Analysing Affective Behavior in the second ABAW2 Competition

Dimitrios Kollias
University of Greenwich, UK
D.Kollias@greenwich.ac.uk
dkollias1@gmail.com

Stefanos Zafeiriou
Imperial College London, UK

Abstract

The Affective Behavior Analysis in-the-wild (ABAW2) 2021 Competition is the second Competition -following the first very successful ABAW Competition held in conjunction with IEEE Conference on Face and Gesture Recognition 2020- that aims at automatically analyzing affect. ABAW2 is split into three Challenges, each one addressing one of the three main behavior tasks of Valence-Arousal Estimation, Seven Basic Expression Classification and Twelve Action Unit Detection. All three Challenges are based on a common benchmark database, Aff-Wild2, which is a large scale in-the-wild database and the first one to be annotated for all these three tasks.

In this paper, we describe this Competition, to be held in conjunction with the International Conference on Computer Vision (ICCV) 2021. We present the three Challenges, with the utilized Competition corpora. We outline the evaluation metrics and present both the baseline systems and the top-5 performing teams' per Challenge; finally we present the obtained results of the baseline systems and of all participating teams. More information regarding the Competition, the leaderboard of each Challenge and details for accessing the utilized database, are provided in the Competition website: https://ibug.doc.ic.ac.uk/resources/iccv-2021-2nd-abaw/.

1. Introduction

The proposed Workshop tackles the problem of affective behavior analysis in-the-wild, which is a major targeted characteristic of HCI systems used in real life applications. The current 5th societal revolution aims at merging the physical and cyber spaces, providing services that contribute to people's well-being. 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.

Affective behavior analysis in diverse environments, such as in people's homes, in their work, operational or industrial environments, will have a positive societal impact. It will provide machines and robots with the ability to interact and assist people in an effective and natural way. Through human affect recognition, the reactions of the machine, or robot, will be consistent with people's expectations and emotions; their verbal and non-verbal interactions will be positively received by humans. Moreover, 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.

Representing human emotions has been a basic topic of research in psychology. The most frequently used emotion representation is the categorical one, including the seven basic categories, i.e., Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral [12]. Discrete emotion representation can also be described in terms of the Facial Action Coding System (FACS) model, in which all possible facial actions are described in terms of Action Units (AUs) [11]. Finally, the dimensional model of affect [55, 49] has been proposed as a means to distinguish between subtly different displays of affect and encode small changes in the intensity of each emotion on a continuous scale. The 2-D Valence and Arousal (VA) Space (valence shows how positive or negative an emotional state is, whereas arousal shows how passive or active it is) is the most usual dimensional emotion representation, depicted in Figure 1.

There are a number of related applications spread across a variety of fields, such as medicine, health, driver fatigue, monitoring, e-learning, marketing, entertainment, lie detection, law [1, 21, 59, 30, 44, 22].

The second ABAW2 Competition 2021 is a continuation

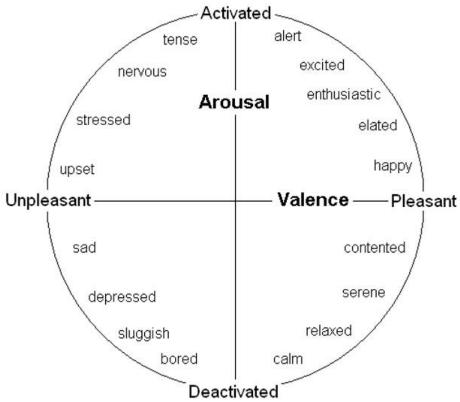

Figure 1. The 2D Valence-Arousal Space

of the first ABAW Competition 2020 [1] held in conjunction with IEEE Conference on Face and Gesture Recognition, which targeted, for the first time, dimensional (in terms of valence and arousal) [7, 40, 64, 9, 32, 4, 3, 37], categorical (in terms of the seven basic emotions) [39, 13, 10, 58, 41, 34] and facial action unit analysis and recognition [46, 19, 28, 16, 27, 6]. The ABAW2 Competition contains three Challenges, which are based on the same in-the-wild database, the (i) Valence-Arousal Estimation Challenge, (ii) Seven Basic Expression Classification Challenge and (iii) Twelve Action Unit Detection Challenge. These Challenges produce a significant step forward when compared to previous events.

In particular, they use the Aff-Wild2 [26, 35, 36, 31, 33], the first comprehensive benchmark for all three affect recognition tasks in-the-wild: the Aff-Wild2 database extends the Aff-Wild [29, 60, 25], with more videos and annotations for all behavior tasks. Aff-Wild consists of 298 videos, displaying reactions of 200 subjects, with a total video duration of about 30 hours, and 1,250,000 video frames, annotated in terms of valence and arousal. It has been used in the Aff-Wild Challenge in CVPR 2017, with participation of more than 10 research groups. To generate Aff-Wild2, we added 266 more videos, displaying the reactions of 266 more subjects, with a duration of more than 18 hours, and 1,500,000 frames. Aff-Wild2 includes extended spontaneous facial behaviors in arbitrary recording conditions and a significantly increased number of different subjects (466; 280 of which are males and 186 females) and frames (around 2,800,000).

The remainder of this paper is organised as follows. We introduce the Competition corpora in Section 2, the Competition evaluation metrics in Section 3, the developed baseline and the top-5 performing teams per Challenge, along with the obtained results in Section 4, before concluding in Section 5.

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

2. Competition Corpora

The second Affective Behavior Analysis in-the-wild (ABAW2) Competition relies on the Aff-Wild2 database [26, 35, 36, 31, 33]. Aff-Wild2 is the first ever database annotated for all three main behavior tasks: valence-arousal estimation, action unit detection and basic expression classification. These three tasks form the three Challenges of this Competition.

Aff-Wild2 consists of 548 videos with $2,813,201$ frames. Sixteen of these videos display two subjects (both have been annotated). All videos have been collected from YouTube. Aff-Wild2 is an extension of Aff-Wild [29, 60, 25]; 260 more YouTube videos, with $1,413,000$ frames, have been added to Aff-Wild. Aff-Wild was the first large scale, captured in-the-wild, dimensionally annotated database, containing 298 YouTube videos that display subjects reacting to a variety of stimuli. Aff-Wild2 shows both subtle and extreme human behaviours in real-world settings. The total number of subjects in Aff-Wild2 is 458; 279 of them are males and 179 females.

The Aff-Wild2 database, in all Challenges, is split into training, validation and test set. 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. The training and validation data contain the videos and their corresponding annotation. Furthermore, to facilitate training, especially for people that do not have access to face detectors/tracking algorithms, we provide bounding boxes and landmarks for the face(s) in the videos (we also provide the aligned faces). At a later stage, the test set without annotations will be given to the participants. Again, we will provide bounding boxes and landmarks for the face(s) in the videos (we will also provide the aligned faces).

In the following, we provide a short overview of each Challenge's dataset and refer the reader to the original work for a more complete description. Finally, we describe the pre-processing steps that we carried out for cropping and aligning the images of Aff-Wild2. The cropped and aligned images have been utilized in our baseline experiments.

2.1. Aff-Wild2: Valence-Arousal Annotation

545 videos in Aff-Wild2 contain annotations in terms of valence-arousal. Sixteen of these videos display two subjects, both of which have been annotated. In total, $2,786,201$ frames, with 455 subjects, 277 of which are male and 178 female, have been annotated by four experts using the method proposed in [5]. The annotators watched each video and provided their (frame-by-frame) annotations through a joystick. A time-continuous annotation was generated for each affect dimension. Valence and arousal values range continuously in $[-1, 1]$. The final label values were the mean of those four annotations. The mean inter-annotation correlation is 0.63 for valence and 0.60 for arousal. Let us note here that all subjects present in each video have been annotated. Figure 2 shows the 2D Valence-Arousal histogram of annotations of Aff-Wild2.

Aff-Wild2 is currently the largest (and audiovisual) in-the-wild database annotated for valence and arousal.

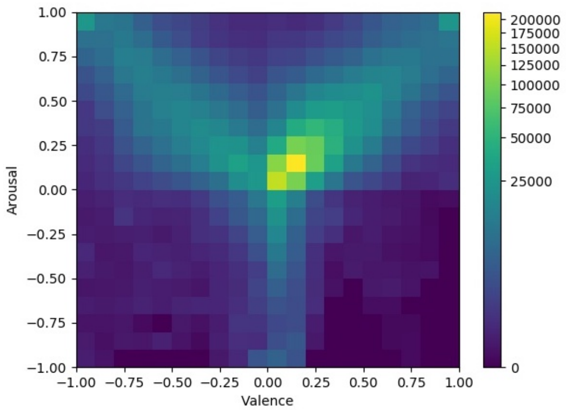

Figure 2. 2D Valence-Arousal Histogram of Aff-Wild2

Aff-Wild2 is split into three subsets: training, validation and test. Partitioning is done in a subject independent manner, in the sense that a person can appear only in one of those three subsets. The resulting training, validation and test subsets consist of 346, 68 and 131 videos, respectively; the resulting training, validation and test subsets contain 5, 3 and 8, respectively, videos that display two subjects.

2.2. Aff-Wild2: Seven Basic Expression Annotation

539 videos in Aff-Wild2 contain annotations in terms of the seven basic expressions. Seven of these videos display two subjects, both of which have been annotated. In total, $2,595,572$ frames, with 431 subjects, 265 of which are male and 166 female, have been annotated by seven experts in a frame-by-frame basis. A platform-tool was developed in order to split each video into frames and let the experts annotate each videoframe. Let us mention that in this platform-tool, an expert could score a videoframe as having either one of the seven basic expressions or none (since there are affective states other than the seven basic expressions).

Due to subjectivity of annotators and wide ranging levels of images' difficulty, there were some disagreements among annotators. We decided to keep only the annotations on which at least six (out of seven) experts agreed. Table 1 shows the distribution of the seven basic expression annotations of Aff-Wild2.

Aff-Wild2 is currently the largest (and audiovisual) in-the-wild database annotated in terms of the seven basic expressions.

Table 1. Number of Annotated Images in Each of the Seven Basic Expressions

Basic Expression	No of Images
Neutral	538,411
Anger	52,005
Disgust	31,138
Fear	26,062
Happiness	395,352
Sadness	173,842
Surprise	99,863

Aff-Wild2 is split into three subsets: training, validation and test. Partitioning is done in a subject independent manner. The resulting training, validation and test subsets consist of 250, 70 and 222 videos, respectively; the resulting training, validation and test subsets contain 3, 0 and 1, respectively, videos that display two subjects.

2.3. Aff-Wild2: Twelve Action Unit Annotation

534 videos in Aff-Wild2 contain annotations in terms of twelve action units. Seven of these videos display two subjects, both of which have been annotated. In total, $2,565,169$ frames, with 426 subjects, 262 of which are male and 164 female, have been annotated in a semi-automatic procedure (that involves manual and automatic annotations). Aff-Wild2 has been annotated for the occurrence of twelve action units 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.

Aff-Wild2 is currently the largest (and audiovisual) in-the-wild database annotated in terms of action units.

Table 2. Distribution of AU annotations in Aff-Wild2

Action Unit #	Action	Total Number of Activated AUs
AU 1	inner brow raiser	294,591
AU 2	outer brow raiser	136,569
AU 4	brow lowerer	384,969
AU 6	cheek raiser	618,929
AU 7	lid tightener	618,929
AU 10	upper lip raiser	845,793
AU 12	lip corner puller	598,699
AU 15	lip corner depressor	62,954
AU 23	lip tightener	77,793
AU 24	lip pressor	61,460
AU 25	lips part	1,579,262
AU 26	jaw drop	202,447

Aff-Wild2 is split into three subsets: training, validation and test. Partitioning is done in a subject independent manner. The resulting training, validation and test subsets consist of 302, 105 and 127 videos, respectively; the resulting training, validation and test subsets contain 3, 0 and 4, respectively, videos that display two subjects.

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

At first, we split all videos into images (frames). Then, the SSH detector [43] based on the ResNet [17] and trained on the WiderFace dataset [57] was used to extract face bounding boxes from all the images. The cropped images according to these bounding boxes were provided to the participating teams. Also, 5 facial landmarks (two eyes, nose and two mouth corners) were extracted and 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.

3. Evaluation Metrics Per 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. Valence-Arousal Estimation Challenge

The Concordance Correlation Coefficient (CCC) is widely used in measuring the performance of dimensional emotion recognition methods, such as in the series of AVEC challenges [48]. 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. In this way, predictions that are well correlated with the annotations but shifted in value are penalized in proportion to the deviation. CCC takes values in the range $[-1, 1]$, where $+1$ indicates perfect concordance and -1 denotes perfect discordance. The highest the value of the CCC the better the fit between annotations and predictions, and therefore 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 mean value of CCC for valence and arousal estimation will be adopted as the main evaluation criterion.

$$\mathcal{E}_{total} = \frac{\rho_a + \rho_v}{2}, \tag{2}$$

3.2. Seven Basic Expression Classification Challenge

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 reaches its best value at 1 and its worst score at 0. The F_1 score is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

The F_1 score for emotions is computed based on a per-frame prediction (an emotion category is specified in each frame).

Total accuracy (denoted as $\mathcal{T}Acc$) is defined on all test samples and is the fraction of predictions that the model got right. Total accuracy reaches its best value at 1 and its worst score at 0. It is defined as:

$$\mathcal{T}Acc = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4}$$

A weighted average between the F_1 score and the total accuracy, $\mathcal{T}Acc$, will be the main evaluation criterion:

$$\mathcal{E}_{total} = 0.67 \times F_1 + 0.33 * \mathcal{T}Acc, \tag{5}$$

3.3. Twelve Action Unit Detection Challenge

To obtain the overall score for the AU detection Challenge, we first obtain the F_1 score for each AU independently, and then compute the (unweighted) average over all 12 AUs (denoted as $\mathcal{A}F_1$) :

$$\mathcal{A}F_1 = \sum_{i=1}^{12} F_1^i \tag{6}$$

The F_1 score for AUs is computed based on a per-frame detection (whether each AU is present or absent).

The average between the $\mathcal{A}F_1$ score and the total accuracy, $\mathcal{T}Acc$, will be the main evaluation criterion:

$$\mathcal{E}_{total} = 0.5 \times \mathcal{A}F_1 + 0.5 * \mathcal{T}Acc \tag{7}$$

4. Baseline & Participating Teams' Systems and Results

All baseline systems rely exclusively on existing open-source machine learning toolkits to ensure the reproducibility of the results. In this Section, we first describe the baseline systems developed for each Challenge, then we present the top-5 performing teams per Challenge and finally report their obtained results.

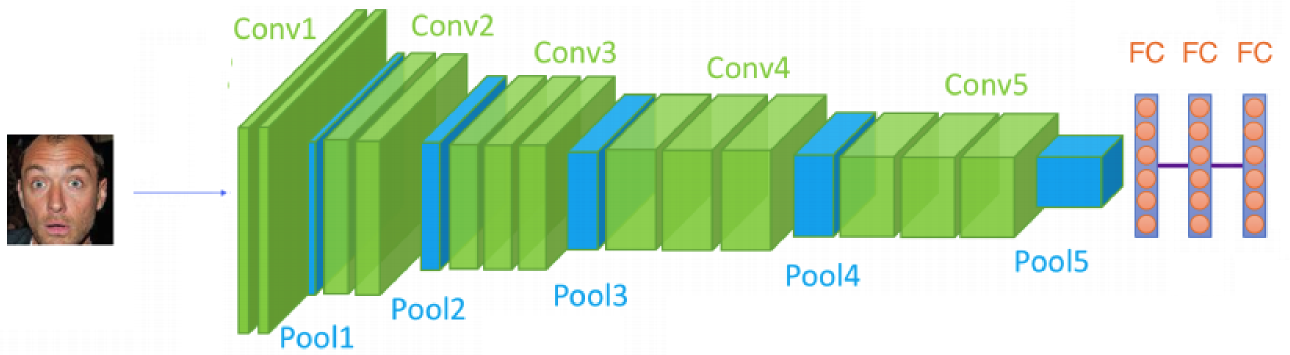

Figure 3. The architecture of the utilized baseline VGG-FACE that has been used in all Challenges; the output is either linear (VA case), or with a softmax unit (in the 7 basic expression case), or with a sigmoid unit (in the 12 AU case)

4.1. Baseline Systems

The architecture that was used in all 3 Challenges was based on the 13 convolutional and pooling layers of VGG-FACE [47] (its fully connected layers are discarded), followed by 2 fully connected layers, each with 4096 hidden units. In the Valence-Arousal Estimation Challenge baseline, a (linear) output layer follows that gives final estimates for valence and arousal. In the Seven Basic Expression Classification Challenge, a final output layer with softmax as activation function follows which gives the 7 basic expression predictions. In the Twelve Action Unit Detection Challenge, a final output layer with sigmoid as activation function follows which gives the 12 action unit predictions. Figure 3 shows this basic architecture of the utilized VGG-FACE.

Let us mention that we utilized the cropped and aligned images from Aff-Wild2, as described in Section 2.4. These images have dimensions $112 \times 112 \times 3$. The pixel intensities are normalized to take values in [-1,1]. No on-the-fly or off-the-fly data augmentation technique [38, 23, 24] was utilized.

The baseline systems have been pre-trained on the VGG-Face dataset; their convolutional layers were fixed (i.e., non-trainable) and only the three fully connected were trained on Aff-Wild2. These systems have been implemented in TensorFlow; training time was around a day on a Titan X GPU, with a learning rate of 10^{-4} and with a batch size of 256.

4.2. Top-5 Performing Teams per Challenge

At first let us mention that in total: 40 Teams participated in the Valence-Arousal Estimation Challenge; 55 Teams participated in the Seven Basic Expression Classification Challenge; 51 Teams participated in the Twelve Action Unit Challenge; 51 Teams participated in the Twelve Action Unit Detection Challenge. These teams come from 44 different universities and 18 companies. In more detail, teams come from different universities: 17 in China, 6 in Korea, 4 in Canada, 5 in India, 2 in Germany, 2 in USA, 1 in Japan, 1 in Singapore, 1 in France, 1 in UK, 1 in Turkey, 1 in Greece, 1 in Bangladesh and 1 in Taiwan. Teams come from different companies: 8 in China, 6 International Companies, 1 in Canada, 1 in France, 1 in India and 1 in Kazakhstan. 20, 30 and 26 Teams submitted their results in the Valence-Arousal Estimation, Seven Basic Expression Classification and Twelve Action Unit Detection Challenges respectively. 10, 13 and 11 Teams scored higher than the baseline and made valid submissions in these Challenges, respectively.

The winner of the Valence-Arousal Estimation Challenge is NISL-2021 (as was the case in the first ABAW Valence-Arousal Estimation Challenge held in conjunction with the IEEE International Conference on Automatic Face and Gesture Recognition 2020) consisting of: Didan Deng and Liang Wu (Hong Kong University of Science and Technology). The runner-up (with a slight difference from the winning team -49.315 vs 49.045-) is Netease Fuxi Virtual Human consisting of: Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang and Yu Ding (Netease Fuxi AI Lab). The Morphoboid team ranks third and consists of: Manh Tu VU and Marie Beurton-AIMAR (Bordeaux University). The STAR team ranks in the fourth place and consists of: Shisen Wang and Linfeng Wang (University of Electronic Science and Technology of China). The Flying-Pigs team ranks in the fifth place and consists of: Su Zhang, Yi Ding and Ziquan Wei (Nanyang Technological University and Huazhong University of Science and Technology).

The winner of the Seven Basic Expression Classification Challenge is Netease Fuxi Virtual Human (Netease Fuxi AI Lab; described above). The runner-up is CPIC-DIR2021 consisting of: Yue Jin, Tianqing Zheng, Chao Gao, Shijie Zhang and Guoqiang Xu (China Pacific Insurance Group Co). The Maybe Next Time team ranks third and consists of: Hoang Manh Hung and Phan Tran Dac Thinh (Chonnam National University). The STAR team ranks in the fourth place (University of Electronic Science and Technology of China; described above). The NISL-2021 team ranks in the

Table 3. Valence-Arousal Challenge Results on the test set of Aff-Wild2; CCC is displayed in % format; the evaluation criterion is the mean valence and arousal CCC; only the best performing submission of each team is shown

Teams	CCC-Valence	CCC-Arousal	Github
NISL-2021 [8]	**53.26**	45.37	link
Netease Fuxi Virtual Human [63]	48.59	49.5	link
Morphoboid [53]	50.51	47.47	link
STAR [54]	47.84	49.75	link
FlyingPigs [62]	46.33	49.24	link
NYCU AIMM [56]	39.66	**49.82**	link
NTUA-CVSP [2]	36.84	46.39	link
Kawakarpo [18]	37.63	37.97	link
FLAB2021 [50]	25.67	35.08	link
IMLAB [45]	26.8	25.61	link
VGG-FACE (baseline)	20	19	-

Table 4. Expression Challenge Results on the test set of Aff-Wild2; all metrics are displayed in % format; the evaluation criterion is $\mathcal{E}_{total} = 0.67 \times F_1 + 0.33 * \mathcal{T}Acc$; only the best performing submission of each team is shown

Teams	F1 Score	Total Accuracy	$\mathcal{E}_{total}$	Github
Netease Fuxi Virtual Human [63]	**76.33**	**80.69**	**77.77**	link
CPIC-DIR2021 [20]	68.34	77.09	71.23	link
Maybe Next Time [52]	60.46	72.89	64.56	link
STAR [54]	47.59	73.21	56.04	link
NISL-2021 [8]	43.11	65.38	50.46	link
FLAB2021 [50]	40.79	67.29	49.53	link
DMACS-SSSIHL [14]	36.1	67.5	46.46	link
Morphoboid [53]	35.11	66.8	45.56	link
NTUA-CVSP [2]	33.67	64.18	43.74	link
SZTU-CityU [42]	30.73	62.34	41.16	link
Kawakarpo [18]	29	64.81	40.82	link
HUST_AUTO1102 [15]	28.09	58.22	38.03	link
Keegs [51]	25.45	50.7	33.78	link
VGG-FACE (baseline)	26	46	32.6	-

fifth place (Hong Kong University of Science and Technology; described above).

The winner of the Twelve Action Unit Detection Challenge Challenge is Netease Fuxi Virtual Human (Netease Fuxi AI Lab; described above). The runner-up (with a small difference from the winning team -69.70 vs 69.04-) is CPIC-DIR2021 (China Pacific Insurance Group Co; described above). The Maybe Next Time team ranks third (Chonnam National University; described above). The TJU-CIC rankes in the fourth place consisting of: Zhilei Liu, Chenggong Zhang, Juan Song, Qiangyang Zhang, Weilong Dong and Ruomeng Ding (Tianjin University). The NISL-2021 team ranks in the fifth place (Hong Kong University of Science and Technology; described above).

4.3. Results

Table 3 presents the CCC evaluation of valence and arousal predictions on the Aff-Wild2 test set, of the baseline network (VGG-FACE) and the 10 participating teams' algorithms that scored higher than the baseline and made valid submissions. Table 3 also includes a link to a Github repository where each team's solution/source code is stored so that the work is reproductible. Let us mention that the VGG-FACE (baseline) results on the Aff-Wild2 validation set are:
CCC-Valence = 0.23 , CCC-Arousal = 0.21.

The NISL-2021 team achieved the overall best performance (in terms of average CCC for valence and arousal). It can be observed that separately for valence, the NISL-2021 team achieved the best CCC score and separately for arousal, the NYCU AIMM team achieved the best CCC score.

Table 4 presents the performance, in the Seven Basic Expression Classification Challenge, on the test set of Aff-Wild2, of the baseline network (VGG-FACE) and the 13 participating teams' algorithms that scored higher than the baseline and made valid submissions. The performance metric is a weighted average between the F1 score and the

Table 5. Action Unit Challenge Results on the test set of Aff-Wild2; all metrics are displayed in % format; the evaluation criterion is $\mathcal{E}_{total} = 0.5 \times \mathcal{A}F_1 + 0.5 * \mathcal{T}Acc$; only the best performing submission of each team is shown

Teams	Average F1 Score	Total Accuracy	$\mathcal{E}_{total}$	Github
Netease Fuxi Virtual Human [63]	**50.59**	88.82	**69.70**	link
CPIC-DIR2021 [20]	48.92	**89.15**	69.04	link
Maybe Next Time [52]	46.14	87.67	66.9	link
TJU-CIC [61]	43.51	88.89	66.2	link
NISL-2021 [8]	45.09	84.65	65.28	link
STAR [54]	39.38	87.47	63.43	link
Defending TAL	38.42	85	61.7	link
Kawakarpo [18]	34.93	87.74	61.34	link
SZTU-CityU [42]	30	88.3	59.14	link
FLAB2021 [50]	33.85	82.5	58.17	link
DETA	28.25	83.07	55.66	link
VGG-FACE (baseline)	36.7	19.3	28	-

total accuracy, as discussed in Section 3.2. Table 4 also includes a link to a Github repository where each team's solution/source code is stored so that the work is reproductible. Let us mention that the VGG-FACE (baseline) results on the Aff-Wild2 validation set are:
F1 Score = 0.30 , Total Accuracy = 0.50 and $\mathcal{E}_{total}$ = 0.366.

The Netease Fuxi Virtual Human team achieved the overall best performance; the team achieved the best performance separately for both the F1 score and the total accuracy.

Table 5 presents the performance, in the Twelve Action Unit Detection Challenge, on the test set of Aff-Wild2, of the baseline network (VGG-FACE) and the 11 participating teams' algorithms that scored higher than the baseline and made valid submissions. The performance metric is the average between the F1 score and the total accuracy, as discussed in Section 3.3. Table 5 also includes a link to a Github repository where each team's solution/source code is stored so that the work is reproductible. Let us mention that the VGG-FACE (baseline) results on the Aff-Wild2 validation set are:
Average F1 Score = 0.40 , Total Accuracy = 0.22 and $\mathcal{E}_{total}$ = 0.31.

The Netease Fuxi Virtual Human team achieved the overall best performance. It can be observed that separately for the F1 score, the Netease Fuxi Virtual Human team achieved the best score and separately for the total accuracy the CPIC-DIR2021 team achieved the best score.

5. Conclusion

In this paper we have presented the second Affective Behavior Analysis in-the-wild Competition (ABAW2) 2021 held in conjunction with the International Conference on Computer Vision (ICCV) 2021. ABAW2 followed the first ABAW Competition held in conjunction with IEEE Conference on Face and Gesture Recognition 2020. ABAW2 comprises three Challenges targeting: i) valence-arousal estimation, ii) seven basic expression classification and iii) twelve action unit detection. The database utilized for this Competition has been derived from the Aff-Wild2, the first and large-scale database annotated for all these three behavior tasks.

The ABAW2 Competition has been a very successful one with the participation of 40 Teams in the Valence-Arousal Estimation Challenge, 55 Teams in the Seven Basic Expression Classification Challenge and 51 Teams in the Twelve Action Unit Detection Challenge; the Teams' solutions were very interesting and creative, providing quite a push from the developed baseline.

References

[1] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, and D Puthankattil Subha. Automated eeg-based screening of depression using deep convolutional neural network. *Computer methods and programs in biomedicine*, 161:103–113, 2018.

[2] 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.

[3] 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.

[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] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Paul Ekman. Facial action coding system (facs). *A human face*, 2002.

[12] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[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] Wei Gong and Hailan Huang. Bayesian convolutional neural networks for seven basic facial expression classifications. *arXiv preprint arXiv:2107.04834*, 2021.

[16] 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.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] Ruian He, Zhen Xing, and Bo Yan. Feature pyramid network for multi-task affective analysis. *arXiv preprint arXiv:2107.03670*, 2021.

[19] 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.

[20] 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.

[21] Junghoe Kim, Vince D Calhoun, Eunsoo Shim, and Jong-Hwan Lee. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage*, 124:127–146, 2016.

[22] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. *arXiv preprint arXiv:2106.07524*, 2021.

[23] Dimitrios Kollias, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic facial synthesis in the dimensional affect space. In *European Conference on Computer Vision*, pages 475–491. Springer, 2018.

[24] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, pages 1–30, 2020.

[25] 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.

[26] 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.

[27] 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.

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

[29] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907–929, 2019.

[30] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and Stefanos D Kollias. Transparent adaptation in deep medical image diagnosis. In *TAILOR*, pages 251–267, 2020.

[31] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[32] Dimitrios Kollias and Stefanos Zafeiriou. A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild. *arXiv preprint arXiv:1805.01452*, 2018.

[33] Dimitrios Kollias and Stefanos Zafeiriou. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv preprint arXiv:1811.07771*, 2018.

[34] Dimitrios Kollias and Stefanos Zafeiriou. Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

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

[36] 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.

[37] Dimitrios Kollias and Stefanos P Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 2020.

[38] Michael Kuchnik and Virginia Smith. Efficient augmentation via data subsampling. *arXiv preprint arXiv:1810.05222*, 2018.

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

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

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

[42] 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.

[43] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry Davis. SSH: Single stage headless face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[44] Ibrahim M Nasser, Mohammed O Al-Shawwa, and Samy S Abu-Naser. Artificial neural network for diagnose autism spectrum disorder. 2019.

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

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

[47] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[48] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019.

[49] James A Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978.

[50] 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.

[51] Satnam Singh and Doris Schicker. Seven basic expression recognition using resnet-18. *arXiv preprint arXiv:2107.04569*, 2021.

[52] Phan Tran Dac Thinh, Hoang Manh Hung, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. Emotion recognition with incomplete labels using modified multi-task learning technique. *arXiv preprint arXiv:2107.04192*, 2021.

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

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

[55] CM Whissel. The dictionary of affect in language, emotion: Theory, research and experience: vol. 4, the measurement of emotions, r. *Plutchik and H. Kellerman, Eds., New York: Academic*, 1989.

[56] 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.

[57] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[58] 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.

[59] Miao Yu, Dimitrios Kollias, James Wingate, Niro Siriwardena, and Stefanos Kollias. Machine learning for predictive modelling of ambulance calls. *Electronics*, 10(4):482, 2021.

[60] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–41, 2017.

[61] Chenggong Zhang, Juan Song, Qingyang Zhang, Weilong Dong, Ruomeng Ding, and Zhilei Liu. Action unit detection with joint adaptive attention and graph relation. *arXiv preprint arXiv:2107.04389*, 2021.

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

[63] 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.

[64] 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.