ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges

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

Panagiotis Tzirakis
Hume AI, USA
panagiotis@hume.ai

Alice Baird
Hume AI, USA
alice@hume.ai

Alan Cowen
Hume AI, USA
alan@hume.ai

Stefanos Zafeiriou
Imperial College London, UK
s.zafeiriou@imperial.ac.uk

Abstract

The 5th ABAW Competition is part of the respective Workshop held in conjunction with IEEE CVPR 2023 and is a continuation of the Competitions held at ECCV 2022, IEEE CVPR 2022, ICCV 2021, IEEE FG 2020 and CVPR 2017 Conferences. It is dedicated at automatically analyzing affect. For this year's Competition, we feature two corpora: i) an extended version of the Aff-Wild2 database and ii) the Hume-Reaction dataset. The former database is an audiovisual (A/V) one of around 600 videos of around 3M frames and is annotated for: a) two continuous affect dimensions, valence (how positive/negative a person is) and arousal (how active/passive a person is); b) basic expressions (e.g. happiness, neutral); and c) action units (i.e., facial muscle actions). The latter dataset is A/V in which reactions of individuals to emotional stimuli have been annotated for seven emotional expression intensities. Thus the 5th ABAW Competition encompasses four Challenges: i) Valence-Arousal Estimation, ii) Expression Classification, iii) Action Unit Detection, and iv) Emotional Reaction Intensity Estimation. In this paper, we present these Challenges and their corpora, we outline the evaluation metrics and present the baseline systems and top performing teams' per Challenge along with their obtained performance. More information for the Competition can be found in: https://ibug.doc.ic.ac.uk/resources/cvpr-2023-5th-abaw.

1. Introduction

The fifth Affective Behavior Analysis in-the-wild (ABAW) Workshop and Competition has a unique aspect of fostering cross-pollination of different disciplines, bringing together experts (from academia, industry, and government) and researchers of mobile and ubiquitous computing, computer vision and pattern recognition, artificial intelligence and machine learning, multimedia, robotics, HCI, ambient intelligence and psychology. The diversity of human behavior, the richness of multi-modal data that arises from its analysis, and the multitude of applications that demand rapid progress in this area ensure that our events provide a timely and relevant discussion and dissemination platform.

The ABAW Workshop tackles the problem of affective behavior analysis in-the-wild, that is a major targeted characteristic of HCI systems used in real life applications. The target is to create machines and robots that are capable of understanding people's feelings, emotions and behaviors; thus, being able to interact in a 'human-centered' and engaging manner with them, and effectively serving them as their digital assistants. This interaction should not be dependent on the respective context, nor the human's age, sex, ethnicity, educational level, profession, or social position. As a result, the development of intelligent systems able to analyze human behaviors in-the-wild can contribute to generation of trust, understanding and closeness between humans and machines in real life environments.

The ABAW Workshop includes the respective Competition which utilizes two corpora: i) an extended version of the Aff-Wild2 database [28–30,32–38,82] and ii) the Hume-Reaction dataset. Aff-Wild2 database is an audiovisual one consisting of around 600 videos of around 3M frames and is annotated with respect to three different models of affect: a) dimensional affect (valence, which characterises an emotional state on a scale from positive to negative, and arousal, which characterises an emotional state on a scale from active to passive); b) categorical affect (six basic expressions -anger, disgust, fear, happiness, sadness, surprise- plus the neutral state); and c) action units (i.e., activations of facial muscles). The Hume-Reaction dataset is an audiovisual one in which reactions of individuals to emotional stimuli have

been annotated with respect to seven emotional expression intensities (i.e., adoration, amusement, anxiety, disgust, empathic pain, fear and surprise).

Using these introduced datasets, the 5th ABAW Competition addresses four contemporary affective computing problems: in the Valence-Arousal (VA) Estimation Challenge, valence and arousal have to be predicted; in the Expression (EXPR) Classification Challenge, 6 basic expressions, the neutral state and a category 'other' (that denotes affective states that do not belong to the categorical model of affect) have to be recognised; in the Action Unit (AU) Detection Challenge, 12 action units (AUs) have to be detected; in the Emotional Reaction Intensity (ERI) Estimation Challenge, seven fine-grained 'in-the-wild' emotions have to be predicted. The 3 former Challenges are based on the Aff-Wild2 database, whereas the latter Challenge is based on the Hume-Reaction dataset.

By providing the mentioned tasks in the 5th ABAW Competition, we aim to address research questions that are of interest to affective computing, machine learning and multimodal signal processing communities and encourage a fusion of their disciplines.

The fifth ABAW Competition, held in conjunction with the IEEE Computer Vision and Pattern Recognition Conference (CVPR), 2023 is a continuation of the successful series of ABAW Competitions held in conjunction with ECCV 2022, IEEE CVPR 2022, ICCV 2021, IEEE FG 2020 and IEEE CVPR 2017, with the participation of many teams coming from both academia and industry, from all across the world [1–4, 7–15, 17–26, 39–42, 45–49, 52–60, 63, 64, 66, 69–72, 75, 77, 79, 80, 83, 84, 86–89].

2. Competition Corpora

In the following, we provide a short overview of each Challenge's dataset. For the first three Challenges, we also describe the pre-processing steps that we carried out for cropping and aligning all provided images. These cropped and aligned images have been utilized in our baseline experiments.

2.1. Valence-Arousal Estimation Challenge

This Challenge's corpora include 594 videos (an augmented version of the Aff-Wild2 database) that contain annotations in terms of valence and arousal. Sixteen of these videos display two subjects, both of which have been annotated. In total, 2,993,081 frames, with 584 subjects have been annotated by four experts using the method proposed in [6]. Valence and arousal values range continuously in $[-1, 1]$. Figure 1 shows the 2D Valence-Arousal histogram of annotations.

Aff-Wild2 is split into training, validation and testing sets. Partitioning is done in a subject independent manner, in the sense that a person can appear strictly in only one of these sets.

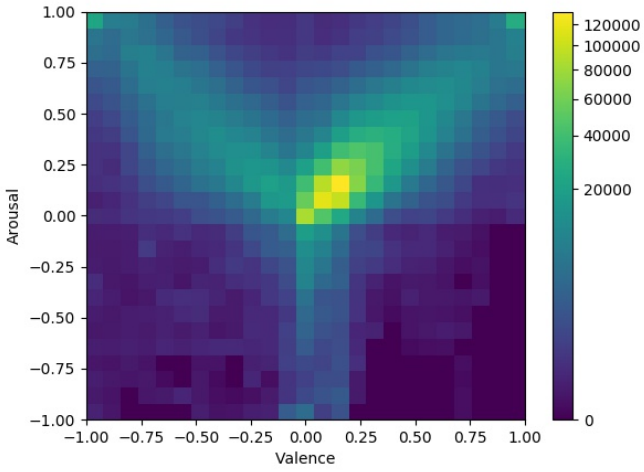

Figure 1. Valence-Arousal Estimation Challenge: 2D Valence-Arousal Histogram of Annotations in Aff-Wild2

2.2. Expression Classification Challenge

This Challenge's corpora include 546 videos in Aff-Wild2 that contain annotations in terms of the the 6 basic expressions, plus the neutral state, plus a category 'other' that denotes expressions/affective states other than the 6 basic ones. Seven of these videos display two subjects, both of which have been annotated. In total, 2,624,160 frames, with 437 subjects, 268 of which are male and 169 female, have been annotated by seven experts in a frame-by-frame basis. Table 1 shows the distribution of the expression annotations of Aff-Wild2.

Table 1. Expression Classification Challenge: Number of Annotated Images for each Expression

Expressions	No of Images
Neutral	468,069
Anger	36,627
Disgust	24,412
Fear	19,830
Happiness	245,031
Sadness	130,128
Surprise	68,077
Other	512,262

Aff-Wild2 is split into training, validation and testing sets, in a subject independent manner.

2.3. Action Unit Detection Challenge

This Challenge's corpora include 541 videos that contain annotations in terms of 12 AUs, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25 and AU26. Seven of these videos display two subjects, both of which have been annotated. In total, $2,627,632$ frames,with 438 subjects, 268 of which are male and 170 female, have been annotated in a semi-automatic procedure (that involves manual and automatic annotations). The annotation has been performed in a frame-by-frame basis. Table 2 shows the name of the twelve action units that have been annotated, the action that they are associated with and the distribution of their annotations in Aff-Wild2.

Table 2. Action Unit Detection Challenge: Distribution of AU Annotations in Aff-Wild2

Action Unit #	Action	Total Number of Activated AUs
AU 1	inner brow raiser	301,102
AU 2	outer brow raiser	139,936
AU 4	brow lowerer	386,689
AU 6	cheek raiser	619,775
AU 7	lid tightener	964,312
AU 10	upper lip raiser	854,519
AU 12	lip corner puller	602,835
AU 15	lip corner depressor	63,230
AU 23	lip tightener	78,649
AU 24	lip pressor	61,500
AU 25	lips part	1,596,055
AU 26	jaw drop	206,535

Aff-Wild2 is split into training, validation and testing sets, in a subject independent manner.

2.4. Emotional Reaction Intensity Estimation Challenge

For the Emotional Reaction Intensity Estimation Challenge Challenge, the large-scale and in-the-wild Hume Reaction dataset is used. The participants of this subchallenge explore a multi-output regression task, utilizing seven, self-annotated, nuanced classes of emotion: 'Adoration,' 'Amusement,' 'Anxiety,' 'Disgust,' 'Empathic-Pain,' 'Fear,' and 'Surprise.' The dataset is multimodal, and the video samples were recorded in uncontrolled environmental conditions in a wide variety of at-home recording settings with varying background and lightning noise conditions. In total, $2,222$ participants from two cultures, South Africa $(1,084)$ and the United States $(1,138)$, aged from 18.5 – 49.0 years old, recorded their facial and vocal reactions to a wide range of emotionally evocative videos via their webcam.

2.5. Aff-Wild2 Pre-Processing: Cropped & Cropped-Aligned Images

At first, all videos are splitted into independent frames. Then they are passed through the RetinaFace detector. so as to extract, for each frame, face bounding boxes and 5 facial landmarks. The images were cropped according the bounding box locations; then the images were provided to the participating teams. The 5 facial landmarks (two eyes, nose and two mouth corners) were used to perform similarity transformation. The resulting cropped and aligned images were additionally provided to the participating teams. Finally, the cropped and aligned images were utilized in our baseline experiments, described in Section 4.

All cropped and cropped-aligned images were resized to $112 \times 112 \times 3$ pixel resolution and their intensity values were normalized to $[-1, 1]$.

3. Evaluation Metrics Per Challenge

3.1. Valence-Arousal Estimation Challenge

The performance measure is the average between the Concordance Correlation Coefficient (CCC) of valence and arousal:

$$\mathcal{P}_{VA} = \frac{CCC_a + CCC_v}{2} \tag{1}$$

CCC evaluates the agreement between two time series (e.g., all video annotations and predictions) by scaling their correlation coefficient with their mean square difference. CCC takes values in the range $[-1, 1]$; high values are desired. CCC is defined as follows:

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

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.

3.2. Expression Classification Challenge

The performance measure is the average F1 Score across all 8 categories (i.e., macro F1 Score):

$$\mathcal{P}_{EXPR} = \frac{\sum_{expr} F_1^{expr}}{8} \tag{3}$$

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:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

3.3. Action Unit Detection Challenge

The performance measure is the average F1 Score across all 12 AUs. Therefore, the evaluation criterion for the Action Unit Detection Challenge is:

$$\mathcal{P}_{AU} = \frac{\sum_{au} F_1^{au}}{12} \tag{5}$$

3.4. Emotional Reaction Intensity Estimation Challenge

The performance measure is the average Pearson's Correlation Coefficient (ρ) across the 7 emotional reactions:

$$\mathcal{P}_{ERI} = \frac{\sum_{i=1}^{7} \rho^i}{7} \tag{6}$$

Pearson's Correlation Coefficient (ρ) takes values in the range $[-1, 1]$; high values are desired.

4. Participating Teams' and Baseline Methods' Results

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.

In this Section, we describe the baseline systems developed for each Challenge, as well as present the top-3 performing teams per Challenge. Finally, we present both participating teams' and baseline methods' obtained results.

4.1. Valence-Arousal Estimation Challenge

In total, 57 Teams participated in the Valence-Arousal Estimation Challenge. 26 Teams submitted their results. 8 Teams made invalid (incomplete) submissions, whilst surpassing the baseline. 8 Teams scored lower than the baseline. 10 Teams scored higher than the baseline and made valid submissions.

Table 3 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline and made valid submissions in the Valence-Arousal Estimation Challenge. Table 3 illustrates the CCC evaluation of valence and arousal predictions on the Aff-Wild2 test set; it further shows the baseline network results. The baseline is a ResNet with 50 layers, pre-trained on ImageNet (ResNet50) and with a (linear) output layer that gives final estimates for valence and arousal.

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 5th ABAW Competition's website.

As can be seen in Table 3, the winner of this Challenge is: *SituTech* consisting of: Chuanhe Liu, Xiaolong Liu, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Yuanyuan Deng, Zhaopei Huang, Liyu Meng, Yuchen Liu (Beijing Seek Truth Data Technology Services Co Ltd).

The runner up is: *Netease Fuxi Virtual Human* consisting of: Wei Zhang, Feng Qiu, Haodong Sun, Suzhen Wang, Zhimeng Zhang, Bowen Ma, Rudong An, Yu Ding (Netease Fuxi AI Lab).

Let us mention that both Teams have participated in our former Competitions at ECCV 2022, IEEE CVPR 2022 and ICCV 2021 and have ranked multiple times in the first, second and third positions in the Valence-Arousal Estimation, Expression Classification, Action Unit Detection and Multi-Task Learning Challenges.

In the third place is: *CBCR* consisting of: Su Zhang, Ziyuan Zhao, Cuntai Guan (Nanyang Technological University).

CBCR also participated in our former Competitions at IEEE CVPR 2022 and ICCV 2021, ranking in the second place in one Valence-Arousal Estimation Challenge.

It can be observed that SituTech's method achieved the overall best performance (evaluation criterion is the mean CCC of valence and arousal) and the best performance in arousal estimation. The method of Netease Fuxi Virtual Human Team although ranked second in overall performance, achieved the best performance in valence estimation. It can be observed that the difference in the performance between the winner and the runner-up is very small (0.6414 vs 0.6372). Finally let us mention that the baseline network performance on the validation set is: 0.24 for valence and 0.20 for arousal.

Table 3. Valence-Arousal Estimation Challenge's Results; Total Score is the average CCC between valence and arousal

Teams	Total Score	CCC-V	CCC-A
SituTech	0.6414	0.6193	0.6634
Netease Fuxi Virtual Human [90]	0.6372	0.6486	0.6258
CBCR [85]	0.5913	0.5526	0.6299
CtyunAI [91]	0.5666	0.5008	0.6325
HFUT-MAC [90]	0.5342	0.5234	0.5451
HSE-NN-SberAI [61]	0.5048	0.4818	0.5279
ACCC [92]	0.4842	0.4622	0.5062
PRL [67]	0.4661	0.5043	0.4279
SCLAB_CNU [51]	0.4640	0.4578	0.4703
USTC-AC [73]	0.2783	0.3245	0.2321
baseline	0.201	0.211	0.191

4.2. Expression Classification Challenge

In total, 67 Teams participated in the Expression Classification Challenge. 43 Teams submitted their results. 17 Teams made invalid (incomplete) submissions, whilst surpassing the baseline. 13 Teams scored lower than the base-

line. 13 Teams scored higher than the baseline and made valid submissions.

Table 4 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline and made valid submissions in the Expression Classification Challenge. Table 4 illustrates the average F1 score evaluation of predictions on the Aff-Wild2 test set; it further shows the baseline network results. The baseline is a VGG16 network with fixed (i.e., non-trainable) convolutional weights (only the 3 fully connected layers were trainable), pre-trained on the VGGFACE dataset and with an output layer equipped with softmax activation function which gives the 8 expression predictions.

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 5th ABAW Competition's website.

It can be seen in Table 4 that the winner of this Challenge is: *Netease Fuxi Virtual Human* consisting of the same members of the Netease Fuxi AI Lab as the ones described previously in the Valence-Arousal Estimation Challenge.

The runner up is: *SituTech* consisting of: Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, Wenqiang Jiang (Beijing Seek Truth Data Technology Services Co Ltd).

In the third place is: *CtyunAI* consisting of: Weiwei Zhou, Jiada Lu, Zhaolong Xiong, Weifeng Wang (Chinatelecom Cloud).

It can be observed that the difference in the performance between this Challenge's winner and the runner-up is quite small (0.4121 vs 0.4072). Finally let us mention that the baseline network performance on the validation set is: 0.23.

Table 4. Expression Classification Challenge's Results

Teams	F1
Netease Fuxi Virtual Human [90]	0.4121
SituTech	0.4072
CtyunAI [91]	0.3532
HFUT-MAC [90]	0.3337
HSE-NN-SberAI [61]	0.3292
AlphaAff [76]	0.3218
USTC-IAT-United [81]	0.3075
SSSIHL DMACS [16]	0.3047
SCLAB_CNU [51]	0.2949
Wall Lab [50]	0.2913
ACCC [92]	0.2846
RT_IAI [62]	0.2834
DGU-IPL [27]	0.2278
baseline	0.2050

4.3. Action Unit Detection Challenge

In total, 60 Teams participated in the Action Unit Detection Challenge. 37 Teams submitted their results. 12 Teams made invalid (incomplete) submissions, whilst surpassing the baseline. 13 Teams scored lower than the baseline. 12 Teams scored higher than the baseline and made valid submissions. Table 5 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline and made valid submissions in the Action Unit Detection Challenge.

Table 5 illustrates the average F1 score evaluation of predictions on the Aff-Wild2 test set; it further shows the baseline network results. The baseline is a VGG16 network with fixed convolutional weights (only the 3 fully connected layers were trained), pre-trained on the VGGFACE dataset and with an output layer equipped with sigmoid activation function which gives the 12 action unit predictions.

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 5th ABAW Competition's website.

In Table 5 can be seen that the winner of this Challenge is: *Netease Fuxi Virtual Human* consisting of the same members of Netease Fuxi AI Lab as the ones described previously in the Valence-Arousal Estimation Challenge.

The runner up is: *SituTech* consisting of: Chuanhe Liu, Wenqiang Jiang, Liyu Meng, Xiaolong Liu, Yuanyuan Deng (Beijing Seek Truth Data Technology Services Co Ltd).

In the third place is: *USTC-IAT-United* consisting of: Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, Wangyuan Zhu (University of Science and Technology of China).

Performance of the baseline on the validation set is: 0.39.

Table 5. Action Unit Detection Challenge's Results

Teams	F1
Netease Fuxi Virtual Human [90]	0.5549
SituTech	0.5422
USTC-IAT-United [81]	0.5144
SZFaceU [74]	0.5128
PRL [67]	0.5101
CtyunAI [91]	0.4887
HSE-NN-SberAI [61]	0.4878
USTC-AC [73]	0.4811
HFUT-MAC [90]	0.4752
SCLAB_CNU [51]	0.4563
USC IHP [78]	0.4292
ACCC [92]	0.3776
baseline	0.365

4.4. Emotional Reaction Intensity Estimation Challenge

The Emotional Reaction Intensity Estimation Challenge baseline results are depicted in Table 6. We also report the results obtained from submission to the Hume-Reaction MuSe 2022 [5] sub-challenge, as the same dataset was used. First, we observe that the audio modality provides low correlation (.0741), with the DEEPSPECTRUM feature set to produce better results than the eGeMAPS. This was expected as the audio is absent in several videos, making it challenging to model the modality.

As expected, the video modality provides a higher correlation than audio, with the baseline results to obtain .2801 ρ_c using Facial Action Units (FAU). There a number of other approaches that were submitted to MuSe 2022, but the best-performing model is obtained by the FaceRNET [31], which is comprised of a convolutional recurrent neural network with a routing mechanism on top.

Combining the audio and visual modalities does not seem to yield better results than the video models. In particular, the performance for the baseline (FAU+DEEPSPECTRUM) and the method of [44] drops. Only the ViPER model seems to see performance gains of .047 when adding the audio modality.

Table 6. Results for emotion reaction estimation sub-challenge. The mean Pearson's Correlation Coefficient (ρ) for the 7 emotional reaction classes is reported, along with the confidence intervals (where possible). The baseline results for the best of 5 fixed seeds are given for each feature and late fusion configuration. The respective mean and standard deviation of the results are provided in parentheses. In addition, the approaches submitted to the MuSe 2022 [5] are presented. A '-' is inserted when results are not available.

Features	[ρ] Development	Test
Audio		
Baseline (eGeMAPS) [5]	.0583 (.0504 ± .0069)	.0552 (.0479 ± .0062)
Baseline (DEEPSPECTRUM) [5]	.1087 (.0945 ± .0096)	.0741 (.0663 ± .0077)
Video		
Baseline (FAU) [5]	.2840 (.2828 ± .0016)	.2801 (.2777 ± .0017)
Baseline (VGGFACE 2) [5]	.2488 (.2441 ± .0027)	.1830 (.1985 ± .0088)
Resnet-18 [44]	.3893 (-)	- (-)
Former-DFER+MLGCN [68]	.3454 (-)	- (-)
ViPER [65]	.2978 (-)	.2859 (-)
FaceRNET [31]	**.3590** (-)	**.3607** (-)
Multimodal		
Baseline [5]	.2382 (.2350 ± 0.0016)	.2029 (.2014 ± .0086)
Resnet-18 + DEEPSPECTRUM [44]	.3968 (-)	- (-)
ViPER [65]	.3025 (-)	.2970 (-)

In total, 18 Teams participated in the Emotional Reaction Intensity Estimation Challenge. 9 Teams submitted their results, with 8 of them surpassing the baseline, and 7 of them making a valid submission. Table 7 presents the leaderboard and results of the latter 7 participating teams' algorithms.

Table 7 illustrates the mean Pearson's Correlation Coefficient results on the Hume Reaction test set. 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 5th ABAW Competition's website.

From Table 7, it can be observerd that the winner of this Challenge is: *HFUT-CVers* consisting of: Jia Li, Yin Chen, Xuesong Zhang, Jiantao Nie, Ziqiang Li, Yangchen Yu, Richang Hong, Meng Wang (Hefei University of Technology, China).

The runner-up is: *USTC-IAT-United* consisting of the same members from the University of Science and Technology of China as the ones described previously in the Action Unit Detection Challenge.

In the third place is: *Netease Fuxi Virtual Human* consisting of the same members of the Netease Fuxi AI Lab as the ones described previously in the Valence-Arousal Estimation Challenge.

Table 7. Emotional Reaction Intensity Estimation Challenge's Results

Teams	ρ
HFUT-CVers [43]	0.4734
USTC-IAT-United [81]	0.4380
Netease Fuxi Virtual Human [90]	0.4046
SituTech	0.3935
CASIA-NLPR	0.3865
USTC-AC [73]	0.3730
NISL-2023 [90]	0.3667
HFUT-MAC [90]	0.2527

5. Conclusion

In this paper we have presented the fifth Affective Behavior Analysis in-the-wild Competition (ABAW) held in conjunction with IEEE CVPR 2023. This Competition is a continuation of the series of ABAW Competitions. This Competition comprises four Challenges targeting: i) Valence-Arousal Estimation, ii) Expression Classification (8 categories), iii) Action Unit Detection (12 action units) and iv) Emotional Reaction Intensity Estimation. The databases utilized for this Competition are an extended version of Aff-Wild2 and the Hume-Reaction dataset.

The fifth ABAW Competition has been a very successful one with the participation of 57 Teams in the Valence-Arousal Estimation Challenge, 67 Teams in the Expression Classification Challenge, 60 Teams in the Action Unit Detection Challenge and 18 Teams in the Emotional Reaction Intensity Estimation Challenge. All teams' solutions were very interesting and creative, providing quite a push from the developed baselines.

References

[1] Panagiotis Antoniadis, Ioannis Pikoulis, Panagiotis P Filntisis, and Petros Maragos. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. *arXiv preprint arXiv:2107.03465*, 2021.

[2] Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. Fatauva-net : An integrated deep learning framework for facial attribute recognition, action unit (au) detection, and valence-arousal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[3] Yanan Chang, Yi Wu, Xiangyu Miao, Jiahe Wang, and Shangfei Wang. Multi-task learning for emotion descriptors estimation at the fourth abaw challenge. *arXiv preprint arXiv:2207.09716*, 2022.

[4] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 19–26. ACM, 2017.

[5] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, et al. The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 5–14, 2022.

[6] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[7] Didan Deng. Multiple emotion descriptors estimation at the abaw3 challenge. *arXiv preprint arXiv:2203.12845*, 2022.

[8] Didan Deng, Zhaokang Chen, and Bertram E Shi. Fau, facial expressions, valence and arousal: A multi-task solution. *arXiv preprint arXiv:2002.03557*, 2020.

[9] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020.

[10] Didan Deng, Liang Wu, and Bertram E Shi. Towards better uncertainty: Iterative training of efficient networks for multitask emotion recognition. *arXiv preprint arXiv:2108.04228*, 2021.

[11] Nhu-Tai Do, Tram-Tran Nguyen-Quynh, and Soo-Hyung Kim. Affective expression analysis in-the-wild using multitask temporal statistical deep learning model. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 624–628. IEEE, 2020.

[12] Denis Dresvyanskiy, Elena Ryumina, Heysem Kaya, Maxim Markitantov, Alexey Karpov, and Wolfgang Minker. An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild. *arXiv preprint arXiv:2010.03692*, 2020.

[13] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. *arXiv preprint arXiv:2009.14440*, 2020.

[14] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using consensual collaborative training. *arXiv preprint arXiv:2107.05736*, 2021.

[15] Darshan Gera, Badveeti Naveen Siva Kumar, Bobbili Veerendra Raj Kumar, and S Balasubramanian. Ss-mfar: Semi-supervised multi-task facial affect recognition. *arXiv preprint arXiv:2207.09012*, 2022.

[16] Darshan Gera, Badveeti Naveen Siva Kumar, Bobbili Veerendra Raj Kumar, and S Balasubramanian. Abaw: Facial expression recognition in the wild. *arXiv preprint arXiv:2303.09785*, 2023.

[17] Irfan Haider, Minh-Trieu Tran, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. An ensemble approach for multiple emotion descriptors estimation using multi-task learning. *arXiv preprint arXiv:2207.10878*, 2022.

[18] Shizhong Han, Zibo Meng, Ahmed-Shehab Khan, and Yan Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in neural information processing systems*, pages 109–117, 2016.

[19] Duy Le Hoai, Eunchae Lim, Eunbin Choi, Sieun Kim, Sudarshan Pant, Guee-Sang Lee, Soo-Huyng Kim, and Hyung-Jeong Yang. An attention-based method for action unit detection at the 3rd abaw competition. *arXiv preprint arXiv:2203.12428*, 2022.

[20] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Yuchul Jung, and Jin-Woo Jeong. Facial expression recognition based on multi-head cross attention network. *arXiv preprint arXiv:2203.13235*, 2022.

[21] Jae-Yeop Jeong, Yeong-Gi Hong, JiYeon Oh, Sumin Hong, Jin-Woo Jeong, and Yuchul Jung. Learning from synthetic data: Facial expression classification based on ensemble of multi-task networks. *arXiv preprint arXiv:2207.10025*, 2022.

[22] Xianpeng Ji, Yu Ding, Lincheng Li, Yu Chen, and Changjie Fan. Multi-label relation modeling in facial action units detection. *arXiv preprint arXiv:2002.01105*, 2020.

[23] Wenqiang Jiang, Yannan Wu, Fengsheng Qiao, Liyu Meng, Yuanyuan Deng, and Chuanhe Liu. Facial action unit recognition with multi-models ensembling. *arXiv preprint arXiv:2203.13046*, 2022.

[24] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021.

[25] Vincent Karas, Mani Kumar Tellamekala, Adria Mallol-Ragolta, Michel Valstar, and Björn W Schuller. Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. *arXiv preprint arXiv:2203.13285*, 2022.

[26] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer. *arXiv preprint arXiv:2203.13472*, 2022.

[27] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Multimodal facial expression recognition with transformer-based fusion networks and dynamic sampling. *arXiv preprint arXiv:2303.08419*, 2023.

[28] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022.

[29] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023.

[30] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1972–1979. IEEE, 2017.

[31] Dimitrios Kollias, Andreas Psaroudakis, Anastasios Arsenos, and Paraskeui Theofilou. Facernet: a facial expression intensity estimation network. *arXiv preprint arXiv:2303.00180*, 2023.

[32] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. IEEE Computer Society, 2020.

[33] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[34] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[35] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, and Georgios Tagaris. Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems*, 4(2):119–131, 2018.

[36] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[37] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[38] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.

[39] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. *arXiv preprint arXiv:2002.03399*, 2020.

[40] Hyungjun Lee, Hwangyu Lim, and Sejoon Lim. Byel: Bootstrap on your emotion latent. *arXiv preprint arXiv:2207.10003*, 2022.

[41] Jie Lei, Zhao Liu, Zeyu Zou, Tong Li, Xu Juan, Shuaiwei Wang, Guoyu Yang, and Zunlei Feng. Mid-level representation enhancement and graph embedded uncertainty suppressing for facial expression recognition. *arXiv preprint arXiv:2207.13235*, 2022.

[42] I Li et al. Technical report for valence-arousal estimation on affwild2 dataset. *arXiv preprint arXiv:2105.01502*, 2021.

[43] Jia Li, Yin Chen, Xuesong Zhang, Jiantao Nie, Yangchen Yu, Ziqiang Li, Meng Wang, and Richang Hong. Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. *arXiv preprint arXiv:2303.09164*, 2023.

[44] Jia Li, Ziyang Zhang, Junjie Lang, Yueqi Jiang, Liuwei An, Peng Zou, Yangyang Xu, Sheng Gao, Jie Lin, Chunxiao Fan, Xiao Sun, and Meng Wang. Hybrid multimodal feature extraction, mining and fusion for sentiment analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, MuSe' 22, page 81–88, New York, NY, USA, 2022. Association for Computing Machinery.

[45] Siyang Li, Yifan Xu, Huanyu Wu, Dongrui Wu, Yingjie Yin, Jiajiong Cao, and Jingting Ding. Facial affect analysis: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.09748*, 2022.

[46] Yifan Li, Haomiao Sun, Zhaori Liu, and Hu Han. Affective behaviour analysis using pretrained model with facial priori. *arXiv preprint arXiv:2207.11679*, 2022.

[47] Hanyu Liu, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Emotion recognition for in-the-wild videos. *arXiv preprint arXiv:2002.05447*, 2020.

[48] Shuyi Mao, Xinqi Fan, and Xiaojiang Peng. Spatial and temporal networks for facial expression recognition in the wild videos. *arXiv preprint arXiv:2107.05160*, 2021.

[49] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Yuanyuan Deng, Ruichen Li, Yannan Wu, Jinming Zhao, et al. Multi-modal emotion estimation for in-the-wild videos. *arXiv preprint arXiv:2203.13032*, 2022.

[50] Onur Cezmi Mutlu, Mohammadmahdi Honarmand, Saimourya Surabhi, and Dennis P Wall. Tempt: Temporal consistency for test-time adaptation. *arXiv preprint arXiv:2303.10536*, 2023.

[51] Dang-Khanh Nguyen, Ngoc-Huynh Ho, Sudarshan Pant, and Hyung-Jeong Yang. A transformer-based approach to video frame-level prediction in affective behaviour analysis in-the-wild. *arXiv preprint arXiv:2303.09293*, 2023.

[52] Dang-Khanh Nguyen, Sudarshan Pant, Ngoc-Huynh Ho, Guee-Sang Lee, Soo-Huyng Kim, and Hyung-Jeong Yang. Multi-task cross attention network in facial behavior analysis. *arXiv preprint arXiv:2207.10293*, 2022.

[53] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial expression analysis in video. *arXiv preprint arXiv:2203.12891*, 2022.

[54] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a facial image sequence. *arXiv preprint arXiv:2107.03886*, 2021.

[55] Jaspar Pahl, Ines Rieger, and Dominik Seuss. Multi-label class balancing algorithm for action unit detection. *arXiv preprint arXiv:2002.03238*, 2020.

[56] Kim Ngan Phan, Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. Expression classification using concatenation of deep neural network for the 3rd abaw3 competition. *arXiv preprint arXiv:2203.12899*, 2022.

[57] Gnana Praveen Rajasekar, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro Koerich, Patrick Cardinal, and

Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. *arXiv preprint arXiv:2203.14779*, 2022.

[58] Junya Saito, Xiaoyu Mi, Akiyoshi Uchida, Sachihiro Youoku, Takahisa Yamamoto, and Kentaro Murase. Action units recognition using improved pairwise deep architecture. *arXiv preprint arXiv:2107.03143*, 2021.

[59] Andrey V Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv preprint arXiv:2203.13436*, 2022.

[60] Andrey V Savchenko. Hse-nn team at the 4th abaw competition: Multi-task emotion recognition and learning from synthetic images. *arXiv preprint arXiv:2207.09508*, 2022.

[61] Andrey V Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition. *arXiv preprint arXiv:2303.09162*, 2023.

[62] Tao Shu, Xinke Wang, Ruotong Wang, Chuang Chen, Yixin Zhang, and Xiao Sun. Mutilmodal feature extraction and attention-based fusion for emotion estimation in videos. *arXiv preprint arXiv:2303.10421*, 2023.

[63] Haiyang Sun, Zheng Lian, Bin Liu, Jianhua Tao, Licai Sun, and Cong Cai. Two-aspect information fusion model for abaw4 multi-task challenge. *arXiv preprint arXiv:2207.11389*, 2022.

[64] Gauthier Tallec, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. Multi-label transformer for action unit detection. *arXiv preprint arXiv:2203.12531*, 2022.

[65] Lorenzo Vaiani, Moreno La Quatra, Luca Cagliero, and Paolo Garza. Viper: Video-based perceiver for emotion recognition. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, page 67–73, New York, NY, USA, 2022. Association for Computing Machinery.

[66] Manh Tu Vu and Marie Beurton-Aimar. Multitask multi-database emotion recognition. *arXiv preprint arXiv:2107.04127*, 2021.

[67] Tu Vu, Van Thong Huynh, and Soo Hyung Kim. Vision transformer for action units detection. *arXiv preprint arXiv:2303.09917*, 2023.

[68] Kexin Wang, Zheng Lian, Licai Sun, Bin Liu, Jianhua Tao, and Yin Fan. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, page 75–80, New York, NY, USA, 2022. Association for Computing Machinery.

[69] Lingfeng Wang, Haocheng Li, and Chunyin Liu. Hybrid cnn-transformer model for facial affect recognition in the abaw4 challenge. *arXiv preprint arXiv:2207.10201*, 2022.

[70] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis. *arXiv preprint arXiv:2107.04225*, 2021.

[71] Lingfeng Wang, Shisen Wang, and Jin Qi. Multi-modal multi-label facial action unit detection with transformer. *arXiv preprint arXiv:2203.13301*, 2022.

[72] Shangfei Wang, Yanan Chang, and Jiahe Wang. Facial action unit recognition based on transfer learning. *arXiv preprint arXiv:2203.14694*, 2022.

[73] Shangfei Wang, Yanan Chang, Yi Wu, Xiangyu Miao, Jiaqiang Wu, Zhouan Zhu, Jiahe Wang, and Yufei Xiao. Facial affective behavior analysis method for 5th abaw competition. *arXiv preprint arXiv:2303.09145*, 2023.

[74] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Weicheng Xie, Linlin Shen, et al. Spatio-temporal au relational graph representation learning for facial action units detection. *arXiv preprint arXiv:2303.10644*, 2023.

[75] Hong-Xia Xie, I Li, Ling Lo, Hong-Han Shuai, Wen-Huang Cheng, et al. Technical report for valence-arousal estimation in abaw2 challenge. *arXiv preprint arXiv:2107.03891*, 2021.

[76] Fanglei Xue, Yifan Sun, and Yi Yang. Exploring expression-related self-supervised learning for affective behaviour analysis. *arXiv preprint arXiv:2303.10511*, 2023.

[77] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. *arXiv preprint arXiv:2203.13052*, 2022.

[78] Yufeng Yin, Minh Tran, Di Chang, Xinrui Wang, and Mohammad Soleymani. Multi-modal facial action unit detection with large pre-trained models for the 5th competition on affective behavior analysis in-the-wild. *arXiv preprint arXiv:2303.10590*, 2023.

[79] Sachihiro Youoku, Yuushi Toyoda, Takahisa Yamamoto, Junya Saito, Ryosuke Kawamura, Xiaoyu Mi, and Kentaro Murase. A multi-term and multi-task analyzing framework for affective analysis in-the-wild. *arXiv preprint arXiv:2009.13885*, 2020.

[80] Jun Yu, Zhongpeng Cai, Peng He, Guocheng Xie, and Qiang Ling. Multi-model ensemble learning method for human expression recognition. *arXiv preprint arXiv:2203.14466*, 2022.

[81] Jun Yu, Zhongpeng Cai, Renda Li, Gongpeng Zhao, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Exploring large-scale unlabeled faces to enhance facial expression recognition. *arXiv preprint arXiv:2303.08617*, 2023.

[82] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017.

[83] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. *arXiv preprint arXiv:2203.13031*, 2022.

[84] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Audio-visual attentive fusion for continuous emotion recognition. *arXiv preprint arXiv:2107.01175*, 2021.

[85] Su Zhang, Ziyuan Zhao, and Cuntai Guan. Multimodal continuous emotion recognition: A technical report for abaw5. *arXiv preprint arXiv:2303.10335*, 2023.

[86] Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, and Fengyuan Zhang. Emotion recognition based on multi-task learning framework in the abaw4 challenge. *arXiv preprint arXiv:2207.09373*, 2022.

[87] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021.

[88] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022.

[89] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. m^3 t: Multi-modal continuous valence-arousal estimation in the wild. *arXiv preprint arXiv:2002.02957*, 2020.

[90] Ziyang Zhang, Liuwei An, Zishun Cui, Tengteng Dong, et al. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge. *arXiv preprint arXiv:2303.09158*, 2023.

[91] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer. *arXiv preprint arXiv:2303.08356*, 2023.

[92] Peng Zou, Rui Wang, Kehua Wen, Yasi Peng, and Xiao Sun. Spatial-temporal transformer for affective behavior analysis. *arXiv preprint arXiv:2303.10561*, 2023.