CMANET: CURVATURE-AWARE SOFT MASK GUIDED ATTENTION FUSION NETWORK FOR 2D+3D FACIAL EXPRESSION RECOGNITION

*Zhaoqing Zhu, Mingzhe Sui, Hanting Li, and Feng Zhao**

University of Science and Technology of China, Hefei 230027, China
{zhaoqingzhu, sa20, ab828658}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859837

ABSTRACT

As 2D texture and 3D structural information can describe facial features complementarily, 2D+3D facial expression recognition (FER) has received widespread attention. Though recent methods for 2D+3D FER have reached excellent performance, they still face two challenges: the way for attending to critical face areas and the strategy for fusing multi-modal information. To address these issues, we propose a curvature-aware soft mask guided attention fusion network (CMANet), which mainly consists of two components: curvature-aware attention module and multi-modal attention fusion module. The former utilizes the curvature-aware soft mask guiding the homo-modal attention mechanism to focus on potentially important areas with soft weights, while the latter applies pixel-level fusion on multi-modal features to retain the significant information from different modalities and also allows multi-modal features to interact in a larger field of view. Extensive experimental results show that our CMANet achieves outstanding accuracies (90.24% on BU-3DFE and 89.36% on Bosphorus) and outperforms the state-of-the-art methods.

Index Terms— 2D+3D FER, Soft mask, Multi-modal fusion

1. INTRODUCTION

Facial expression is one of the most important human interaction languages, helping us distinguish emotions from other people. Given an image, the goal of facial expression recognition (FER) is to understand human facial emotions and automatically distinguish different expressions. The development of FER has facilitated various downstream tasks, e.g., human-computer interaction, security, and psychology [1].

Tremendous efforts have been devoted to 2D FER [2]. However, these works only rely on texture maps, thus the illumination and pose changes may decrease the recognition performance significantly. To solve this problem, researchers developed various illumination and pose normalization methods. Nevertheless, the normalized precision of them is not enough for FER, and extra auxiliary preprocessing [3] makes it extremely complicated and time-consuming. Considering the above limitations and benefiting from the robustness of 3D structural information to illumination and pose changes, the combination of 2D texture and 3D depth features begins to be used in FER [1], leading to the 2D+3D FER.

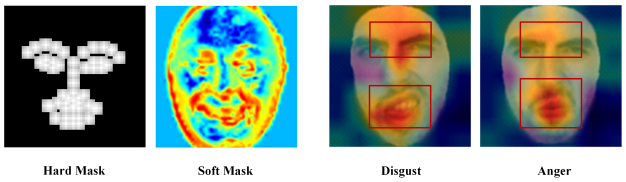
The techniques of 2D+3D FER can be simply divided into two categories: traditional methods and deep learning approaches. The former employ hand-crafted feature extractors or generic 3D deformable models for FER, and the latter apply flexible neural networks to automatically train feature extractors. For example, Li *et al*. [4] first introduced the convolutional neural network (CNN) for 2D+3D FER.

There exist two challenges in 2D+3D FER. First, the critical areas (i.e., eyebrows, eyes, nose, and mouth) that are relevant to the expressions of the face should be effectively focused on. Sui *et al*. [5] and Yang *et al*. [6] suggested cropping the surrounding areas around the landmark points extracted in advance as hard masks, guiding the network to pay attention to the salient regions. These methods depend on the precision of landmarks, and may ignore the potential areas (e.g., the areas between eyebrows) that are crucial for FER. Additionally, they only apply the small-size kernels of CNNs to locate local patches without highlighting non-local interaction that reflects the relationship among significant local regions.

Another important challenge in the 2D+3D FER field is how to integrate the multi-modal information of the same instance. Most existing fusion methods can be usually classified into two types: fusion at the data level and the final concatenation after the fully-connected (FC) layers [7]. The data-level fusion is called "early fusion", which cannot benefit from various advantages of the hetero-modal information due to the premature interaction of the two modalities. On the other hand, the "late fusion" after FC in the high-level semantics is not beneficial for free attention flow between two modalities [8]. Fortunately, Li *et al*. [9] and Nagrani *et al*. [8] proved the effectiveness of the multi-modal mid-to-late fusion, which makes early layers specialize in learning homo-modal patterns, and late layers interact in a relatively loose feature space. Inspired by this, we consider performing direct pixel-level attention fusion of the "mid-to-late" features.

In this work, we propose a curvature-aware soft mask guided attention fusion network (CMANet), as illustrated in

*The corresponding author is Feng Zhao.



(a) (b)

Fig. 1. (a) Compared with the hard mask, our soft mask can pay attention to potentially non-typically significant areas (e.g., the area between eyebrows), and (b) visualization of the heat maps of the disgust and anger expressions.

Fig. 2. First, to extract more effective features, we design the curvature-aware attention (CA) module to focus on the salient areas and the long-range relations between them. As shown in Fig. 1(a), compared with the hard mask, our curvature-aware soft mask (C-Mask) focuses on the salient regions with soft weights and explores other non-typical potentials represented by key areas, such as the area between eyebrows. Besides, we use the non-local module [10] to obtain the long-range enhanced representations. As depicted in Fig. 1(b), our model performs well on classifying easily-confused expression pairs (e.g., disgust and anger), since it focuses on both local and non-local information at the same time. Second, we introduce a novel long-range multi-modal attention fusion (MAF) module to fuse the mid-to-late features, which can not only extract the superior information of hetero-modal effectively, but also make features interact freely in a larger receptive field.

In summary, this work makes the following contributions:

- We propose an attention fusion network named CMANet for 2D+3D FER, achieving state-of-the-art results on both BU-3DFE and Bosphorus datasets.
- We design the curvature-aware attention module that mainly utilizes the curvature-aware soft mask to focus salient regions with soft weights. Then, non-local attention is used to enhance the long-range features further.
- We introduce the multi-modal attention fusion module that performs fusion at the pixel level to fully interact with the multi-modal features, and retains unique information of each modality simultaneously.

2. METHODOLOGIES

2.1. Framework Overview

The architecture of our CMANet is shown in Fig. 2, which can achieve local saliency enhancement and long-range attention at the same time. In order to represent a 3D scan, similar to Sui *et al*. [5], we first use *gridfit* [11] to generate three well-aligned maps: depth map, curvature map, and texture map. Then, a series of preprocessing on these maps is performed including outlier elimination, hole filling, and noise removal. At the feature extraction stage, we choose the widely used VGG16-BN [12] and F3DNet [5] as the backbones. On this basis, we propose the CA and MAF modules, where the former utilizes the curve-aware soft mask to focus on the homo-modal salient areas related to facial expressions, and the latter uses the long-range attention mechanism for integrating hetero-modal features at the pixel level.

2.2. Curvature-aware Attention

Given an image, how to find the most conducive salient areas is a challenge for FER. A common solution is to use masks prepared in advance and multiply them with the original features to locate the key regions. In order to obtain hard masks, it is necessary to apply a landmark detector to get the landmark points, and then crop a small area around these landmarks, as shown in Fig. 1(a). However, this method highly depends on the precision of the detector, and can only label limited areas. To overcome such shortcomings, we design the CA module, which includes C-Mask and non-local attention.

Curvature-aware Soft Mask Module. To provide prior knowledge for the salient areas of the face, Zhang *et al*. [13] found that the areas with high facial curvatures are more likely to have key information about facial features. Fortunately, the development of 3D acquisition equipment allows us to easily obtain 3D scans of human faces, including geometric information. Based on the acquired 3D scan image, we attain the curvature map I_c as the prior information guiding the network to focus on the salient areas, without directly ignoring the potentially non-typically important ones (e.g., the area between eyebrows). As depicted in Fig. 3, we first extract the curvature feature F_c through the convolution with a large-scale kernel that is suitable for extracting geometric information [7], and then take the maximum among all the channels to produce a soft mask M_{soft},

$$M_{soft} = max\{F_{c_1}(x,y), F_{c_2}(x,y), ..., F_{c_C}(x,y)\}, \quad (1)$$

where F_{c_i} represents the features of the i-th channel of $F_c \in R^{C\times H\times W}$, $1 \le i \le C$, $1 \le x \le H$, and $1 \le y \le H$.

After that, the original features are multiplied by M_{soft} to obtain the curve-aware attention features F_{C-Mask},

$$F_{C\text{-}Mask} = F^{(0)} \otimes M_{soft} + f_{d/t}\left(F^{(0)}\right), \quad (2)$$

where $f_{d/t}(\cdot)$ denotes a series of operations including CNNs, activation function (i.e., Softmax and ReLU), and BN (see Fig. 3), and $\otimes$ means matrix multiplication. These features can softly represent both salient areas and potential areas. However, the shallow network can easily extract the contour information, resulting in interference. So, we use several CNN layers to filter it out, and finally get the contour-filtered C-Mask features with soft weights.

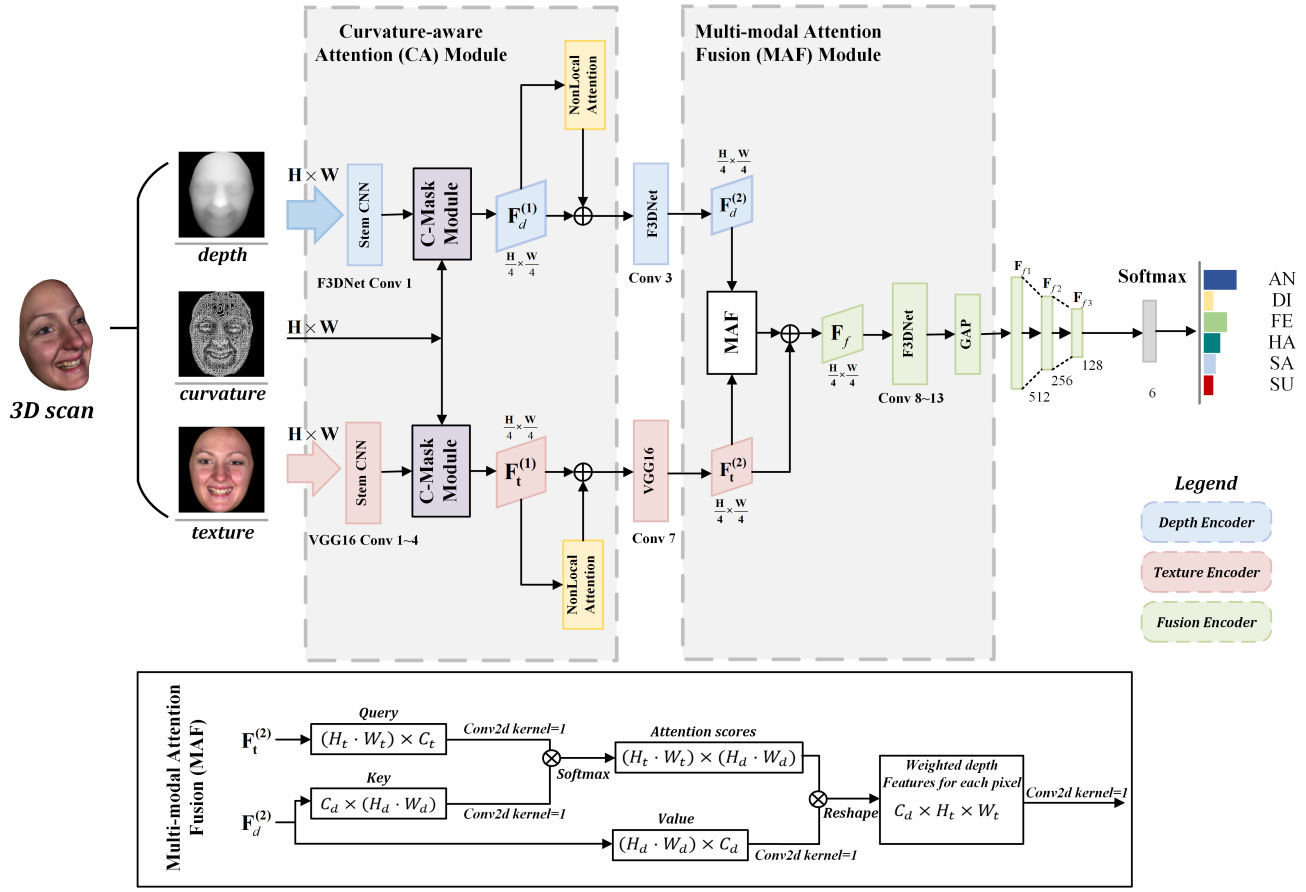


Fig. 2. Overall architecture of our CMANet consists of a curvature-aware attention (CA) module and a multi-modal attention fusion (MAF) module. The details of the MAF module are given at the figure bottom.

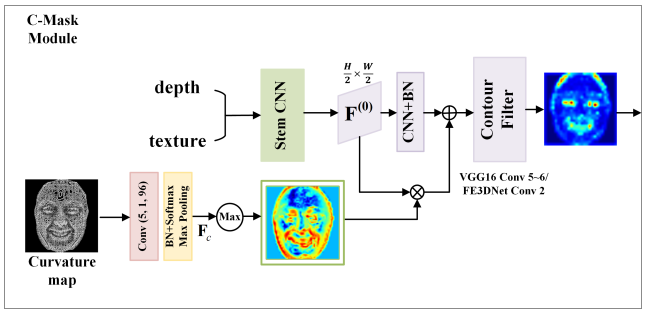


Fig. 3. Architecture of the C-Mask feature generation module.

Moreover, to enhance the characteristics of the homo-modal, considering the limitation of the small kernel size of CNN, we further introduce the non-local module to expand the network view and add the long-range information for promoting the interaction between local regions.

2.3. Multi-modal Attention Fusion

In order to allow multi-modal features to maintain their own unique information and remove redundant information, we consider the pixel-level fusion of mid-to-late features before the FC layers. In addition, to interact between the two modalities from a distant view, we propose a long-range MAF module. Since small feature maps are hard to concern long-range reliance on the location information [10], and considering that fewer channels can reduce the number of parameters, we choose the 28×28 texture feature map F_t and the depth feature map F_d as inputs of the MAF. In addition, the texture features contain certain depth information (texture map can reconstruct depth map, but not vice versa), so F_t is used as the main dictionary to guide the fusion of them.

As described at the bottom of Fig. 2, we first reshape F_t and F_d to obtain query $q \in R^{(H_t \cdot W_t) \times C_t}$, key $k \in R^{C_d \times (H_d \cdot W_d)}$, and value $v \in R^{(H_d \cdot W_d) \times C_d}$. Then, the multi-modal long-range attention scores $A_{td} \in R^{(H_t \cdot W_t) \times (H_d \cdot W_d)}$ are computed as follows:

$$A_{td} = \text{Softmax}\left(\varphi_q(q) \otimes \varphi_k(k) / \sqrt{C_d}\right), \tag{3}$$

where $\varphi_q(\cdot)$ and $\varphi_k(\cdot)$ represent 2D convolution with 1×1 kernel size.

Finally, the weighted depth features for each pixel of the texture feature map can be employed to obtain the final long-

range attention features F_{MAF} by:

$$F_{MAF} = \psi\left(A_{td} \otimes \phi(v)\right), \tag{4}$$

where $\psi(\cdot)$ and $\phi(\cdot)$ denote the 2D convolutions with 1×1 kernel size, and $\psi(\cdot)$ is used to adjust the weights of the MAF module.

3. EXPERIMENTAL RESULTS

In this section, we conduct a series of experiments to prove the effectiveness of our model and compare it with other state-of-the-art methods.

3.1. Datasets and Protocols

Datasets. We choose BU-3DFE [14] and Bosphorus [15] datasets, which are widely used to evaluate 2D+3D FER. BU-3DFE has 100 subjects aging from 18 to 70. Each subject has 25 scans consisting of one neutral expression and six basic expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise) with four different intensities. Bosphorus has 4,666 scans from 105 subjects, and only 63 subjects have complete six basic facial expressions with one intensity.

Protocols. We follow the protocols mentioned in most previous works [5, 7]. In both BU-3DFE and Bosphorus, we randomly select 60 subjects with six basic expression scans (BU-3DFE chooses each expression scan with the third and fourth intensities) as the experimental samples for 10-fold cross-validation. In the experiments, all the subjects are evenly divided into 10 subsets, where one subset is selected as the testing set, and the remaining nine subsets are used as the training set. The cross-validation is repeated 100 times to obtain the final credible result.

3.2. Implementation Details

Given a 3D scan, we resize the generated depth maps, curvature maps, and texture maps into $1\times112\times112$, $1\times112\times112$, and $3\times112\times112$, respectively, and they are used as the inputs of our model. We then use the model parameters pre-trained on the ImageNet [16] to initialize VGG16-BN [12], and train F3DNet [5] from scratch. Following that, the Adam optimizer with betas (0.9, 0.999) is adopted and the learning rate is set to 0.0001 during 70 epochs for optimizing our model. All the experiments are conducted on two NVIDIA GeForce RTX 3070 cards using PyTorch.

3.3. Results

Comparison with Other Methods. Tables 1 and 2 show the comparison results with previous methods on BU-3DFE and Bosphorus datasets, respectively. It can be seen that remarkable performance improvement has been obtained by utilizing multi-modal data and deep learning techniques. In addition,

Table 1. Comparison on BU-3DFE.

Method	Modality	Feature	Acc (%)
Gong *et al*. (2009) [17]	3D	Hand-crafted	76.22
Li *et al*. (2012) [18]	3D	Hand-crafted	80.14
Yang *et al*. (2015) [19]	3D	Hand-crafted	84.80
Li *et al*. (2015) [20]	2D+3D	Hand-crafted	**86.32**
Zhen *et al*. (2016) [21]	3D	Hand-crafted	84.50
Fu *et al*. (2019) [22]	3D	Hand-crafted	82.89
Li *et al*. (2017) [4]	2D+3D	Deep learning	86.86
Chen *et al*. (2018) [23]	2D+3D	Deep learning	86.67
Wei *et al*. (2018) [24]	2D+3D	Deep learning	88.03
Jan *et al*. (2018) [25]	2D+3D	Deep learning	88.54
Zhu *et al*. (2019) [26]	2D+3D	Deep learning	88.35
Ly *et al*. (2019) [27]	2D+3D	Deep learning	87.66
Lin *et al*. (2020) [7]	2D+3D	Deep learning	89.05
Sui *et al*. (2021) [5]	2D+3D	Deep learning	**89.82**
CMANet (Ours)	2D+3D	Deep learning	**90.24**

Table 2. Comparison on Bosphorus.

Method	Data	Feature	Acc (%)
Li *et al*. (2012) [18]	3D	Hand-crafted	75.83
Yang *et al*. (2015) [19]	3D	Hand-crafted	77.50
Li *et al*. (2015) [20]	2D+3D	Hand-crafted	**79.72**
Fu *et al*. (2019) [22]	3D	Hand-crafted	75.93
Li *et al*. (2017) [4]	2D+3D	Deep learning	80.28
Wei *et al*. (2018) [24]	2D+3D	Deep learning	82.50
Vo *et al*. (2019) [28]	2D+3D	Deep learning	82.40
Tian *et al*. (2019) [29]	2D+3D	Deep learning	79.17
Lin *et al*. (2020) [7]	2D+3D	Deep learning	**89.28**
Sui *et al*. (2021) [5]	2D+3D	Deep learning	87.65
CMANet (Ours)	2D+3D	Deep learning	**89.36**

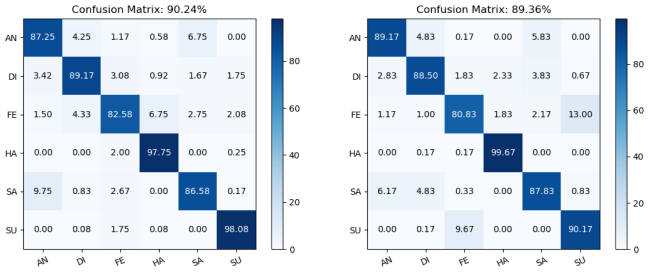
Sui *et al*. [5] achieved the best performance among all previous works on BU-3DFE (see Table 1), while Lin *et al*. [7] performed the best on Bosphorus (see Table 2). Compared with the best two state-of-the-art methods, our CMANet achieves better performance on both datasets. It is worthy to note that the CMANet is 0.42% higher on BU-3DFE and 1.71% higher on Bosphorus than the latest FFNet-M [5], which demonstrates the effectiveness of our model.

Parameter Analysis. We also compare the number of parameters with other state-of-the-art methods. As shown in Table 3, our CMANet attains the best performance with the least number of parameters, indicating that our model is both effective and efficient.

Confusion Matrix. Fig. 4 shows the confusion matrices of the proposed CMANet on BU-3DFE and Bosphorus. We can see that our CMANet achieves excellent performance on happiness and surprise expressions, but poor performance on fear and sadness expressions. One possible reason could be that happiness and surprise can drive substantial changes in

Table 3. Comparison of parameters on BU-3DFE.

Method	Modality	Parameter (MB)	Acc (%)
Jan *et al.* (2018) [25]	2D+3D	$\approx$ 327	88.54
Zhu *et al.* (2019) [26]	2D+3D	$\approx$ 463	88.35
Lin *et al.* (2020) [7]	2D+3D	$\approx$ 137	89.05
Sui *et al.* (2021) [5]	2D+3D	$\approx$ **93**	**89.82**
CMANet (Ours)	2D+3D	$\approx$ **77**	**90.24**



(a) BU-3DFE (b) Bosphorus

Fig. 4. Confusion matrices on BU-3DFE and Bosphorus.

facial muscles that are convenient to be distinguished. However, fear and sadness may be confused easily due to their inconspicuous changes.

3.4. Ablation Studies

To verify the effectiveness of each component incorporated in the CMANet, we evaluate the performance of its two critical modules with different models on both datasets and show the results in Tables 4 and 5. In Table 4, we can see that the non-local module improves the performance of all the models. Based on this, we can find that C-Mask further improves the performance with large margins among 2D texture, 3D depth, and multi-modal features (using VGG16-BN, F3DNet, and CMANet, respectively). It indicates that our C-Mask can pay attention to all the potential information regions that could benefit FER significantly.

In addition, to explore the performance of different fusion strategies, we conduct experiments with several strategies, as reported in Table 5. The early fusion strategy integrates multiple modalities at the data level, which extracts features in a single branch, so we examine the fusion strategy in our single-branch backbone, i.e., VGG16 or F3DNet. On the other hand, late fusion means the concatenation of the multi-modal features after FC. In order to compare these fusion strategies fairly, we add our CA module to each comparison baseline. It can be easily seen that the late fusion strategy performs better than the early fusion strategy. Moreover, our MAF achieves the best performance among all the strategies. The reason is that our fusion strategy can retain the unique advantages from different modalities, while fully interacting with each other in a looser feature space.

Table 4. Ablation study of the CA module on BU-3DFE / Bosphorus.

Model	Non Local	C-Mask	BU-3DFE / Bosphorus (%)
VGG16-BN			85.97 / 81.67
VGG16-BN	✓		86.25 / 82.78
VGG16-BN	✓	✓	**86.81 / 83.61**
F3DNet			82.99 / 80.56
F3DNet	✓		83.68 / 80.97
F3DNet	✓	✓	**84.03 / 81.25**
CMANet (Ours)			88.75 / 86.67
CMANet (Ours)	✓		89.03 / 87.83
CMANet (Ours)	✓	✓	**90.24 / 89.36**

Table 5. Ablation study of different fusion strategies on BU-3DFE / Bosphorus.

Model	Fusion Strategy	BU-3DFE / Bosphorus (%)
VGG16 + CA	Early Fusion	88.19 / 87.92
F3DNet + CA	Early Fusion	88.06 / 87.89
CMANet (Ours)	Late Fusion	**89.72 / 88.06**
CMANet (Ours)	MAF	**90.24 / 89.36**

4. CONCLUSION

In this work, we develop a curvature-aware soft mask guided attention fusion network for 2D+3D FER, which consists of two main modules. The curvature attention module is designed for attending to all the potential regions softly, which may be ignored by traditional methods using hard mask attention. The multi-modal attention fusion module is introduced for combining the pixel-level mid-to-late features, which can retain the unique virtues of different modalities and also promote the fusion of various features in the long-range field. Extensive experimental results demonstrate that our method achieves state-of-the-art performance on both BU-3DFE and Bosphorus datasets, reaching 90.24% and 89.36%, respectively.

5. ACKNOWLEDGMENTS

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12, and the JKW Research Funds under Grant 20-163-14-LZ-001-004-01. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. The authors also would like to thank Yuanhang Zhou at Renmin University of China and Huatian Zhang at University of Science and Technology of China for their discussions on the experiment design.

6. REFERENCES

[1] E. Ahmed et al., "Deep learning advances on different 3D data representations: A survey," *arXiv e-prints*, p. arXiv:1808.01462, 2018.

[2] A.F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: A survey," *Pattern Recognit. lett.*, vol. 28, no. 14, pp. 1885–1906, 2007.

[3] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3601–3610.

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

[6] H. Yang and L. Yin, "CNN based 3D facial expression recognition using masking and landmark features," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 556–560.

[7] S. Lin et al., "Orthogonalization-guided feature fusion network for multimodal 2D+3D facial expression recognition," *IEEE Trans. on Multimedia*, vol. 23, pp. 1581–1591, 2020.

[8] A. Nagrani et al., "Attention bottlenecks for multimodal fusion," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[9] J. Li and G. Lee, "DeepI2P: Image-to-point cloud registration via deep classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15960–15969.

[10] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[11] S. Zulqarnain Gilani and A. Mian, "Learning from millions of 3D scans for large-scale 3D face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1896–1905.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[13] Z. Zhang, F. Da, and Y. Yu, "Data-free point cloud network for 3D face recognition," *arXiv e-prints*, p. arXiv:1911.04731, 2019.

[14] 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.

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

[16] J. Deng et al., "ImageNet: A large-scale hierarchical image database," vol. 25, pp. 1097–1105, 2012.

[17] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3D face by exploring shape deformation," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 569–572.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] Z. Chen, D. Huang, Y. Wang, and L. Chen, "Fast and light manifold CNN based 3D facial expression recognition across pose variations," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 229–238.

[24] X. Wei, H. Li, J. Sun, and L. Chen, "Unsupervised domain adaptation with regularized optimal transport for multimodal 2D+3D facial expression recognition," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2018, pp. 31–37.

[25] 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.

[26] 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.

[27] T.S. Ly et al., "A novel 2D and 3D multimodal approach for in-the-wild facial expression recognition," *Image Vis. Comput.*, vol. 92, pp. 103817, 2019.

[28] Q. N. Vo, K. Tran, and G. Zhao, "3D facial expression recognition based on multi-view and prior knowledge fusion," in *Proc. IEEE Int. Workshop Multimed. Signal Process.*, 2019, pp. 1–6.

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