Displays 74 (2022) 102245

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Facial expression recognition in facial occlusion scenarios: A path selection multi-network☆

Liheng Ruan [a], Yuexing Han [a,b,*], Jiarui Sun [a], Qiaochuan Chen [a], Jiaqi Li [a]

[a] School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai 200444, People's Republic of China
[b] Zhejiang Laboratory, Hangzhou 311100, People's Republic of China

ARTICLE INFO

Keywords:
Facial occlusion
FER
CNN

ABSTRACT

The occlusion scenarios including wearing mask, wearing sunglasses, wearing hat, etc. Thus, parts of face are occluded, e.g. nose and mouth are obscured if a mask is worn. Facial expression recognition (FER) tasks have been widely researched. However, less attention has been paid to FER in occlusion scenarios, which are not uncommon in the real world. In this paper, we propose a method that structures a path selection multi-network model to achieve the FER in the above three types of facial occlusion scenarios. The method contains two parts. For the multi-network, we segment the labels in one database which results in three new sub-databases to train three Subnets, respectively. For the path selection, which is an integration method of multi-network, we merge groups of labels in one database to train an initial network called BeginNet. The prediction of BeginNet selects one of the Subnets to make the final prediction. We concatenate four popular databases Fer2013, JAFFE, KDEF and Raf-DB databases into one larger database, and use the combined database to verify our method effectiveness. The experimental results show that our method has better results in coping with the expression recognition task in multiple types of facial occlusion scenarios.

1. Introduction

Facial expressions are crucial for human communication, which can convey some emotions and thoughts that may be difficult to express verbally through visual information [1]. In 1968, psychologist Mehrabian found that the amount of information that can be conveyed through the human face during communication occupies 55% of the total information [2], so it is possible to obtain useful information through facial expression recognition(FER). FER has practical applications in many scenarios. In smart medical care, by collecting facial information from patients areas and capturing abnormal expressions in a timely manner, the patient's current physical condition can be accurately and efficiently evaluated, and feedback can be submitted promptly and rapidly for specific abnormalities [3]. In the field of intelligent security, the emotional state of pedestrians can be collected and warned of potential risks such as the presence of suspicious people. Thus, protection can be done in advance to avoid unnecessary losses and dangers [4].

Deep convolutional neural networks (DCNNs) have recently achieved remarkable results on unoccluded FER tasks, but a variety of occlusion exist in real-world. For instance, during corona virus disease(COVID-19) outbreak, almost everyone wears masks in crowded places [5], which completely cover two important facial features i.e., the mouth and nose. The accuracy of existing methods for recognizing facial expressions in unoccluded or small-area occlusion scenarios are seriously affected by the occlusion. Therefore, it is of practical significance to study the expression recognition technology in large area facial occlusion scenes.

The existing methods for recognizing facial expressions in unoccluded scenarios have difficulty in accurately recognizing in large-area facial occlusion scenarios mainly because of the following two situations.

The first one is that the features extracted by the FER models trained with the unoccluded facial expression databases(standard databases) cannot be directly applied to the occlusion scenarios in most cases, because the part of the region where the important features are extracted may be masked due to various circumstances [6]. The methods can extract a large number of invalid features to ineffectively do FER in the occlusion scenarios.

☆ This paper was recommended for publication by Guangtao Zhai.

* Corresponding author at: School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai 200444, People's Republic of China.

E-mail address: Han_yx@i.shu.edu.cn (Y. Han).

https://doi.org/10.1016/j.displa.2022.102245
Received 22 September 2021; Received in revised form 24 May 2022; Accepted 28 May 2022
Available online 7 June 2022

The other situation is that the occlusion of the face dramatically changes the visual appearance of the face, making it more difficult to extract features. Masking one or more of the five senses has a huge impact on expression recognition because it leads to high similarity between different expression classes [7]. Among the five senses, the eyes and mouth are the most important facial regions. Therefore, obscuring these two parts can have a significant impact on the most FER tasks [6].

To overcome the difficulties, we adopt the popular method of training models with masked processed database [8,9]. By masking, the recognition models do not pay attention to the masked regions and learn effective features from other unmasked regions instead. The main contributions in this work are described as follows: Unlike the traditional single-network model structure where a single network needs to learn all the features of expressions, we propose a multi-network model structure, where each network only needs to learn features of a small portion of expressions. We use this structure to reduce the difficulty of networks to learn expression features in occluded scenarios. And by extracting the nuances between a small portion of different expressions the FER is more accurately dealt. Finally, the expression recognition models are integrated by the expression features learned by each network for facial expression with different parts through path selection.

The rest of the paper is organized as follows: the method of FER in facial occlusion scenarios is explained in Section 3, which includes creation of masked databases in Section 3.1, the path selection multi-network architecture in Section 3.2. The experiments as well as their results in different occlusion scenarios are presented and discussed in Section 4. Finally, Section 5 concludes the research work.

2. Related work

The research on FER has become an independent research area. Most of the early traditional methods use hand-designed features or shallow learning, such as local binary patterns [10], local binary patterns of tri-orthogonal planes [11], non-negative matrix decomposition [12] and sparse learning [13]. Some researchers also proposed a mathematical model that can represent facial features based on the unique structure of the face [14]. Most of the current FER is based on the complete human face which can provide enough information to accurately recognize the face expression. However, in some situations, the whole face is difficult to be obtained, and partial facial occlusion is quite common in practical applications. In recent years, DCNNs have been dominating the field of face recognition, and a large number of researches has also focused on partial face occlusion scenarios.

One class of approaches reconstructs the occluded face and extracts the relevant features from the reconstructed face. For instance, Yang et al. [15] proposed an identity adaptation generation model with two parts. The first part used conditional generative adversarial networks to generate different expression images for each person separately. The second part performed FER for each person individually in turn. Zhao et al. [16] proposed an LSTM autoencoder that could effectively recover the occluded face and ensure the accuracy of face recognition for the recovered face. Pan et al. [17] described an adversarial learning method that first trained two CNNs on unoccluded and occluded images separately, and then used the unoccluded network to guide the learning of the occluded network. Dong et al. [18] presented a two-stage GAN structure, where the object causing the occlusion was resolved in the first stage and used to be input the second stage along with the original image, and finally an unobscured image was output.

The other class of approaches concentrate on extracting sufficient features from the remaining unoccluded regions [19,20]. An end-to-end trainable patch-gated CNN (PG-CNN) with an attention mechanism has been proposed [21], which divides the complete facial features into multiple small patches and calculates the corresponding weights of the patches separately. The occluded regions of the face are automatically sensed and the most distinguishable non-occluded regions are focused. Based on the idea, a patch-based attentional convolutional neural network pACNN was proposed [22]. Furthermore, a gACNN based on both local representations extracted from the patches and global representations extracted from the complete image was presented. Daniel et al. [23] proposed a data augmentation method that detected during the training process mask the face to enforce the model to extract features from the remaining regions. Moreover, a MaskNet module has been added into the existing CNN model, by ignoring the features that are distorted by occlusion and focus on learning high-fidelity image features [24]. Ding et al. [25] proposed a landmark-guided attention branch which was used to find and discard the corrupted features from occluded regions. So that they are not used for recognition. Moreover, an end-to-end architecture including a spatial channel attention network with a branch for complementing it has been proposed, which does not require the feature point information extracted by a landmark detector [26]. Since occlusion in images increases intra-class differences and decreases inter-class differences in expression classification, the second class of most of approaches addresses this problem by improving existing CNN models that seek to learn sufficient features for expression classification from the unoccluded parts of the occluded images.

Previous studies have shown that integrating multiple different networks can achieve better performance than a single network [27]. There are three types of integration of multiple networks: majority voting, simple averaging, and weight averaging. Majority voting takes the predicted labels of the sub-networks as their votes to select the label with the largest number of votes. Simple averaging selects the label with the highest average of posterior probabilities instead of label yielded from sub-networks [28]. Weight averaging is an improvement of simple averaging, which adds different weights to each sub-network.

Since weight averaging takes into account the difference in importance and confidence of different sub-networks, many studies adopt weight averaging methods to find the optimal set of weights for network integration. Kahou et al. [29] proposed a random search method to weight the prediction results for each class of expressions assigning different weights to each network. Kim et al. [28] proposed an exponentially weighted average based on validation accuracy to emphasize various sub-networks. On the other hand, Pons et al. [30] used convolutional networks to learn the weights of each sub-model autonomously.

Most of multi-network integration methods are applied to unoccluded FER tasks, and the sub-networks need to learn all the classification features before using an integrated approach. In this paper, we propose a method that integrate multi-network to expression recognition tasks in facial occlusion scenarios based on path selection, where each sub-network needs to learn only a part of the classification features.

3. Methodology

3.1. Creation of masked database

We mainly consider three types of facial occlusion scenarios, namely upper face occlusion as well as lower face occlusion and eye occlusion. Since it is difficult to find expression databases that publicly contain the above three types of occlusions, we select several existing standard expression recognition databases Fer2013 [31], JAFFE [32], KDEF [33], RAF-DB [34,35] and merge them into a larger database to increase the training volume, which occluded the upper and lower half of the face, the eyes of the images, respectively, as training, validating and testing databases. Among the databases, the KDEF database is only used to create the frontal face image database.

In order to simulate the masking of the upper and the lower half of the face, we use the simpler and more effective way of directly blacken the upper or lower half of all images, instead of using a geometric model as in [9] to precisely achieve the exact masking of

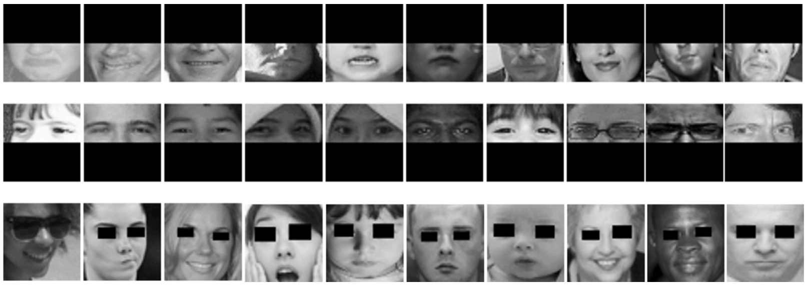


Fig. 1. The first line shows the simulation of the upper face mask; the second line shows the simulation of the lower face mask; the third shows the simulation of the eyes mask.

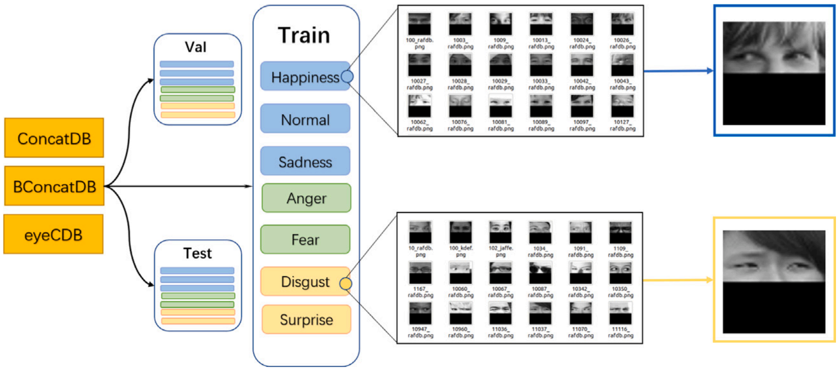


Fig. 2. Each processed database is divided into three parts: training set, validation set, and test set, which contains seven identical expression labels.

Table 1
Number of images with different expression labels for each database concatenated.

Database	Quantity							
	Happiness	Normal	Sadness	Anger	Fear	Disgust	Surprise	Total
Fer2013	8 989	6 198	6 077	4 953	5 121	547	4 002	35 887
JAFFE	31	30	31	30	32	29	30	213
KDEF	140	140	140	140	140	140	140	980
RAF-DB	5 957	3 204	2 460	867	355	877	1 619	15 339
ConcatDB BConcatDB eyeCDB	15 117	9 572	8 708	5 990	5 648	1 593	5 791	52 419

some specific areas of the face. To simulate the eye occlusion, we firstly search the location of both eyes using the classifier of the Haarcascade_eye.xml [36,37] and blacken the area of them. The simulated masking effect is shown in Fig. 1.

The new databases of the occluded images are called ConcatDB, BConcatDB and eyeCDB, respectively. The difference among the three databases is the occlusion areas. Thus, the methods of operations and descriptions are equally performed for the three types of databases. Each database includes the same numbers of facial expression including happiness, normal, sadness, anger, fear, disgust, surprise. The number of images contained in each expressions labels of the database is shown in Table 1. It can be observed that the number of images of Happiness is the most, followed by Normal and Sadness. Disgust is the least and the number of remaining emotional images is relatively average.

ConcatDB, BConcatDB and eyeCDB all consists of 52419 face expression images. 80% of the images from each expression label are taken as training set and 10% of the images are with validation set, and the last 10% are taken as test set. Each face is a grayscale image with a fixed size of 48 * 48, with the labels of expressions from 0 to 6, as follows: [0, Happiness], [1, Normal], [2, Sadness], [3, Anger], [4, Fear], [5, Disgust], [6, Surprise]. The overall picture of the databases is shown in Fig. 2.

3.2. Path selection multi-network

Before diving into the details, some notations are firstly introduced here. ConcatDB contains images with upper face occlusion, while BConcatDB includes images with lower face occlusion. EyeCDB consists of images with eye occlusion. BeginNet, SubnetX, SubnetY, SubnetZ are networks which receive image as input and output expression label. And Begin, SepA, SepB and SepC are datasets used for training networks mentioned above. We take BConcatDB as an example. Due to the large masking area, the features of the mouth and nose are almost completely unlearnable, to make the FER more difficult. Considering that in classification tasks, tasks with fewer classifications such as binary and tertiary classifications are considered simpler, we design some experiments and the results are shown in Section 4.

Inspired by the results, we propose a method to first extract some of the classifications from the original database and create three new databases, then assign the corresponding number of network models to be trained separately at the same time. Finally, we integrate the individual networks based on the path selection method, which can be used to cope with the FER tasks in facial occlusion scenarios. For example, an expression image corresponding to anger is input into BeginNet to get the initial prediction result B, and B corresponds to SubnetY, so this image is input into SubnetY again to get the final

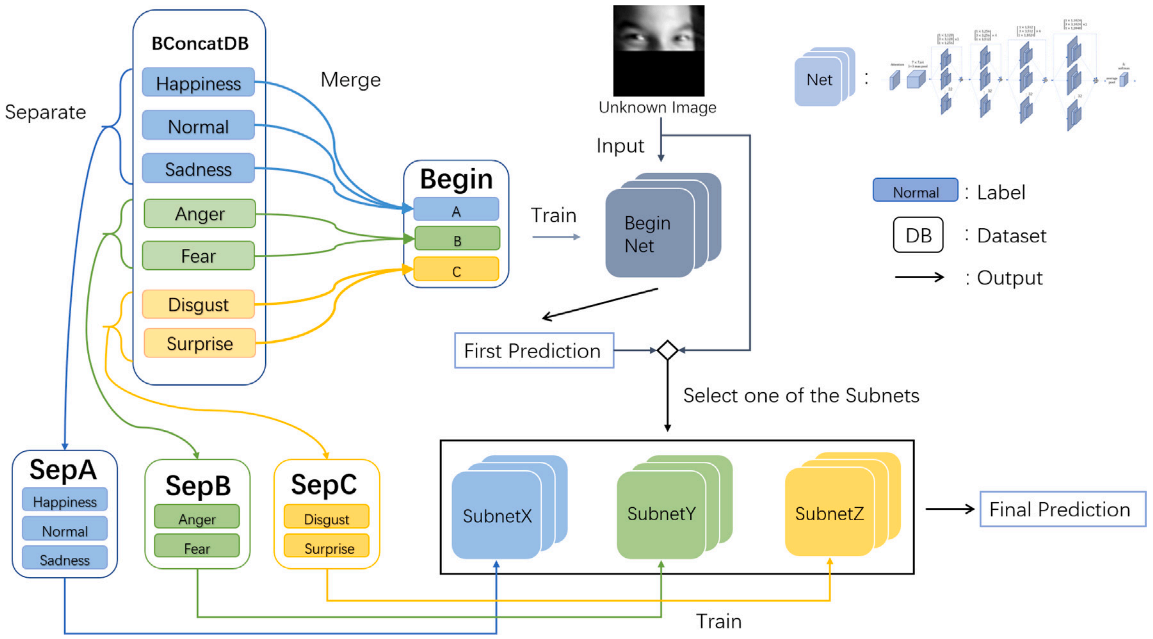


Fig. 3. The datasets generated after the segmentation and merging method are used to train BeginNet and three subnets. The BeginNet receives as input an unknown image then outputs the initial prediction, which selects one of the subnet to output the result. The selected subnet receives the same image.

predicted expression label: anger. The overall flowchart is shown in Fig. 3.

We segment the seven expression labels of the ConcatDB, BConcatDB, eyeCDB and their corresponding expression images: the three expression labels [“Happiness”, “Normal”, “Sadness”] are made into a separate sub-database called SepA. Similarly [“Anger”, “Fear”], [“Disgust”, “Surprise”] into separate sub-databases SepB and SepC, respectively. The three sub-databases are used to train the three models separately, i.e., SepA is used to train SubnetX for “Happiness”, “Normal”, “Sadness”, and SepB is used to train SubnetY for “Anger”, “Fear”, and SepC is used to train SubnetZ for “Disgust”, “Surprise” as shown in the left half of Fig. 3.

An unknown image is fed into the subnets, i.e., SubnetX, SubnetY and SubnetZ. The three Subnets are trained with the different sub-databases, i.e., SepA, SepB and SepC, which output one of the labels in the corresponding sub-database, respectively. The three expression labels obtained come from three sub-databases, respectively. For an input image including a face, only one Subnet can output a correct result. Thus, we need to select the correct label from the output of the subnets, which is a problem we need to consider.

After experimenting with several integration methods, the final results were poor, as described in Section 4. We propose a method to integrate multiple Subnets based on path selection. What path selection doing is to choose a path that is most likely to lead to the correct expression label. For doing this, we need to add another initial network, BeginNet, using its output to decide which Subnet the unknown image should be fed into for the next step of prediction. In order to train BeginNet, we need to merge some labels in BConcatDB as well as ConcatDB and eyeCDB. We merge all the images under the expression labels “Happiness”, “Normal” and “Sadness” into a new path label A. Similarly “Anger”, “Fear” are merged into path label B and “Disgust”, “Surprise” into path label C. Containing three path labels mentioned above, the database Begin is used to train the BeginNet. Path labels A, B, and C corresponds to SubnetX, SubnetY, and SubnetZ, respectively.

Our proposed method is divided into two steps. Firstly, the unknown image is input into BeginNet for the initial prediction. Based on the result of the initial prediction, i.e., one of the three path labels A, B and C, The correspondent subnet is selected to make the final prediction result. Two or three expressions are predicted with each subnet. For a total of seven expressions, we convert the seven classifications FER task in facial occlusion scenarios into binary and tertiary classifications tasks.

Table 2
Prediction accuracy of models under validation set with the merged ConcatDB, which contains three classifications instead of seven.

Model	Accuracy/%
Resnet18	63
Resnet101	73
Resnet152	75
wide_resnet101	64

We use ResNeXt [38] as the structure of both BeginNet and Subnets, and add a simple spatial attention module [39] before the ResNeXt structure to better extract local features. ResNeXt is an optimization based on ResNet, replacing the original 3-layer convolutional module of ResNet with a small parallel-stacked module. With the replacement, the parameters do not grow significantly in order of magnitude and are easy for the model portability. The structure of both BeginNet and Subnets is shown in Fig. 4, and the final fully connected layer differs due to the different number of output results for each network.

4. Experiments and discussion

Experimental environment is as follows: Nvidia GeForce GTX 1060 Max-Q GPU and Nvidia Tesla K80 GPU.

Before the three occlusion databases, ConcatDB, BConcatDB and eyeCDB are used for training, a few final processing steps need to be done. First, the images are normalized to the size 224 * 224 for model and converted to gray. Then randomly selected images are flipped horizontally for data augmentation. Finally, the original images with values in the range [0,255] are transformed into floating point tensors with values in the range [0.0, 1.0] by min–max normalization.

To observe what the existing popular network structure without any modification can achieve with the occlusion database, we first try to train Resnet18 [40] directly on the ConcatDB training set. It is observed that the prediction accuracy of the Resnet18 on the test set is 19% and without increasing trend, which is low and far from being usable.

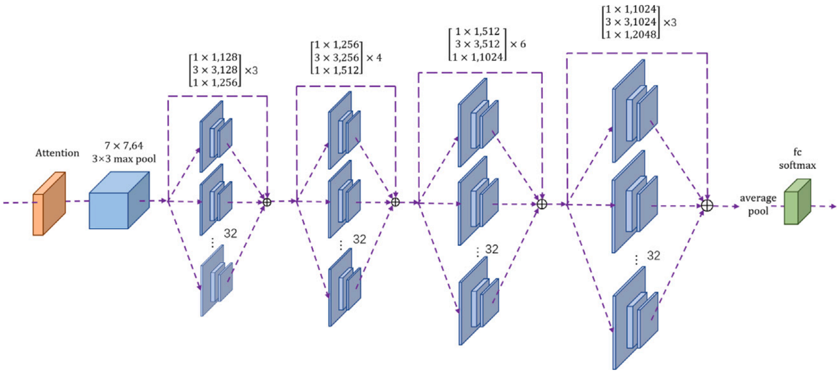


Fig. 4. Structure of both BeginNet and Subnets.

Table 3
Prediction accuracy of each network under test sets with different occlusion databases.

Networks	Accuracy/%		
	ConcatDB	BConcatDB	eyeCDB
BeginNet	76.7	77.3	78.4
SubnetX	77.1	76.6	79.9
SubnetY	74.2	73.1	73.1
SubnetZ	89.7	88.2	92.3

Table 4
Comparison with other state-of-arts methods on test sets of three occlusion databases.

Model	Accuracy/%		
	ConcatDB	BConcatDB	eyeCDB
Ours	**59.8**	**60.6**	63.7
ResNeXt-50	27.0	30.2	30.4
SCAN-CCI [26]	55.6	55.8	**74.2**
OADN [25]	51.7	59.6	60.7
gACNN [22]	50.8	51.2	56.2

Considering that in classification tasks, tasks with fewer classifications such as binary and tertiary classifications are considered simpler, we merge several expression classifications in ConcatDB based on what we observe above and reduced ConcatDB from seven classifications to a database containing only three classifications, keeping the total number of images the same. Then we train Resnet18 again using the reduced number of classifications on the ConcatDB training set, and observe a larger improvement in prediction accuracy. We further train Resnet101, Resnet152 and wide_Resnet101 [41], two larger and deeper networks, on the training set, with higher prediction accuracy as shown in Table 2. Inspired by the above observation, we randomly segment the seven expression labels of the ConcatDB, BConcatDB and eyeCDB.

After softmax function, the Subnets output an array of two or three values, and the label corresponding to the largest value is taken as the prediction result of the Subnets. Thus, for the method to integrate multiple Subnets, we first directly concatenate the outputs of three Subnets into a 1 * 7 array and consider taking the label corresponding to the largest value as the overall prediction. However, the results are poor. The accuracy after applying this method is comparable to the prediction accuracy of the Resnet18 trained on the original ConcatDB. Then we add a small network to select the label from the concatenated array instead of simply selecting the label corresponding to the largest value. Similarly, the results of which are poor. Based on above experiments, we have an idea of the path selection method for integrating Subnets.

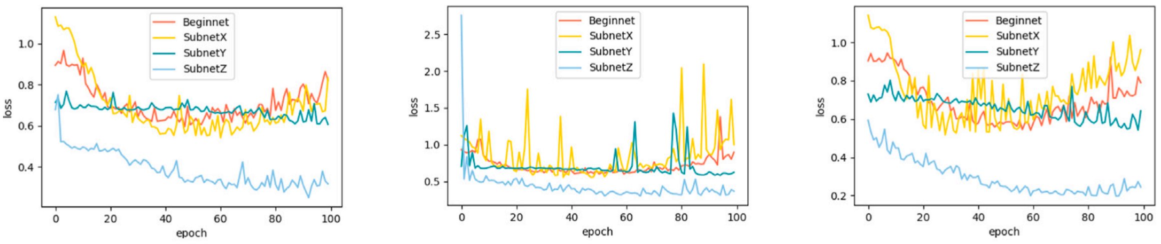
We use ResNeXt50 as the backbone structure for both the BeginNet and all Subnets, without pre-trained parameters from ImageNet. The minimum batch size for the BeginNet training is set to 64; the minimum batch size for the Subnets is set to 128; the momentum is set to 0.9; the learning rate for the BeginNet and Subnets Y and Z is set to 0.2; the learning rate for SubnetX is set to 0.1. The loss is calculated by the cross-entropy loss function separately for each network. SubnetY is trained for 300 epochs while the others are 100 epochs. The optimization algorithm uses the stochastic gradient descent(SGD) method. The prediction accuracy of each network for the test sets of ConcatDB, eyeCDB, and BConcatDB is shown in Table 3, in which the highest accuracy is the SubnetZ, while the accuracy of SubnetY is lower.

The value of loss indicates the gap between the predicted label and actual label. The lower the loss, the closer the gap. As shown in Fig. 5, BeginNet and SubnetX achieve the minimal loss fast during the 100 epochs of training, while SubnetZ is slower. While the other networks have a downward trend followed by an upward trend, the minimal loss of SubnetY seems not to be reached in 100 epochs of training. Thus we train SubnetY for 300 epochs instead of 100. Moreover, the loss of SubnetZ is much lower than those of the remaining three networks in the model. The minimum of loss of SubnetZ is 0.182 during the training with eyeCDB. The losses of BeginNet, SubnetX, Y, Z under eyeCDB are 0.067, 0.032, 0.032, 0.094 lower in average than those of the other two occlusion databases.

The confusion matrix for ConcatDB is shown in Fig. 6(a). From Fig. 6(a), Anger, Disgust and Fear expressions are the most easily confused with the SubnetY and the SubnetZ, respectively. Disgust is most easily confused with Normal as well as Anger, Happiness and Sadness. The confusion relationship of Fear is similar to that of Disgust with the difference of Surprise replacing Happiness. Due to the similarity of the remaining unoccluded parts of the faces in expressing different expressions, the accuracy of the BeginNet is improved by increasing the accuracy of the lower Subnets. From the confusion matrix for BConcatDB shown in Fig. 6(b), we can observe that it is basically the same as that for ConcatDB. The confusion matrix for eyeCDB is shown in Fig. 6(c). Compared with other matrices, the accuracy of other 5 expressions is increased more or less apart from Normal and Happiness. Anger is easy to be confused with Fear, Sadness and Normal. Disgust is most easily confused with Normal and then Anger, Sadness, Surprise. The confusion relationship of Fear is similar to that of ConcatDB.

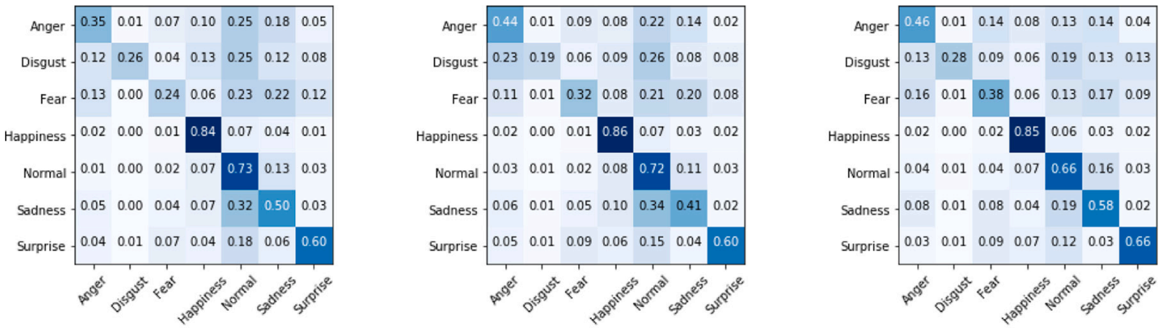
The common point observed from the above confusion matrices for the three occlusion databases is that Happiness, Normal, Surprise, and Sadness are less likely to be confused. They are more likely to be correctly predicted when important regions get occluded. While the remaining three expressions Anger, Disgust, and Fear can easily be incorrectly predicted.

We train our model and ResNeXt-50 from scratch on the training sets of ConcatDB, eyeCDB, and BConcatDB. SCAN-CCI [26] is first initialized with ResNet-50 model pre-trained on the VGGFace2 [42] and it is trained on the above training sets as well. OADN and gACNN are also initialized with the ResNet-50 model pre-trained with the VGGFace2 database but trained on the FERPlus [43] database without masking process. All methods are tested on ConcatDB, eyeCDB, and BConcatDB



(a) ConcatDB (b) BConcatDB (c) eyeCDB

Fig. 5. Losses of each network under validation sets with different occlusion databases.



	Anger	Disgust	Fear	Happiness	Normal	Sadness	Surprise
Anger	0.35	0.01	0.07	0.10	0.25	0.18	0.05
Disgust	0.12	0.26	0.04	0.13	0.25	0.12	0.08
Fear	0.13	0.00	0.24	0.06	0.23	0.22	0.12
Happiness	0.02	0.00	0.01	0.84	0.07	0.04	0.01
Normal	0.01	0.00	0.02	0.07	0.73	0.13	0.03
Sadness	0.05	0.00	0.04	0.07	0.32	0.50	0.03
Surprise	0.04	0.01	0.07	0.04	0.18	0.06	0.60

(a) ConcatDB

	Anger	Disgust	Fear	Happiness	Normal	Sadness	Surprise
Anger	0.44	0.01	0.09	0.08	0.22	0.14	0.02
Disgust	0.23	0.19	0.06	0.09	0.26	0.08	0.08
Fear	0.11	0.01	0.32	0.08	0.21	0.20	0.08
Happiness	0.02	0.00	0.01	0.86	0.07	0.03	0.02
Normal	0.03	0.01	0.02	0.08	0.72	0.11	0.03
Sadness	0.06	0.01	0.05	0.10	0.34	0.41	0.02
Surprise	0.05	0.01	0.09	0.06	0.15	0.04	0.60

(b) BConcatDB

	Anger	Disgust	Fear	Happiness	Normal	Sadness	Surprise
Anger	0.46	0.01	0.14	0.08	0.13	0.14	0.04
Disgust	0.13	0.28	0.09	0.06	0.19	0.13	0.13
Fear	0.16	0.01	0.38	0.06	0.13	0.17	0.09
Happiness	0.02	0.00	0.02	0.85	0.06	0.03	0.02
Normal	0.04	0.01	0.04	0.07	0.66	0.16	0.03
Sadness	0.08	0.01	0.08	0.04	0.19	0.58	0.02
Surprise	0.03	0.01	0.09	0.07	0.12	0.03	0.66

(c) eyeCDB

Fig. 6. (a) is the confusion matrix for ConcatDB; (b) is the confusion matrix for BConcatDB; (c) is the confusion matrix for eyeCDB, the darker the color, the higher the accuracy.

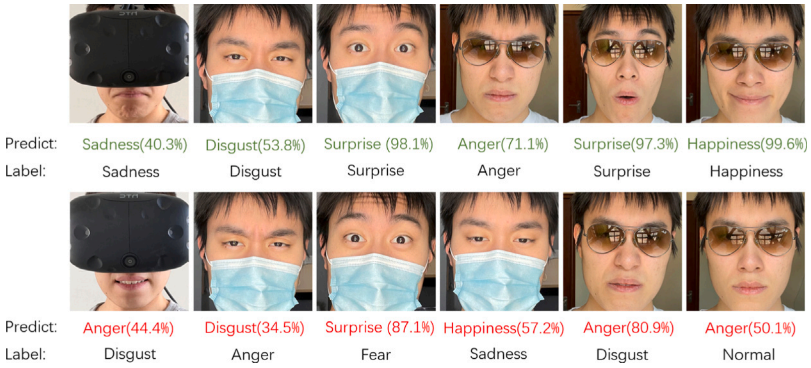

Predict: Sadness(40.3%) Disgust(53.8%) Surprise (98.1%) Anger(71.1%) Surprise(97.3%) Happiness(99.6%)
Label: Sadness Disgust Surprise Anger Surprise Happiness
Predict: Anger(44.4%) Disgust(34.5%) Surprise (87.1%) Happiness(57.2%) Anger(80.9%) Anger(50.1%)
Label: Disgust Anger Fear Sadness Disgust Normal

Fig. 7. The prediction results of our method in three facial occlusion scenarios. The correct predictions and incorrect predictions are represented in green and red, respectively. The value following prediction indicates the product of BeginNet prediction probability and Subnet prediction probability.

test sets for prediction accuracy. We compare our method with others in Table 4. From the table, we can observe that our method achieves 59.6% and 60.6% accuracy on the two databases with large occlusion area, ConcatDB and BConcatDB, respectively, while other SOA models achieve up to 55.6% and 59.6%. ResNeXt-50, i.e., backbone network of our method, has average 32.2% increasing after applying our method compared to those without our method. Our method does not perform as well as SCAN-CCI on eyeCDB, a database with a small masking area, and is 16% lower in accuracy. However, our method still outperforms other models. Combined with the Table 1, there are more images of label A in the test set, so the overall accuracy of our method also converges to the product of the BeginNet accuracy and SubnetX accuracy.

Furthermore, we take some photos for each of the three facial occlusion scenarios. In the photos, Wearing a VR headset is for upper face occlusion; wearing a mask is for lower face occlusion; wearing sunglasses is for eye occlusion. The photos correspond to all seven classifications and are appropriately cropped into squares containing only faces. The prediction results using our method are shown in Fig. 7. The top half of some of the predicted results match the actual expressions, while the bottom half mismatch. It can be observed that some expressions look very similar after masking part of the face e.g. Surprise and Fear, Anger and Disgust. Thus, the prediction probability is still not low on the second and fourth photos in the bottom half of Fig. 7 even though they are incorrectly predicted. On the contrary, there are some expressions that are more distinguishable from others, such as Surprise and Happiness, which can be better predicted by our method.

5. Conclusion

In this paper, we propose a path selection multi-network to tackle FER tasks in upper, lower face and eye occlusion scenarios. We evaluate our method on both masking processed in-the-wild and in-lab datasets, respectively. Our method achieves better performance comparing to the state-of-the art results on upper and lower face occlusion scenarios. Our method allows each of the subnets to classify less labels. Therefore, the subnets extract more efficient features from the limited input. In addition, the training of each subnets is independent, which means we can train them in parallel.

However, the proposed method still has several points for improvement. For instance, more occluded facial expression ought to be recognized by improving the proposed method. Moreover, the overall accuracy needs still to be improved, which is affected by the path selection method on the both layers in the networks. The path selection method will be researched to select the right sub-networks in the future work.

CRediT authorship contribution statement

Liheng Ruan: Methodology, Software, Investigation, Writing – original draft. **Yuexing Han:** Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition. **Jiarui Sun:** Software, Investigation. **Qiaochuan Chen:** Validation, Investigation. **Jiaqi Li:** Validation, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is sponsored by the Natural Science Foundation of Shanghai, China (Grant No. 20ZR1419000), Key Research Project of Zhejiang Laboratory (No. 2021PE0AC02).

References

[1] S. Hickson, N. Dufour, A. Sud, V. Kwatra, I. Essa, Eyemotion: Classifying facial expressions in VR using eye-tracking cameras, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1626–1635.

[2] A. Mehrabian, Communication without words, Commun. Theory 6 (2008) 193–200.

[3] D. Kong, M. Zhu, J. Yu, Research on the application and method of facial expression recognition in assistive medical care, Life Sci. Instrum. 2 (2019).

[4] Z. Jian, Key Techniques of Content-based Intelligent Video Surveillance and the Applications in Public Security, Zhejiang University, 2007.

[5] N. Ottakath, O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Mohamed, T. Khattab, K. Abualsaud, Vidmask dataset for face mask detection with social distance measurement, Displays (2022) 102235.

[6] L. Zhang, B. Verma, D. Tjondronegoro, V. Chandran, Facial expression analysis under partial occlusion: A survey, ACM Comput. Surv. 51 (2) (2018) 1–49.

[7] M. Mehdipour Ghazi, H. Kemal Ekenel, A comprehensive analysis of deep learning based representation for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 34–41.

[8] L. Song, D. Gong, Z. Li, C. Liu, W. Liu, Occlusion robust face recognition based on mask learning with pairwise differential siamese network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 773–782.

[9] B. Houshmand, N.M. Khan, Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning, in: 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), IEEE, 2020, pp. 70–75.

[10] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.

[11] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[12] R. Zhi, M. Flierl, Q. Ruan, W.B. Kleijn, Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition, IEEE Trans. Syst. Man Cybern. B 41 (1) (2010) 38–52.

[13] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D.N. Metaxas, Learning active facial patches for expression analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2562–2569.

[14] L. Zhang, W. Li, L. Yu, L. Sun, X. Dong, X. Ning, Gmface: An explicit function for face image representation, Displays 68 (2021) 102022.

[15] H. Yang, Z. Zhang, L. Yin, Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 294–301.

[16] F. Zhao, J. Feng, J. Zhao, W. Yang, S. Yan, Robust LSTM-autoencoders for face de-occlusion in the wild, IEEE Trans. Image Process. 27 (2) (2017) 778–790.

[17] B. Pan, S. Wang, B. Xia, Occluded facial expression recognition enhanced through privileged information, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 566–573.

[18] J. Dong, L. Zhang, H. Zhang, W. Liu, Occlusion-aware GAN for face de-occlusion in the wild, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.

[19] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, IEEE Trans. Image Process. 29 (2020) 4057–4069.

[20] B. Yang, W. Jianming, G. Hattori, Face mask aware robust facial expression recognition during the COVID-19 pandemic, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 240–244.

[21] Y. Li, J. Zeng, S. Shan, X. Chen, Patch-gated cnn for occlusion-aware facial expression recognition, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 2209–2214.

[22] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, IEEE Trans. Image Process. 28 (5) (2018) 2439–2450.

[23] D.S. Trigueros, L. Meng, M. Hartnett, Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss, Image Vis. Comput. 79 (2018) 99–108.

[24] W. Wan, J. Chen, Occlusion robust face recognition based on mask learning, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3795–3799.

[25] H. Ding, P. Zhou, R. Chellappa, Occlusion-adaptive deep network for robust facial expression recognition, in: 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2020, pp. 1–9.

[26] D. Gera, S. Balasubramanian, Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition, Pattern Recognit. Lett. 145 (2021) 58–66.

[27] D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3642–3649.

[28] B.-K. Kim, H. Lee, J. Roh, S.-Y. Lee, Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 427–434.

[29] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al., Combining modality specific deep neural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 543–550.

[30] G. Pons, D. Masip, Supervised committee of convolutional neural networks in automated facial expression analysis, IEEE Trans. Affect. Comput. 9 (3) (2017) 343–350.

[31] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: A report on three machine learning contests, in: International Conference on Neural Information Processing, Springer, 2013, pp. 117–124.

[32] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 1998, pp. 200–205.

[33] D. Lundqvist, A. Flykt, A. Öhman, The karolinska directed emotional faces (KDEF), in: CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet, Vol. 91, (630) 1998, p. 2.

[34] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Trans. Image Process. 28 (1) (2018) 356–370.

[35] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2852–2861.

[36] Itseez, The OpenCV Reference Manual, second ed., 2014.

[37] Itseez, Open source computer vision library, 2015, https://github.com/itseez/opencv.

[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[39] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[41] S. Zagoruyko, N. Komodakis, Wide residual networks, 2016, arXiv preprint arXiv:1605.07146.

[42] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 67–74.

[43] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283.