2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9666982

Toward Personalized Emotion Recognition: A Face Recognition Based Attention Method for Facial Emotion Recognition

Mostafa Shahabinejad[1], Yang Wang[1,2], Yuanhao Yu[1], Jin Tang[1], Jiani Li[1]
[1]Noah's Ark Lab, Huawei Technologies [2]University of Manitoba, Canada

***Abstract*— This paper aims to address the subject-dependent challenge of the facial emotion recognition (FER) task. To accomplish this, we propose a novel face recognition based attention FER (FRA-FER) framework which propagates subtle face recognition (FR) features through the FER network. Particularly, first a spatial attention map from the feature maps of an FR convolutional neural network (CNN) is created and then it is fused into the FER-CNN. By doing this FR feature propagation, the FER network is personalized as it takes the advantage of the FR features learned from large-scale face recognition datasets. Experiments on the two challenging datasets AffectNet and AFEW demonstrate the superiority of our proposed FRA-FER network to the state-of-the-art work.**

I. INTRODUCTION

Emotion recognition has a variety of applications, e.g., in human-computer interaction [1], [2], mental diseases diagnosis [3], avatars and computer animations [4], and more [5]. Several automatic expression recognition methods based on speech [6], face [7], audio-visual [8], context [9], and multimodal sensor data [5] have been proposed. The focus of this work is on the expression recognition from facial features.

Convolutional neural network (CNN) based methods have made a lot of progress in automatic facial expression recognition (FER) [10], [11], [12], [13], [14], [15], [16], yet accurate facial expression recognition remains challenging because different individuals may have different ways of expressing the same emotion. There are many attempts in the FER literature to address the subjective notion of the emotion expression, e.g., see [17], [7], [18], [19], [20], [21] and references therein. To tackle the subject variation challenge, these methods use either a loss function to learn identity-related information [18], [22], [23] or generative adversarial networks (GANs) to generate expressions from facial images [20], [21]. The loss-based methods are dependent on the identity-related images and perform weakly when identities are limited, while the GAN-based methods depend largely on the unrealistic generated images for the classifier which result in a poorer expression recognition [17].

Unlike the previously proposed identity-robust FER methods, we consider fusing face recognition (FR) features into the FER network in order to personalize FER systems (Fig. 1). Because FR facial features learned from enormous identities from in-the-wild datasets contain subtle, rich information about individuals' faces, we incorporate facial features from the FR-CNN and propagate them through the FER-CNN. More specifically, as shown in Fig. 1, a face image is passed through a trainable FER network and a frozen FR network. The features from the frozen FR-CNN are then used to create a spatial attention map which is later applied to the features of the FER-CNN. This leads to an improvement of the FER system with a negligible computation cost.

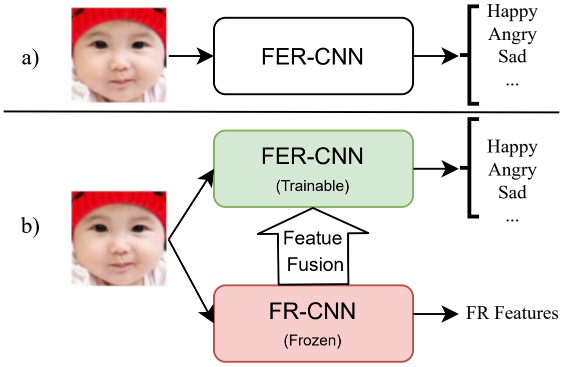

Fig. 1. a) A non-personalized facial expression recognition model vs. b) our proposed personalized facial expression recognition model composed of a trainable FER network and a frozen FR network. The personalized FER model leverages features of a frozen face recognition network which leads to the state-of-the-art performance.

To the best of our knowledge, our work is the first to incorporate FR features to personalize the FER system. We show in the experiment section that this leads to a performance boost of 1.98% for the challenging AffectNet dataset [24] leading to the state-of-the-art results. Note that this performance improvement is achieved without increasing the capacity of the backbone network as the newly added FR-CNN parameters are frozen. It is also worth mentioning that the state-of-the-art results achieved by our method require only a minimal increase in computational cost compared to the backbone network because FR features are extracted from a very early layer of the FR-CNN.

The rest of the paper is organized as follows. In Section II, we summarize the related work. In Section III, we present details of our personalized FER networks. The experimental evaluation is conducted in Section IV. The conclusion is discussed in Section V of the paper.

II. RELATED WORK

Three general approaches can be considered in the research area of automatic facial emotion recognition. First, categorical models assign discrete emotion labels to the subject. Second, dimensional models describe emotions with two

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

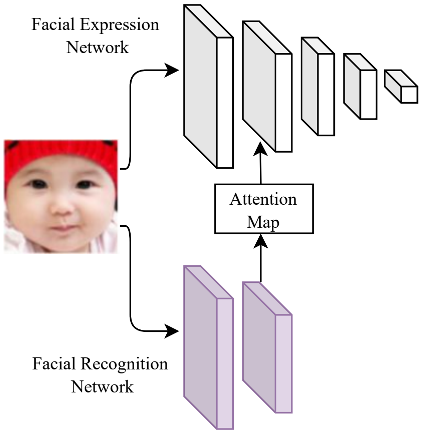

Fig. 2. The proposed network: upper and lower networks represent a trainable FER-CNN and a frozen FR-CNN, respectively. Features of the FR-CNN are used to generate a spatial attention map for the FER network which leads to a personalized FER system and improves the FER system performance.

(valence-arousal [25]) or three (pleasure-arousal-dominance [26]) continuous values. Third, facial action coding system (FACS) models detect the presence and the strength of different action units defined by FACS [27]. For recent surveys, please refer to [7], [28].

Here, we briefly review the literature that is most relevant to our work, namely categorical FER and FR. We first overview categorical FER. We then mention the FR state-of-the-art methods and the advantage of using them in the FER system.

A. Categorical FER

Categorical FER models assign emotion labels such as *angry*, *happy*, *sad*, etc. to the face image. The subject-dependent nature of the expression recognition makes the task challenging. There are several works addressing the subject-dependent issue of the FER which can be categorized into two general groups. The first group tries to learn identity along with expression through an identity loss via multi-task learning, while the second group tries to synthesize identity-based images through generative adversarial networks (GANs). In the following, we briefly review these methods.

In [18], an identity-sensitive contrastive loss is used to learn identity information. Their proposed method has two identical sub-networks, where one network is responsible to learn expression-discriminative features and the other one is responsible to learn identity-related features. The contrastive loss of this method suffers from drastic data expansion during image pairs construction from training data [17]. In [22], a multisignal CNN is proposed where the training is done under supervision of both face verification and expression recognition tasks in order to make the model focus on the expression information. In [23], an adaptive deep metric learning method is proposed where an adaptive triplet loss along with an identity-aware hard-negative mining and an online positive mining method is used to make the facial expression recognition identity-robust. The triplet loss of this method depends on identity-related image pairs and its performance degrades in case of limited identities.

There is also work on using GAN to address the subject-dependent challenge of FER. In [17], an adversarial feature extraction method is proposed to resolve the issue of disturbance corresponding with identity and pose variation. However, this method is largely dependent on the label in disturbing factors formation. In [20], an identity-adaptive based on conditional GAN is proposed to generate images of the same person with different expressions. Then a regular CNN is used to fine-tune for the expression recognition task. In [21], a generative adversarial recognition model is proposed to enlarge the facial images similar to the target domain with comparable action unit patterns of the source domain. These GAN-based networks depend largely on the unrealistic generated images for the classifier which in turn results in a weaker expression recognition.

Although all these approaches have shown great performance, they do not take into consideration the face recognition feature as a tool to satisfy the subjective-dependent nature of the facial expression problem. In Section IV, we show that propagating face recognition feature through the FER network leads to the state-of-the-art performance.

B. Face Recognition:

Face recognition (FR) is a problem related to FER. Current FR systems have achieved a very high accuracy. For example for the in-the-wild Megaface dataset which contains 1M images from 690K individuals with no constraint on expression, pose, and exposure [29], Arcface network [30] achieves an accuracy of $\%98.47$. In this paper, we take advantage of the pretrained FR models learned from large-scale face recognition datasets to improve FER.

III. PROPOSED METHOD

In this section, we present our proposed personalized FR-based attention FER (FRA-FER) model. We first formally define the problem and overview the architecture of the FRA-FER in Section III.A. Then, we present details of the FRA-FER network in Section III.B.

A. Architecture Overview

Here, we overview the architecture of our proposed personalized FRA-FER CNN. Given a facial emotion recognition dataset $\mathcal{D} = \{(I_i, y_i)\}_{i=1}^{N}$ with N images, where each image $I_i \in \mathcal{I} = \{I_1, I_1, \cdots, I_N\}$ has a spatial resolution of $3 \times H \times W$ and is labeled from C predefined emotion classes $y_i \in \mathcal{Y} = \{0, 1, \cdots, C-1\}$, our goal is to personalize the FER classifier $\mathcal{F}_{FER} : \mathcal{I} \to \mathcal{Y}$ by propagating subtle FR features through the classifier. To this end, as depicted in Fig. 2, first the face image I is passed through an $\mathcal{F}_{FR}$ network, where $\mathcal{F}_{FR}$ represents a frozen CNN which has learned face embedding through an FR database. Then a 3D feature map is extracted from some hidden layers of the FR-CNN. From the extracted 3D feature map, a 2D spatial attention map is constructed which is then applied to the feature map of a trainable FER-CNN. Using this method, features of the FR-CNN are propagated through the FER-CNN which in turn improves the accuracy of the FER classifier. In the following, we present details of our methodology.

B. *FRA-FER Network*

Here, we present the details of our proposed personalized FRA-FER CNN. For a face image I, let

$$H_{FR_1}, \cdots, H_{FR_L} = \mathcal{F}_{FR}(I; \theta_{FR}), \tag{1}$$

where $\mathcal{F}_{FR}$ with the frozen parameters θ_{FR} is an FR-CNN with L layers and H_{FR_l} is the output of the l-th layer of the network with $l \in \{1, \cdots, L\}$. Similarly, let

$$H_{FER_1}, \cdots, H_{FER_{L'}} = \mathcal{F}_{FER}(I; \theta_{FER}), \tag{2}$$

where $\mathcal{F}_{FER}$ with trainable parameters θ_{FER} represents a FER classifier with L' layers and $H_{FER_{l'}}$ is the output of the l'-th network layer with $l' \in \{1, \cdots, L'\}$.

As depicted in Fig. 3, in order to propagate the FR features through the FER networks, we create a spatial attention map from the p-th feature map of the FR-CNN and apply it to the q-th feature map of the FER-CNN as follows. Let $H_{FR_p} \in \mathbb{R}^{C^p \times H^p \times L^p}$ be the p-th feature map of FR-CNN, where C^p, H^p, and W^p represent number of channels, height, and width of the feature map H_{FR_p} respectively. Similarly, let $H_{FER_q} \in \mathbb{R}^{C^q \times H^q \times L^q}$ be the q-th feature map of FER-CNN, where C^q, H^q, and W^q represent number of channels, height, and width of the feature map H_{FER_q} respectively. First, the spatial attention map $S_{FR_p^{sq}} \in \mathbb{R}^{1 \times H^q \times L^q}$ is generated as follows:

$$S_{FR_p^{sq}} = Interpolate(Sigmoid(Conv(H_{FR_p}))), \tag{3}$$

where $Conv$, $Sigmoid$, and $Interpolate$ stand for a 2D convolution with C^p kernels of size 1×1, the sigmoid function, and a 2D resizing operation, respectively. This spatial attention map is then applied channel-wise to the q-th feature map of the FER-CNN as follows:

$$\hat{H}_{FER_q^{att}} = Mult(S_{FR_p^{sq}}, H_{FER_q}), \tag{4}$$

where $Mult$ represents the spatial channel-wise multiplication and $\hat{H}_{FER_q^{att}} \in \mathbb{R}^{C^q \times H^q \times L^q}$ represents the feature map H_{FER_q} with the applied attention map. The final feature map $H_{FER_q^{att}} \in \mathbb{R}^{C^q \times H^q \times L^q}$ is obtained as

$$H_{FER_q^{att}} = Max(\hat{H}_{FER_q^{att}}, H_{FER_q}), \tag{5}$$

where Max represents the element-wise maximum operation. Then the FER-CNN proceeds with $H_{FER_q^{att}}$ to generate FER features.

We note that the parameters θ_{FR} are learned offline using an FR databases and frozen in our FRA-FER network, while the parameters θ_{FER} are trainable. The pretrained FR models capture information about a person's identify. By propagating the learned FR features through the FER network as shown in Eq. (4), we achieve a FER-CNN specific to each person.

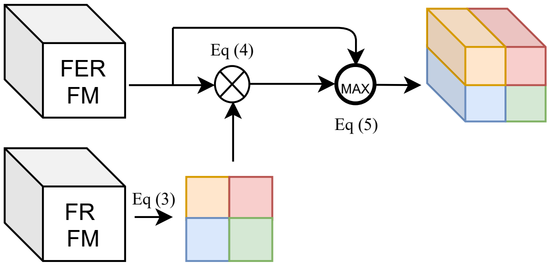

Fig. 3. The proposed method for personalizing a FER feature map (FER-FM): a squeeze map is generated from the FR feature map (FR-FM) and applied to the FER-FM spatially.

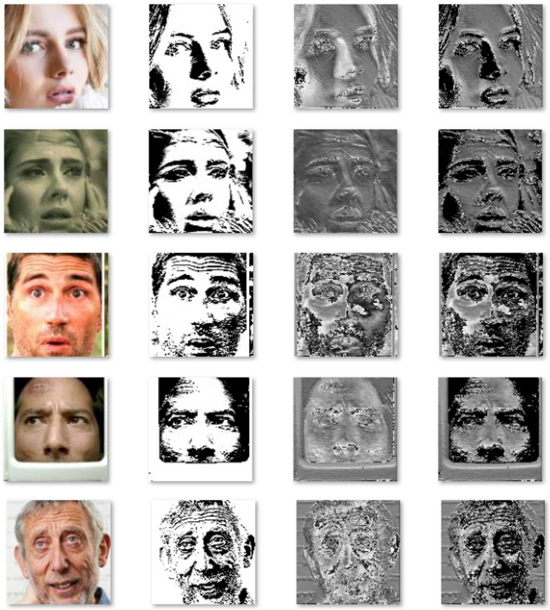

Fig. 4. Visualization of the spatial attention map and feature map of our proposed model. The first, the second, the third, and the forth columns show the input image, the spatial attention mask obtained from FR feature maps, the 2D mean of FER feature maps before applying the attention map, and the 2D mean of FER feature maps after applying the attention map, respectively.

Fig. 4 shows five image samples of the AffectNet dataset [24] along with their correponding features. In this figure, the first and the second columns show the input image and the attention map, respectively. And the third and the forth columns show the spatial average of the FER feature map before and after applying the spatial attention map, respectively. As this figure shows, FER feature maps with the spatial attention (the forth column) provide more subtle features compared to the FER features with no spatial attention (the third column).

IV. EXPERIMENTS

In this section, we present our experiment results. To have a fair comparison, we adapt the EfficientNet-B0 network from the state-of-the-art work in [14] as the backbone network and follow their training and evaluation procedures.

For both FR and FER networks, we used the same EfficientNet-B0 network [31], [14]. We have found that getting the attention map from an early layer of the FR network and applying it to an early layer of the FER network

TABLE I

Comparison of Accuracy Percentage for the Validation Set of the AffectNet Dataset for 8 and 7 Classes.

Method	8 Classes	7 Classes
PSR [10]	60.68	-
DACL [11]	-	65.20
PAENet [15]	-	65.29
ARM (ResNet-18) [33]	61.33	65.2
Distilled Student [8]	61.60	65.4
SL + SSL in-panting-pl (B0) [16]	61.72	-
EfficientNet-B0 [14]	61.32	65.74
FRA-FER (ours)	**63.28**	**66.4**

is more beneficial for the expression recognition task. Getting features from an early layer of the FR network has also the advantage of less computation. We get the $H_{FER_q^{att}}$ from (5) by setting p to 1 in (3) and apply it using (4) by setting q to 1. This results in a spatial attention map $S_{FR_1^{sq}}$ of size $1 \times 112 \times 112$ and a FER feature map $H_{FER_1^{att}}$ of size $32 \times 112 \times 112$ for EfficientNet-B0 with the face image of size $3 \times 224 \times 224$. In our experiments, we use the same train parameters as in [14]: 6 epochs, batch of size 8, and 3×10^{-5} flat learning rate.

We report accuracy of the validation sets corresponding to two challenging datasets AffectNet [24] and acted facial expression in-the-wild (AFEW) [32]. In the following, we first present our results on these two datasets and then show the ablation study on the AffectNet dataset.

A. Results of AffectNet Dataset

AffectNet is the largest emotion recognition database containing more than 1M facial images gathered from the Internet. For the categorical emotion model of this dataset, there are more than 287K images for the train set and 8K images for the validation set (500 images per class) with 8 discrete classes: *Anger*, *Contempt*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad*, and *Surprise*. Table I shows the validation accuracy results for the AffectNet dataset where our proposed personalized FRA-FER network achieves the state-of-the-art results for both 8 and 7 classes[1].

B. Results of AFEW Dataset

The AFEW dataset contains 773 training and 383 validation samples collected from TV series and movies. There are 7 emotion classes corresponding with this dataset: *Anger*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad*, *Surprise*. For the temporal feature extraction, we follow the same steps as in Section 3.3 of [14]. Table II shows the validation accuracy results for the AFEW dataset where our proposed personalized FRA-FER network achieves the state-of-the-art performance.

TABLE II

Comparison of Accuracy Percentage for the Validation Set of the AFEW Dataset for 7 Classes.

Method	7 Classes
DenseNet-161 [12]	51.44
Noisy student w/o iterative training [13]	52.49
Noisy student with iterative training [13]	55.17
MobileNet-v1 [14]	55.35
RexNet-150 [14]	57.27
EfficientNet-B0 [14]	59.27
FRA-FER (ours)	**60.83**

TABLE III

Ablation Study: Comparison of Accuracy Percentage for the Validation Set of the AffectNet Dataset for 8 Classes.

Method	8 Classes
FR-FER feature concatination	34.18
No FR fusion [14]	61.32
Self-attention-CSCSSE [34]	58.8
Self-attention-CSSE [34]	59.83
Self-attention-SSCE [35]	61.10
FRA-FER (ours)	**63.28**

C. Ablation Study

Here, we present our ablation study by reporting the accuracy of the validation set of the AffectNet dataset with 8 emotion classes. More specifically, we present results of no FR fusion, FR-FER feature concatination and three self attention methods of the spatial squeeze and channel excitation (SSCE) [35], channel squeeze and spatial excitation (CSSE) [34], and concurrent spatial and channel squeeze and spatial excitation (CSCSSE) [34]. As table III shows, creating the attention map from the features of the FR network using our proposed method outperforms other fusion and self-attention methods.

V. Conclusion

In this work, we have proposed a novel way of applying attention mechanism to the facial expression recognition convolutional neural network (FER-CNN). More specifically, rather than getting the attention map from the same FER-CNN, we propose to generate the attention map from a face recognition (FR) CNN. By doing this, we can make the FER network personalized by propagating the subtle FR features through the FER network. We have demonstrated that this approach results in the state-of-the-art performance for the two challenging datasets AffectNet and AFEW.

[1] The 7 classes of AffectNet dataset are defined as *Anger*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad*, and *Surprise*.

References

[1] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[2] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639, 2015.

[3] Simon Wallace, Michael Coleman, and Anthony Bailey. An investigation of basic facial expression recognition in autism spectrum disorders. *Cognition and Emotion*, 22(7):1353–1380, 2008.

[4] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer Vision*, pages 143–157. Springer, 2014.

[5] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors*, 19(8):1863, 2019.

[6] Md Shah Fahad, Ashish Ranjan, Jainath Yadav, and Akshay Deepak. A survey of speech emotion recognition in natural environment. *Digital Signal Processing*, page 102951, 2020.

[7] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[8] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 2021.

[9] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019.

[10] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020.

[11] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2402–2411, 2021.

[12] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 630–634, 2018.

[13] Vikas Kumar, Shivansh Rao, and Li Yu. Noisy student training using body language dataset improves facial expression recognition. In *European Conference on Computer Vision*, pages 756–773. Springer, 2020.

[14] A. Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. *ArXiv*, abs/2103.17107, 2021.

[15] Steven CY Hung, Jia-Hong Lee, Timmy ST Wan, Chein-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 339–343, 2019.

[16] Mahdi Pourmirzaei, Farzaneh Esmaili, and Gholam Ali Montazer. Using self-supervised co-training to improve facial representation. *arXiv preprint arXiv:2105.06421*, 2021.

[17] Can Wang, Shangfei Wang, and Guang Liang. Identity-and pose-robust facial expression recognition through adversarial feature learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 238–246, 2019.

[18] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE, 2017.

[19] Yanwei Li, Xingang Wang, Shilei Zhang, Lingxi Xie, Wenqi Wu, Hongyuan Yu, and Zheng Zhu. Identity-enhanced network for facial expression recognition. In *Asian Conference on Computer Vision*, pages 534–550. Springer, 2018.

[20] Huiyuan Yang, Zheng Zhang, and Lijun 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)*, pages 294–301. IEEE, 2018.

[21] Can Wang and Shangfei Wang. Personalized multiple facial action unit recognition through generative adversarial recognition network. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 302–310, 2018.

[22] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.

[23] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 20–29, 2017.

[24] 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*, 10(1):18–31, 2017.

[25] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[26] Albert Mehrabian. Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies. 1980.

[27] W. V. Friesen P. Ekman and J. C. Hager. Facial action coding system - manual. 2002.

[28] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019.

[29] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[30] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[31] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[32] Abhinav Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pages 546–550, 2019.

[33] Jiawei Shi and Songhao Zhu. Learning to amend facial expression representation via de-albino and affinity. *arXiv preprint arXiv:2103.10189*, 2021.

[34] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 421–429. Springer, 2018.

[35] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.