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

Irene Kotsia
Cogitat, UK

Alice Baird
Hume AI, USA

Chris Gagne
Hume AI, USA

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 and top performing teams' per Challenge, 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 are 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 [34, 36–42, 44, 106] 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 are 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 [33]), or any affine transformations, or in general data augmentation techniques (e.g., [67]) 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 are 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., [67]) 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 [35] database is 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 final fourth challenge is the Emotional Mimicry Intensity (EMI) Estimation challenge where emotional mimics are explored. Participants are asked to predict six emotional dimensions using a multi-output regression approach. The following emotional expressions have been used: "Admiration", "Amusement", "Determination", "Empathic Pain", "Excitement", and "Joy".

For the purposes of this challenge, we use the audiovisual and in-the-wild HUME-Vidmimic2 dataset, a comprehensive collection derived from 'in-the-wild' settings, which contains more than 17 000 videos totaling over 30 hours from the United States, similar to our first version [7].

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, 4–6, 9–14, 16–20, 22–31, 43, 45, 47–52, 54–56, 58–61, 63–65, 68, 70–73, 73, 75–77, 83, 84, 87–92, 94–100, 108–113, 115, 116, 116, 117, 119, 119].

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 [8]. Valence and arousal values are continuous and range in $[-1, 1]$. The 2D Valence-Arousal histogram of annotations is visualized in Figure 1.

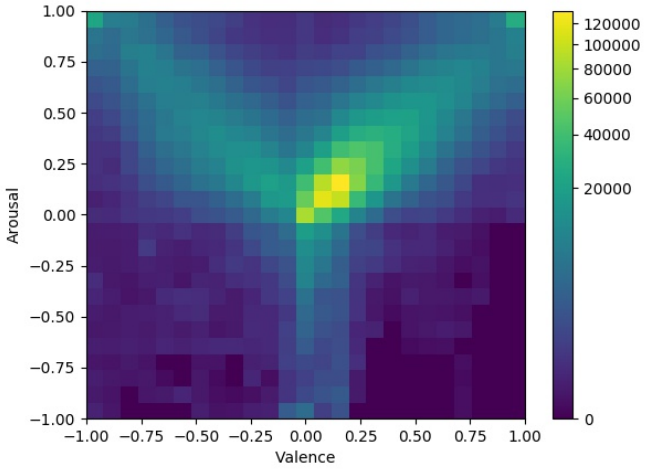

Figure 1. Valence-Arousal Estimation Challenge: 2D Valence-Arousal Histogram of Annotations in Aff-Wild2

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

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.

Table 1. Expression Recognition 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

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 corresponding to the inner and outer brow raiser, the brow lowerer, the cheek raiser, the lid tightener, the upper lip raiser, the lip corner puller and depressor, the lip tightener and pressor, lips part and jaw drop. The exact utulized AUs along with their corresponding actions can be seen in Table 2. 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 further details the annotated AUs distribution.

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 is 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 follow-

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

ing 7 compound expressions will be considered: Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful and Sadly Angry.

Participants are provided with a part of C-EXPR-DB database (56 videos in total with around 26,500 frames), which is unannotated, and are 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

In the Emotional Mimicry Intensity Challenge (EMI-Challenge), we investigate the study of emotional mimicry by presenting a large-scale and in-the-wild dataset, HUME-Vidmimic2, featuring 557 participants and over 30 hours of audiovisual content. This dataset was collected in naturalistic settings, with participants using their webcams to record their facial and vocal responses by mimicking a "seed" video and rating it in the range from 0 to 100.

The data preparation process involved a speaker-independent partitioning of the dataset into training, validation, and test sets. Table 3 statistics of the dataset for each partition. 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.

Along with the data, the participants are provided the faces of individuals within the videos that were detected with the use of MTCNN [107] at a frequency of 6 frames per second. In addition, features extracted from the raw signals and thus enabling participants to use end-to-end approaches [78–82] are provided. Specifically, the feature sets provided are the Vision Transformer (ViT) [3] for the faces and Wav2Vec 2.0 [2] for the audio signals.

Partition	Duration (HH:MM:SS)	# Samples
Train	15:07:03	8072
Validation	9:12:02	4588
Test	9:04:05	4586

Table 3. HUME-Vidmimic2 partition statistics.

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:

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

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. 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:

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

where ρ_{xy} is the Pearson's Correlation Coefficient, s_x and s_y are the variances of all video valence/arousal annotations and predicted values, respectively and s_{xy} is the corresponding covariance value.

3.2. Expression Recognition Challenge

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

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

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:

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

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:

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

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:

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

3.5. Emotional Mimicry Intensity Estimation Challenge

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

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

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, 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, 60 Teams participated in the VA Estimation Challenge. 23 Teams submitted their results. 10 Teams made invalid (incomplete) submissions, whilst surpassing the baseline. 3 Teams scored lower than the baseline. 10 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 Valence-Arousal Estimation Challenge. Table 4 illustrates the CCC evaluation of valence and arousal predictions on the Aff-Wild2 test set; it further shows the baseline network results. The baseline 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.

For the sake of reproducibility, links to Github repositories detailing each participating team's methodology are available on the leaderboard published on the official website of the 6th ABAW Competition.

Table 4. Valence-Arousal Estimation Challenge Results; 'Total' is the average CCC between valence and arousal

Teams	Total	CCC-V	CCC-A
Netease Fuxi AI Lab [114]	0.6721	0.6873	0.6569
DeepAVER [66]	0.5807	0.5418	0.6196
CtyunAI [118]	0.564	0.5223	0.6057
SUN_CE [15]	0.5608	0.5355	0.5861
USTC-IAT-United [103]	0.5478	0.5208	0.5748
HSEmotion [74]	0.5193	0.4925	0.5461
KBS-DGU [32]	0.5077	0.4836	0.5318
ETS-LIVIA [85]	0.4434	0.4198	0.4669
CAS-MAIS[93]	0.3830	0.4245	0.3414
IMLAB [57]	0.2684	0.2912	0.2456
baseline [46]	0.201	0.211	0.191

As can be seen in Table 4, the winner of this Challenge is: *Netease Fuxi AI Lab*. It can be observed this method achieved the overall best performance, as well as the best performance in both valence and arousal estimation. In their developed methodology, they employ a Masked Autoen-

coder (MAE) for visual data, pre-trained on a large facial dataset with a "mask-then-reconstruct" strategy to enhance feature generalizability, followed by fine-tuning. Audio features are extracted using a pre-trained VGGish model, and textual features through the LoRA model. These modalities are fused using a transformer-based approach. Finally, an ensemble learning strategy is applied, by training separate classifiers for data subsets and combining their outputs through a voting mechanism.

The runner up of the Challenge is: *DeepAVER*. In their developed methodology, they fine-tune a pre-trained Resnet-50 for visual inputs, a pre-trained VGG-Net for audio inputs, and a BERT encoder for textual data. All modalities are processed through Temporal Convolutional Networks (TCNs) to capture dynamic changes over time. The core element of the method is the Recursive Joint Cross-Modal Attention mechanism that fuses features from each modality. This process involves pre-processing and concatenating the modality inputs, then applying cross-modal attention to enhance semantic integration and refine the feature representations recursively.

In the third place is: *CtyunAI*. In their developed methodology, they employ a MAE pre-trained on a vast facial dataset for initial feature extraction and then fine-tune it on Aff-Wild2. Temporal dynamics are addressed by segmenting videos and processing these segments through a pre-trained ViT-Base encoder and a TCN, which captures the temporal information effectively. A Transformer Encoder further enhances this by modeling within-segment temporal details, while overlapping segments help capture inter-segment relationships.

Finally let us mention that the baseline network's 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).

4.2. Expression Recognition Challenge

In total, 70 Teams participated in the Expression Recognition Challenge. 40 Teams submitted their results. 14 Teams made invalid (incomplete) submissions, whilst surpassing the baseline. 16 Teams scored lower than the baseline. 10 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 Expression Recognition Challenge. Table 5 illustrates the F1 score evaluation of predictions on the Aff-Wild2 test set; it further shows the baseline network results. The baseline adopts a VGG16 architecture with fixed convolutional weights (i.e., non-trainable), while only the three fully connected layers are trainable. It is pre-trained on the VGGFACE dataset and equipped with an output layer featuring a softmax activation function, facilitating the prediction of eight expressions. Mixaugment [67] has been used as data augmentation technique.

MixAugment is a simple and data-agnostic data augmentation routine that trains a method on convex combinations of pairs of examples and their labels. It extends the training distribution by incorporating the prior knowledge that linear interpolations of feature vectors should lead to linear interpolations of the associated targets. MixAugment constructs virtual training examples $(\tilde{x}, \tilde{y})$ as follows:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda) x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{8}$$

where x_i and x_j are two random raw inputs (i.e., images), y_i and $y_j \in \{0,1\}^8$ are their corresponding one-hot label encodings and $\lambda \sim \mathrm{B}(\alpha, \alpha) \in [0,1]$ (i.e., Beta distribution) for $\alpha \in (0, \infty)$.

During each training iteration, the baseline network is trained concurrently on both real (r) and virtual (v) examples. Specifically, in each training iteration, the network is fed with both x_i and x_j, and the generated image $\tilde{x}$ (of Eq. 8).

For the sake of reproducibility, links to Github repositories detailing each participating team's methodology are available on the leaderboard published on the official website of the 6th ABAW Competition.

Table 5. Expression Recognition Challenge Results

Teams	F1
Netease Fuxi AI Lab [114]	0.5005
CtyunAI [118]	0.3625
USTC-IAT-United [101]	0.3534
HSEmotion [74]	0.3414
M2-Lab-Purdue [53]	0.3228
KBS-DGU [32]	0.3005
SUN_CE [15]	0.2877
AIOBT [62]	0.2797
CAS-MAIS [93]	0.265
IMLAB [57]	0.2296
baseline [46] (with MixAugment [67])	0.2250
baseline [46] (without MixAugment [67])	0.2050

As can be seen in Table 5, the winner of this Challenge is: *Netease Fuxi AI Lab*. Their method is the same as described in the Valence-Arousal Estimation Challenge.

The runner-up of the Challenge is: *CtyunAI*. Their method is the same as described in the Valence-Arousal Estimation Challenge.

In the third place is: *USTC-IAT-United*. Their developed methodology consists of two phases. The first phase is the spatial pre-training one; they use a semi-supervised learning approach with a student-teacher model for de-biasing. The

second phase is the temporal refinement one; the trained student network extracts image features, which are then analyzed by a temporal encoder using a transformer-based architecture to capture temporal relationships in video sequences, enhancing the dynamic recognition of facial expressions. They also apply a post-processing sliding window technique to ensure consistent and accurate labeling of expressions across frames.

Finally let us mention that the baseline network's average F1 score performance on the validation set is 0.25 (when MixAugment is used) and 0.23 (when MixAugment is not used).

4.3. Action Unit Detection Challenge

In total, 63 Teams participated in the Action Unit Detection Challenge. 40 Teams submitted their results. 16 Teams made invalid (incomplete) submissions, whilst surpassing the baseline. 17 Teams scored lower than the baseline. 7 Teams scored higher than the baseline and made valid submissions.

Table 6 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline and made valid submissions in the AU Detection Challenge. Table 6 illustrates the F1 score evaluation of predictions on the Aff-Wild2 test set; it further shows the baseline network results. The baseline adopts a VGG16 architecture with fixed convolutional weights (i.e., non-trainable), while only the three fully connected layers are trainable. It is pre-trained on the VGGFACE dataset and equipped with an output layer featuring a sigmoid activation function, facilitating the detection of the twelve AUs.

For the sake of reproducibility, links to Github repositories detailing each participating team's methodology are available on the leaderboard published on the official website of the 6th ABAW Competition.

Table 6. Action Unit Detection Challenge Results

Teams	F1
Netease Fuxi AI Lab [114]	0.5601
CtyunAI [118]	0.4941
HSEmotion [74]	0.4878
USTC-IAT-United [102]	0.484
KBS-DGU [32]	0.4652
M2-Lab-Purdue [53]	0.3832
baseline [46]	0.365

As can be seen in Table 6, the winner of this Challenge is: *Netease Fuxi AI Lab*. Their method is the same as described in the Valence-Arousal Estimation Challenge.

The runner-up of the Challenge is: *CtyunAI*. Their method is the same as described in the Valence-Arousal Estimation Challenge.

In the third place is: *HSEmotion*. Their developed methodology utilizes a two-phase, multi-task learning strategy. Initially, lightweight neural network architectures are pre-trained for facial recognition and are then fine-tuned for recognizing eight expressions as well as valence and arousal. The embeddings of the penultimate layer of these architectures are extracted and fed to a MLP which is further fine tuned on the Aff-Wild2.

Finally let us mention that the baseline network's average F1 score performance on the validation set is 0.39.

4.4. Compound Expression Recognition Challenge

In total, 40 Teams participated in the Compound Expression Recognition Challenge. 17 Teams submitted their results. 12 Teams made invalid (incomplete) submissions. 5 Teams made valid submissions.

Table 7 presents the leaderboard and results of the participating teams' algorithms that made valid submissions in the Compound Expression Recognition Challenge. Table 7 illustrates the F1 score evaluation of predictions on C-EXPR-DB. No baseline network and results are provided for this Challenge due to the nature of the Challenge.

For the sake of reproducibility, links to Github repositories detailing each participating team's methodology are available on the leaderboard published on the official website of the 6th ABAW Competition.

Table 7. Compound Expression Recognition Challenge Results

Teams	F1
Netease Fuxi AI Lab [114]	0.5526
HSEmotion [74]	0.2708
USTC-IAT-United [104]	0.2240
SUN_CE [69]	0.2201
USTC-AC [86]	0.1845

As can be seen in Table 7, the winner of this Challenge is: *Netease Fuxi AI Lab*. Their method is the same as described in the Valence-Arousal Estimation Challenge.

The runner-up of the Challenge is: *HSEmotion*. Their method is the same as described in the AU Detection Challenge.

In the third place is: *USTC-IAT-United*. Their developed methodology is a late-fusion ensemble model that combines three architectures: Vision Transformer (ViT), Multi-scale and Focal Attention Network (MANet), and ResNet. Features extracted from each architectures are concatenated and fed to an MLP that produces the final predictions.

4.5. Emotional Mimicry Intensity Estimation Challenge

In total, 7 Teams participated in the Emotional Mimicry Intensity Estimation Challenge. 4 Teams scored higher than

the baseline and made valid submissions.

Table 8 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline and made valid submissions in the Emotional Mimicry Intensity Estimation Challenge. Table 8 illustrates the PCC score evaluation of predictions on the HUME-Vidmimic2 test set. It further shows the baseline networks' results. We set initial baseline results with two distinct sets of features. Initially, we utilized features derived from a pre-trained Vision Transformer (ViT), which were then processed by a three-layer Gated Recurrent Unit (GRU) network. Subsequently, we leveraged features extracted from Wav2Vec2, paired with a linear processing layer. Furthermore, by averaging the predictions from both of these unimodal techniques, we pursued a multimodal strategy.

For the sake of reproducibility, links to Github repositories detailing each participating team's methodology are available on the leaderboard published on the official website of the 6th ABAW Competition.

Table 8. Emotional Mimicry Intensity Estimation Challenge Results

Teams	PCC
Netease Fuxi AI Lab [114]	0.7185
HCAI-VIS [21]	0.5536
USTC-IAT-United [105]	0.3594
HSEmotion [74]	0.3316
audio baseline [46]	0.2705
vision baseline [46]	0.1318
multimodal baseline [46]	0.2926

As can be seen in Table 6, the winner of this Challenge is: *Netease Fuxi AI Lab*. Their method is the same as described in the Valence-Arousal Estimation Challenge.

The runner-up of the Challenge is: *HCAI-VIS*. Their developed methodology utilizes a pre-trained Wav2Vec model enhanced with a Valence-Arousal-Dominance (VAD) prediction module and a global pooling, followed by a LSTM model.

In the third place is: *USTC-IAT-United*. Their developed methodology standardizes video frame rates, employs a pre-trained ViT and ResNet18 for visual feature extraction and Wav2Vec2.0 for audio feature extraction, a TCN and a Transformer Encoder for integrating these features, and a late fusion strategy for averaging them.

Finally let us mention that: i) the vision baseline network's average PCC score performance on the validation set is 0.09; ii) the audio baseline network's average PCC score performance on the validation set is 0.24; and iii) the multimodal baseline network's average PCC score performance on the validation set is 0.25.

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.

The sixth ABAW Competition has been a very successful one with the participation of 60 Teams in the Valence-Arousal Estimation Challenge, 70 Teams in the Expression Recognition Challenge, 63 Teams in the Action Unit Detection Challenge, 40 Teams in the Compound Expression Recognition Challenge, and 7 Teams in the Emotional Mimicry 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

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 4

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4

[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. 2

[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] Lukas Christ, Shahin Amiriparian, Alice Baird, Alexander Kathan, Niklas Müller, Steffen Klug, Chris Gagne, Panagiotis Tzirakis, Lukas Stappen, Eva-Maria Meßner, et al. The muse 2023 multimodal sentiment analysis challenge: Mimicked emotions, cross-cultural humour, and personalisation. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, pages 1–10, 2023. 2

[8] 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. 3

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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. 2

[15] Denis Dresvyanskiy, Maxim Markitantov, Jiawei Yu, Peitong Li, Heysem Kaya, and Alexey Karpov. Sun team's contribution to abaw 2024 competition: Audiovisual valence-arousal estimation and expression recognition. *arXiv preprint arXiv:2403.12609*, 2024. 5, 6

[16] 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. 2

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

[18] 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.

[19] 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.

[20] 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. 2

[21] Tobias Hallmen, Fabian Deuser, Norbert Oswald, and Elisabeth André. Unimodal multi-task fusion for emotional mimicry prediciton. *arXiv preprint arXiv:2403.11879*, 2024. 8

[22] 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. 2

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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

[32] Jun-Hwa Kim, Namho Kim, Minsoo Hong, and Cheesun Won. Cca-transformer: Cascaded cross-attention based transformer for facial analysis in multi-modal data. 2024. 5, 6, 7

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

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

[35] 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

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

[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] 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.

[40] 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.

[41] 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.

[42] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2

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

[44] 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

[45] 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

[46] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 5, 6, 7, 8

[47] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. *arXiv preprint arXiv:2002.03399*, 2020. 2

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

[49] 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.

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

[51] 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.

[52] Yifan Li, Haomiao Sun, Zhaori Liu, and Hu Han. Affective behaviour analysis using pretrained model with facial priori. *arXiv preprint arXiv:2207.11679*, 2022. 2

[53] Li Lin, Sarah Papabathini, Xin Wang, and Shu Hu. Robust light-weight facial affective behavior recognition with clip. *arXiv preprint arXiv:2403.09915*, 2024. 6, 7

[54] Hanyu Liu, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Emotion recognition for in-the-wild videos. *arXiv preprint arXiv:2002.05447*, 2020. 2

[55] 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.

[56] 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. 2

[57] Seongjae Min, Junseok Yang, Sangjun Lim, Junyong Lee, Sangwon Lee, and Sejoon Lim. Emotion recognition using transformers with masked learning. *arXiv preprint arXiv:2403.13731*, 2024. 5, 6

[58] Onur Cezmi Mutlu, Mohammadmahdi Honarmand, Saimourya Surabhi, and Dennis P Wall. Tempt: Temporal consistency for test-time adaptation. *arXiv preprint arXiv:2303.10536*, 2023. 2

[59] 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.

[60] 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.

[61] 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. 2

[62] Bach Nguyen-Xuan, Thien Nguyen-Hoang, and Nhu Tai-Do. Emotic masked autoencoder with attention fusion for facial expression recognition. *arXiv preprint arXiv:2403.13039*, 2024. 6

[63] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a facial image sequence. *arXiv preprint arXiv:2107.03886*, 2021. 2

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

[65] 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

[66] R Gnana Praveen and Jahangir Alam. Recursive cross-modal attention for multimodal fusion in dimensional emotion recognition. *arXiv preprint arXiv:2403.13659*, 2024. 5

[67] 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, 6

[68] 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

[69] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, Heysem Kaya, and Alexey Karpov. Audio-visual compound expression recognition method based on late modality fusion and rule-based decision. *arXiv preprint arXiv:2403.12687*, 2024. 7

[70] 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. 2

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

[72] 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.

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

[74] Andrey V Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, 2024. 5, 6, 7, 8

[75] 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. 2

[76] 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.

[77] Gauthier Tallec, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. Multi-label transformer for action unit detection. *arXiv preprint arXiv:2203.12531*, 2022. 2

[78] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309, 2017. 4

[79] Panagiotis Tzirakis, Stefanos Zafeiriou, and Bjorn W Schuller. End2you–the imperial toolkit for multimodal profiling by end-to-end learning. *arXiv preprint arXiv:1802.01115*, 2018.

[80] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

[81] Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53, 2021.

[82] Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, and Björn W Schuller. Speech emotion recognition using semantic information. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283. IEEE, 2021. 4

[83] Manh Tu Vu and Marie Beurton-Aimar. Multitask multi-database emotion recognition. *arXiv preprint arXiv:2107.04127*, 2021. 2

[84] Tu Vu, Van Thong Huynh, and Soo Hyung Kim. Vision transformer for action units detection. *arXiv preprint arXiv:2303.09917*, 2023. 2

[85] Paul Waligora, Osama Zeeshan, Haseeb Aslam, Soufiane Belharbi, Alessandro Lameiras Koerich, Marco Pedersoli, Simon Bacon, and Eric Granger. Joint multimodal transformer for dimensional emotional recognition in the wild. *arXiv preprint arXiv:2403.10488*, 2024. 5

[86] Jiahe Wang, Jiale Huang, Bingzhao Cai, Yifan Cao, Xin Yun, and Shangfei Wang. Zero-shot compound expression recognition with visual language model at the 6th abaw challenge. *arXiv preprint arXiv:2403.11450*, 2024. 7

[87] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis. *arXiv preprint arXiv:2107.04225*, 2021. 2

[88] 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.

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

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

[91] 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.

[92] 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. 2

[93] Zhuofan Wen, Fengyu Zhang, Siyuan Zhang, Haiyang Sun, Mingyu Xu, Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Multimodal fusion with pre-trained model features in affective behaviour analysis in-the-wild. *arXiv preprint arXiv:2403.15044*, 2024. 5, 6

[94] 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. 2

[95] 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.

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

[97] 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.

[98] 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.

[99] 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.

[100] 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

[101] Jun Yu, Zhihong Wei, and Zhongpeng Cai. Exploring facial expression recognition through semi-supervised pretraining and temporal modeling. *arXiv preprint arXiv:2403.11942*, 2024. 6

[102] Jun Yu, Zerui Zhang, Zhihong Wei, Gongpeng Zhao, Zhongpeng Cai, Yongqi Wang, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Aud-tgn: Advancing action unit detection with temporal convolution and gpt-2 in wild audiovisual contexts. *arXiv preprint arXiv:2403.13678*, 2024. 7

[103] Jun Yu, Gongpeng Zhao, Yongqi Wan, Zhihong Wei, Yang Zheng, Zerui Zhang, Zhongpeng Cai, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Multimodal fusion method with spatiotemporal sequences and relationship learning for valence-arousal estimation. *arXiv preprint arXiv:2403.12425*, 2024. 5

[104] Jun Yu, Jichao Zhu, and Wangyuan Zhu. Compound expression recognition via multi model ensemble. *arXiv preprint arXiv:2403.12572*, 2024. 7

[105] Jun Yu, Wangyuan Zhu, and Jichao Zhu. Efficient feature extraction and late fusion strategy for audiovisual emotional mimicry intensity estimation. *arXiv preprint arXiv:2403.11757*, 2024. 8

[106] 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

[107] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 4

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

[109] 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.

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

[111] 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.

[112] 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.

[113] 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. 2

[114] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge. *arXiv preprint arXiv:2403.10825*, 2024. 5, 6, 7, 8

[115] 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. 2

[116] 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

[117] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer. *arXiv preprint arXiv:2303.08356*, 2023. 2

[118] Weiwei Zhou, Jiada Lu, Chenkun Ling, Weifeng Wang, and Shaowei Liu. Boosting continuous emotion recognition with self-pretraining using masked autoencoders, temporal convolutional networks, and transformers. *arXiv preprint arXiv:2403.11440*, 2024. 5, 6, 7

[119] 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