FAMGAN: Fine-grained AUs Modulation based Generative Adversarial Network for Micro-Expression Generation

Yifan Xu
University of Science and Technology of China
Hefei, AnHui, China
xyf0103@mail.ustc.edu.cn

Sirui Zhao
University of Science and Technology of China
Hefei, AnHui, China
sirui@mail.ustc.edu.cn

Huaying Tang
University of Science and Technology of China
Hefei, AnHui, China
iamthy@mail.ustc.edu.cn

Xinglong Mao
University of Science and Technology of China
Hefei, AnHui, China
maoxl@mail.ustc.edu.cn

Tong Xu*
University of Science and Technology of China
Hefei, AnHui, China
tongxu@ustc.edu.cn

Enhong Chen*
University of Science and Technology of China
Hefei, AnHui, China
cheneh@ustc.edu.cn

ABSTRACT

Micro-expressions (MEs) are significant and effective clues to reveal the true feelings and emotions of human beings, and thus MEs analysis is widely used in different fields such as medical diagnosis, interrogation and security. However, it is extremely difficult to elicit and label MEs, resulting in a lack of sufficient MEs data for MEs analysis. To address this challenge and inspired by the current face generation technology, in this paper we introduce Generative Adversarial Network based on fine-grained Action Units (AUs) modulation to generate MEs sequence (FAMGAN). Specifically, after comprehensively analyzing the factors that lead to inaccurate AU values detection, we performed fine-grained AUs modulation, which includes carefully eliminating the various noises and dealing with the asymmetry of AUs intensity. Additionally, we incorporate super-resolution into our model to enhance the quality of the generated images. Through experiments, we show that the system achieves very competitive results on the Micro-Expression Grand Challenge (MEGC2021).

CCS CONCEPTS

• **Computing methodologies → Computer vision problems**.

KEYWORDS

micro-expression, action units, generative adversarial network, Micro-Expression Grand Challenge

ACM Reference Format:
Yifan Xu, Sirui Zhao, Huaying Tang, Xinglong Mao, Tong Xu, and Enhong Chen. 2021. FAMGAN: Fine-grained AUs Modulation based Generative Adversarial Network for Micro-Expression Generation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, 5 pages.

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00

1 INTRODUCTION

Micro-expressions (MEs) are involuntary movements of the face that occur spontaneously in a high-stakes environment [6]. Since MEs contain large amounts of significant and effective information about the genuine emotions, automatic MEs analysis has many potential applications such as treatment of depression, business negotiation, interrogations, and security [5, 16, 21]. Recently, automatic MEs analysis has attracted increasing attention of computer vision researchers. Especially in MEs recognition task and MEs spotting task, various models [8, 11, 12, 15, 18–20, 26, 27] based on deep neural networks have been proposed.

Generally, MEs analysis is data-driven, which means we need sufficient MEs samples to implement efficient deep learning models with good generalization ability. However, the existing public MEs datasets [4, 9, 23, 24] usually contain few-shot samples, which restricts the development of both MEs spotting and recognition tasks. Therefore, the construction of large-scale MEs dataset is urgent and significant.

As a matter of fact, there are three major challenges in obtaining large-scale MEs samples. At first, it's difficult to induce MEs, which is mainly because different individuals usually have different reactions to same eliciting mechanism. Secondly, due to the short duration and subtle facial movements of MEs, manually annotating MEs data is an extremely time-consuming and laborious task. Last but not least, the emotion labels of MEs are usually difficult to be accurately labeled, since some facial reactions are difficult to perceive even by well-trained experts.

To address the above challenges, data augmentation is naturally a low-cost and effective approach, which is widely used to enrich the dataset when facing the scarcity of samples. For data augmentation, generating new data is proved to be effective, and the generative adversarial network (GAN) is usually used to generate new samples for data augmentation in many other computer vision tasks. Typical tasks include face generation, facial macro-expression generation, which are simliar to MEs generation. Thereby, a natural idea is to use models associated with face generation for MEs generation.

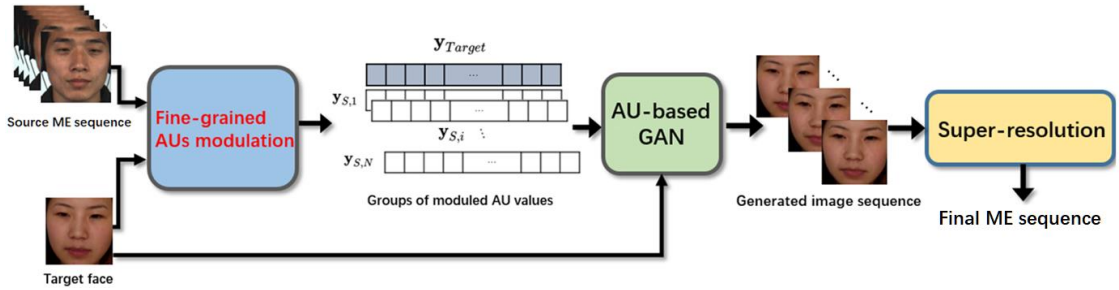

Figure 1: The overall framework of our FAMGAN, including Fine-grained AUs Modulation, AU-based GAN, and Super-resolution.

However, for the face generation, although many effective models such as StarGan [2], RSGAN [13], FSGAN [14] and IPGAN [1] are proposed, they are not good at generating expressions. For the facial macro-expression generation, typical methods include GANimation [17] and AU-ICGAN [22]. GANimation focuses on a continuous manifold of anatomical facial movements. It leans a generator to edit AU intensity of the target face. AU-ICGAN proposes a generation method for data augmentation based on Action Units(AUs). Inspired by GANimation, it enriches the discriminator with video-quality evaluating, allowing it to generate nearly indistinguishable image sequences. However, these models' performance still relies highly on accurate AU values, and without experts' annotation, AU values often have noises.

In this paper, we introduce a GAN based on fine-grained AUs modulation to generate MEs sequences (FAMGAN) with different magnitudes of AU values. Specifically, to generate MEs sequence with continuous movement change, we propose a fine-grained modulation method to fine-tune the AU values. Generally, our FAMGAN mainly includes three key modules: fine-grained AUs modulation, AU-based GAN, and super-resolution. Through experiment, we show that the proposed FAMAGN could generate robust and natural facial micro-expression with fine-tuned AU values. Our code for this challenge is available on https://github.com/QAQFrank/MEGC2021.

2 METHODOLOGY

2.1 Framework Overview

To achieve the objective we stated before and inspired by the GANimation [17], we propose a combined system named FAMGAN, which is based on the fine-grained AUs modulation and AU-based GAN to generate the final MEs sequence. The overall structure is illustrated in Fig. 1, including three key modules: fine-grained AUs modulation, AU-based GAN, and super-resolution. Specifically, the fine-grained AUs modulation aims at extracting the precise AUs based on the OpenFace [1]. The AU-based GAN module is based on GANimation [17], which could generate human face with new AU intensity. For super-resolution network, we introduce it to make our system generate much clearer face image with few blur.

[1]https://github.com/TadasBaltrusaitis/OpenFace

2.2 AUs extraction and modulation

To obtain the precise AU values of each frame in the MEs video, we first use MTCNN [25] and CLM [3], which are integrated into the Openface tool to detect the 68 landmarks of human face in each frame of MEs sequence and calculate the AU values. In this paper, we select 17 AU values outputted by Openface as a vector $\mathbf{y}_r = (y_1, \ldots, y_N)^\top$. Since the resolution of the facial image is always not high and the cropped images drop some boundaries of human face, the AU values we get are usually not accurate, which reflects in:

- A small change on image dimension sometimes could cause huge change on AU values.
- Similar images have totally different AU values.
- For neutral faces of different individuals, AU value obtained by AU detector is often not zero, which does not square with the facts that neural faces' AU values should be zero.
- Facial movement is usually asymmetrical, while Openface can not record both left and right AU values.

Therefore, fine-grained AUs modulation is necessary. To achieve this mind, we successively give our solutions and methods, as follows:

- In order to eliminate the error caused by the image size, we first resize the original image to four similar sizes, then detect their AU values. In detail, for an image with size $H \times W$, we get four different dimension images by adjusting the height and width with $\Delta \in \{-10, -5, 0, 5, 10\}$, and obtain the corresponding AU values. Formally, we could calculate the average AU values as the new one:

$$y_{oi} = \frac{\sum_{j=1}^{5} y_{ji}}{5} \tag{1}$$

  , where $i \in \{0, 1, \ldots, 16\}$, and y_{ji} means the $i-th$ value of AU vector in the $j-th$ adjusted image.
- As mentioned above, in some worse cases, a very subtle change on face could cause unacceptable error on AU values. This could break the continuity of the video and AU values of expression. What we need to do is to reduce the noise, and keep the entire trend of AU values. Accordingly, we use the mean filter to eliminate the noise in the AU sequence:

$$x_i = \frac{\sum_{j=i-n}^{i+n} y_i}{2n+1} \tag{2}$$

$i \in \{n, n+1, \ldots, 16-n\}$, and y_i represents initial AU values.

- Ideally, the AU value of the neural face image should be all zeros, or values near zero. We first multiple the AU value of target face by a small coefficient α, which makes the AU value close to zero. Then, noticed that the onset frame of a ME sequence can be regarded as a neural frame, we can get relative AU changes from $frame_1$ to $frame_i$. Formally, we calculate the new source AU sequence by:

$$y'_{s,i} = \alpha \cdot y_t + (y_{s,i} - y_{s,i-1}) \tag{3}$$

where $y_{s,i}$ denotes the AU value of the i^{th} frame of source sequence, and y_t represents the AU value of target face. By calculating the new AU sequence, we eliminate the impropriety caused by the identity information.

- Since facial movement is usually asymmetrical, it is necessary to record AUs of left face and right face independently. To achieve this mind, we generate two new images with left and right part of human face independently. Two series of new images form two AU sequences, and we use them to generate new face sequences respectively. We will concatenate two results in the end. Fig.2 shows how we generate two faces and some subsequent management.

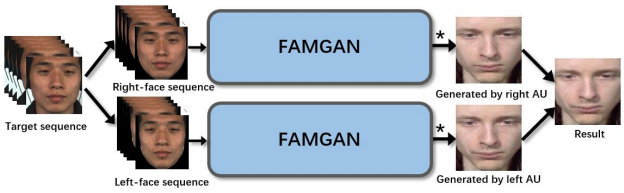

Figure 2: Solution for asymmetry of human face. We use the apex frame to illustrate our method.

2.3 AU-based GAN for MEs generation

For MEs generation, we choose Ganimation [17] as our backbone network, which could get impressive results in the facial expression generation task. In detail, GANimation leverages the auto encoder-decoder structure with attention mechanism, and only requires facial images annotated with their activated AUs. Although its performance on video generation is not satisfactory, Ganimation is still a suitable choice for MEs sequence generation.

2.4 Super-resolution

Although the generated facial MEs images of our AU-based GAN have high performance on realism and consistency, the modified parts of the human face are still slightly blurred. Besides, we need to resize the facial images' dimension to fit the model's input, which will inevitably lose some facial information. To recover the image back to the original size and eliminate the blurred part of generated images, we adopt a state-of-the-art super-resolution method, i.e., RealSR [7]. RealSR designs a novel degradation framework by estimating various blur kernels as well as real noise distributions, and outputs high-resolution images with lower noise and better visual quality.

3 EXPERIMENTS

To verify the effectiveness of our proposed method, we carried out a series of comprehensive experiments. In this section, we first introduce the experimental datasets, which are provided by Facial Micro-Expression Challenge 2021. Then, we discuss how to evaluate the generated MEs sequences. At last, We present our experimental results from a subjective perspective.

3.1 Experimental Datasets

For this challenge, three spontaneous facial MEs datasets are used, including CASME II [23], SAMM [4], and SMIC [10] datasets. Furthermore, three videos with three different kinds of emotion labels (positive, negative, surprised) are specified from each dataset.

3.2 Evaluation Metric

Evaluating GAN models objectively is hard, especially in MEs generation task. In face generation and face reenactment tasks, new methods are evaluated by the quality of the generated picture and whether the results meet requirements subjectively. In our generation task, generated images will be evaluated based on the quality and action units. The quality of the generated image indicates whether the result looks like a real human face, and AU values show whether the generated image has the same ME with the source video. Furthermore, to judge the overall image quality, noise will also be taken into consideration. Each image will be separated into the upper part and the lower part, and results will be evaluated by three experts.

3.3 Experimental Results and comparison

In this section, we will intuitively illustrate our generated results and discuss them. Due to the limited paper space, here, we present only three main results as shown in Fig.3,4,5, and the other results are provided in the github[2].

Intuitively, the generated MEs sequence in Fig.3 is an Asian woman with a 'negative' expression. Focusing on the apex frame of the generated sequence, the generated image shows that the upper face has a clear symbol of feeling disgusted with AU4 (Brow lowerer) and AU7 (Lid tightener), and the lower face has AU9 (Nose wrinkler) and AU10 (Upper lip raiser), which are also symbols of disgust. Overall, the changes on frames in this generated sequence are clear and smooth. In addition, the intensity of our generated expressions highly matches with those in source videos, and the generated face looks natural and continuous.

Fig.4 and Fig.5 are also the results of our method. Specifically, in Fig.5, the results show the superiority of our method dealing with asymmetry of expression. In this case, the source video shows that when this woman laugh, the left part of her face has more intense expression than her right face, which means the values of AU14 (Dimpler) and AU12 (Lip corner puller) are larger in her left face. Fig.5 shows the result generated separately from her left and right part AU values, and as we expect, the one generated by left part AU values has stronger expression. By combining them together, we get our final result. More details can be checked in the videos we uploaded to the github.

[2]https://github.com/QAQFrank/MEGC2021

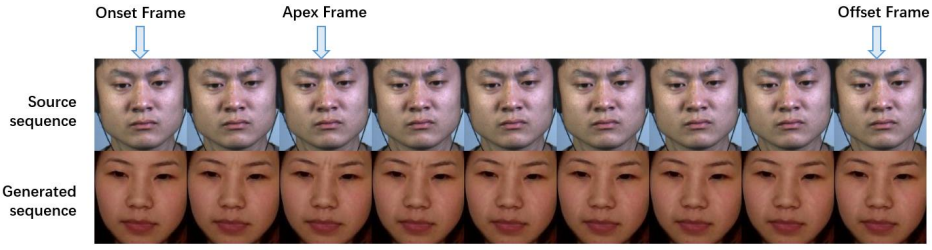

Figure 3: Generation result of female face with negative expression from CASME II. In this case, the left AU values are almost the same with the right ones, so we just show the sequence generated by left AU.

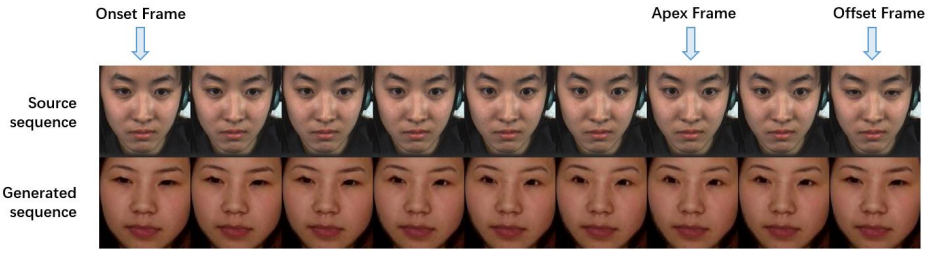

Figure 4: Generation result of female face with surprise expression from CASME II. We just show the sequence generated by left AU values in this case.

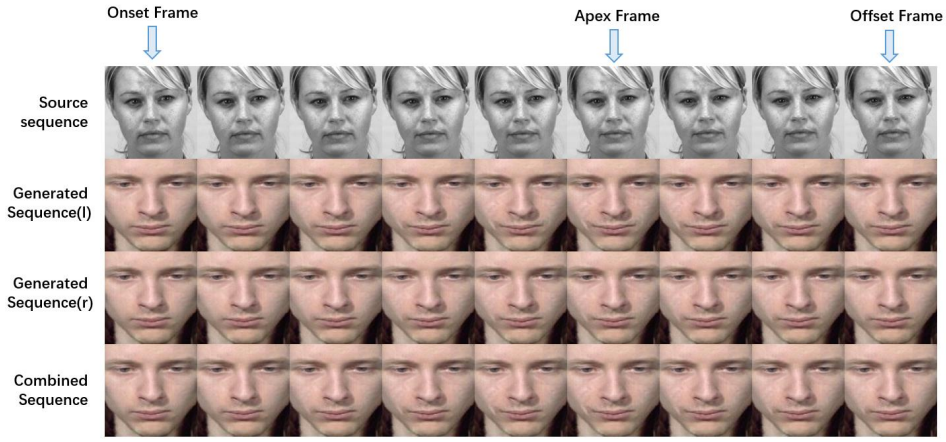

Figure 5: Generation result of male face with positive expression from SAMM.

4 CONCLUSION

In this paper, we proposed a combined system named FAMGAN, which is based on the fine-grained AUs modulation, and adopts AU-based GAN with super-resolution network to generate the final MEs sequences. Specifically, after comprehensively analyzing the factors that lead to inaccurate AU values detection, we performed fine-grained AUs modulation, which includes carefully eliminating the various noises and dealing with the symmetry of AU intensity. By modulating AU values, we could erase the identity information in AU, and solve the problem caused by asymmetry of facial expression. Finally, with the fine-tuned AU values, we could adopt the AU-based GAN to generate a robust, clear and natural face image with specific ME. The experiment shows that our results meet the requirement of quality, noise and expression well.

5 ACKNOWLEDGMENTS

This work was partially supported by the grants from the National Key Research and Development Program of China (Grant No. 2018YFB1402600), and the National Natural Science Foundation of China (No.61727809, 62072423)

REFERENCES

[1] Jianmin Bao, D. Chen, Fang Wen, Houqiang Li, and G. Hua. 2018. Towards Open-Set Identity Preserving Face Synthesis. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 6713–6722.

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[3] David Cristinacce, Timothy F Cootes, et al. 2006. Feature detection and tracking with constrained local models.. In *Bmvc*, Vol. 1. Citeseer, 3.

[4] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2016. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing* 9, 1 (2016), 116–129.

[5] Paul Ekman. 2009. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company.

[6] Paul Ekman and Wallace V Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry* 32, 1 (1969), 88–106.

[7] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. 2020. Real-World Super-Resolution via Kernel Estimation and Noise Injection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[8] Huai-Qian Khor, John See, Sze-Teng Liong, Raphael CW Phan, and Weiyao Lin. 2019. Dual-stream Shallow Networks for Facial Micro-expression Recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 36–40.

[9] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. 2013. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE, 1–6.

[10] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen. 2013. A Spontaneous Micro-expression Database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*.

[11] Gen-Bing Liong, John See, and Lai-Kuan Wong. 2021. Shallow Optical Flow Three-Stream CNN for Macro-and Micro-Expression Spotting from Long Videos. *arXiv preprint arXiv:2106.06489* (2021).

[12] Sze-Teng Liong, YS Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. 2019. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–5.

[13] R. Natsume, Tatsuya Yatagawa, and S. Morishima. 2018. RSGAN: face swapping and editing using face and hair representation in latent spaces. *ACM SIGGRAPH 2018 Posters*, Article 69, 2 pages.

[14] Yuval Nirkin, Y. Keller, and Tal Hassner. 2019. FSGAN: Subject Agnostic Face Swapping and Reenactment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 7183–7192.

[15] Min Peng, Chongyang Wang, Tong Chen, Guangyuan Liu, and Xiaolan Fu. 2017. Dual temporal scale convolutional neural network for micro-expression recognition. *Frontiers in psychology* 8 (2017), 1745.

[16] Stephen Porter and Leanne Ten Brinke. 2008. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological science* 19, 5 (2008), 508–514.

[17] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. 2019. GANimation: One-Shot Anatomically Consistent Facial Animation. (2019).

[18] Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. 2020. Dynamic Micro-Expression Recognition Using Knowledge Distillation. *IEEE Transactions on Affective Computing* (2020).

[19] Su-Jing Wang, Ying He, Jingting Li, and Xiaolan Fu. 2021. MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos. *IEEE Transactions on Image Processing* 30 (2021), 3956–3969.

[20] Su-Jing Wang, Bing-Jun Li, Yong-Jin Liu, Wen-Jing Yan, Xinyu Ou, Xiaohua Huang, Feng Xu, and Xiaolan Fu. 2018. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing* 312 (2018), 251–262.

[21] Sharon Weinberger. 2010. Intent to deceive? Can the science of deception detection help to catch terrorists? Sharon Weinberger takes a close look at the evidence for it. *Nature* 465, 7297 (2010), 412–416.

[22] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2020. AU-Assisted Graph Attention Convolutional Network for Micro-Expression Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. 2871–2880.

[23] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one* 9, 1 (2014), e86041.

[24] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–7.

[25] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[26] Tong Zhang, Yuan Zong, Wenming Zheng, CL Philip Chen, Xiaopeng Hong, Chuangao Tang, Zhen Cui, and Guoying Zhao. 2020. Cross-database micro-expression recognition: A benchmark. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[27] Sirui Zhao, Hanqing Tao, Yangsong Zhang, Tong Xu, Kun Zhang, Zhongkai Hao, and Enhong Chen. 2021. A two-stage 3D CNN based learning method for spontaneous micro-expression recognition. *Neurocomputing* 448 (2021), 276–289.