MASK-BASED ATTENTION PARALLEL NETWORK FOR IN-THE-WILD FACIAL EXPRESSION RECOGNITION

Lingzhao Ju, Xu Zhao

Department of Automation, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Facial expression recognition suffers big pose and occlusion in real world and attention mechanism is deployed widely to cope with these challenges. But most previous attention-based methods are inadequate in locating crucial expression-related regions precisely and capturing useful facial expression features comprehensively. For these reasons, we present a novel mask-based attention parallel network (MAPNet). Firstly, mask-based attention module that locates expression-related regions is constructed from binary mask extracted by key landmark detection. Secondly, the designed parallel network embeds mask-based attention modules into its different layers to acquire comprehensive facial expression features. Thirdly, the extracted parallel features are divided into several detached blocks from spatial dimension to predict facial expression independently. Finally, the expression label is acquired by combining two predictions of the parallel network and a new loss function is designed to weigh unbalanced facial expression distribution. We validate our method on three popular in-the-wild datasets and the results demonstrate that our MANPnet outperforms previous state-of-the-art methods among RAFDB, AffectNet and FEDRO.

Index Terms— Facial expression recognition, mask-based attention parallel network, facial expression features

1. INTRODUCTION

Facial expression, a primary form of human communication, varies between different people and difference in background, pose, lighting. One kind of public facial expression dataset contains mainly lab-controlled images acquired by acting different expressions by subjects in the laboratory. Another type is in-the-wild dataset captured in uncontrolled environments, in which facial images are more realistic but more difficult to recognize than that collected in lab condition.

Under in-the-wild scenarios, facial expression recognition (FER) now has many primary applications such as social robot [1], offline education [2] and safe driving [3]. But these applications still meet some disruptive factors like occlusion and big pose variation, which hinder helpful features extraction and reduces FER prediction performance greatly.

This work has been funded in part by the NSFC grants 62176156, 61673269, and Shanghai Engineering Research Center of Intelligent Control and Management.

Inspired by the attention mechanism of human, many methods have applied attention mechanism to alleviate these disruptive factors for in-the-wild facial expression datasets, which focus on reserving useful facial expression features and eliminating irrelevant information [4, 5].

However, on utilizing this attention mechanism, there are some limitations in previous works. Firstly, facial expressions are related to whole facial area and have no relevance with background and other disturbance. Therefore, it is inaccurate to generate attention module by neutral network automatically like in [6]. Secondly, several studies [7, 8] adopted region-level attention to examine the importance of different regions. But these regional attention modules sometimes are mislead and learn useless features due to regional occlusion and big pose. Thirdly, attention module is always applied into one deep layer of the network backbone [9], which means that critical high level features are enhanced only once while low level discriminative features are neglected. Although one approach [10] applies attention modules into its both low and high layers in a cascaded way, features can not complement each other like in a parallel network structure. Additionally, some of these methods transport whole acquired features to global adaptive pooling layer directly [11].But each spatial split block of the final features has different significance to final emotion label, some blocks are more crucial.

To alleviate the above problems, this paper presents a novel mask-based attention parallel network (MAPNet) to enhance features extraction and expression classification. The main contributions are summarized as follows:

(1) Parallel network structure is built to learn features differently and combine independent prediction of each branch together, which improves the robustness of model.

(2) Mask-based attention modules are embedded into both the shallow layer and the deep layer of the MAPNet to concentrate on expression-related features extraction precisely.

(3) Features of each parallel branch are split from height and width dimension and different weights are applied to each split block to strengthen expression classification.

(4) A new loss function is designed to alleviate the negative impact of unevenly distributed expression categories.

The rest of this paper is organized as follows. Section 2 delivers the proposed method of MAPNet. Section 3 presents the experimental setup and result. Section 4 gives conclusion.

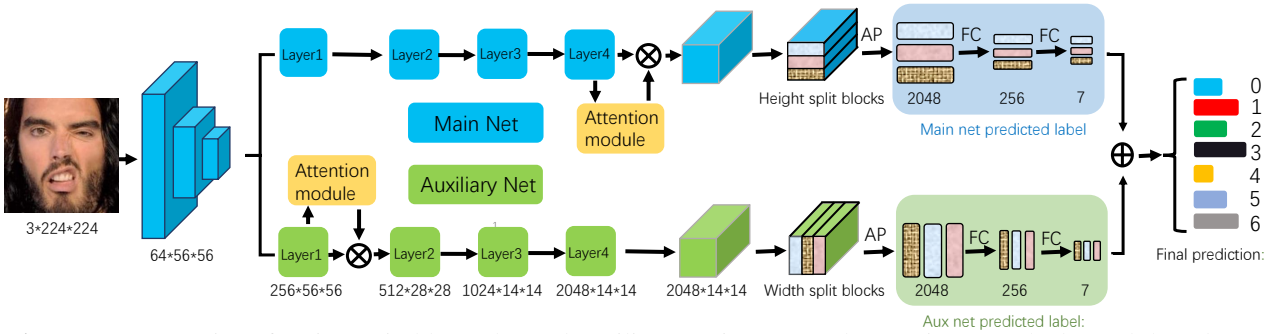


Fig. 1. MAPNet contains of main net in blue color and auxiliary net in green color. Main net embeds mask-based attention module into its fourth layer to strengthen high level features extraction while auxiliary net applies attention module into its first layer to enhance low level features extraction. These two branches are then split from height and width dimension to predict expression label independently. The final prediction is obtained by combining label of each branch together.

2. PROPOSED APPROACH

2.1. Framework overview

The MAPNet architecture is built based on the backbone of the residual net (ResNet) [12], which is constructed from lower-level convolutional layers, four cascaded structured CNN blocks and a fully connected layer.

To extract more expression-related high level features, mask-based attention module is embedded into the fourth CNN block of main net to strengthen deep layer features extraction. After that, features are split into several blocks from the height dimension and adaptive average pooling (AP) is applied to each block. Two cascaded fully connected (FC) layers are further used to reduce features dimension from 2048 to 7 and dropout is utilized. Then, the output of each block is fed into a Softmax layer to predict the expression label. Finally, predicted labels of these split blocks are integrated together through applying different weight coefficients. The formula is as follow:

$$P = \sum_{i=1}^{N} \delta_i p_i, \quad \sum_{i=1}^{N} \delta_i = 1 \tag{1}$$

where i is the i-th split block, N is the number of total blocks, P is the final label prediction of this branch, p_i is the predicted label and δ_i is the coefficient of each block. δ_i is usually set to a larger value of the inner block while small coefficient is given to the outer block.

Different from the main net, auxiliary net embeds mask-based attention module into the first CNN blocks to reinforce extracting low level useful features. Then, the extracted features are divided into several blocks from the width dimension to predict emotional label separately. The predicted label of auxiliary net is also acquired by the Equation (1).

The final predicted label P_F is obtained by combining the main net predicted label P_M and the auxiliary net predicted label P_A. The equation can be denoted as follow, where a and b are set to 0.5 respectively because two branches have same significance and they complement each other.

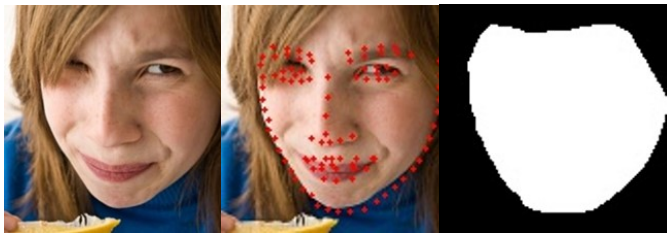
$$P_F = a * P_M + b * P_A \tag{2}$$

2.2. Mask-based attention module

Facial expression is closely related to whole face regions. It is easily influenced by the background information and other regional occlusions. To make our parallel model focus on the expression-related areas, we design mask-based attention module to pay more attention to the whole face area. First, we use the face landmarks detection method PFLD [13] to get 98 key facial landmarks as shown in Fig.2 (b). Then, we get one general mask M which highlights the whole face area as shown in Fig.2 (c). Furthermore, we generate mask-based attention module through Equation (3).

$$F = Relu(X * (\alpha * M + \beta)), \quad \alpha + \beta = 2 \tag{3}$$

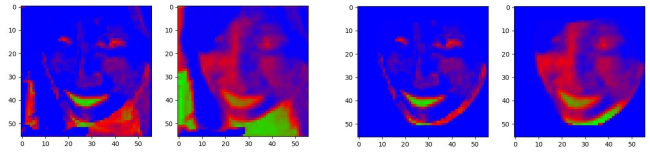
where X is the feature generated by the former layer of the backbone, M is the obtained binary mask, α and β are coefficients of the mask M and extracted feature X respectively. The strengthened feature is activated by the rectified linear unit (ReLU) and transported to next layer.



(a) Original image (b) 98 key points (c) Binary mask M

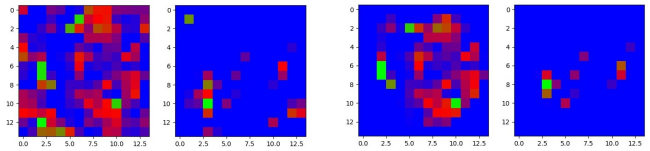
Fig. 2. PFLD is used to locate 98 key facial landmarks (b) to further generate binary mask (c) of whole facial region.

To visualize the function of mask-based attention module, the original and masked features of different layers are compared through heatmaps. As shown in Fig.3, the original shallow extracted features after layer 1 contains disturbing background information while the masked shallow features precisely focus on the expression-related area. In Fig.4, it is found that the masked high level features after layer 4 pay close attention to the inner pixels while the outer useless pixels are weaken.



(a) Original feature of layer 1 (b) Masked feature of layer 1

Fig. 3. Low level feature comparison



(a) Original feature of layer 4 (b) Masked feature of layer 4

Fig. 4. High level feature comparison

2.3. Loss function

Facial expression datasets collected in the real world are mostly imbalanced. To weight the inhomogeneous distribution of facial expression and achieve competitive performance, we design a novel weighted cross-entropy loss:

$$L(\hat{y}, i) = \omega_i * (-log(\frac{exp(\hat{y}_i)}{\sum_{j=1}^{K} exp(\hat{y}_j)})) \quad (4)$$

where $\hat{y}$ is the output probability of the expression, i is the ground-truth label of the expression, K is the number of total classes. ω_i is the weight of the i-th class, which demotes the penalty factor of each class and is calculated as

$$\omega_i = \sqrt[\gamma]{\frac{N_{avg}}{N_i}}, \quad i \in [1, K], \quad 0.5 \leqq \gamma \leq 3 \quad (5)$$

where N_{avg} is the average number of the training samples and N_i is the number of samples of each i-th class. γ depends on the undistributed level. By applying the new weighted cross-entropy loss, the unbalanced distribution level of training samples of each class can be alleviated and better prediction performance can be achieved.

Because of the special parallel structure design, the final loss L_F is combined by the main net loss L_M and the auxiliary net loss L_A. The formula is as follow:

$$L_F(\hat{y}, i) = u * L_M(\hat{y}, i) + v * L_A(\hat{y}, i) \quad (6)$$

where u is the loss coefficient of the main net and v is the loss coefficient of the auxiliary net. Both of them are set to 0.5.

3. EXPERIMENTS

3.1. Implementation details

Our MAPNet was trained on Pytorch framework and was initialized with the weights pre-trained on ImageNet. Stochastic Gradient Descent (SGD) was adopted as optimizer with the momentum of 0.9 and the weight-decay of 0.0005. The learning rate was initialized as 0.01 and was decreased by multiplying with 0.1 after 4 epochs to discover a suitable learning rate. The mini-batch size was set to 128. During training, only random horizontal flipping was used for data augmentation.

3.2. Performance comparison

To verify the effectiveness of our proposed method under natural scenarios, we evaluate MAPNet on 3 manually annotated in-the-wild datasets. All the validation results are compared with the previous state-of-the-art methods.

Previous Methods	RAFDB	AffectNet	FEDRO
WGAN [14]	83.49%	59.73%	/
DLP-CNN[15]	84.13%	/	/
gACNN[16]	85.07%	58.78%	66.50%
DFER-Net[17]	85.13%	/	/
OAENet[18]	85.69%	/	/
VSAN [11]	86.20%	/	/
RAN[8]	86.90%	59.50%	67.98%
SCN [5]	87.03%	60.23%	/
OADN[19]	87.16%	61.89%	71.17%
HERO [20]	/	62.11%	/
LLHF [21]	/	63.31%	/
LAENet-SA [10]	/	64.09%	68.25%
MAPNet(ours)	**87.26%**	62.91%	**71.50%**
MAPNet*(ours)	/	**64.09%**	/

Table 1. Test set performance comparison on three datasets

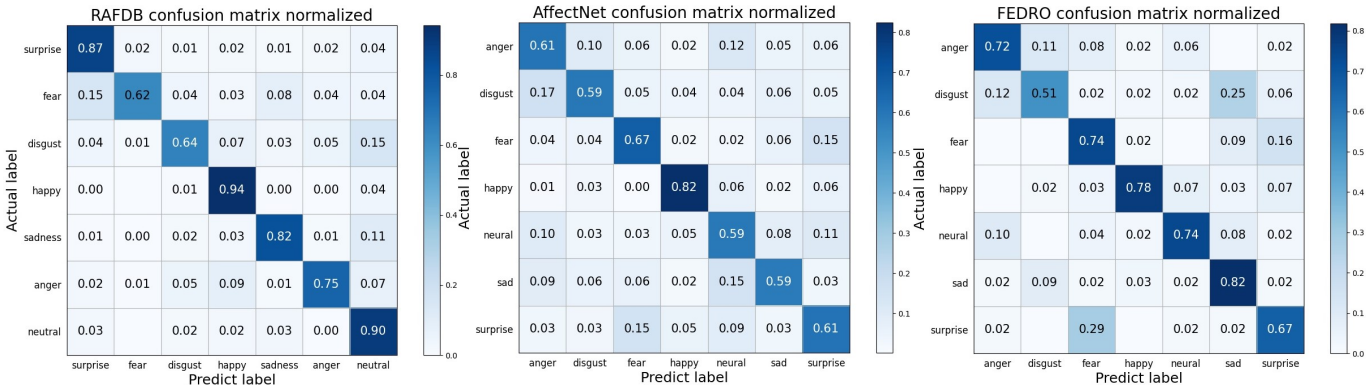
RAFDB [15] is one of the widely used expression datasets that contains 29672 in-the-wild facial images. We choose 12271 as training samples and 3068 as testing samples annotated with 7 emotional categories singly.

Table 1 shows the performance comparison in RAFDB dataset. MAPNet achieves 87.26% in terms of total accuracy, improving over OADN by 0.10% and RAN by 0.36%. Fig.5 (a) shows the confusion matrix of RAFDB. It is observed that *Fear* and *Disgust* are the two most confusing expressions, where *Fear* is easily confused with *Surprise* because of similar facial appearance while *Disgust* is mainly confused by *Neutral* due to subtle difference of these two expressions.

AffectNet [22] is the largest in-the-wild dataset for FER which has around 400000 manually annotated facial expression images. We only choose 33803 images as the training set and 3500 images in 7 categories as the validation set.

Table 1 also presents the performance of MANPnet in AffectNet dataset. Our model achieves the accuracy of 62.91%, outperforms some competitive methods while using far less samples as training set. When we train our MAPnet* on both AffectNet and RAFDB, its performance rises to 64.09%, which is equal to the performance of LAENet-SA. Fig.5 (b) presents the confusion matrix of the AffectNet. The *Happy* expression has the highest prediction rate, followed by *Fear*. *Sad*, *Neutral* and *Disgust* are three difficult expressions to classify with the accuracy of 59%.

FEDRO [16] has 400 in-the-wild images totally and every image in this dataset contains real occlusions and varied in position. We train our model on the joint training data of AffectNet and RAFDB and test on the FEDRO.



(a) RAFDB (b) AffectNet (c) FEDRO

Fig. 5. Confuse matrix analysis

Table 1 delivers the experimental results on the FEDRO. Our MAPNet achieves the best performance of 71.50%, surpassing the OADN by 0.37% and LAENet-SA by 3.25% greatly. From the confusions matrix in Fig.5 (c), we can see that *Sad* has the highest accuracy of 82%, while *Disgust* has the lowest accuracy of 51% due to the lack of training samples. 29% of *Surprise* are easily confused with *Fear*.

3.3. Abalation study

In this section, we conduct ablation study on three in-the-wild datasets to validate different functional blocks of our MAPNet as shown in Table 2 and make an analysis of the number of split blocks and the weight of attention module.

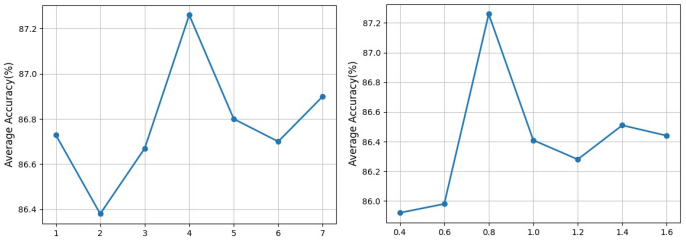
Baseline	PS	MAM	WCL	RAFDB	AffectNet	FEDRO
✓				84.48%	61.77%	64.25%
✓	✓			84.91%	62.60%	66.50%
✓	✓	✓		86.34%	63.69%	70.00%
✓	✓	✓	✓	87.26%	64.09%	71.05%

Table 2. Abalation study

Parallel Structure (PS): We first study the parallel structure design of our MAPNet without attention module and weighted cross-entropy loss. It is observed that PS improves the baseline by 0.43%, 0.83% and 2.25% of RAFDB, AffectNet and FEDRO respectively, which results from complementary features extraction of two parallel branches.

Mask-based Attention Module (MAM): Mask-based attention module is added to parallel structure. It is obvious that MAM embedded in two parallel structures significantly improves accuracy by 1.43% on RAFDB, 1.09% on AffectNet and 3.50% on FEDRO, which validates that MAM not only focus on critical high level features in main branch but also extracts more useful low level features in auxiliary branch.

Weighted Cross-entropy Loss (WCL): Our newly designed weighted cross-entropy loss is deployed as model loss function. From Table 2, we can find that WCL increases prediction accuracy by 0.92%, 0.40%, 1.05% on RAFDB, AffectNet and FEDRO respectively, due to its practical function of balancing uneven category distribution.



(a) Blocks number K (b) Mask weight α

Fig. 6. Influencing factors analysis

Analysis of split blocks: In parallel network structure, we divide the extracted feature maps into K blocks and train expression classifier of each block independently. K=1 means that the extracted feature is not split. Best accuracy appears at K=4 when higher coefficients are given to two inner blocks as shown in Fig.6 (a). Low accuracy occurs at K=2 because its two split blocks have same weight coefficients.

Analysis of attention module weight: Our mask-based attention module combines the original features and masked features as defined in Equation (3). The weight α controls the relative importance of masked features. From Fig.6 (b), we can see that mask-based attention module obtains the best performance at α=0.8 since more expression-related features are activated at this point.

4. CONCLUSION

In this paper, we proposes a novel mask-based attention parallel network (MAPNet) for in-the wild facial expression recognition to settle regional occlusion and big pose problems. The network is constructed of two parallel networks which are embedded with mask-based attention modules in different CNN layers to extract expression-related features. Feature splitting block is utilized to reinforce expression classification and a new loss function is designed to weight unbalanced category distribution. The proposed method MAPNet has achieved state-of-the-art accuracy with 87.26%, 64.09% and 71.50% on RAFDB, AffectNet and FEDRO respectively.

5. REFERENCES

[1] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction.," in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, 2003, vol. 5, pp. 53–53.

[2] Güray Tonguç and Betul Ozaydın Ozkara, "Automatic recognition of student emotions from facial expressions during a lecture," *Computers and Education*, vol. 148, pp. 103797, 2020.

[3] M. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors*, vol. 18, no. 12, pp. 4270, 2018.

[4] Wenhao Cao, Zhuoyu Feng, Dongyao Zhang, and Yisiyuan Huang, "Facial expression recognition via a cbam embedded network," *Procedia Computer Science*, vol. 174, pp. 463–477, 2020, 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.

[5] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6896–6905.

[6] Yanling Gan, Jingying Chen, Zongkai Yang, and Luhui Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020.

[7] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2209–2214.

[8] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[9] Siyue Xie, Haifeng Hu, and Yongbo Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognition*, vol. 92, pp. 177–191, 2019.

[10] Cong Wang, Jian Xue, Ke Lu, and Yanfu Yan, "Light attention embedding for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[11] Naigong Yu and Deguo Bai, "A visual self-attention network for facial expression recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[13] Xiaojie Guo, Siyuan Li, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling, "PFLD: A practical facial landmark detector," vol. abs/1902.10859, 2019.

[14] Yang Lu, Shigang Wang, Wenting Zhao, and Yan Zhao, "Wgan-based robust occluded facial expression recognition," *IEEE Access*, vol. 7, pp. 93594–93610, 2019.

[15] Shan Li and Weihong Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[16] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[17] Yumin Tian, Mengqi Li, and Di Wang, "Dfer-net: Recognizing facial expression in the wild," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2334–2338.

[18] Zhengning Wang, Fanwei Zeng, Shuaicheng Liu, and Bing Zeng, "Oaenet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognition*, vol. 112, pp. 107694, 2021.

[19] Hui Ding, Peng Zhou, and Rama Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," *arXiv: 2005.06040 [cs.CV]*, 2021.

[20] Wentao Hua, Fei Dai, Liya Huang, Jian Xiong, and Guan Gui, "Hero: Human emotions recognition for realizing intelligent internet of things," *IEEE Access*, vol. 7, pp. 24321–24332, 2019.

[21] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.