The 6th Affective Behavior Analysis in-the-wild (ABAW) Competition

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

Panagiotis Tzirakis
Hume AI, USA
panagiotis@hume.ai

Alan Cowen
Hume AI, USA

Stefanos Zafeiriou
Imperial College London, UK

Chunchang Shao
Queen Mary University of London, UK

Guanyu Hu
Queen Mary University of London, UK
Xi'an Jiaotong University, China

Abstract

This paper describes the 6th Affective Behavior Analysis in-the-wild (ABAW) Competition, which is part of the respective Workshop held in conjunction with IEEE CVPR 2024. The 6th ABAW Competition addresses contemporary challenges in understanding human emotions and behaviors, crucial for the development of human-centered technologies. In more detail, the Competition focuses on affect related benchmarking tasks and comprises of five sub-challenges: i) Valence-Arousal Estimation (the target is to estimate two continuous affect dimensions, valence and arousal), ii) Expression Recognition (the target is to recognise between the mutually exclusive classes of the 7 basic expressions and 'other'), iii) Action Unit Detection (the target is to detect 12 action units), iv) Compound Expression Recognition (the target is to recognise between the 7 mutually exclusive compound expression classes), and v) Emotional Mimicry Intensity Estimation (the target is to estimate six continuous emotion dimensions). In the paper, we present these Challenges, describe their respective datasets and challenge protocols (we outline the evaluation metrics) and present the baseline systems as well as their obtained performance. More information for the Competition can be found in: https://affective-behavior-analysis-in-the-wild.github.io/6th.

1. Introduction

The 6th Affective Behavior Analysis in-the-wild (ABAW) Workshop and Competition continues its tradition of fostering interdisciplinary collaboration by bringing together experts from various fields including academia, industry, and government. This workshop, held in conjunction with IEEE Computer Vision and Pattern Recognition Conference (CVPR) 2024, aims to delve into the analysis of affective behavior in real-world settings, a critical aspect for the development of human-centered technologies such as HCI systems and intelligent digital assistants. By understanding human emotions and behaviors, machines can better engage with users irrespective of contextual factors like age, gender, or social background, thereby enhancing trust and interaction in real-life scenarios.

The ABAW Competition, an integral part of the workshop, is split into five Challenges.

The first Challenge is the Valence-Arousal (VA) Estimation one; the target of this Challenge is to estimate the two continuous affect dimensions of valence and arousal in each frame of the utilized Challenge corpora. Valence characterises an affective state on a continuous scale from positive to negative (in other words from -1 to 1). Arousal characterises an affective state on a continuous scale from active to passive (in other words from -1 to 1).

Only uni-task solutions will be accepted for this Challenge. Teams are allowed to use any -publicly or not- available pre-trained model (as long as it has not been pre-trained on the utilized in this Challenge corpora, Aff-Wild2). The pre-trained model can be pre-trained on any task (e.g., VA estimation, Expression Recognition, AU detection, Face Recognition). However when the teams are refining the model and developing the methodology they should not use any other annotations (expressions or AUs): the methodology should be purely uni-task, using only the VA annotations. This means that teams are allowed to use other databases' VA annotations, or generated/synthetic data, or any affine transformations, or in general data augmentation

techniques for increasing the size of the training dataset.

For this Challenge, an augmented version of the Aff-Wild2 [30, 31, 37, 39, 40, 44, 47, 48, 51, 54, 59, 105] is used. This corpora is audiovisual (A/V), in-the-wild and in total consists of 594 videos of around 3M frames of 584 subjects.

The second Challenge is the Expression (Expr) Recognition one; the target of this Challenge is to recognise between eight mutually exclusive classes in each frame of the utilized Challenge corpora; these classes are the 6 basic expressions (i.e., anger, disgust, fear, happiness, sadness and surprise), the neutral state and a category 'other' that denotes affective states that are not included in the neutral state or in the 6 basic expressions.

Only uni-task solutions will be accepted for this Challenge. Teams are allowed to use any -publicly or not- available pre-trained model (as long as it has not been pre-trained on the utilized in this Challenge corpora, Aff-Wild2). The pre-trained model can be pre-trained on any task (e.g., VA estimation, Expression Recognition, AU detection, Face Recognition). However when the teams are refining the model and developing the methodology, they should not use any other annotations (VA or AUs): the methodology should be purely uni-task, using only the Expr annotations. This means that teams are allowed to use other databases' Expr annotations, or generated/synthetic data (e.g. the data provided in the ECCV 2022 run of the ABAW Challenge [29]), or any affine transformations, or in general data augmentation techniques (e.g., [60, 78]) for increasing the size of the training dataset.

For this Challenge, the Aff-Wild2 is used. This corpora is audiovisual (A/V), in-the-wild and in total consists of 548 videos of around 2.7M frames.

The third Challenge is the Action Unit (AU) Detection one; the target of this Challenge is to detect which of the 12 Action Units are activated in each frame of the utilized Challenge corpora. Action Units refer to a set of facial muscle movements or configurations. The action units that have been selected for the purposes of this Challenge are the following: AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25 and AU26.

Only uni-task solutions will be accepted for this Challenge. Teams are allowed to use any -publicly or not- available pre-trained model (as long as it has not been pre-trained on the utilized in this Challenge corpora, 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 should not use any other annotations (VA or Expr): the methodology should be purely uni-task, using only the AU annotations. This means that teams are allowed to use other databases' AU annotations, or generated/synthetic data, or any affine

transformations, or in general data augmentation techniques (e.g., [60, 78]) for increasing the size of the training dataset.

For this Challenge, the Aff-Wild2 is used. This corpora is audiovisual (A/V), in-the-wild and in total consists of 542 videos of around 2.7M frames.

The fourth Challenge is the Compound Expression (CE) Recognition one; the target of this Challenge is to recognise between the 7 mutually exclusive classes in each frame of the utilized Challenge corpora. These classes are the following compound expressions: Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful and Sadly Angry.

Teams are allowed to use any -publicly or not- available pre-trained model and any -publicly or not- available database (that contains any annotations, e.g. VA, basic or compound expressions, AUs).

For this Challenge, a part of C-EXPR-DB [32] database will be used, which consists of 56 videos in total. C-EXPR-DB is an audiovisual (A/V) in-the-wild database and in total consists of 400 videos of around 200K frames.

The sixth ABAW Competition, which is part of the respective Workshop to be held in conjunction with the IEEE Computer Vision and Pattern Recognition Conference (CVPR) 2024 is a continuation of the successful series of ABAW Competitions held in conjunction with IEEE CVPR 2023, 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 [1–6, 8–28, 33–36, 38, 41–43, 45, 46, 49, 50, 52, 53, 55–58, 61–77, 79–84, 84–104, 106–113, 113–115, 115].

2. Competition Corpora

In the following, we present a brief synopsis of each Challenge's dataset. For a more comprehensive understanding, readers are encouraged to consult the original documentation. Additionally, we detail the pre-processing steps undertaken for the first three Challenges, which involve cropping and aligning all image-frames. These have been utilized in our baseline experiments.

2.1. Valence-Arousal Estimation Challenge

This Challenge's dataset comprises 594 videos, an expansion of the Aff-Wild2 database, annotated in terms of valence and arousal. Notably, sixteen videos feature two subjects, both of whom are annotated. In total, annotations are provided for 2,993,081 frames from 584 subjects; these annotations have been conducted by four experts using the methodology outlined in [7]. Valence and arousal values are continuous and range in $[-1, 1]$. The 2D Valence-Arousal histogram of annotations is visualized in Figure 1.

Aff-Wild2 is split into training, validation and testing sets, in a subject independent manner, ensuring each subject
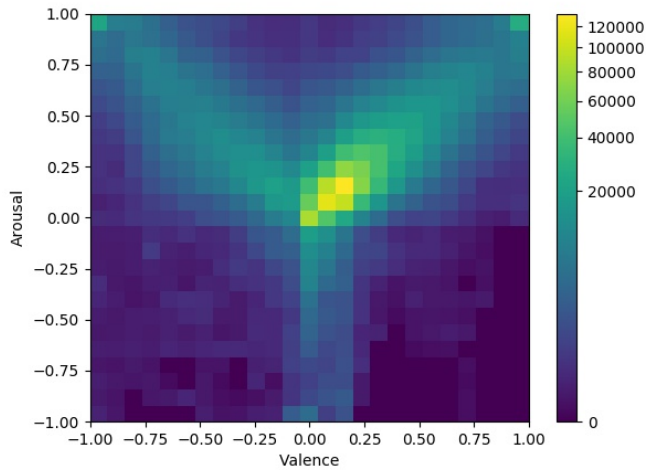
Figure 1. Valence-Arousal Estimation Challenge: 2D Valence-Arousal Histogram of Annotations in Aff-Wild2

appears exclusively in one set. The training, validation and testing sets consist of 356, 76 and 162 videos, respectively.

The Train and Validation data along with their corresponding annotations are provided to the participating teams. The unlabeled test data are provided to the participating teams who will upload their test set predictions to an evaluation server. Participating teams are allowed to submit up to five sets of predictions for this challenge.

2.2. Expression Recognition Challenge

In this Challenge, the dataset consists of 548 videos from Aff-Wild2, annotated for the six basic expressions, neutral state, and an 'other' category representing non-basic expressions. Seven videos feature two subjects, both of whom are annotated. In total, annotations are provided for 2,624,160 frames from 437 subjects. Annotation is conducted by seven experts on a frame-by-frame basis. Table 1 presents the distribution of expression annotations.

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. Aff-Wild2 is split into training, validation and testing sets, in a subject independent manner. The training, validation and testing sets consist of 248, 70 and 230 videos, respectively.

The Train and Validation data along with their corresponding annotations are provided to the participating teams. The unlabeled test data are provided to the participating teams who will upload their test set predictions to an evaluation server. Participating teams are allowed to submit up to five sets of predictions for this challenge.

2.3. Action Unit Detection Challenge

The dataset for this Challenge comprises 542 videos annotated for 12 AUs, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25 and AU26. Annotations are provided for 2,627,632 frames from 438 subjects. A semi-automatic annotation procedure, involving both manual and automatic annotations, is employed. Table 2 details the annotated AUs and their distribution.

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. The training, validation and testing sets consist of 295, 105 and 142 videos, respectively.

The Train and Validation data along with their corresponding annotations are provided to the participating teams. The unlabeled test data are provided to the participating teams who will upload their test set predictions to an evaluation server. Participating teams are allowed to submit up to five sets of predictions for this challenge.

2.4. Compound Expression Recognition Challenge

For this Challenge, a part of C-EXPR-DB database will be used (56 videos in total). C-EXPR-DB is audiovisual (A/V)

in-the-wild database and in total consists of 400 videos of around 200K frames; each frame is annotated in terms of 12 compound expressions. For this Challenge, the following 7 compound expressions will be considered: Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful and Sadly Angry.

Goal of the Challenge and Rules Participants will be provided with a part of C-EXPR-DB database (56 videos in total), which will be unannotated, and will be required to develop their methodologies (supervised/self-supervised, domain adaptation, zero-/few-shot learning etc) for recognising the 7 compound expressions in this unannotated part, in a per-frame basis.

2.5. Emotional Mimicry Intensity Estimation Challenge

For this Challenge, the multimodal Hume-Vidmimic2 dataset is used which consists of more than 15,000 videos totaling over 25 hours. In this dataset, every participant was tasked with imitating a 'seed' video that showcased an individual displaying a particular emotion. Following the mimicry, they were then asked to assess the emotional intensity of the seed video by selecting from a range of predefined emotional categories. The following emotion dimensions are targeted: 'Admiration', 'Amusement', 'Determination', 'Empathic Pain', 'Excitement', and 'Joy'. A normalized score from 0 to 1 is provided as a ground truth value.

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

Initially, all videos are segmented into individual frames, after which they undergo processing using the RetinaFace detector. This step aims to extract face bounding boxes and five facial landmarks for each frame. Subsequently, the images are cropped based on the bounding box coordinates, and these cropped images are provided to the participating teams.

Using the five facial landmarks (representing two eyes, the nose, and two mouth corners), a similarity transformation is applied. This transformation ensures alignment, resulting in cropped and aligned images, which are also shared with the participating teams. Ultimately, these cropped and aligned images are utilized in our baseline experiments.

All cropped and cropped-aligned images are resized to dimensions of $112 \times 112 \times 3$ pixels and their intensity values are normalized to fall within the range of $[-1, 1]$.

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:

$$\mathcal{P}_{VA} = \frac{CCC_a + CCC_v}{2} \qquad (1)$$

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:

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

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):

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

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:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \qquad (4)$$

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:

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

3.4. Compound Expression Recognition Challenge

The performance measure is the average F1 Score across all 7 compound expressions. Therefore, the evaluation criterion for the Compound Expression Recognition Challenge is:

$$\mathcal{P}_{CE} = \frac{\sum_{expr} F_1^{expr}}{7} \qquad (6)$$

3.5. Emotional Mimicry Intensity Estimation Challenge

The performance measure is the average Pearson's Correlation Coefficient (ρ) across the 6 emotion dimensions:

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

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 are built solely on existing open-source machine learning toolkits to maintain result reproducibility. TensorFlow is the chosen framework for implementing all systems.

In this Section, we describe the baseline systems developed for each Challenge and present their obtained results.

4.1. Valence-Arousal Estimation Challenge

The baseline network comprises a ResNet architecture with 50 layers, initially trained on ImageNet (ResNet50). It incorporates a linear output layer responsible for providing the final estimations for valence and arousal. Its average CCC performance on the validation set is 0.22 (the CCC for valence is 0.24 and the CCC for arousal is 0.20), as can be seen in Table 3.

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
baseline	0.22	0.24	0.20

4.2. Expression Classification Challenge

The baseline network adopts a VGG16 architecture with fixed convolutional weights (i.e., non-trainable), while only the three fully connected layers are trainable. It is pretrained on the VGGFACE dataset and equipped with an output layer featuring a softmax activation function, facilitating the prediction of eight expressions. Mixaugment has been used as data augmentation technique. Its performance on the validation set is 0.25, as can be seen in Table 6.

Table 4. Expression Recognition Challenge's Results; the evaluation criterion is the average F1 Score

Teams	Total Score
baseline	0.25

4.3. Action Unit Detection Challenge

The baseline network adopts a VGG16 architecture with fixed convolutional weights (i.e., non-trainable), while only the three fully connected layers are trainable. It is pretrained on the VGGFACE dataset and equipped with an output layer featuring a sigmoid activation function, facilitating the detection of the twelve AUs. Its performance on the validation set is 0.39, as can be seen in Table 6.

Table 5. Action Unit Detection Challenge's Results; the evaluation criterion is the average F1 Score

Teams	Total Score
baseline	0.39

4.4. Emotional Mimicry Intensity Estimation Challenge

We established baseline results using two different feature sets.

First, we employed pre-trained Vision Transformer (ViT) features, which were further processed through a three-layer Gated Recurrent Unit (GRU) network. This approach achieved a performance score of 0.09.

Secondly, we utilized features extracted from Wav2Vec2, combined with a linear processing layer, which resulted in a performance score of 0.24.

Additionally, we explored a multimodal approach by averaging the predictions from both unimodal methods, leading to a combined performance score of 0.48.

Table 6. Emotional Mimicry Intensity Estimation Challenge's Results; the evaluation criterion is the average Pearson's Correlation

Teams	Total Score
ViT-GRU baseline	0.09
Wav2Vec2-FC	0.24
ViT-GRU + Wav2Vec2-FC	0.48

5. Conclusion

In this paper we have presented the sixth Affective Behavior Analysis in-the-wild Competition (ABAW) held in conjunction with IEEE CVPR 2024. This Competition is a continuation of the series of ABAW Competitions. This Competition comprises five Challenges targeting: i) Valence-Arousal Estimation, ii) Expression Recognition (8 categories), iii) Action Unit Detection (12 action units), iv) Compound Expression Recognition (7 categories) and v) Emotional Mimicry Intensity Estimation (6 emotion dimensions). The databases utilized for this Competition are an extended version of Aff-Wild2, the C-EXPR-DB and the Hume-Vidmimic2 dataset.

References

[1] Panagiotis Antoniadis, Ioannis Pikoulis, Panagiotis P Filntsis, and Petros Maragos. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. *arXiv preprint arXiv:2107.03465*, 2021. 2

[2] Anastasios Arsenos, Dimitrios Kollias, and Stefanos Kollias. A large imaging database and novel deep neural architecture for covid-19 diagnosis. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.

[3] Anastasios Arsenos, Andjoli Davidhi, Dimitrios Kollias, Panos Prassopoulos, and Stefanos Kollias. Data-driven covid-19 detection through medical imaging. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, page 1–5. IEEE, 2023.

[4] 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.

[5] 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.

[6] 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. 2

[7] 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. 2

[8] Didan Deng. Multiple emotion descriptors estimation at the abaw3 challenge. *arXiv preprint arXiv:2203.12845*, 2022. 2

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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. 2

[29] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2022. 2

[30] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 2

[31] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 2

[32] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023. 2

[33] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 2

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

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

[36] 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. 2

[37] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2

[38] Dimitrios Kollias and Stefanos Zafeiriou. Va-stargan: Continuous affect generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 227–238. Springer, 2020. 2

[39] 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. 2

[40] 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. 2

[41] 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. 2

[42] Dimitris Kollias, George Marandianos, Amaryllis Raouzaiou, and Andreas-Georgios Stafylopatis. Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction. In *2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–6. IEEE, 2015.

[43] Dimitrios Kollias, Athanasios Tagaris, and Andreas Stafylopatis. On line emotion detection using retrainable deep neural networks. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, pages 1–8. IEEE, 2016. 2

[44] 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. 2

[45] Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis, and Stefanos Kollias. Adaptation and contextualization of deep neural network models. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–8. IEEE, 2017. 2

[46] 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. 2

[47] 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. 2

[48] 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. 2

[49] Dimitrios Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044*, 2020. 2

[50] 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. 2

[51] 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. 2

[52] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Transparent adaptation in deep medical image diagnosis. In *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, pages 251–267. Springer, 2020. 2

[53] 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. 2

[54] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-

task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2

[55] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-mia: Covid-19 detection and severity analysis through medical imaging. In *European Conference on Computer Vision*, page 677–690. Springer, 2022. 2

[56] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-enabled analysis of 3-d ct scans for diagnosis of covid-19 & its severity. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, page 1–5. IEEE, 2023.

[57] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing*, 542:126244, 2023.

[58] Dimitrios Kollias, Andreas Psaroudakis, Anastasios Arsenos, and Paraskeui Theofilou. Facernet: a facial expression intensity estimation network. *arXiv preprint arXiv:2303.00180*, 2023. 2

[59] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023. 2

[60] Dimitrios Kollias, Karanjot Vendal, Priyankaben Gadhavi, and Solomon Russom. Btdnet: A multi-modal approach for brain tumor radiogenomic classification. *Applied Sciences*, 13(21):11984, 2023. 2

[61] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. *arXiv preprint arXiv:2401.01219*, 2024. 2

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

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

[64] 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.

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

[66] 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.

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

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

[69] 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.

[70] 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.

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

[72] 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.

[73] 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.

[74] 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.

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

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

[77] 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. 2

[78] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2375, 2022. 2

[79] 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. 2

[80] 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.

[81] Natalia Salpea, Paraskevi Tzouveli, and Dimitrios Kollias. Medical image segmentation: A review of modern architectures. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 691–708. Springer, 2023.

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

[83] 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.

[84] Andrey V Savchenko. Emotieffnet facial features in unitask emotion recognition in video at abaw-5 competition. *arXiv preprint arXiv:2303.09162*, 2023. 2

[85] 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.

[86] 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.

[87] Athanasios Tagaris, Dimitrios Kollias, and Andreas Stafylopatis. Assessment of parkinson's disease based on deep neural networks. In *International Conference on Engineering Applications of Neural Networks*, pages 391–403. Springer, 2017.

[88] Athanasios Tagaris, Dimitrios Kollias, Andreas Stafylopatis, Georgios Tagaris, and Stefanos Kollias. Machine learning for neurodegenerative disorder diagnosis—survey of practices and launch of benchmark dataset. *International Journal on Artificial Intelligence Tools*, 27(03):1850011, 2018.

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

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

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

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

[93] 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.

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

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

[96] 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.

[97] 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.

[98] 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.

[99] 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.

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

[101] 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.

[102] 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.

[103] 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.

[104] 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. 2

[105] 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. 2

[106] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Audio-visual attentive fusion for continuous emotion recognition. *arXiv preprint arXiv:2107.01175*, 2021. 2

[107] 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.

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

[109] 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.

[110] 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.

[111] 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.

[112] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. m^3t: Multi-modal continuous

valence-arousal estimation in the wild. *arXiv preprint arXiv:2002.02957*, 2020.

[113] 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. 2

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

[115] Peng Zou, Rui Wang, Kehua Wen, Yasi Peng, and Xiao Sun. Spatial-temporal transformer for affective behavior analysis. *arXiv preprint arXiv:2303.10561*, 2023. 2