Facial Expression Recognition In-the-Wild with Deep Pre-trained Models

Siyang Li[1], Yifan Xu[1], Huanyu Wu[1], Dongrui Wu[1(✉)], Yingjie Yin[2], Jiajiong Cao[2], and Jingting Ding[2]

[1] Ministry of Education Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
{syoungli,yfxu,m202173087,drwu}@hust.edu.cn
[2] Ant Group, Hangzhou, China
{gaoshi.yyj,jiajiong.caojiajio,yimou.djt}@antgroup.com

Abstract. Facial expression recognition (FER) is challenging, when transiting from the laboratory to in-the-wild situations. In this paper, we present a general framework for the Learning from Synthetic Data Challenge in the 4th Affective Behavior Analysis In-The-Wild (ABAW4) competition, to learn as much knowledge as possible from synthetic faces with expressions. To cope with four problems in training robust deep FER models, including uncertain labels, class imbalance, mismatch between pretraining and downstream tasks, and incapability of a single model structure, our framework consists of four respective modules, which can be utilized for FER in-the-wild. Experimental results on the official validation set from the competition demonstrated that our proposed approach outperformed the baseline by a large margin.

Keywords: Facial expression recognition · Affective computing · Learning from synthetic data · ABAW · Affective behavior analysis in-the-wild

1 Introduction

Facial expression is one powerful signal for human beings to convey their emotional states and intentions [28]. Automatic facial expression recognition (FER) is a challenging task in various interactive computing domains, including depression diagnostics/treatment and human-computer/-machine interaction [19]. Although affect models based on the six/seven basic emotions are limited in their ability to represent universal human emotions [31], such easy-to-comprehend pioneering categorical models are still popular in FER.

FER databases are very important for model development. Some FER databases are collected from controlled environments, e.g., inside a laboratory with constant light conditions and angles, including CK+ [20] and Oulu-CASIA

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13806, pp. 181–190, 2023.
https://doi.org/10.1007/978-3-031-25075-0_14

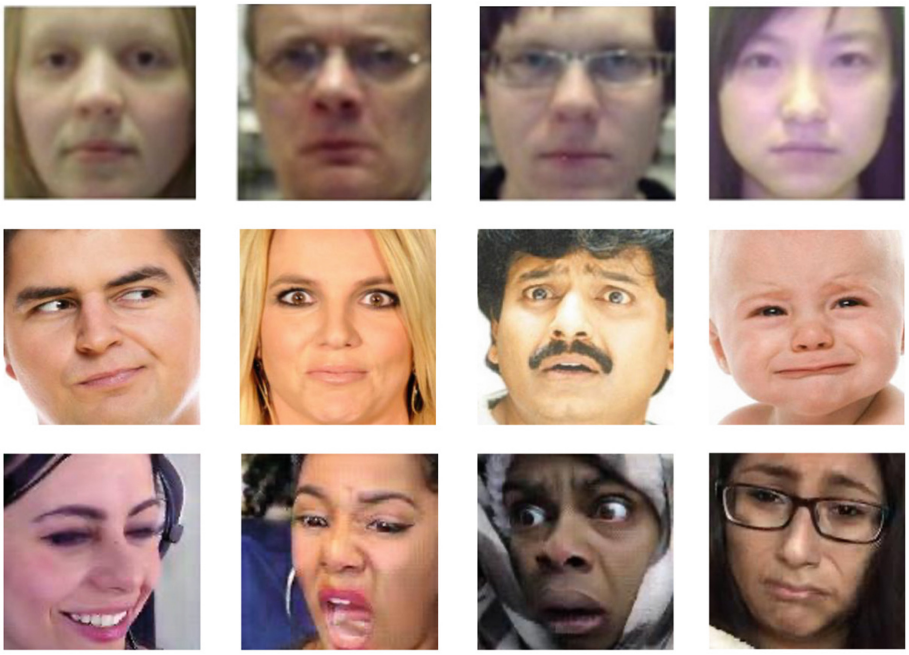

Fig. 1. Sample faces from some representative FER databases. Top to bottom: Oulu-CASIA, AffectNet, and ABAW4 Synthetic training data.

[33]. Others are collected from uncontrolled or wild environments in real-world settings [1], including popular ones such as AffectNet [22] and EmotioNet [7]. Sample images of some representative FER databases, including ABAW4 used in this competition, are shown in Fig. 1.

Recently, deep neural networks have been gaining increasingly popularity, compared with traditional methods which use handcrafted features [34] or shallow learning [30]. With the availability of large-scale databases, such deep models have demonstrated their capability of learning robust deep features [21]. However, the effectiveness of data-driven approaches also come with limitations. Uncertain annotations, imbalanced distributions and dataset biases are common in most databases [19], which impact the performances of deep learning.

At least four challenges should be taken into consideration when training robust models for facial expression classification:

1. *Uncertain labels.* Usually each face is given only one label, which may not be enough to adequately reflect the true expression. Emotions can co-occur on one face [23], e.g., one can demonstrate surprise and fear simultaneously.
2. *Class imbalance.* The number of training samples from different classes often differ significantly. In ABAW4, the class with the most samples (sadness) has 10 times more samples than the class with the fewest samples (fear). Biases could be introduced during learning, if class imbalance is not carefully addressed.

3. *Mismatch between pretraining and downstream tasks.* Pretrained models are often fixed as feature-extractors, with only the fully-connected layers updated during fine-tuning. When there is discrepancy between the pretraining task and the down-stream task, e.g., general image classification and facial affect classification, features extracted with a fixed pretrained model may not be optimal.
4. *Incapability of a single model structure.* Different deep learning architectures have different characteristics and hence different applications. Convolutional networks [26] are more inclined towards local features, whereas Transformers [5] have better ability to extract global features. Both local and global features are needed in FER.

Our main contributions are:

1. We identify four common challenges across FER databases for in-the-wild deep facial expression classification.
2. We propose a framework consisting of four modules, corresponding to the four challenges. They can be applied separately or simultaneously to better train or fine-tune deep models.
3. Experiments on the official Learning from Synthetic Data (LSD) challenge validation dataset of Aff-Wild2 verified the effectiveness of our proposed framework.

The remainder of this paper is organized as follows: Sect. 2 introduces related work. Section 3 describes our proposed framework. Section 4 presents the experimental results. Finally, Sect. 5 draws conclusions.

2 Related Work

ABAW. To promote facial affect analysis, Kollias et al. [11] organized four Affective Behavior in-the-wild (ABAW) Competitions. The previous three were held in conjunction with IEEE FG 2021 [12], ICCV 2022 [18], and IEEE CVPR 2022 [8], respectively. The 4th ABAW (ABAW4) is held in conjunction with ECCV 2022 [16], which includes a Multi-Task-Learning (MTL) Challenge and a LSD Challenge. The large scale in-the-wild Aff-Wild2 database [14,29] was used.

Three tasks are considered in the ABAW MTL challenge [13,15,17]: basic expression classification, estimation of continuous affect (usually valence and arousal [24]), and detection of facial action units (AUs) based on the facial action coding system (FACS) [6].

This paper focuses on the LSD challenge.

Synthetic Facial Affect. Facial affect could be synthesized through either traditional graph-based methods or data-driven generative models [10]. GAN-based approaches [2,4] can add subtle affect changes to a neutral face in the dimensional space [9,16], from an emotion category [4], or on AUs [25].

In the LSD challenge, synthetic data were given without disclosing the data generation approach. The goal was to extract knowledge from such synthetic faces and the train models suitable for real facial affect analysis.

3 Method

This section introduces our proposed framework, which is shown in Fig. 2.

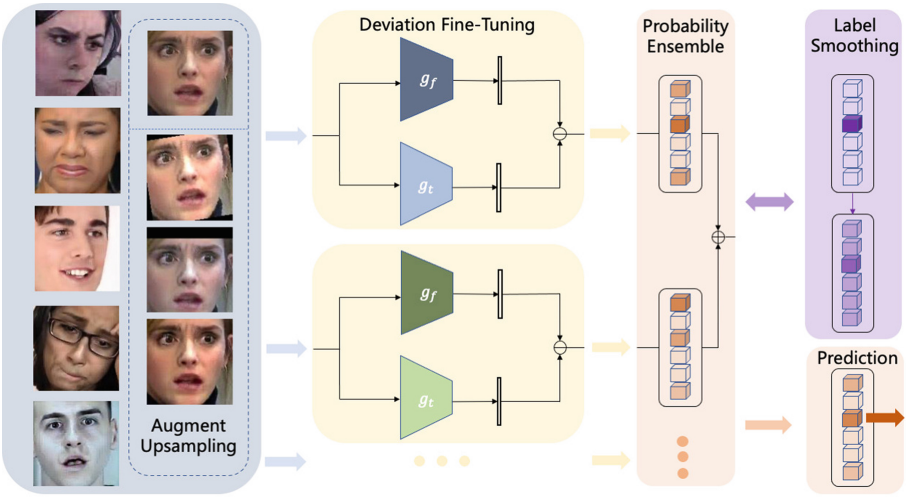

Fig. 2. The proposed framework for FER in-the-wild with four modules.

Label Smoothing. Label smoothing [27] was used in the cross-entropy loss to cope with uncertain labels. The label of the k^{th} face is modified to:

$$y_k^{LS} = y_k(1-\alpha) + \frac{\alpha}{K}, \tag{1}$$

where $y_k \in \{0, 1\}$ is the given binary label, K is the number of classes, and $\alpha \in [0, 1]$ is the smoothing amount. $\alpha = 0.2$ were used in our experiments.

Label smoothing promotes the co-existence of different emotions in a single face, and prevents the model from being over-confident on its labels.

Data Augmentation and Weighted Cross-Entropy Loss. RandAugment [3] was adopted to cope with class imbalance. Specifically, we upsampled the minority classes with RandAugment, so that all classes had the same size. For each image, two of 12 transformations (Solarize and Equalize were excluded, since they are not suitable for faces) in RandAugment were randomly selected and combined.

Weighted cross-entropy loss was also applied to further force the learning on the minority classes. The weights were empirically set to [1, 3, 5, 1, 1, 1] for the six classes.

Fine-Tuning with Deviation Module. In order to utilize the full potential of models with millions of parameters, while preventing overfitting, we propose a more robust method for fine-tuning pretrained models, inspired by the deviation module of DLN [32].

Specifically, a pair of siamese feature extractor with the same architecture of pretrained weights were used. One's parameters were fixed, while the other's were trainable. The actual feature vector used was their tensor discrepancy. In this way, all parameters of feature extractors were involved during fine-tuning for the downstream expression classification task. The actual feature vector representation is

$$\bar{x} = g_f(x) - g_t(x) + \epsilon, \tag{2}$$

where $g_f : \mathcal{X} \rightarrow \mathbb{R}^d$ is the frozen feature encoding module, $g_t : \mathcal{X} \rightarrow \mathbb{R}^d$ is the feature encoding module being trained, d is the dimensionality of the input feature, and $\epsilon = 10^{-6}$ is used to prevent features being all zero.

Ensemble Learning. Each deep learning architecture has its unique characteristics, which may be more suitable for certain applications. For example, transformers are better for extracting global features, whereas CNNs focus more on local ones. Both are important in FER.

Four backbones, namely ViT (ViT_B_16), ResNet (ResNet50), EfficientNet (EfficientNet_B0), and GoogleNet (InceptionV1), were used and separately trained on the synthetic data. The feature extractors were pretrained on ImageNet, and the classifier of fully-connected layer was trained during fine-tuning. These four models were ensembled at the test stage. Specifically, the softmax scores of logits after the fully-connected layers were treated as each model's prediction probabilities. These four sets of probabilities were averaged as the final probability.

Implementation Details. Images were normalized using mean and standard deviation of the synthetic training set, and resized to 224×224 before being input into the networks. Batch size was 64. Adam optimizer of learning rate 10^{-6} was used. Fine-tuning took less than 20 epochs.

All models were implemented using PyTorch. All computations were performed on a single GeForce RTX 3090 GPU. The code is available online.

4 Experimental Results

This section shows our experimental results in the LSD challenge of ABAW4. The performance measure was F1 score.

Table 1 shows the performances of different label smoothing amounts.

Table 2 shows the performances of different approaches for coping with class imbalance.

Table 3 shows the performances of different model architectures under deviation fine-tuning.

Table 1. F1 on the LSD validation set, with different label smoothing amounts using the ViT backbone.

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.
LS ($\alpha = 0$)	0.5224	0.6559	0.3199	0.6533	0.4970	0.5506	0.5332
LS ($\alpha = 0.1$)	0.5557	0.6751	0.3333	0.6718	0.5365	0.5848	0.5596
LS ($\alpha = 0.2$)	0.5719	0.6854	0.3487	0.6795	0.5568	0.6010	0.5739
LS ($\alpha = 0.4$)	**0.6040**	**0.6872**	**0.4366**	**0.6923**	**0.5848**	**0.6208**	**0.6043**

Table 2. F1 on the LSD validation set, with different approaches for handling class imbalance. The ViT backbone was used.

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.
Baseline	0.5224	**0.6559**	0.3199	0.6533	0.4970	0.5506	0.5332
Weighted cross-entropy	0.6037	0.5915	**0.3765**	0.6437	**0.5461**	0.5587	**0.5534**
RandAugment on-the-fly	0.4961	0.6432	0.3110	0.6414	0.4863	0.5520	0.5217
RandAugment upsampling	**0.5727**	0.6087	0.3636	**0.6631**	0.5353	**0.5675**	0.5518

Table 3. F1 on the LSD validation set, with different model architectures under deviation fine-tuning (DFT).

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.
ViT	0.5870	0.6933	0.3858	0.6859	0.6047	0.6652	0.6037
ViT-DFT	0.4580	0.7454	0.2025	0.8245	0.5534	**0.7190**	0.5838
ResNet	0.6120	0.5953	0.5662	0.6482	0.5425	0.6509	0.6025
ResNet-DFT	0.6952	**0.7655**	0.3138	**0.8312**	**0.6357**	0.6873	**0.6548**
EfficientNet	0.5611	0.5790	0.5711	0.6313	0.5586	0.6293	0.5884
EfficientNet-DFT	0.5709	0.7612	0.3840	0.8104	0.5644	0.6710	0.6270
GoogleNet	0.5283	0.5482	**0.6316**	0.6026	0.5126	0.5419	0.5609
GoogleNet-DFT	**0.7070**	0.6736	0.2082	0.8185	0.6091	0.6534	0.6116

Compared to the usual method of only fine-tuning classifier layer, all parameters are involved in downstream training in DFT. Since feature extractors are tuned as well, models are able to reach better performance on classes that are comparatively easier to distinguish, i.e., happiness and disgust. However, the risk of overfitting also rises because of the magnitude of trainable parameters. We noticed that model performances on classes that are relatively harder to distinguish would drop under DFT. Empirically, the overall average performance gain of DFT is around 0.03 average F1 score.

For more consistency between training and test, we pretrained the feature extractors on AffectNet with six expression classes, instead of ImageNet. The final experimental results of different architectures with ensemble prediction is shown in Table 4.

Table 4. F1 on the LSD validation set, with four modules integrated.

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.
ViT	0.6937	**0.7767**	0.1829	0.8591	**0.6584**	0.6970	0.6447
ResNet	0.7257	0.6010	0.2320	0.8435	0.6218	0.6917	0.6193
EfficientNet	0.6099	0.7683	**0.3732**	0.8363	0.5814	0.6614	0.6384
GoogleNet	**0.7333**	0.7039	0.2631	0.8440	0.6371	0.6906	0.6453
Ensemble	0.7331	0.7730	0.2486	**0.8640**	0.6532	**0.7229**	**0.6658**

Qualitative results of the final ensemble model on the LSD validation set are shown in Fig. 3.

Fig. 3. Qualitative results including correct and wrong predictions of our final ensemble model on the LSD validation set. P: model prediction; T: true label.

The final ensemble achieved an F1 score of 0.3483 on the official test set of the LSD challenge, ranked among the top 5. The baseline had an F1 score of 0.30.

5 Conclusion

Facial expression recognition is challenging in-the-wild. This paper has presented a general framework for the LSD Challenge in the ABAW4 competition, to learn

from synthetic faces with expressions and then apply to real faces. To cope with four problems in training robust deep FER models, including uncertain labels, class imbalance, mismatch between pretraining and downstream tasks, and incapability of a single model structure, our framework has four respective modules. Experimental results on the official validation set from the competition demonstrated that our proposed approach outperformed the baseline by a large margin.

Acknowledgment. This research was supported by CCF-AFSG Research Fund (RF20210007).

References

1. Canedo, D., Neves, A.J.: Facial expression recognition using computer vision: a systematic review. Appl. Sci. **9**(21), 4678 (2019)
2. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 8789–8797, June 2018
3. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)
4. Ding, H., Sricharan, K., Chellappa, R.: ExprGAN: facial expression editing with controllable expression intensity. In: Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, vol. 32, February 2018
5. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR), May 2021
6. Ekman, P., Friesen, W.V.: Facial action coding system. Environ. Psychol. Nonverbal Behav. (1978)
7. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 2016
8. Kollias, D.: ABAW: valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2328–2336, June 2022
9. Kollias, D., Cheng, S., Pantic, M., Zafeiriou, S.: Photorealistic facial synthesis in the dimensional affect space. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11130, pp. 475–491. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11012-3_36
10. Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.: Deep neural network augmentation: generating faces for affect analysis. Int. J. Comput. Vis. **128**(5), 1455–1484 (2020). https://doi.org/10.1007/s11263-020-01304-3
11. Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.: Recognition of affect in the wild using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 26–33, July 2017

12. Kollias, D., Schulc, A., Hajiyev, E., Zafeiriou, S.: Analysing affective behavior in the first ABAW 2020 competition. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 637–643 (2020)
13. Kollias, D., Sharmanska, V., Zafeiriou, S.: Distribution matching for heterogeneous multi-task learning: a large-scale face study. arXiv preprint arXiv:2105.03790 (2021)
14. Kollias, D., et al.: Deep affect prediction in-the-wild: aff-wild database and challenge, deep architectures, and beyond. Int. J. Comput. Vis. **127**, 907–929 (2019). https://doi.org/10.1007/s11263-019-01158-4
15. Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: aff-wild2, multi-task learning and ArcFace. arXiv preprint arXiv:1910.04855 (2019)
16. Kollias, D., Zafeiriou, S.: VA-StarGAN: continuous affect generation. In: Blanc-Talon, J., Delmas, P., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2020. LNCS, vol. 12002, pp. 227–238. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-40605-9_20
17. Kollias, D., Zafeiriou, S.: Affect analysis in-the-wild: valence-arousal, expressions, action units and a unified framework. arXiv preprint arXiv:2103.15792 (2021)
18. Kollias, D., Zafeiriou, S.: Analysing affective behavior in the second ABAW2 competition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 3652–3660, October 2021
19. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **13**(3), 1195–1215 (2022)
20. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, pp. 94–101, June 2010
21. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10 (2016)
22. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2019)
23. Navarretta, C.: Mirroring facial expressions and emotions in dyadic conversations. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp. 469–474, May 2016
24. Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. Affect. Comput. **2**(2), 92–105 (2011)
25. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: GANimation: anatomically-aware facial animation from a single image. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 835–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_50
26. Shin, M., Kim, M., Kwon, D.S.: Baseline CNN structure analysis for facial expression recognition. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 724–729 (2016)
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, pp. 2818–2826, June 2016

28. Tian, Y.I., Kanade, T., Cohn, J.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 97–115 (2001)
29. Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: valence and arousal 'in-the-wild' challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, pp. 34–41, July 2017
30. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643–649 (2018)
31. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2008)
32. Zhang, W., Ji, X., Chen, K., Ding, Y., Fan, C.: Learning a facial expression embedding disentangled from identity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6759–6768, June 2021
33. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
34. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)