *IEEE ICCV*, 2019. 1

[32] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *IEEE CVPR*, 2021. 2, 3, 4

[33] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *NIPS*, 2014. 1

[34] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. *IEEE CVPR*, 2014. 1

[35] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, 2019. 2

[36] Daniel Saez Trigueros, Li Meng, and Margaret Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 2018. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017. 2

[38] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *IEEE CVPR*, 2018. 2, 6

[39] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 2018. 1, 2

[40] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face. evoLVe: A high-performance face recognition library. *arXiv: 2107.08621*, 2021. 7

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *IEEE CVPR*, 2018. 2, 4

[42] Renliang Weng, Jiwen Lu, , and Yap-Peng Tan. Robust point set matching for partial face recognition. *IEEE TIP*, 2016. 3

[43] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. *arXiv: 2103.15808*, 2021. 2

[44] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. *arXiv: 2103.11816*, 2021. 2

[45] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv: 2101.11986*, 2021. 2, 3

[46] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *arXiv: 2106.13112*, 2021. 2, 3

[47] Dan Zeng, Raymond Veldhuis, and Luuk Spreeuwers. A survey of face recognition techniques under occlusion. *arXiv: 2006.11366*, 2020. 1, 3

[48] Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE TIP*, 2018. 3

[49] Tianyue Zheng and Weihong Deng. Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Technical Report 18-01*, 2018. 7

[50] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv: 1708.08197*, 2017. 7

[51] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. *IEEE CVPR*, 2021. 1