ABAW: Learning from Synthetic Data & Multi-task Learning Challenges

Dimitrios Kollias(✉)

Queen Mary University of London, London, UK
d.kollias@qmul.ac.uk

Abstract. This paper describes the fourth Affective Behavior Analysis in-the-wild (ABAW) Competition, held in conjunction with European Conference on Computer Vision (ECCV), 2022. The 4th ABAW Competition is a continuation of the Competitions held at IEEE CVPR 2022, ICCV 2021, IEEE FG 2020 and IEEE CVPR 2017 Conferences, and aims at automatically analyzing affect. In the previous runs of this Competition, the Challenges targeted Valence-Arousal Estimation, Expression Classification and Action Unit Detection. This year the Competition encompasses two different Challenges: i) a Multi-Task-Learning one in which the goal is to learn at the same time (i.e., in a multi-task learning setting) all the three above mentioned tasks; and ii) a Learning from Synthetic Data one in which the goal is to learn to recognise the six basic expressions from artificially generated data and generalise to real data.

The Aff-Wild2 database is a large scale in-the-wild database and the first one that contains annotations for valence and arousal, expressions and action units. This database is the basis for the above Challenges. In more detail: i) s-Aff-Wild2 - a static version of Aff-Wild2 database- has been constructed and utilized for the purposes of the Multi-Task-Learning Challenge; and ii) some specific frames-images from the Aff-Wild2 database have been used in an expression manipulation manner for creating the synthetic dataset, which is the basis for the Learning from Synthetic Data Challenge. In this paper, at first we present the two Challenges, along with the utilized corpora, then we outline the evaluation metrics and finally present both the baseline systems and the top performing teams' per Challenge, as well as their derived results. More information regarding the Competition can be found in the competition's website: https://ibug.doc.ic.ac.uk/resources/eccv-2023-4th-abaw/.

Keywords: Multi-task learning · Learning from synthetic data · ABAW · Affective behavior analysis in-the-wild · Aff-Wild2 · s-Aff-Wild2 · Valence and arousal estimation · Expression recognition and classification · Action unit detection · Facial expression transfer · Expression synthesis

1 Introduction

Automatic facial behavior analysis has a long history of studies in the intersection of computer vision, physiology and psychology and has applications spread

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13806, pp. 157–172, 2023.
https://doi.org/10.1007/978-3-031-25075-0_12

across a variety of fields, such as medicine, health, or driver fatigue, monitoring, e-learning, marketing, entertainment, lie detection and law. However it is only recently, with the collection of large-scale datasets and powerful machine learning methods such as deep neural networks, that automatic facial behavior analysis started to thrive. When it comes to automatically recognising affect in-the-wild (i.e., in uncontrolled conditions and unconstrained environments), there exist three iconic tasks, which are: i) recognition of basic expressions (anger, disgust, fear, happiness, sadness, surprise and the neutral state); ii) estimation of continuous affect (valence -how positive/negative a person is- and arousal -how active/passive a person is-); iii) detection of facial action units (coding of facial motion with respect to activation of facial muscles, e.g. upper/inner eyebrows, nose wrinkles).

Ekman [13] defined the six basic emotions, i.e., Anger, Disgust, Fear, Happiness, Sadness, Surprise and the Neutral State, based on a cross-culture study [13], which indicated that humans perceive certain basic emotions in the same way regardless of culture. Nevertheless, advanced research on neuroscience and psychology argued that the model of six basic emotions are culture-specific and not universal. Additionally, the affect model based on basic emotions is limited in the ability to represent the complexity and subtlety of our daily affective displays. Despite these findings, the categorical model that describes emotions in terms of discrete basic emotions is still the most popular perspective for Expression Recognition, due to its pioneering investigations along with the direct and intuitive definition of facial expressions.

The dimensional model of affect, that is appropriate to represent not only extreme, but also subtle emotions appearing in everyday human-computer interactions, has also attracted significant attention over the last years. According to the dimensional approach [14,67,80], affective behavior is described by a number of latent continuous dimensions. The most commonly used dimensions include valence (indicating how positive or negative an emotional state is) and arousal (measuring the power of emotion activation).

Detection of Facial Action Units (AUs) has also attained large attention. The Facial Action Coding System (FACS) [4,13] provides a standardised taxonomy of facial muscles' movements and has been widely adopted as a common standard towards systematically categorising physical manifestation of complex facial expressions. Since any facial expression can be represented as a combination of action units, they constitute a natural physiological basis for face analysis. Consequently, in the last years, there has been a shift of related research towards the detection of action units. The presence of action units is typically brief and unconscious, and their detection requires analyzing subtle appearance changes in the human face. Furthermore, action units do not appear in isolation, but as elemental units of facial expressions, and hence some AUs co-occur frequently, while others are mutually exclusive.

The fourth Affective Behavior Analysis in-the-wild (ABAW) Competition, held in conjunction with the European Conference on Computer Vision (ECCV), 2022, is a continuation of the first[1] [37], second[2] [46] and third [33][3] ABAW Competitions held in conjunction with the IEEE Conference on Face and Gesture Recognition (IEEE FG) 2021, with the International Conference on Computer Vision (ICCV) 2022 and the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2022, respectively. The previous Competitions targeted dimensional (in terms of valence and arousal) [1,7,9,11,28–30,51,57,58,60–62,66,69,74,77,81,86–88,90,92,93], categorical (in terms of the basic expressions) [2,12,15,16,22,23,31,47,54–56,60,64,69,82–84,91,92] and facial action unit analysis-recognition [6,19,20,25–27,48,61,63,68,69,73,74,76,78,79,92]. The third ABAW Challenge further targeted Multi-Task Learning for valence and arousal estimation, expression recognition and action unit detection [5,8,21,22,69,71,92,92].

The fourth ABAW Competition contains two Challenges (i) the Multi-Task-Learning (MTL) one in which the goal is to create a system that learns at the same time (i.e., in a multi-task learning setting) to estimate valence and arousal, classify eight expressions (6 basic expressions plus the neutral state plus a category ‘other’ which denotes expressions/affective states other than the 6 basic ones) and detect twelve action units; ii) the Learning from Synthetic Data (LSD) one in which the goal is to create a system that learns to recognise the six basic expressions (anger, disgust, fear, happiness, sadness, surprise) from artificially generated data (i.e., synthetic data) and generalise its knowledge to real-world (i.e., real) data.

Both Challenges’ corpora are based on the Aff-Wild2 database [33,36–43,45,46,85], which is the first comprehensive in-the-wild benchmark for all the three above-mentioned affect recognition tasks; the Aff-Wild2 database is an extensions of the Aff-Wild database [36,40,85], with more videos and annotations for all behavior tasks. The MTL Challenge utilises a static version of the Aff-Wild2 database, named s-Aff-Wild2. The LSD Challenge utilizes a synthetic dataset which has been constructed after manipulating the displayed expressions in some frames of the Aff-Wild2 database.

The remainder of this paper is organised as follows. The Competition corpora is introduced in Sect. 2, the Competition evaluation metrics are mentioned and described in Sect. 3, the developed baselines and the top performing teams in each Challenge are explained and their obtained results are presented in Sect. 4, before concluding in Sect. 5.

2 Competition Corpora

The fourth Affective Behavior Analysis in-the-wild (ABAW) Competition relies on the Aff-Wild2 database, which is the first ever database annotated in terms

[1] https://ibug.doc.ic.ac.uk/resources/fg-2020-competition-affective-behavior-analysis/.

[2] https://ibug.doc.ic.ac.uk/resources/iccv-2021-2nd-abaw/.

[3] https://ibug.doc.ic.ac.uk/resources/cvpr-2022-3rd-abaw/.

of the tasks of: valence-arousal estimation, action unit detection and expression recognition. These three tasks constitute the basis of the two Challenges.

In the following, we provide a short overview of each Challenge's dataset along with a description of the pre-processing steps that we carried out for cropping and/or aligning the images of Aff-Wild2. These images have been utilized in our baseline experiments.

2.1 Multi-task Learning Challenge

A static version of the Aff-Wild2 database has been generated by selecting some specific frames of the database; this Challenge's corpora is named s-Aff-Wild2. In total, 221,928 images are used that contain annotations in terms of: i) valence and arousal; ii) 6 basic expressions (anger, disgust, fear, happiness, sadness, surprise), plus the neutral state, plus the 'other' category (which denotes expressions/affective states other than the 6 basic ones); 12 action units.

Figure 1 shows the 2D Valence-Arousal histogram of annotations of s-Aff-Wild2. Table 1 shows the distribution of the 8 expression annotations of s-Aff-Wild2. Table 2 shows the name of the 12 action units that have been annotated, the action that they correspond to and the distribution of their annotations in s-Aff-Wild2.

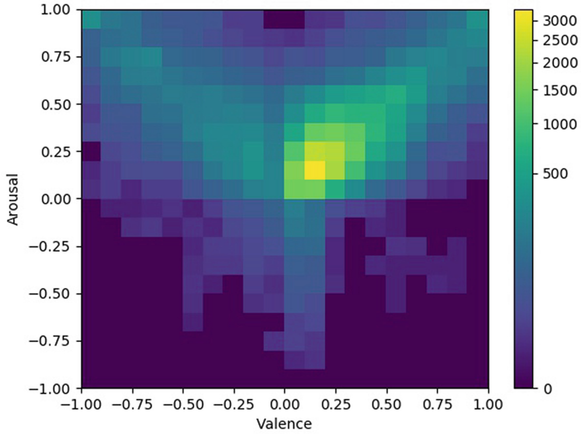

Fig. 1. Multi-task-learning challenge: 2D valence-arousal histogram of annotations in s-Aff-Wild2

The s-Aff-Wild2 database is split into training, validation and test sets. At first the training and validation sets, along with their corresponding annotations, are being made public to the participants, so that they can develop their own methodologies and test them. At a later stage, the test set without annotations is given to the participants.

Table 1. Multi-task-learning challenge: number of annotated images for each of the 8 expressions

Expressions	No of images
Neutral	37,073
Anger	8,094
Disgust	5,922
Fear	6,899
Happiness	32,397
Sadness	13,447
Surprise	9,873
Other	39,701

Table 2. Multi-task-learning challenge: distribution of AU annotations in Aff-Wild2

Action unit #	Action	Total number of activated AUs
AU 1	Inner brow raiser	29,995
AU 2	Outer brow raiser	14,183
AU 4	Brow lower	31,926
AU 6	Cheek raiser	49,413
AU 7	Lid tightener	72,806
AU 10	Upper lip raiser	68,090
AU 12	Lip corner puller	47,820
AU 15	Lip corner depressor	5,105
AU 23	Lip tightener	6,538
AU 24	Lip pressor	8,052
AU 25	Lips part	122,518
AU 26	Jaw drop	19,439

The participants are given two versions of s-Aff-Wild2: the cropped and cropped-aligned ones. At first, all images/frames of s-Aff-Wild2 are passed through the RetinaFace [10] to extract, for each image/frame, face bounding boxes and 5 facial landmarks. The images/frames are then cropped according the bounding box locations. All cropped-aligned images have the same dimensions $112 \times 112 \times 3$. These cropped images/frames constitute the cropped version of s-Aff-Wild2 that is given to the participants. The 5 facial landmarks (two eyes, nose and two mouth corners) have then been used to perform similarity transformation. The resulting cropped-aligned images/frames constitute the cropped-aligned version of s-Aff-Wild2 that is given to the participants. The cropped-aligned version has been utilized in our baseline experiments, described in Sect. 4.

Let us note that for the purposes of this Challenge, all participants are allowed to use the provided s-Aff-Wild2 database and/or any publicly available or private database; the participants are not allowed to use the audiovisual (A/V) Aff-Wild2 database (images and annotations). Any methodological solution will be accepted for this Challenge.

2.2 Learning from Synthetic Data Challenge

Some specific cropped images/frames of the Aff-Wild2 database have been selected; these images/frames, which show a face with an arbitrary expression/affective state, have been used -in a facial expression manipulation manner [34,35,44]- so as to synthesize basic facial expressions of the same person. Therefore a synthetic facial dataset has been generated and used for the purposes of this Challenge. In total, 277,251 images that contain annotations in terms of the 6 basic expressions (anger, disgust, fear, happiness, sadness, surprise) have been generated. These images constitute the training set of this Challenge. Table 3 shows the distribution of the 6 basic expression annotations of these generated images.

Table 3. Learning from synthetic data challenge: number of annotated images for each of the 6 basic expressions

Expressions	No of images
Anger	18,286
Disgust	15,150
Fear	10,923
Happiness	73,285
Sadness	144,631
Surprise	14,976

The validation and test sets of this Challenge are real images of the Aff-Wild2 database. Let us note that the synthetic data have been generated from subjects of the validation set, but not of the test set.

At first the training (synthetic data) and validation (real data) sets, along with their corresponding annotations, are being made public to the participants, so that they can develop their own methodologies and test them. At a later stage, the test set (real data) without annotations is given to the participants.

Let us note that for the purposes of this Challenge, all participants are allowed to use any -publicly or not- available pre-trained model (as long as it has not been pre-trained on Aff-Wild2). The pre-trained model can be pre-trained on any task (e.g. VA estimation, Expression Classification, AU detection, Face Recognition). However when the teams are refining the model and developing the methodology they must only use the provided synthetic data. No real data should be used in model training/methodology development.

3 Evaluation Metrics for Each Challenge

Next, we present the metrics that will be used for assessing the performance of the developed methodologies of the participating teams in each Challenge.

3.1 Multi-task Learning Challenge

The performance measure is the sum of: the average between the Concordance Correlation Coefficient (CCC) of valence and arousal; the average F1 Score of the 8 expression categories (i.e., the macro F1 Score); the average binary F1 Score over the 12 action units (when detecting each AU, we are interested in the binary F1 Score - following the literature in which the positive class is of particular interest and is thus measured and reported-).

CCC takes values in the range $[-1, 1]$; high values are desired. CCC is defined as follows:

$$\rho_c = \frac{2 s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}, \tag{1}$$

where s_x and s_y are the variances of all video valence/arousal annotations and predicted values, respectively, $\bar{x}$ and $\bar{y}$ are their corresponding mean values and s_{xy} is the corresponding covariance value.

The F_1 score is a weighted average of the recall (i.e., the ability of the classifier to find all the positive samples) and precision (i.e., the ability of the classifier not to label as positive a sample that is negative). The F_1 score takes values in the range $[0, 1]$; high values are desired. The F_1 score is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{2}$$

Therefore, the evaluation criterion for the Multi-Task-Learning Challenge is:

$$\begin{aligned} \mathcal{P}_{MTL} &= \mathcal{P}_{VA} + \mathcal{P}_{EXPR} + \mathcal{P}_{AU} \\ &= \frac{\rho_a + \rho_v}{2} + \frac{\sum_{expr} F_1^{expr}}{8} + \frac{\sum_{au} F_1^{au}}{12} \end{aligned} \tag{3}$$

3.2 Learning from Synthetic Data Challenge

The performance measure is the average F1 Score of the 6 basic expression categories (i.e., the macro F1 Score):

$$\mathcal{P}_{LSD} = \frac{\sum_{expr} F_1^{expr}}{6} \tag{4}$$

4 Baseline Networks and Performance

All baseline systems rely exclusively on existing open-source machine learning toolkits to ensure the reproducibility of the results. All systems have been implemented in TensorFlow; training time was around five hours on a Titan X GPU, with a learning rate of 10^{-4} and with a batch size of 128.

In this Section, we first present the top-performing teams per Challenge as well as describe the baseline systems developed for each Challenge; then we report their obtained results, also declaring the winners of each Challenge.

4.1 Multi-task Learning Challenge

In total, 55 Teams participated in the Multi-Task Learning Challenge. 25 Teams submitted their results. 11 Teams scored higher than the baseline and made valid submissions.

The winner of this Challenge is *Situ-RUCAIM3* (that has was the winner of the Valence-Arousal Estimation Challenge of the 3rd ABAW Competition) consisting of: Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, and Fengyuan Zhang (Renmin University of China; Beijing Seek Truth Data Technology Services Co Ltd).

The runner up is *ICT-VIPL* (that took the 2nd and 3rd places in the Expression Classification and Valence-Arousal Estimation Challenges of the 1st ABAW Competition, respectively) consisting of: Hu Han (Chinese Academy of Science), Yifan Li, Haomiao Sun, Zhaori Liu, Shiguang Shan and Xilin Chen (Institute of Computing Technology Chinese Academy of Sciences, China).

In the third place is *HSE-NN* (that took the 3rd place in the corresponding Multi-Task Learning Challenge of the 3rd ABAW Competition) consisting of: Andrey Savchenko (HSE University, Russia).

Baseline Network. The baseline network is a VGG16 network with fixed convolutional weights (only the 3 fully connected layers were trained), pre-trained on the VGGFACE dataset. The output layer consists of 22 units: 2 linear units that give the valence and arousal predictions; 8 units equipped with softmax activation function that give the expression predictions; 12 units equipped with sigmoid activation function that give the action unit predictions. Let us mention here that no data augmentation techniques [65] have been utilized when training this baseline network with the cropped-aligned version of s-Aff-Wild2 database. We just normalised all images' pixel intensity values in the range $[-1, 1]$.

Table 4 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline in the Multi-Task Learning Challenge. Table 4 illustrates the evaluation of predictions on the s-Aff-Wild2 test set (in terms of the sum of the average CCC between valence and arousal, the average F1 score of the expression classes and the average F1 score of the action units); it further shows the baseline network results (VGG16). For reproducibility reasons, a link to a Github repository for each participating team's methodology exists and can be found in the corresponding leaderboard published in the official 4th ABAW Competition's website. Finally let us mention that the baseline network performance on the validation set is: 0.30.

4.2 Learning from Synthetic Data Challenge

In total, 51 Teams participated in the Learning from Synthetic Data Challenge. 21 Teams submitted their results. 10 Teams scored higher than the baseline and made valid submissions.

The winner of this Challenge is *HSE-NN* (that took the 3rd place in the Multi-Task Learning Challenge of the 3rd ABAW Competition) consisting of: Andrey Savchenko (HSE University, Russia).

Table 4. Multi-Task Learning Challenge results of participating teams' methods and baseline model; overall metric is in %; in bold is the best performing submission on s-Aff-Wild2 test set

Teams	Overall metric	Github
Situ-RUCAIM3 [89]	141.05 131.89 137.17 134.53 **143.61**	link
ICT-VIPL [53]	114.24 107.16 119.03 **119.45** 108.04	link
HSE-NN [70]	111.23 **112.99** 111.89 110.96 80.86	link
CNU_Sclab [59]	110.56 **111.35** 108.01 107.75 108.55	link
STAR-2022 [75]	**108.55**	link
HUST-ANT [52]	**107.12**	link
SSSIHL-DMACS [17]	**104.06** 96.66 101.72 93.21 98.34	link
DL_ISIR	**101.87** 93.57 99.46	link
USTC-AC [3]	87.11 92.45 92.69 93.29 **93.97**	link
CASIA-NLPR [72]	**91.38**	link
ITCNU [18]	59.40 57.22 65.27 57.52 **68.54**	link
Baseline [32]	28.00	-

Table 5. Learning from Synthetic Data Challenge results of participating teams' methods and baseline mode: metric is in %; in bold is the best performing submission on the test set, which consists of only real data of the Aff-Wild2 database.

Teams	Performance metric	Github
HSE-NN [70]	35.19 11.05 31.66 2.32 **37.18**	link
PPAA [50]	36.13 31.54 36.33 **36.51** 36.35	link
IXLAB [24]	33.74 34.96 **35.87** 34.93 35.84	link
ICT-VIPL [53]	32.51 32.81 34.60 **34.83** 31.42	link
HUST-ANT [52]	33.46 30.82 **34.83** 27.75 31.40	link
SSSIHL-DMACS [17]	32.66 33.35 **33.64**	link
STAR-2022 [75]	**32.40** 30.00 26.17	link
USTC-AC [3]	25.93 25.34 28.06 30.83 **30.92**	link
IMLAB [49]	**30.84** 29.76	link
Baseline [32]	30.00	

The runner up is *PPAA* consisting of: Jie Lei, Zhao Liu, Tong Li, Zeyu Zou, Xu Juan, Shuaiwei Wang, Guoyu Yang and Zunlei Feng (Zhejiang University of Technology; Ping An Life Insurance Of China Ltd).

In the third place is *IXLAB* consisting of: Jae-Yeop Jeong, Yeong-Gi Hong, JiYeon Oh, Sumin Hong, Jin-Woo Jeong (Seoul National University of Science and Technology, Korea), Yuchul Jung (Kumoh National Institute of Technology, Korea).

Baseline Network. The baseline network is a ResNet with 50 layers, pre-trained on ImageNet (ResNet50); its output layer consists of 6 units and is equipped with softmax activation function that gives the basic expression predictions. Let us mention here that no data augmentation techniques have been utilized when training this baseline network with the synthetic images. We just normalised all images' pixel intensity values in the range $[-1, 1]$.

Table 5 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline in the Learning from Synthetic Data Challenge. Table 5 illustrates the evaluation of predictions on the test set -which consists of only real images of the Aff-Wild2 database- (in terms of the average F1 score of the expression classes); it further shows the baseline network results (ResNet50). For reproducibility reasons, a link to a Github repository for each participating team's methodology exists and can be found in the corresponding leaderboard published in the official 4th ABAW Competition's website. Finally let us mention that the baseline network performance on the validation set is: 0.50.

5 Conclusion

In this paper we have presented the fourth Affective Behavior Analysis in-the-wild Competition (ABAW) 2022 held in conjunction with ECCV 2022. This Competition is a continuation of the first, second and third ABAW Competitions held in conjunction with IEEE FG 2020, ICCV 2021 and IEEE CVPR 2022, respectively. This Competition differentiates from the previous Competitions by including two Challenges: i) the Multi-Task- Learning (MTL) Challenge in which the goal is to create a system that learns at the same time (i.e., in a multi-task learning setting) to estimate valence and arousal, classify eight expressions (6 basic expressions plus the neutral state plus a category 'other' which denotes expressions/affective states other than the 6 basic ones) and detect twelve action units; ii) the Learning from Synthetic Data (LSD) Challenge in which the goal is to create a system that learns to recognise the six basic expressions (anger, disgust, fear, happiness, sadness, surprise) from artificially generated data (i.e., synthetic data) and generalise its knowledge to real-world (i.e., real) data. Each Challenge's corpora is derived from the Aff-Wild2 database.

The fourth ABAW Competition has been a very successful one with the participation of 55 Teams in the Multi-Task Learning Challenge and 51 Teams in the Learning from Synthetic Data Challenge. All teams' solutions were very interesting and creative, providing quite a push from the developed baselines.

References

1. Antoniadis, P., Pikoulis, I., Filntisis, P.P., Maragos, P.: An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. arXiv preprint arXiv:2107.03465 (2021)
2. Caridakis, G., Raouzaiou, A., Karpouzis, K., Kollias, S.: Synthesizing gesture expressivity based on real sequences. In: Workshop Programme, vol. 10, p. 19
3. Chang, Y., Wu, Y., Miao, X., Wang, J., Wang, S.: Multi-task learning for emotion descriptors estimation at the fourth ABAW challenge. arXiv preprint arXiv:2207.09716 (2022)
4. Darwin, C., Prodger, P.: The Expression of the Emotions in Man and Animals. Oxford University Press, Oxford (1998)
5. Deng, D.: Multiple emotion descriptors estimation at the ABAW3 challenge. arXiv preprint arXiv:2203.12845 (2022)
6. Deng, D., Chen, Z., Shi, B.E.: FAU, facial expressions, valence and arousal: a multi-task solution. arXiv preprint arXiv:2002.03557 (2020)
7. Deng, D., Chen, Z., Shi, B.E.: Multitask emotion recognition with incomplete labels. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 592–599. IEEE (2020)
8. Deng, D., Shi, B.E.: Estimating multiple emotion descriptors by separating description and inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2392–2400, June 2022
9. Deng, D., Wu, L., Shi, B.E.: Towards better uncertainty: iterative training of efficient networks for multitask emotion recognition. arXiv preprint arXiv:2108.04228 (2021)
10. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: RetinaFace: single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5203–5212 (2020)
11. Do, N.T., Nguyen-Quynh, T.T., Kim, S.H.: Affective expression analysis in-the-wild using multi-task temporal statistical deep learning model. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 624–628. IEEE (2020)
12. Dresvyanskiy, D., Ryumina, E., Kaya, H., Markitantov, M., Karpov, A., Minker, W.: An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild. arXiv preprint arXiv:2010.03692 (2020)
13. Ekman, P.: Facial action coding system (FACS). A human face (2002)
14. Frijda, N.H., et al.: The Emotions. Cambridge University Press, Cambridge (1986)
15. Gera, D., Balasubramanian, S.: Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. arXiv preprint arXiv:2009.14440 (2020)
16. Gera, D., Balasubramanian, S.: Affect expression behaviour analysis in the wild using consensual collaborative training. arXiv preprint arXiv:2107.05736 (2021)
17. Gera, D., Kumar, B.N.S., Kumar, B.V.R., Balasubramanian, S.: SS-MFAR: semi-supervised multi-task facial affect recognition. arXiv preprint arXiv:2207.09012 (2022)
18. Haider, I., Tran, M.T., Kim, S.H., Yang, H.J., Lee, G.S.: An ensemble approach for multiple emotion descriptors estimation using multi-task learning. arXiv preprint arXiv:2207.10878 (2022)
19. Han, S., Meng, Z., Khan, A.S., Tong, Y.: Incremental boosting convolutional neural network for facial action unit recognition. In: Advances in Neural Information Processing Systems, pp. 109–117 (2016)

20. Hoai, D.L., et al.: An attention-based method for action unit detection at the 3rd ABAW competition. arXiv preprint arXiv:2203.12428 (2022)
21. Jeong, E., Oh, G., Lim, S.: Multitask emotion recognition model with knowledge distillation and task discriminator. arXiv preprint arXiv:2203.13072 (2022)
22. Jeong, J.Y., Hong, Y.G., Kim, D., Jeong, J.W., Jung, Y., Kim, S.H.: Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2353–2358, June 2022
23. Jeong, J.Y., Hong, Y.G., Kim, D., Jung, Y., Jeong, J.W.: Facial expression recognition based on multi-head cross attention network. arXiv preprint arXiv:2203.13235 (2022)
24. Jeong, J.Y., Hong, Y.G., Oh, J., Hong, S., Jeong, J.W., Jung, Y.: Learning from synthetic data: facial expression classification based on ensemble of multi-task networks. arXiv preprint arXiv:2207.10025 (2022)
25. Ji, X., Ding, Y., Li, L., Chen, Y., Fan, C.: Multi-label relation modeling in facial action units detection. arXiv preprint arXiv:2002.01105 (2020)
26. Jiang, W., Wu, Y., Qiao, F., Meng, L., Deng, Y., Liu, C.: Facial action unit recognition with multi-models ensembling. arXiv preprint arXiv:2203.13046 (2022)
27. Jiang, W., Wu, Y., Qiao, F., Meng, L., Deng, Y., Liu, C.: Model level ensemble for facial action unit recognition at the 3rd ABAW challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2337–2344, June 2022
28. Jin, Y., Zheng, T., Gao, C., Xu, G.: A multi-modal and multi-task learning method for action unit and expression recognition. arXiv preprint arXiv:2107.04187 (2021)
29. Karas, V., Tellamekala, M.K., Mallol-Ragolta, A., Valstar, M., Schuller, B.W.: Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. arXiv preprint arXiv:2203.13285 (2022)
30. Karas, V., Tellamekala, M.K., Mallol-Ragolta, A., Valstar, M., Schuller, B.W.: Time-continuous audiovisual fusion with recurrence vs attention for in-the-wild affect recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2382–2391, June 2022
31. Kim, J.H., Kim, N., Won, C.S.: Facial expression recognition with swin transformer. arXiv preprint arXiv:2203.13472 (2022)
32. Kollias, D.: ABAW: learning from synthetic data & multi-task learning challenges. arXiv preprint arXiv:2207.01138 (2022)
33. 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 (2022)
34. 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
35. Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.: Deep neural network augmentation: generating faces for affect analysis. Int. J. Comput. Vis. **128**, 1455–1484 (2020). https://doi.org/10.1007/s11263-020-01304-3
36. Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.: Recognition of affect in the wild using deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1972–1979. IEEE (2017)

37. 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. 794–800. IEEE Computer Society (2020)
38. Kollias, D., Sharmanska, V., Zafeiriou, S.: Face behavior a la carte: expressions, affect and action units in a single network. arXiv preprint arXiv:1910.11111 (2019)
39. Kollias, D., Sharmanska, V., Zafeiriou, S.: Distribution matching for heterogeneous multi-task learning: a large-scale face study. arXiv preprint arXiv:2105.03790 (2021)
40. Kollias, D., et al.: Deep affect prediction in-the-wild: aff-wild database and challenge, deep architectures, and beyond. Int. J. Comput. Vis. **127**(6–7), 907–929 (2019). https://doi.org/10.1007/s11263-019-01158-4
41. Kollias, D., Zafeiriou, S.: Aff-wild2: extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)
42. Kollias, D., Zafeiriou, S.: A multi-task learning & generation framework: valence-arousal, action units & primary expressions. arXiv preprint arXiv:1811.07771 (2018)
43. Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855 (2019)
44. 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
45. Kollias, D., Zafeiriou, S.: Affect analysis in-the-wild: valence-arousal, expressions, action units and a unified framework. arXiv preprint arXiv:2103.15792 (2021)
46. Kollias, D., Zafeiriou, S.: Analysing affective behavior in the second ABAW2 competition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3652–3660 (2021)
47. Kuhnke, F., Rumberg, L., Ostermann, J.: Two-stream aural-visual affect analysis in the wild. arXiv preprint arXiv:2002.03399 (2020)
48. Le Hoai, D., et al.: An attention-based method for multi-label facial action unit detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2454–2459, June 2022
49. Lee, H., Lim, H., Lim, S.: BYEL: bootstrap on your emotion latent. arXiv preprint arXiv:2207.10003 (2022)
50. Lei, J., et al.: Mid-level representation enhancement and graph embedded uncertainty suppressing for facial expression recognition. arXiv preprint arXiv:2207.13235 (2022)
51. Li, I., et al.: Technical report for valence-arousal estimation on affwild2 dataset. arXiv preprint arXiv:2105.01502 (2021)
52. Li, S., et al.: Facial affect analysis: learning from synthetic data & multi-task learning challenges. arXiv preprint arXiv:2207.09748 (2022)
53. Li, Y., Sun, H., Liu, Z., Han, H.: Affective behaviour analysis using pretrained model with facial priori. arXiv preprint arXiv:2207.11679 (2022)
54. Liu, H., Zeng, J., Shan, S., Chen, X.: Emotion recognition for in-the-wild videos. arXiv preprint arXiv:2002.05447 (2020)
55. Malatesta, L., Raouzaiou, A., Karpouzis, K., Kollias, S.: Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. Appl. Intell. **30**(1), 58–64 (2009). https://doi.org/10.1007/s10489-007-0076-9

56. Mao, S., Fan, X., Peng, X.: Spatial and temporal networks for facial expression recognition in the wild videos. arXiv preprint arXiv:2107.05160 (2021)
57. Meng, L., et al.: Multi-modal emotion estimation for in-the-wild videos. arXiv preprint arXiv:2203.13032 (2022)
58. Meng, L., et al.: Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2345–2352, June 2022
59. Nguyen, D.K., Pant, S., Ho, N.H., Lee, G.S., Kim, S.H., Yang, H.J.: Multi-task cross attention network in facial behavior analysis. arXiv preprint arXiv:2207.10293 (2022)
60. Nguyen, H.H., Huynh, V.T., Kim, S.H.: An ensemble approach for facial behavior analysis in-the-wild video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2512–2517, June 2022
61. Nguyen, H.H., Huynh, V.T., Kim, S.H.: An ensemble approach for facial expression analysis in video. arXiv preprint arXiv:2203.12891 (2022)
62. Oh, G., Jeong, E., Lim, S.: Causal affect prediction model using a facial image sequence. arXiv preprint arXiv:2107.03886 (2021)
63. Pahl, J., Rieger, I., Seuss, D.: Multi-label class balancing algorithm for action unit detection. arXiv preprint arXiv:2002.03238 (2020)
64. Phan, K.N., Nguyen, H.H., Huynh, V.T., Kim, S.H.: Expression classification using concatenation of deep neural network for the 3rd ABAW3 competition. arXiv preprint arXiv:2203.12899 (2022)
65. Psaroudakis, A., Kollias, D.: MixAugment & Mixup: augmentation methods for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2367–2375 (2022)
66. Rajasekar, G.P., et al.: A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. arXiv preprint arXiv:2203.14779 (2022)
67. Russell, J.A.: Evidence of convergent validity on the dimensions of affect. J. Pers. Soc. Psychol. **36**(10), 1152 (1978)
68. Saito, J., Mi, X., Uchida, A., Youoku, S., Yamamoto, T., Murase, K.: Action units recognition using improved pairwise deep architecture. arXiv preprint arXiv:2107.03143 (2021)
69. Savchenko, A.V.: Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. arXiv preprint arXiv:2203.13436 (2022)
70. Savchenko, A.V.: HSE-NN team at the 4th ABAW competition: multi-task emotion recognition and learning from synthetic images. arXiv preprint arXiv:2207.09508 (2022)
71. Savchenko, A.V.: Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2359–2366, June 2022
72. Sun, H., Lian, Z., Liu, B., Tao, J., Sun, L., Cai, C.: Two-aspect information fusion model for ABAW4 multi-task challenge. arXiv preprint arXiv:2207.11389 (2022)
73. Tallec, G., Yvinec, E., Dapogny, A., Bailly, K.: Multi-label transformer for action unit detection. arXiv preprint arXiv:2203.12531 (2022)
74. Vu, M.T., Beurton-Aimar, M.: Multitask multi-database emotion recognition. arXiv preprint arXiv:2107.04127 (2021)
75. Wang, L., Li, H., Liu, C.: Hybrid CNN-transformer model for facial affect recognition in the ABAW4 challenge. arXiv preprint arXiv:2207.10201 (2022)

76. Wang, L., Qi, J., Cheng, J., Suzuki, K.: Action unit detection by exploiting spatial-temporal and label-wise attention with transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2470–2475, June 2022
77. Wang, L., Wang, S.: A multi-task mean teacher for semi-supervised facial affective behavior analysis. arXiv preprint arXiv:2107.04225 (2021)
78. Wang, L., Wang, S., Qi, J.: Multi-modal multi-label facial action unit detection with transformer. arXiv preprint arXiv:2203.13301 (2022)
79. Wang, S., Chang, Y., Wang, J.: Facial action unit recognition based on transfer learning. arXiv preprint arXiv:2203.14694 (2022)
80. Whissel, C.: The dictionary of affect in language. In: Plutchik, R., Kellerman, H. (eds.) Emotion: Theory, Research and Experience: Volume 4, The Measurement of Emotions. Academic, New York (1989)
81. Xie, H.X., Li, I., Lo, L., Shuai, H.H., Cheng, W.H., et al.: Technical report for valence-arousal estimation in ABAW2 challenge. arXiv preprint arXiv:2107.03891 (2021)
82. Xue, F., Tan, Z., Zhu, Y., Ma, Z., Guo, G.: Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. arXiv preprint arXiv:2203.13052 (2022)
83. Youoku, S., et al.: A multi-term and multi-task analyzing framework for affective analysis in-the-wild. arXiv preprint arXiv:2009.13885 (2020)
84. Yu, J., Cai, Z., He, P., Xie, G., Ling, Q.: Multi-model ensemble learning method for human expression recognition. arXiv preprint arXiv:2203.14466 (2022)
85. 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, pp. 34–41 (2017)
86. Zhang, S., An, R., Ding, Y., Guan, C.: Continuous emotion recognition using visual-audio-linguistic information: a technical report for ABAW3. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2376–2381, June 2022
87. Zhang, S., An, R., Ding, Y., Guan, C.: Continuous emotion recognition using visual-audio-linguistic information: a technical report for ABAW3. arXiv preprint arXiv:2203.13031 (2022)
88. Zhang, S., Ding, Y., Wei, Z., Guan, C.: Audio-visual attentive fusion for continuous emotion recognition. arXiv preprint arXiv:2107.01175 (2021)
89. Zhang, T., et al.: Emotion recognition based on multi-task learning framework in the ABAW4 challenge. arXiv preprint arXiv:2207.09373 (2022)
90. Zhang, W., Guo, Z., Chen, K., Li, L., Zhang, Z., Ding, Y.: Prior aided streaming network for multi-task affective recognitionat the 2nd ABAW2 competition. arXiv preprint arXiv:2107.03708 (2021)
91. Zhang, W., et al.: Prior aided streaming network for multi-task affective analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 3539–3549, October 2021
92. Zhang, W., et al.: Transformer-based multimodal information fusion for facial expression analysis. arXiv preprint arXiv:2203.12367 (2022)
93. Zhang, Y.H., Huang, R., Zeng, J., Shan, S., Chen, X.: M^{3}T: multi-modal continuous valence-arousal estimation in the wild. arXiv preprint arXiv:2002.02957 (2020)