Intra- and Inter-Contrastive Learning for Micro-expression Action Unit Detection

Yante Li and Guoying Zhao*

University of Oulu

yante.li@oulu.fi and guoying.zhao@oulu.fi

Abstract

Encoding facial expressions via Action Units (**AUs**) has been found effective for resolving the ambiguity issue among different expressions. In the literature, AU detection has extensive researches in macro-expressions. However, there is limited research about AU analysis for micro-expressions **(MEs)**. ME AU detection becomes a challenging problem because of the subtle facial motion. To alleviate this problem, in this paper, we study the contrastive learning for modeling subtle AUs and propose a novel ME AU detection method by learning the intra- and inter-contrastive information among MEs. Through the intra-contrastive learning module, the difference between the onset and apex frames is enlarged and utilized to obtain the discriminative representation for low-intensity AU detection. In addition, considering the subtle difference between ME AUs, the inter-contrastive learning is designed to automatically explore and enlarge the difference between different AUs to enhance the ME AU detection robustness. Intensive experiments on two widely used ME databases have demonstrated the effectiveness and generalization ability of our proposed method.

1 Introduction

Micro-expressions **(MEs)** are rapid and subtle involuntary facial movements that reveal people's hidden emotions [5, 4]. Recent researches demonstrate that ME analysis has potential and emerging applications in different fields, such as clinical diagnosis, national security and interrogations [6, 25]. However, direct ME recognition can be very challenging because of ambiguities between several MEs [3, 24, 12]. One of the effective methods in resolving the ambiguity issue is employing the Facial Action Coding System (**FACS**) to represent individual expressions [7]. Ekman declared that he would never ever discover the MEs without FACS [7], in which each facial expression is identified as a specific configuration of multiple basic Action Units (**AUs**) [14]. Therefore, a robust AU detection system is important for the analysis of MEs [14].

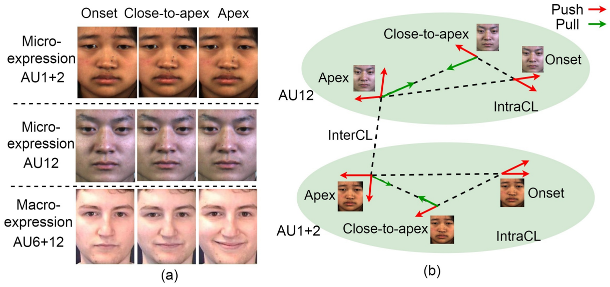

Figure 1. (a) Examples of AUs in micro- and macro-expressions. In each facial expression clip, the onset is the starting frame, the apex is the frame with the largest intensity, and the close-to-apex is around the apex; (b) Illustration of the distribution of the AU features learned by intra- and inter-contrastive learning.

Mostly existing researches focus on the analysis of strong AUs in macro-expressions [30, 20]. To the best of our knowledge, few work is conducted on analyzing AUs for MEs. Compared with macro-expression AU detection, ME AU detection is more difficult. This is explained by the fact that ME AU detection suffers from much lower intensity of AU occurrence, shown as Figure 1 (a). In Figure 1, the onset frame is the first frame which changes from the baseline (usually neutral facial expressions) in a particular facial expression clip. The apex frame is the frame that the facial muscle movement expresses the largest intensity. The frames around the apex frame are regarded as close-to-apex frames. From Figure 1 (a), it can be seen that the AUs in ME

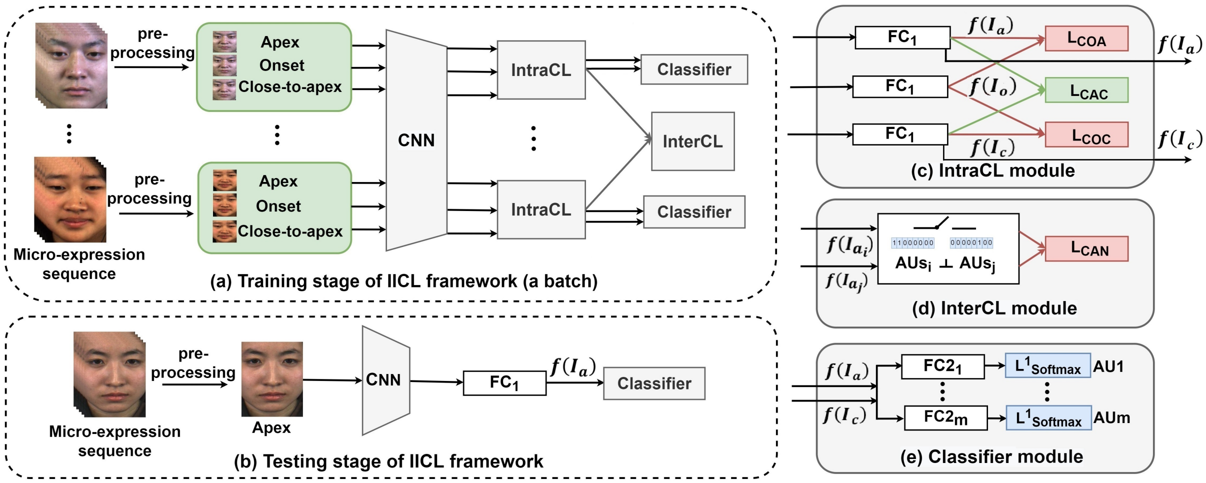

Figure 2. Illustration of the training and testing stages of IICL for ME AU detection. During training, IICL takes the processed onset, apex, and close-to-apex images as input. After passing the images through several convolutional layers, the features can be obtained for onset, apex, and close-to-apex images, respectively. The IntraCL drives the AU-related features of apex and close-to-apex away from the features of onset. The InterCL module automatically selects the negative pairs (a pair of apex images without the same AU presence) in a batch and enlarges the distance between different AUs. During testing, the IICL takes the apex as input, outputting the predicted probabilities for all ME AUs.

apex frames are very subtle. It is hard to distinguish the facial muscle movements between the onset and apex frames in MEs, which are referred as AUs, let alone identifying different AUs. The subtle change of ME leads to a hard AU detection. Li et al. [18] firstly proposed to detect subtle ME AUs through spatial and channel attention (SCA). However, the performance of SCA is far from satisfactory.

In this paper, to address the low intensity problem of MEs, an Intra- and Inter-Contrastive Learning (**IICL**) framework is proposed to increase the discrimination and robustness of features to represent ME AUs effectively. Different from current methods employing the apex frame [17, 19] or sequences [1, 28] for ME analysis, we study the contrastive information in MEs for better modeling subtle AUs. As Figure 1 (b) shows, Intra-contrastive learning (**IntraCL**) utilizes contrastive learning to push the onset which is usually neutral apart from AU-related apex frames. Thus, IntraCL can enlarge the difference between them and obtain discriminative AU representation form apex frames. Furthermore, considering the subtle difference between different AUs in MEs, Inter-contrastive learning (**InterCL**) is developed to enlarge the difference between different AUs. In this way, InterCL is able to improve the robustness of ME AU detection.

Our main contribution is in three-fold:

- We propose a novel intraCL module for ME AU detection, which obtains discriminative AU representations through driving the features of apex and close-to-apex away from the features of onset.
- The interCL learning module is designed to explore and enlarge the difference between different AUs to enhance the AU detection robustness.
- To the best of our knowledge, this is the first work studying contrastive learning for ME analysis and achieving subtle AU detection through learning and reinforcing the between-frame and between-AU distance. We conduct intensive experiments on two widely used ME databases CASME II and SAMM. The results demonstrate the effectiveness of our method.

2 METHODOLOGY

To learn the discriminative and robust representation for AU detection in MEs, we propose a simple yet effective intra- and inter- contrastive learning network, as shown in Figure 2. In this section, we introduce the details of IntraCL and InterCL modules. Then, the loss of multi-label AU detection is elaborated.

2.1 Intra-contrastive learning

ME AU detection suffers from low intensity. It is difficult to identify the subtle facial motion. Even the apex frame with the highest intensity does not have much difference compared with the onset frame. To cope with the problem, an IntraCL module composed of three contrastive losses is constructed to make sure that the AU-related features of apex and close-to-apex far way from the features of onset in the feature space, as illustrated in Figure 1 (b).

Firstly, we locate the apex frame based on the frequency representation of facial muscle change in the frequency domain [16]. Then, a contrastive loss [9, 26, 13, 8] of the onset and apex frames L_{OA} is developed to maximize the difference between the onset and apex frames in the feature space to obtain the discriminative representation of ME AUs. Moreover, considering the limited number of apex frames may restrict the learning ability for ME AUs, IntraCL further explores the relationship with weaker close-to-apex frames which are more commonly displayed in MEs. Specifically, L_{OC} loss is designed to push the close-to-apex apart from the onset, as shown in Figure 2 (c).

The L_{OA} and L_{OC} are defined as follows:

$$L_{OA} = \frac{1}{N}\sum_{i=1}^{N} max\left\{0, \delta - \parallel f(I_{o_i}) - f(I_{a_i}) \parallel_2^2\right\}, \quad (1)$$

$$L_{OC} = \frac{1}{N}\sum_{i=1}^{N} max\left\{0, \delta - \parallel f(I_{o_i}) - f(I_{c_i}) \parallel_2^2\right\}, \quad (2)$$

where $f(I_o)$, $f(I_a)$, and $f(I_c)$ represent the normalized features of the onset, apex, and close-to-apex frames. The objectives of L_{OC} and L_{OA} are learning representations with a greater distance for onset and apex frames, and onset and close-to-apex frames, respectively. In this way, the AUs in apex and close-to-apex frames can be differentiated from onset frames. Specifically, when the distance is not bigger than δ, the loss will be positive and the net parameters will be updated to generate more discriminative features for subtle ME AU representation. δ is a margin and set to 1 in our experiment following [23]. N is the training batch size.

Furthermore, the recent research [31] demonstrated that considering the intrinsic correlations between weak and strong expressions can achieve better results on weak expressions. Inspired by [31], AUs in both apex and close-to-apex frames are classified during training, as shown in Figure 2 (e). Moreover, a loss termed as L_{AC} is designed to pull apex and close-to-apex frames towards each other for robust ME AU detection. The L_{AC} is formulated as:

$$L_{AC} = \frac{1}{N}\sum_{i=1}^{N} \parallel f(I_{a_i}) - f(I_{c_i}) \parallel_2^2, \quad (3)$$

where $f(I_a)$ and $f(I_c)$ are the normalized apex and close-to-apex features, respectively.

2.2 Inter-contrastive learning

As discussed in the introduction section, AUs in ME apex frames still have low intensity. This makes it difficult to distinguish different AUs in MEs. InterCL designs a strategy automatically choosing the negative ME AU pairs in a batch and enlarges their difference during training to improve the AU detection robustness, as shown in Figure 2 (d).

As AUs can co-exist in MEs, only the pairs without the same AU presence can be treated as the negative pairs. Thus, the negative AU pairs should be orthogonal. In practical, the negative ME pairs are decided following the equation below:

$$C_{negative} = Aus_i \cdot Aus_j, \quad (4)$$

where Aus_i and Aus_j are the i-th and j-th ME AU labels in a batch and $1 \leq i<j \leq N$. N represents the batch size. If $C_{negative} = 0$, it is the negative pair.

Then the contrastive loss L_{NA} is employed to maximize the apex feature distance of negative AU pairs for the improvement of AU detection robustness.

$$L_{NA} = \frac{1}{K}\sum^{K} max\left\{0, \delta - \parallel f(I_{a_i}) - f(I_{a_j}) \parallel_2^2\right\}, \quad (5)$$

where $f(I_{a_i})$ and $f(I_{a_j})$ are i-th and j-th normalized apex features belonging to a negative pair. K represents the number of negative AU pairs in a batch. Similar with L_{OC} and L_{OA}, L_{NA} will be positive when the distance is smaller than the margin value δ and network will be updated for negative AU pairs, so that the training can focus on more difficult AU pairs. The margin δ is set to 1 following [23].

2.3 ME AU detection objective

For multi-label ME AUs, each AU can be treated as a specific task, as shown in Figure 2 (e). The loss for each task is defined as a binary cross-entropy loss. Thus, the loss of M AUs is formulated as:

$$L_{AUs} = \frac{1}{M}\sum_{m=1}^{M} y_m log(\hat{y_m}) + (1-y_m) log(1-\hat{y_m}), \quad (6)$$

where y_m is the ground truth for the m-th AU in the ME, with 1 denoting occurrence of the AU and 0 denoting absence. $\hat{y_m}$ is the predicted probability of the occurrence of m-th AU. M is number of AU categories.

The overall loss of the IICL framework is defined as:

$$L_{total} = L_{AUsa} + L_{AUsc} + \lambda(L_{OA} + L_{OC} + L_{AC} + L_{NA}), \quad (7)$$

where L_{AUsa} and L_{AUsc} are the losses for AUs in apex and close-to-apex frames, respectively. λ is the hyper-parameter that balance the influences of contrastive loss.

3 Experiments

3.1 Database and annotation.

We perform experiments about the ME AU detection on the CASME II and SAMM databases. The common AUs with the number of samples no less than 15 are utilized in the experiments. The sample numbers are denoted as 'Number' in Tables 2 and 3. For CASME II [29], we evaluate 243 videos with eight AUs: 'AU1', 'AU2', 'AU4', 'AU7', 'AU12', 'AU14', 'AU15', and 'AU17' occurred in *disgust*, *happy*, *surprise*, *angry*, and *others* emotions. In SAMM [2] database, the 101 samples including 'AU2', 'AU4', 'AU7', and 'AU12' are used to verify the effectiveness of the proposed method.

3.2 Metrics

AU detection is a multi-label binary classification problem. In our evaluation, F1-scores are computed for eight AUs on CASME II and four AUs on SAMM, according to the AU samples quantity and importance following [18]. The overall performance of the algorithm is described by the average F1-score.

3.3 Implementation

In the experiments, the aligned face images provided by the databases are employed. The three frames before and after the apex frame are regarded as the close-to-apex frames. The popular SEnet-50 [11] is employed as the backbone. All models were pre-trained on VGG-FACE database [22]. The images are resized to 256×256 and then randomly cropped to 224×224 for training [21]. In the pre-processing step, the MEs are magnified with ratio 10, according to the research of [27, 15]. All of the methods are evaluated with magnified MEs. During training, the networks are optimized using stochastic gradient descent (SGD). The initial learning rate is set to 0.01, divided by 10 every 40 epochs until 80 epochs. The weight of the contrastive losses λ is set to 0.1 for balancing the losses. Following experimental settings for AU detection [18], the subject independent four-fold cross validation is used in our experiments.

3.4 Ablation study

In this section, we provide ablation study on CASME II database to investigate the effectiveness of each part in the IICL network. To verify the effectiveness of IntraCL and InterCL modules, we add IntraCL and InterCL modules on the baseline, separately.

Table 1. Abalation study on the CASME II database. The baseline is SEnet [11].

Methods	AU1	AU2	AU4	AU7	AU12	AU14	AU15	AU17	AVG
Baseline	0.43	0.48	0.89	0.20	0.60	0.59	0.20	0.17	0.45
IntraCL	0.85	0.66	0.90	0.26	0.63	0.49	0.20	0.25	0.53
InterCL	0.70	0.58	0.90	0.22	0.48	0.48	0.15	0.43	0.49
IICL	0.78	0.67	0.89	0.30	0.56	0.53	0.33	0.33	0.55

Table 2. F1-scores on the CASME II database. The best is indicated using bold. The baseline is SEnet [11].

Methods	AU1	AU2	AU4	AU7	AU12	AU14	AU15	AU17	AVG
Number	26	21	129	58	34	21	16	25	
Baseline	0.43	0.48	0.89	0.20	0.60	0.59	0.20	0.17	0.45
SEseq	0.47	0.41	**0.91**	0.20	0.36	**0.60**	0.26	0.13	0.42
RESnet [10]	0.51	0.35	0.90	0.11	**0.62**	0.51	0.22	0.00	0.40
RESseq	0.24	0.21	0.86	0.09	0.43	0.49	0.20	0.27	0.35
SCA [18]	0.29	0.45	0.89	0.25	0.48	0.33	0.40	0.52	0.45
CL [9]	0.65	0.58	0.90	**0.30**	0.56	0.50	0.20	0.25	0.49
TL [23]	0.71	0.62	0.88	0.28	0.50	0.54	0.32	0.32	0.52
IICL	**0.78**	**0.67**	0.89	**0.30**	0.56	0.53	**0.33**	**0.33**	**0.55**

As shown in Table 1, the framework with InterCL outperforms the baseline by 0.04 in terms of the average F1-score on CASME II. The framework with IntraCL reaches higher F1-scores in six out of eight AUs on CASME II, compared with the baseline. The results demonstrate the effectiveness of InterCL and IntraCL modules. Furthermore, IICL consisted of InterCL and IntraCL modules achieves the best performance and improves the average F1-score by 22.22%, in comparison with the baseline. The results indicate that contrastive learning can explore discriminative representation for subtle ME AUs. From Table 1, it can be seen that the F1-scores declines on AU12 (Lip corner puller) and AU14 (Dimpler). This may caused by the similar appearance changes on the same region lip corner. It is hard to distinguish them. In general, the results suggest that enlarging the between-frame and between-AU differences can improve the discriminative ability of subtle AUs and is useful for most ME AUs.

3.5 Comparisons of methods

Tables 2 and 3 show the AU detection results of different methods on the CASME II and SAMM databases, respectively. The proposed IICL and the baseline methods based on SEnet [11] are tested on apex images in MEs. Compared with the baseline (SEnet), the IICL reached higher F1-scores in five out of eight AUs on the CASME II database and all AUs on the SAMM database. Furthermore, IICL outperforms SEseq which employs temporal in-

Table 3. F1-scores on the SAMM database. The best is indicated using bold. The baseline is SEnet [11].

Methods	AU2	AU4	AU7	AU12	AVG
Number	18	23	46	30	
Baseline	0.23	0.37	0.40	0.35	0.34
SEseq	0.24	0.31	0.45	0.41	0.35
RESnet [10]	0.21	0.37	0.32	0.37	0.32
RESseq	0.27	0.36	0.42	0.30	0.34
SCA [18]	0.33	0.13	**0.49**	0.42	0.34
CL [9]	0.28	0.30	0.42	0.38	0.35
TL [23]	0.30	0.35	0.47	0.39	0.38
IICL	**0.34**	**0.40**	0.42	**0.45**	**0.40**

formation through aggregating the onset, close-to-apex, and apex frame features by 0.13 and 0.05 in terms of the average F1-score on CASME II and SAMM, respectively. Moreover, IICL achieves large improvements on the challenging AUs containing unclear motions and very few samples, compared with the baseline on CASME II (0.43 vs. 0.78 on AU1, 0.20 vs. 0.33 on AU15, and 0.17 vs 0.33 on AU17 in terms of F1-score). The results demonstrate that learning the between-frame and between-AU contrastive information can improve the discriminative ability for ME AU detection.

In order to further verify the IICL effectiveness of ME AU detection, the IICL is compared with the methods contrastive learning based on Contrastive loss (**CL**) [9] and Triplet loss (**TL**) [23] and SCA [18]. In Tables 2 and 3, it can be seen that IICL improves the average F1-score by around 12% and 6% on CASME II, and 14% and 5% on SAMM, in comparison with CL and TL, respectively. Furthermore, IICL outperforms SCA by 0.10 and 0.06 in terms of average F1-score on CASME II and SAMM, respectively. The results indicates that the IICL has a good generalization ability and can learn discriminative AU representation from the subtle movements in MEs effectively.

Figure 3 shows some example class activation maps on CASME II. It can be seen that the IICL can focus more on the accurate region of ME AUs. For example, the baseline and CL networks focus on the wrong nose and eye regions for AU2 (Outer Brow Raiser), respectively. The TL roughly learns features from the whole forehead. The proposed IICL enlarging the difference of multiple frames and different AUs can accurately focus on the outer brow region for AU2 detection. This further verifies the effectiveness of IICL for ME AU detection.

4 Conclusion

Micro-expression AU detection becomes an important and challenging task, as micro-expression has subtle facial muscle changes. In this paper, we design an Intra- and Inter-Contrastive Learning (**IICL**) network for ME AU detection to improve the discriminative and robust ability for subtle AUs. The IICL network is able to efficiently identify subtle AUs by exploring the between-frame and between-AU difference. Intensive experiments demonstrate the effectiveness and generalization ability of our IICL network. In future, we will consider the study on an end-to-end contrastive learning network for micro-expression AU detection and exploring ME recognition enhanced by AU information.

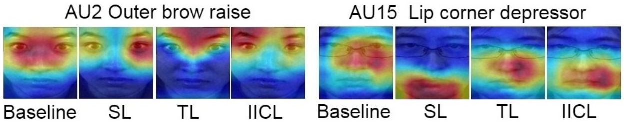

Figure 3. ME AU visualization.

5 ACKNOWLEDGMENT

This work was supported by Infotech Oulu, National Natural Sci-ence Foundation of China (Grant No: 61772419), Ministry of Edu-cation and Culture of Finland for AI forum project, and Academyof Finland for ICT 2023 project (grant 328115). As well, the au-thors wish to acknowledge CSC IT Center for Science, Finland, forcomputational resources.

References

[1] M. Bai and R. Goecke. Investigating lstm for micro-expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 7–11, 2020.

[2] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. SAMM: A spontaneous micro-facial movement dataset. *IEEE Trans. on Affect. Comput.*, 9(1):116–129, 2018.

[3] A. K. Davison, W. Merghani, and M. H. Yap. Objective classes for micro-facial expression recognition. *Journal of Imaging*, 4(10):119, 2018.

[4] P. Ekman. Lie catching and micro-expressions. *Phil. Decept.*, pages 118–133, 2009.

[5] P. Ekman and W. V. Friesen. Nonverbal leakage and clues to deception. *Study Interpers*, 32:88–106, 1969.

[6] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Personal. Soc. Psychol.*, 17(2), 1971.

[7] W. V. Friesen and P. Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978.

[8] R. Gao, R. Yang, W. Yang, and Q. Liao. Margin loss: Making faces more separable. *IEEE Signal Process. Lett.*, 25(2):308–312, 2018.

[9] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Comput. Society Conf. on Comput. Vis. and Pattern Recognit.*, volume 2, pages 1735–1742. IEEE, 2006.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pages 770–778, 2016.

[11] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pages 7132–7141, 2018.

[12] P. Jiang, B. Wan, Q.Wang, and J. Wu. Fast and efficient facial expression recognition using a gabor convolutional network. *IEEE Signal Process. Lett.*, 27:1954–1958, 2020.

[13] P. Last, H. A. Engelbrecht, and H. Kamper. Unsupervised feature learning for speech using correspondence and siamese networks. *IEEE Signal Process. Lett.*, 27:421–425, 2020.

[14] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recog.*, pages 1841–1850, 2017.

[15] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Trans. Affect. Comput.*, 9(4):563 – 577, 2018.

[16] Y. Li, X. Huang, and G. Zhao. Can micro-expression be recognized based on single apex frame? In *IEEE Int. Conf. on Image Process.*, pages 3094–3098. IEEE, 2018.

[17] Y. Li, X. Huang, and G. Zhao. Joint local and global information learning with single apex frame detection for micro-expression recognition. *IEEE Transactions on Image Processing*, 30:249–263, 2020.

[18] Y. Li, X. Huang, and G. Zhao. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing*, 436:221–231, 2021.

[19] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018.

[20] W. Michiel. Judgments of facial expressions of emotion predicted from facial behavior. *Journal of Nonverbal Behav.*, 7(2):101–116, 1982.

[21] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pages 11917–11926, 2019.

[22] P. M. Omkar, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Brit. Mach. Vis. Conf.*, pages 41.1–41.12, 2015.

[23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pages 815–823, 2015.

[24] Y. Tian, J. Cheng, Y. Li, and S. Wang. Secondary information aware facial expression recognition. *IEEE Signal Process. Lett.*, 26(12):1753–1757, 2019.

[25] S. Wang, W. Yan, X. Li, G. Zhao, and X. Fu. Micro-expression recognition using dynamic textures on tensor independent color space. In *Int. Conf. on Pattern Recognit.*, pages 4678–4683. IEEE, 2014.

[26] B. Wu, Y. Wei, B. Wu, and C. Lin. Contrastive feature learning and class-weighted loss for facial action unit detection. In *IEEE Int. Conf. on Syst. Man and Cybern.*, pages 2478–2483, 2019.

[27] H. Wu, E. Shih, E. Shih, J. Guttag, and W. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graphics*, 31(4):1–8, 2012.

[28] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29:8590–8605, 2020.

[29] W. Yan, X. Li, S. Wang, G. Zhao, Y. Liu, Y. Chen, and X. Fu. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. *Plos One*, 9(1):e86041, 2014.

[30] K. Zhao, W. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pages 3391–3399, 2016.

[31] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. Peak-piloted deep network for facial expression recognition. In *Eur. Conf. on Comput. Vis.*, pages 425–442. Springer, 2016.