2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222441

AFNET-M: ADAPTIVE FUSION NETWORK WITH MASKS FOR 2D+3D FACIAL EXPRESSION RECOGNITION

Mingzhe Sui, Hanting Li, Zhaoqing Zhu, Feng Zhao

University of Science and Technology of China, Hefei 230027, China

ABSTRACT

2D+3D facial expression recognition (FER) can effectively cope with illumination and pose changes by merging texture and robust depth information. Most deep learning-based approaches employ the simple fusion strategy that concatenates the multimodal features directly after fully-connected layers, without considering the different degrees of significance for each modality. Meanwhile, how to focus more on both 2D and 3D local features is still a great challenge. In this paper, we propose the adaptive fusion network with masks (AFNet-M) for 2D+3D FER. To enhance 2D and 3D local features, we take the masks annotating salient regions of the face as prior knowledge and design the mask attention module (MA) which can automatically learn two modulation vectors to scale the feature maps. We also introduce an adaptive fusion module (AF) at convolutional layers through the computed importance weights. Experimental results demonstrate that our AFNet-M achieves the state-of-the-art performance on BU-3DFE and Bosphorus datasets and requires fewer parameters in comparison with other models.

Index Terms— 2D+3D facial expression recognition, mask attention, adaptive fusion, AFNet-M.

1. INTRODUCTION

Facial expression is a significant means of nonverbal communication since it can express human cognition and emotions. In general, facial expression recognition (FER) aims to help machines infer six basic expressions and figures prominently in human-computer interaction areas [1]. Although previous studies based on hand-crafted features [2] or deep learning [3] have achieved excellent performance in 2D FER, it is under the premise of good image quality. The drastic variations in illumination and poses can still have a great impact [1].

3D scans containing depth information perform better robustness to illumination and pose changes, and can also capture subtle muscle deformations [1]. Therefore, complementary multimodal 2D+3D FER has gradually attracted increasing attention in recent years. Li *et al*. [4] first introduced CNN to 2D+3D FER, where they represented each 3D scan as six attribute maps, and fed them into a deep fusion CNN with six branches for classification. Benefiting from the powerful learning ability of networks, the accuracy of deep learning-based models [4–12] has comprehensively surpassed the methods based on hand-crafted features [13–20].

However, there still exist two issues in 2D+3D FER. First, most algorithms do not take good advantage of local features in salient regions (e.g., the neighborhoods of the mouth , nose, and eyes). Jiao *et al*. [7] proposed the FA-CNN to localize the discriminative facial parts, while the receptive fields will also focus on irrelevant areas such as the forehead, and the distribution is not stable enough from their visualization of heat maps. Sui *et al*. [11] designed the masks to directly enhance the local features in the whole salient regions, however, diverse components make various contributions to the judgment of one expression. For example, the features of the eyes and mouth are more critical than those of the nose. Thus, learning the distribution of salient regions discriminately from the masks is necessary. Another is that many deep models [7,8,11] employ the simple fusion strategy, which concatenates the multimodal features directly and equally after fully-connected layers. At this position, the resolution of the features with each modality is too low and the ability to perceive local geometric details is poor, which is not conducive to the free attention flow among modalities [21]. Furthermore, each modality places a different emphasis on the current classification task. We need to consider it before fusion.

To address the above problems, we propose the adaptive fusion network with masks (AFNet-M) for 2D+3D FER. We design the mask attention module (MA), which can learn two modulation vectors from the masks annotating salient regions to enhance local features discriminatively. Considering the contribution rates of the features with two modalities are different, we introduce an adaptive fusion strategy that incorporates depth features into texture features with the computed importance weights. Our AFNet-M achieves great performance on BU-3DFE and Bosphorus datasets and demands fewer parameters compared with state-of-the-art methods.

2. METHODOLOGIES

2.1. Overview of AFNet-M

The framework is illustrated in Fig. 1. The whole network is based on dual-branch ResNet18s. In preprocessing, we ex-

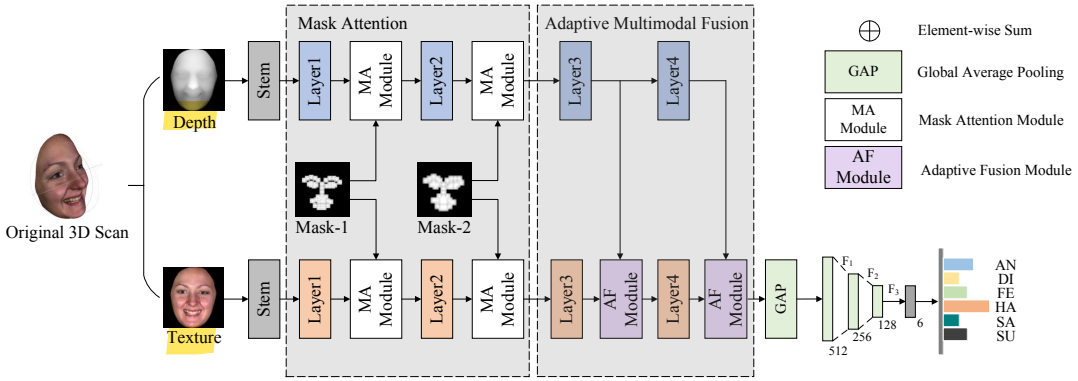


Fig. 1. The pipeline of our AFNet-M. Stem means all the operations before the first residual block in ResNet18, including a 7×7 convolutional layer and a max pooling layer.

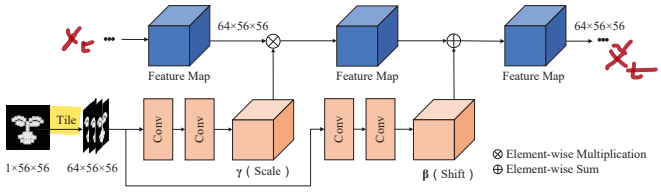


Fig. 2. The structure of the proposed MA module at Layer1.

ecute gridfit algorithm to generate aligned texture and depth images from each 3D scan and perform surface processing containing outlier removal, hole filling, and noise removal to improve their quality. We also generate the masks annotating salient regions at two scales (Mask-1 is 56×56 and Mask-2 is 28×28) as [11] did. In feature extraction, we design the MA module at Layer1 and Layer2 and introduce the generated masks as prior knowledge to enhance depth and texture local features separately in the spatial dimension. At Layer3 and Layer4, we use the devised AF module to perform the adaptive fusion of depth and texture features with their weights.

2.2. Mask Attention

Different from directly enhancing all salient regions [11], we consider learning the distribution of salient regions discriminately from the masks and propose the MA module, as shown in Fig. 2. The MA module consists of four convolutions. Taking the texture branch as an example, we first reshape the single-channel mask to 64×56×56 at Layer1 to match the texture feature maps. Two modulation vectors ($\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$) are automatically learned separately through two independent convolution groups, which represent the distribution of salient regions in the masks. To reduce extra parameters, the kernel size of all convolutions is 1×1. The process of enhancing the texture feature maps can be expressed as:

$$\tilde{\mathbf{X}}_t = \boldsymbol{\gamma} \otimes \mathbf{X}_t + \boldsymbol{\beta} \tag{1}$$

where $\tilde{\mathbf{X}}_t$ and $\mathbf{X}_t$ represent the texture feature maps after and before enhancement, respectively. $\otimes$ represents the element-wise multiplication. The shapes of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are the same as $\mathbf{X}_t$. It is equivalent that we scale and shift the feature map of each channel of $\mathbf{X}_t$ in the spatial dimension. We do the same process for the depth images. During the training stage, the two modulation vectors will be continuously adjusted to make the network discriminately enhance the depth and texture local features in salient regions of the face. We do not use the MA module in the second half in that the landmarks cannot be detected to generate the mask and the the receptive field of each pixel has almost covered the entire input image when the spatial size is too small.

2.3. Adaptive Multimodal Fusion

To fully utilize the multiscale features at convolutional layers and form more comprehensive multimodal representations, we incorporate depth features into texture features, as shown in Fig. 1. Since the contribution rates of the features with two modalities are different, inspired by ACM block [22], we introduce the AF module to perform adaptive fusion, as depicted in Fig. 3. We first compute the importance weight for each channel of the feature maps, formulated as:

$$\mathbf{t}_{iw} = Sigmoid(Conv(AvgPool(\mathbf{X}_t)) + Conv(MaxPool(\mathbf{X}_t))) \tag{2}$$

where $\mathbf{X}_t$ is the texture feature maps at Layer3 or Layer4, $\mathbf{t}_{iw}$ ranging from 0-1 represents the texture importance weights. $Conv$ represents the shared convolution with the kernel size of 1×1, which can mine the correlations among channels. Similarly, we also compute the depth importance weights $\mathbf{d}_{iw}$. The adaptive fusion can be expressed as:

$$\begin{cases} \hat{\mathbf{X}}_d = \mathbf{d}_{iw} \otimes \mathbf{X}_d \\ \hat{\mathbf{X}}_t = \mathbf{t}_{iw} \otimes \mathbf{X}_t \\ \mathbf{M} = \hat{\mathbf{X}}_t + \hat{\mathbf{X}}_d \end{cases} \tag{3}$$

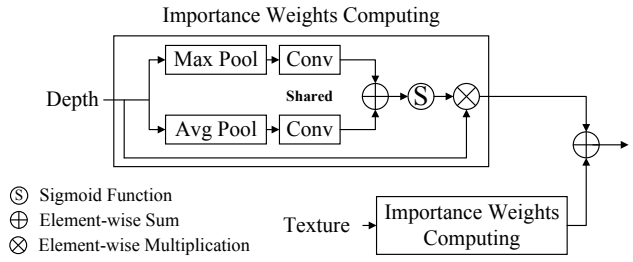


Fig. 3. The structure of the AF module.

where M represents the formed multimodal features after adaptive fusion. Therefore, the channel features with some modality that are more crucial to the results will also account for a larger proportion of the formed multimodal representations, while those that have a negative impact on the classification will be suppressed to a certain extent. The reason why we choose to perform adaptive fusion at Layer3 and Layer4 is that the resolution and the ability to perceive local geometric details of the features with each modality are appropriate. We also give the ablation studies to prove this in the next section.

3. EXPERIMENTAL RESULTS

3.1. Dataset, Protocol and Implementation Details

BU-3DFE dataset. The BU-3DFE [23] comprises 100 subjects with ages from 18 to 70. Each subject contains six prototypical expressions with four levels of expression intensity.

Bosphorus dataset. The Bosphorus [24] consists of 4666 3D scans from 105 subjects with ages from 25 to 35. Only 63 subjects contain six prototypical expressions.

Evaluation Protocol. We follow the standard protocol applied in [4,7–12]. In this protocol, 60 subjects in BU-3DFE and 60 subjects subjects in Bosphorus are selected randomly, which are fixed in the whole experiments. Then, the average accuracy of 100 times of 10-fold cross-validation is executed to evaluate the model for more stable and reliable results.

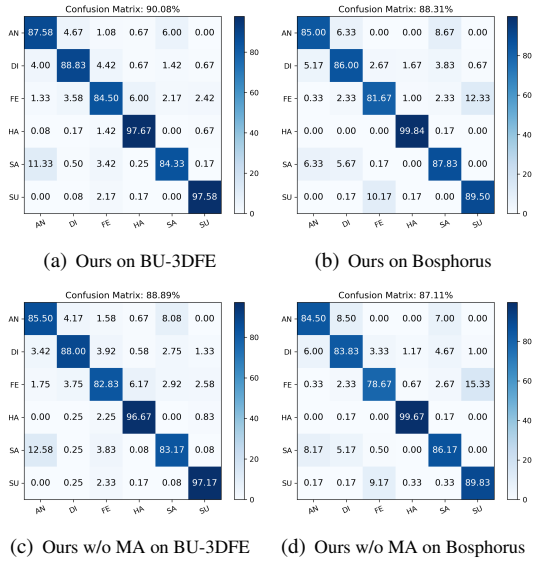
Implementation Details The depth and texture images are resized to $3\times224\times224$ after preprocessing. We initialize two ResNet18s in the AFNet-M with the pre-trained parameters on ImageNet. All the convolutions in the MA and AF module follow a normal distribution. The Adam optimizer with betas $(0.9, 0.999)$ is adopted and the learning rate is fixed at 0.0001. All the experiments are conducted on two NVIDIA GeForce RTX3070 cards with Pytorch.

3.2. Results

Comparisons with the state-of-the-art methods. Table 1 shows the performance comparisons of our model with other approaches on BU-3DFE and Bosphorus. We can see that our AFNet-M outperforms state-of-the-art methods with the highest accuracy of 90.08% and 88.31%, whether compared with hand-crafted features or deep networks.

Table 1. Comparisons for recognition accuracy(%). HC and DL represent hand-crafted and deep learning-based features.

Method	Year	Data	Feature	BU-3DFE	Bosphorus
Li *et al*. [13]	2012	3D	HC	80.14	75.83
Yang *et al*. [14]	2015	3D	HC	84.80	77.50
Li *et al*. [15]	2015	2D+3D	HC	86.32	79.72
Fu *et al*. [16]	2019	2D+3D	HC	82.89	75.93
Li *et al*. [4]	2017	2D+3D	DL	86.86	80.28
Jan *et al*. [5]	2018	2D+3D	DL	88.54	-
Tian *et al*. [6]	2019	2D+3D	DL	-	79.17
Jiao *et al*. [7]	2019	2D+3D	DL	89.11	-
Zhu *et al*. [8]	2019	2D+3D	DL	88.35	-
Jiao *et al*. [9]	2020	2D+3D	DL	89.72	83.63
Zhu *et al*. [10]	2020	2D+3D	DL	88.75	-
Sui *et al*. [11]	2021	2D+3D	DL	89.82	87.65
Ni *et al*. [12]	2022	2D+3D	DL	88.91	85.16
Ours	-	2D	DL	87.68	85.42
Ours	-	3D	DL	86.97	82.06
Ours	-	2D+3D	DL	**90.08**	**88.31**



(a) Ours on BU-3DFE (b) Ours on Bosphorus

(c) Ours w/o MA on BU-3DFE (d) Ours w/o MA on Bosphorus

Fig. 4. Confusion matrices. AN, DI, FE, HA, SA, and SU represent anger, disgust, fear, happiness, sadness, and surprise expression, respectively.

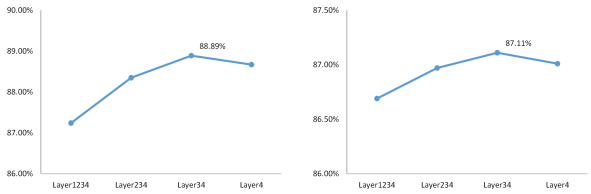
Confusion matrix. To reflect the performance for each category, we also give the confusion matrices, as shown in Fig. 4(a) and (b). We can find that happiness and surprise have better results, which is because that the features of them with exaggerated muscle deformations are more discriminate, and the others may confuse with each other.

3.3. Ablation Studies

Evaluation of the fusion strategy. We evaluate different fusion strategies without the MA module, as shown in Table 2.

Table 2. Ablation experiments of the fusion strategy.

Method	Fusion Strategy			AF	BU-3DFE	Bosphorus
	Data	Feature	Decision			
S1	✓				86.54	85.08
S2			✓		86.28	84.36
S3(FC)		✓			87.61	85.54
S4		✓			88.12	86.17
Ours		✓		✓	**88.89**	**87.11**



(a) Fusion positions on BU-3DFE (b) Fusion positions on Bosphorus

Fig. 5. Ablation experiments of choosing the fusion positions.

Table 3. Ablation experiments of the MA module.

Data	MA	Fusion	BU-3DFE	Bosphorus
2D			85.06	82.69
2D	✓		87.68	85.42
3D			84.37	80.58
3D	✓		86.97	82.06
2D+3D		✓	88.89	87.11
2D+3D	✓	✓	**90.08**	**88.31**

S1, S2, and S3(FC) represent the fusion strategy at data level, decision level, and fully-connected layer at feature level. S4 is our proposed fusion strategy without computing the importance weights. Compared with the first three strategies, S4 achieves the highest accuracy, indicating that fusing the features of the convolutional layers can obtain multimodal features with better representations. From the last two rows in Table 3, we can see that the AF module calculates the importance weights for the features with each modality to perform adaptive fusion, which can further improve the performance. Moreover, we also give the ablation experiments of choosing the fusion positions, as pictured in Fig. 5. The results in both subfigures prove the correctness of choosing Layer3 and Layer4 for fusion. For Layer1234, it begins since the first residual block. The receptive field of the underlying convolution kernel is relatively small, too much edge information may be extracted, which is not conducive to multimodal fusion.

Evaluation of the MA module. We evaluate the MA module with different modalities, as illustrated in Table 3. From the last two rows, we can see that using the proposed MA module to enhance local features discriminatively can improve the accuracy by 1.19% and 1.2% on BU-3DFE and Bosphorus under multimodal input. Furthermore, it can also boost the performance for a single 2D or 3D modality. From the comparisons of Fig. 4(c) and (a), (d) and (b), we can find that the masks annotating salient regions can help improve

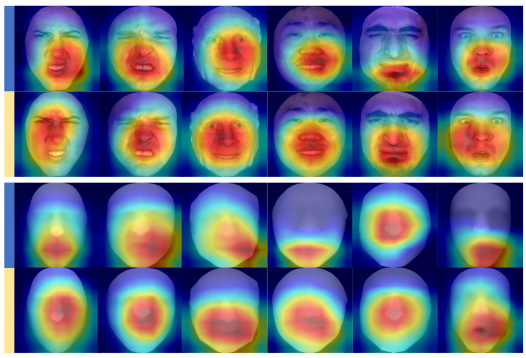


Fig. 6. Visualization of heat maps. The 1st and 2nd rows are texture heat maps w/o and with the MA module. The 3rd and 4th rows are depth heat maps w/o and with the MA module.

Table 4. Comparisons of the parameters

Method	Year	Data	Parameters (MB)
VGG-M-DF [5]	2018	2D+3D	≈ 327
DA-CNN [8]	2019	2D+3D	≈ 463
FFNet-M [11]	2021	2D+3D	≈ 93
Ours w/o MA	-	2D+3D	90.51
Ours w/o AF	-	2D+3D	86.53
AFNet-M	-	2D+3D	91.54

the recognition accuracies for almost all six categories, which fully shows the effectiveness of the designed MA module.

To reflect the distribution of the region of interest after employing the MA module, we also visualize the depth and texture heat maps with Grad-CAM, as given in Fig. 6. The 2nd and 4th rows obviously show that our AFNet-M has a more stable and concentrated region of interest for all expressions. More importantly, the masks enable the network to utilize the features of multiple facial parts (such as the eyes and mouth) to jointly determine the classification, rather than depending on a single part (see the anger expression in the 1st column in Fig. 6).

Parameters Analysis. To analyze the scale of the network, Table 4 illustrates the comparison results of parameters. We can see that our AFNet-M achieves the highest accuracy with relatively minimal parameters (91.54 MB) compared with other methods. And the incorporation of the designed MA and AF module only need tiny extra parameters.

4. CONCLUSION

In this paper, we propose the AFNet-M for 2D+3D FER. Based on the masks annotating salient regions, we design the MA module which can learn two modulation vectors to enhance 2D and 3D local features. To form better multimodal representations, we introduce an adaptive fusion module (AF) at convolutional layers through the computed importance weights. Experimental results show that our AFNet-M has superior performance and requires fewer parameters.

5. REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, pp. 1–20, 2020.

[2] S. L. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," in *Proc. Int. Conf. on Intell. Human Comput. Interaction*, 2012, pp. 1–5.

[3] Y. Xia et al., "Local and global perception generative adversarial network for facial expression synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–10, 2021.

[4] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[5] A. Jan et al., "Accurate facial parts localization and deep learning for 3D facial expression recognition," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2018, pp. 466–472.

[6] K. Tian et al., "3D facial expression recognition using deep feature fusion CNN," in *Proc. Ir. Signals Syst. Conf.*, 2019, pp. 1–6.

[7] Y. Jiao et al., "Facial attention based convolutional neural network for 2D+3D facial expression recognition," in *Proc. IEEE Vis. Commun. Image Process.*, 2019, pp. 1–4.

[8] K. Zhu et al., "Discriminative attention-based convolutional neural network for 3D facial expression recognition," in *Proc. 14th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2019, pp. 1–8.

[9] J. Yang et al., "2D+3D facial expression recognition via discriminative dynamic range enhancement and multi-scale learning," *arXiv preprint arXiv:2011.08333*, 2020.

[10] K. Zhu et al., "Intensity enhancement via gan for multimodal facial expression recognition," in *Proc. IEEE Int. Conf. Inf. Process.*, 2020, pp. 1346–1350.

[11] M. Sui, Z. Zhu, F. Zhao, and F. Wu, "FFNet-M: Feature fusion network with masks for multimodal facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[12] R. Ni et al., "Facial expression recognition through cross-modality attention fusion," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 45, no. 12, pp. 64–75.

[13] H. Li et al., "3D facial expression recognition via multiple kernel learning of multi-scale local normal patterns," in *Proc. Int. Conf. on Pattern Recog.*, 2012, pp. 2577–2580.

[14] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3D facial expression recognition using geometric scattering representation," in *Proc. 11th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2015, vol. 1, pp. 1–6.

[15] H. Li et al., "An efficient multimodal 2D + 3D feature-based approach to automatic facial expression recognition," *Comput. Vis. Image Understand.*, vol. 140, pp. 83–92, 2015.

[16] Y. Fu et al., "FERLrTc: 2D+3D facial expression recognition via low-rank tensor completion," *Signal Processing*, vol. 161, pp. 74–88, 2019.

[17] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-D face and facial expression recognition," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 498–511, 2008.

[18] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model-based automatic 3D/4D facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1438–1450, 2016.

[19] S. Berretti et al., "A set of selected SIFT features for 3D facial expression recognition," in *Proc. Int. Conf. on Pattern Recog.*, 2010, pp. 4125–4128.

[20] P. Lemaire, L. Chen, M. Ardabilian, and M. Daoudi, "Fully automatic 3D facial expression recognition using differential mean curvature maps and histograms of oriented gradients," in *Proc. 10th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2013, pp. 1–7.

[21] N. Arsha et al., "Attention bottlenecks for multimodal fusion," *Adv. Neural Inf. Proces. Syst.*, vol. 34, pp. 1–17.

[22] X. Hu, K. Yang, F. Lei, and K. Wang, "ACNet: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. Int. Conf. Image Process.*, 2019, pp. 1440–1444.

[23] L. Yin et al., "A 3D facial expression database for facial behavior research," in *Proc. 7th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2006, pp. 211–216.

[24] A. Savran et al., "Bosphorus database for 3D face analysis," in *Proc. European Workshop on Biometrics and Identity Management*, 2008, pp. 47–56.