An Effective Ensemble Learning Framework for Affective Behaviour Analysis

Wei Zhang[1,*], Feng Qiu[1,*], Chen Liu[1,2], Lincheng Li[1, †], Heming Du[2], Tianchen Guo[2], Xin Yu[2]
[1] Netease Fuxi AI Lab
[2] The University of Queensland
{zhangwei05, qiufeng, lilincheng}@corp.netease.com, chen.liu7@uqconnect.edu.au,
{Heming.du, tianchen.guo, xin.yu}@uq.edu.au

Abstract

Affective Behavior Analysis aims to facilitate technology emotionally smart, creating a world where devices can understand and react to our emotions as humans do. To comprehensively evaluate the authenticity and applicability of emotional behavior analysis techniques in natural environments, the 6th competition on Affective Behavior Analysis in-the-wild (ABAW) utilizes the Aff-Wild2, Hume-Vidmimic2, and C-EXPR-DB datasets to set up five competitive tracks, i.e., Valence-Arousal (VA) Estimation, Expression (EXPR) Recognition, Action Unit (AU) Detection, Compound Expression (CE) Recognition, and Emotional Mimicry Intensity (EMI) Estimation. In this paper, we present our method designs for VA estimation, expression recognition, and AU detection tracks. Specifically, our framework mainly includes three aspects: 1) To achieve high-quality facial feature representations, we employ Masked-Auto Encoder as the visual features extraction model and fine-tune it with our facial dataset. 2) Utilizing a transformer-based feature fusion module to fully integrate emotional information provided by audio signals, visual images, and transcripts, offering high-quality expression features for the downstream tasks. 3) Considering the complexity of the video collection scenes, we conduct a more detailed dataset division based on scene characteristics and train the classifier for each scene. Extensive experiments demonstrate the superiority of our designs. Our work won the championship in the AU, EXPR, and VA tracks at the ABAW6 competition.

1. Introduction

Affective Behavior Analysis is dedicated to enhancing the emotional intelligence of artificial intelligence systems by analyzing and understanding human emotional behavior [25, 27–34, 52, 58, 75, 84, 89]. It involves identifying and interpreting the emotions and feelings people express through facial expressions, voice, body language, *etc*. The goal is to enable computers and robots to better understand human emotional states for more natural and effective human-machine interactions, support mental monitoring, and improve applications in education, entertainment, and social interactions [14, 16, 56, 64, 65].

The 6th Affective Behavior Analysis competition (ABAW6) has set up the following five tasks to analyze various aspects of human emotions and expressions. Action Unit (**AU**) Detection aims to identify facial action types from the Facia Action Coding System based on facial muscle movements [2, 32, 41, 66]. Expression Recognition (**EXPR**) identifies basic emotional expressions like happiness, sadness, and anger [13, 39, 57, 69, 94]. Valence-arousal (**VA**) estimation determines people's emotional states on continuous emotional dimensions, where "valence" refers to the positivity or negativity of the emotion, and "arousal" refers to the level of emotional activation [22, 26, 32, 44, 54]. Compound Expression (**CE**) Recognition requires recognizing complex expressions that combine two or more basic expressions [9, 18, 61, 69]. Emotional Mimicry Intensity (**EMI**) Estimation evaluates the intensity of an individual's emotional mimicry [15, 20, 36, 70]. This work mainly focuses on the tasks of AU, EXPR, and VA.

ABAW6 assesses the method performance on Aff-Wild2 [24], C-EXPR-DB [23], and Hume-Vidmimic2 [34], in which videos are captured in uncontrolled natural environments. The AU, EXPR, and VA tracks utilize the Aff-Wild2 dataset, which is a large-scale multi-modal video dataset annotated with AU, basic expression categories, and VA. Aff-Wild2 features individuals of diverse skin tones, ages, and genders, captured in various lighting conditions, backgrounds, and head poses, adding richness to its diversity and complexity. This close resemblance to practical application scenarios facilitates the development of human affective behavior analysis applications.

Based on the characteristics of the Aff-Wild2 dataset, our objectives are to fully utilize the emotional information provided in multimodal data and improve the applicability of our method in real-world scenarios. In this paper, we outline our method designs in three aspects. Firstly, we integrate a large-scale facial image dataset and utilize the self-supervised model Masked Auto Encoder (MAE) [17, 89] to learn deep feature representations from these emotional data, enhancing the performance of downstream tasks. It is worth noting that the amount of data used in the MAE pre-training for this competition is nearly double that of the previous ABAW5 [89], further boosting the representation capacity of facial visual features encoded by MAE.

Secondly, we leverage a transformer-based model to fuse the multi-modal information. This architecture facilitates the interactions across modalities (*i.e.,* audio and visual) and provides scalable, efficient, and effective solutions for integrating multimodal information [71]. Thirdly, we adopt an ensemble learning strategy to enhance the robustness of our method in various complex scenes. This involves dividing the entire dataset into multiple sub-datasets based on distinct background characteristics and assigning them to different classifiers. After that, we integrate the outputs of these classifiers to obtain the final prediction results.

Experiments conducted on the three datasets demonstrate the effectiveness of our design choices. Overall, our contributions are three-fold:

- We integrate a large-scale facial expression dataset and fine-tune MAE on it to obtain an effective facial expression feature extractor, enhancing the performance for downstream tasks.
- We employ a transformer-based multi-modal integration model to facilitate the interactions of multi-modalities, enriching the expression features extracted from multi-modal data.
- We adopt an ensemble learning strategy, which trains multiple classifiers on sub-datasets with different scene characteristics and ensemble the results of these classifiers to attain the final results. This strategy enables our method to generalize better in various environments.

2. Related Work

2.1. Action Unit Detection

Detecting Action Units (AU) in the wild is a challenging yet crucial advancement task in facial expression analysis, pushing the boundaries of applicability from controlled laboratory settings to real-world environments [2, 41, 66, 78–83]. This endeavor addresses the inherent variability in lighting, pose, occlusion, and emotional context encountered in natural environments [32]. Recent works highlight the effectiveness of multi-task frameworks in leveraging extra regularization, such as the extra label constraint, to enhance detection performance. Zhang *et al.* [85] introduce a streaming model to concurrently execute AU detection, expression, recognition, and Valence-Arousal (VA) regression based on the fine-grained expression embedding [86]. Some works [87, 88] also employ this expression embedding to facilitate AU detection. Cui *et al.* [7] present a biomechanics-guided AU detection approach to explicitly incorporate facial biomechanics for AU detection. Moreover, to achieve robust and generalized AU detection, some works take generic knowledge (*i.e.* static spatial muscle relationships) into account [6], while others consider integrating multi-modal knowledge to obtain rich expression features [92].

2.2. Expression Recognition

Expression Recognition has witnessed substantial growth, driven by the integration of psychological insights and advanced deep learning techniques [39, 45, 57, 69]. Recently, the adaptation of transformer-based models from natural language processing (NLP) [68] to computer vision tasks [10] has led to their application in extracting spatial and temporal features from video sequences for emotion recognition. Notably, Zhao *et al.* [93] introduce a transformer model specifically for dynamic facial expression recognition, the Former-DFER, which includes CSFormer [73] and T-Former [73] modules to learn spatial and temporal features, respectively. Ma *et al.* [49] developed a Spatio-Temporal Transformer (STT) that captures both spatial and temporal information through a transformer-based encoder. Additionally, Li *et al.* [38] proposed the NR-DFERNet, designed to minimize the influence of noisy frames within video sequences. While these advancements represent significant progress in addressing the challenges of dynamic facial expression recognition (DFER) with discrete labels, they overlook the interference from the background in images. To address this, we incorporate ensemble learning into our method.

2.3. Valence-arousal Estimation

Valence-arousal estimation focuses on mapping emotional states onto a two-dimensional space, where valence represents the positivity or negativity of emotion, and arousal indicates its intensity or activation level [22, 26, 32]. Conventional approaches mainly relied on physiological signals, such as heart rate or skin conductance, to estimate these dimensions [3, 35, 37]. However, with advancements in deep learning, researchers shift towards leveraging visual and auditory cues from facial expressions, voice tones, and body language. Notably, convolutional neural networks and recurrent neural networks have been extensively applied to capture the nuanced and dynamic aspects of emotions from images, videos, and audio data [4, 50, 72].

Recent studies introduce transformer models to better handle the sequential and contextual nature of emotional expressions in multi-modal data [5, 21, 42, 43, 55, 62]. These improvements have not only improved the accu-

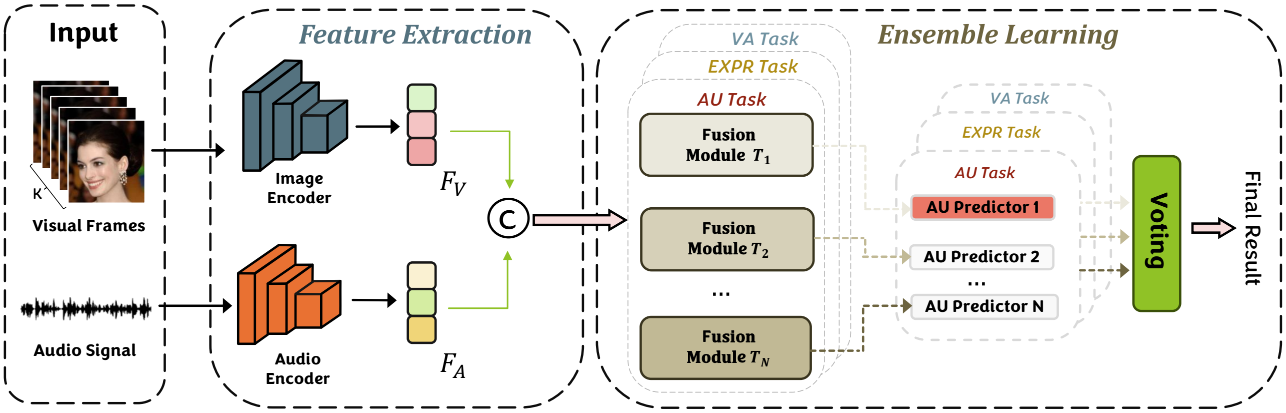

Figure 1. The overview of our proposed framework. We first utilize the images in the facial image datasets to train the Image Encoder in a self-supervised manner. thus obtaining the visual feature F_I. We first utilize our self-trained Image Encoder to generate the visual feature F_V. Meanwhile, we leverage the pre-trained audio encoder to attain the audio feature F_A. Then, we concat these features and feed them into the Fusion Modules. Here, we train N Fusion Modules and predictors on sub-datasets divided based on background characteristics. In the inference stage, we adopt a voting strategy to integrate the results predicted by all branches. Note that only the Fusion Modules and Predictors are trainable in each task.

racy and efficiency of valence-arousal estimation but also broadened its applicability in real-world scenarios, such as human-computer interaction and mental health assessment [12, 48, 63]. Despite progress, challenges remain in capturing the complex and subjective nature of emotions, necessitating further research into model interpretability and the integration of diverse data sources.

3. Method

In this section, we describe our method for analyzing human affective behavior. The architecture flow is illustrated in Fig. 1. The proposed approach addresses two critical problems: 1) the emotional information in the multimodal data is not fully explored and 2) the model has poor generalization ability for videos recorded genuine emotional responses to various forms of media. For a clear exposition, we first introduce how we utilize the encoders to extract features from multi-modal data in Sec. 3.1. Then we detail the transformer-based multi-modal feature fusion method in Sec. 3.2. Finally, in Sec. 3.3, we present the ensemble learning strategy that is leveraged to enhance the model generalization ability.

3.1. Feature Extraction

Image Encoder. In this work, we employ MAE as the image encoder since its self-supervised training manner enables the extracted features more generalizable. To further attain powerful and expressive features, we construct a large-scale facial image dataset which consists of AffectNet [51], CASIA-WebFace [74], CelebA [46], IMDB-WIKI [59], and WebFace260M [96]. After removing low-quality images with unclear faces, we preserve nearly 4.5M high-quality facial images, which is double the amount utilized in the 5th competition [89]. Based on the integrated facial dataset, we finetune MAE through facial image reconstruction. Specifically, in the pre-training phase, our method adopts the "mask-then-reconstruct" strategy. Here, images are dissected into multiple patches (measuring 16×16 pixels), with a random selection of 75% being obscured. These masked images are then input into the encoder, while the decoder restores them to the corresponding original. We adopt the pixel-wise L2 loss to optimize the model, ensuring the reconstructed facial images closely mirror the originals.

After the pre-training, we modify the model for specific downstream tasks by detaching the MAE decoder and incorporating a fully connected layer to the end of the encoder. This alternation facilitates the model to better adapt to the downstream tasks.

Audio Encoder. Considering that the tone and intonation of the speech can also reflect certain emotional information, we leverage different pre-trained audio feature extraction models, *e.g.* Vggish [19], Hubert [67] and Wav2vec2 [1], as our audio encoder to generate the audio representation. Given that these models are trained on large-scale datasets and with the ability to capture a wide range of audio features, we directly utilize them as the feature extractor without training on our dataset.

3.2. Transformer-based Multi-modal Fusion

We fuse features across different modalities to obtain more reliable emotional features and utilize the fused feature for downstream tasks. By combining information from visual F_V and audio F_A, we achieve a more comprehensive and accurate emotion representation.

To align the three modalities at the temporal dimension, we trim each video into multiple clips with k frames. For each frame, we employ our image encoder to extract the visual feature f_V. In this fashion, we attain the visual feature $F_V^{K\times d}$ for the whole clip. Here, d represents the feature dimension. Meanwhile, we employ the audio to generate the features for the whole clip, and the feature is expressed by $F_A^{1\times d}$. Subsequently, we concat these features and input them into the Transformer Encoder. Specifically, our transformer encoder consists of four encoder layers with a dropout rate of 0.3. The output is then fed into a fully connected layer to adjust the final output dimension according to the task requirements. Note that, at the feature fused stage, the image encoder and audio encoder are fixed, while only the fusion modules as well as predictors (*i.e.,* the fully connected layers) are trainable.

3.3. Ensemble Learning

To improve the applicability of affective behavior analysis methods, the 6th ABAW leverages the datasets that record human real-emotion reactions as the official test data. We observe that a significant proportion of the Aff-Wild2 dataset consists of "reaction videos", capturing genuine emotional responses to various forms of media. These videos often display similar emotional expressions, with surprise and happiness being prevalent. Therefore, we train separate fusion modules and predictors for the reaction videos and the entire training dataset.

In the inference stage, we manually pick out the reaction videos from the test set and utilize the corresponding model to predict their emotional labels. For the remaining test videos, we leverage the model trained on the entire training set to obtain their labels. Moreover, we train multiple models on the five-fold random split sub-datasets and the whole dataset. In this fashion, we collect the results from these models and devise a vote ensemble strategy to integrate the final result. Notably, we choose the predicted label with the highest number of votes as the final classification result for AU and EXPR tasks. As for the VA task, we calculate the average value predicted by the different models. Our voting method effectively minimizes errors due to biases in classifiers from individual subsets, thereby improving the overall classification performance.

3.4. Training Objectives

Objectives for Image Encoder. To enhance the adaptability of the Image Encoder across various tasks, we fine-tune it for each downstream task. Specifically, when dealing with AU and EXPR, we optimize the model via cross-entropy loss $\mathcal{L}_{AU\_CE}$ and $\mathcal{L}_{EXPR-CE}$, respectively. They are defined as follows:

$$\mathcal{L}_{AU\_CE} = -\frac{1}{12}\sum_{j=1}^{12} W_{au_j}\left[y_j \log \hat{y}_j + (1-y_j)\log\left(1-\hat{y}_j\right)\right], \tag{1}$$

$$\mathcal{L}_{EXPR-CE} = -\frac{1}{8}\sum_{j=1}^{8} W_{exp-j} z_j \log \hat{z}_j, \tag{2}$$

where $\hat{y}$ and $\hat{z}$ represent the predicted results for the action unit and expression category respectively, whereas y and z denote the ground truth values for the action unit and expression category.

In the VA task, to better capture the correlation between valence and arousal and thus improve the accuracy of emotion recognition, we leverage the consistency correlation coefficient as the model optimization function, defined as:

$$\mathrm{CCC}(\mathcal{X},\hat{\mathcal{X}}) = \frac{2\rho_{\mathcal{X}\hat{\mathcal{X}}}\delta_{\mathcal{X}}\delta_{\hat{\mathcal{X}}}}{\delta_{\mathcal{X}}^2+\delta_{\hat{\mathcal{X}}}^2+\left(\mu_{\mathcal{X}}-\mu_{\hat{\mathcal{X}}}\right)^2}, \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{VA\_CCC}} = 1 &- \mathrm{CCC}(\hat{v}_{batch_i}, v_{batch_i}) \\ &+1 - \mathrm{CCC}(\hat{a}_{batch_i}, a_{batch_i}).\end{aligned} \tag{4}$$

Here, $\hat{v}$ and $\hat{a}$ represent the predicted valence and arousal value. $\delta_{\mathcal{X}}$ and $\delta_{\hat{\mathcal{X}}}$ indicate the ground-truth sample set and the predicted sample set. $\rho_{\mathcal{X}\hat{\mathcal{X}}}$ is the Pearson correlation coefficient between $\mathcal{X}$ and $\hat{\mathcal{X}}$, $\delta_{\mathcal{X}}$ and $\delta_{\hat{\mathcal{X}}}$ are the standard deviations of $\mathcal{X}$ and $\hat{\mathcal{X}}$, and $\mu_{\mathcal{X}}$, $\mu_{\hat{\mathcal{X}}}$ are the corresponding means. The numerator $2\rho_{\mathcal{X}\hat{\mathcal{X}}}\delta_{\mathcal{X}}\delta_{\hat{\mathcal{X}}}$ represents the covariance between the $\delta_{\mathcal{X}}$ and $\delta_{\hat{\mathcal{X}}}$ sample sets.

Objectives for Transformer-based Multi-modal Fusion Model. In the training stage for Transformer-based Multi-modal Fusion Model (TMF), we first convert image data into sequence data by combining fixed-length adjacent frames. The related audio features are also aligned frame by frame with the temporal order of the image sequence. Then the sequence multi-modal data is sent to TMF and outputs the predictions for the sequence. When computing loss, we expand the sequence to each frame and calculate the frame-wise function as Equ. 1,2,3.

Post processing. In the inference stage, we leverage a Gaussian filter to refine the likelihood estimations for AU, EXPR, as well as VA. It is formulated as:

$$\mathcal{L}_{smooth} = \int_{-\infty}^{\infty}\left(y - \frac{\sqrt{2}\cdot f(x)\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{2\sqrt{\pi}\sigma}\right)^2 dx, \tag{5}$$

where y represents the predicted value of the downstream tasks, $f(x)$ is the predicted likelihood estimation before applying the Gaussian filter, and e is the base of the natural logarithm. x and μ represent the input value and the mean of the distribution, respectively. σ indicates the standard deviation of the distribution, determining the width of the Gaussian curve. The Gaussian filter's sigma parameter is tuned specifically for each task.

4. Experiment

In this section, we will first introduce the evaluation metrics datasets as well as the implementation details. Then we evaluate our model on the ABAW6 competition metrics.

4.1. Evaluation Metrics

To assess the model performance on each track, ABAW set a specific evaluation metric for each track.

Valence-Arousal Estimation. The performance measure (P) is the mean Concordance Correlation Coefficient (CCC) of valence and arousal, as follows:

$$P = \frac{\mathrm{CCC}_{arousal} + \mathrm{CCC}_{valence}}{2} \tag{6}$$

Here, the calculation of CCC is defined in Eq. 3.

Expression Recognition. The performance assessment is conducted by averaging F1 score across all 8 categories, defined as:

$$\begin{cases} F1 = \dfrac{2 \times Precision \times Recall}{Precision + Recall}; \\ Precision = \dfrac{TP}{TP + FP}; \\ Recall = \dfrac{TP}{TP + FN}, \end{cases} \tag{7}$$

$$P = \frac{\sum_{c=1}^{8} F1_c}{8}. \tag{8}$$

Here, c represents the category ID, TP represents True Positives, FP represents False Positives, and FN represents False Negatives.

Action Unit Detection. The performance is evaluated by averaging the F1 score across all 12 categories, formulated as:

$$P = \frac{\sum_{c=1}^{12} F1_c}{12} \tag{9}$$

Here, the calculation way of $F1$ is the same as the Eq. 7.

4.2. Datasets

The first tracks of ABAW6 are based on Aff-wild2 which contains around 600 videos annotated with AU, base expression category, and VA. The AU detection track utilizes 547 videos of around 2.7M frames that are annotated in terms of 12 action units, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26. The performance measure is the average F1 Score across all 12 categories. The expression recognition track utilizes 548 videos of around 2.7M frames that are annotated in terms of the 6 basic expressions (*i.e.,* anger, disgust, fear, happiness, sadness, surprise), plus the neutral state, plus a category 'other' that denotes expressions/affective states other than the 6 basic ones. The performance measure is the average F1 Score across all 8 categories. The VA estimation track utilizes 594 videos of around 3M frames of 584 subjects annotated in terms of valence and arousal. The performance measure is the mean Concordance Correlation Coefficient (CCC) of valence and arousal.

In addition to the official datasets mentioned above, we also used some additional data from the open-source and private datasets. In the pre-training for MAE self-supervised, we collect large number of facial images from the available public datasets (i.e. AffectNet [51], CASIA-WebFace [74], CelebA [46], IMDB-WIKI [59], and WebFace260M [96]) and private facial image datasets that from the Internet. Our private images are mainly from film and television works, and public video platforms. This data relies on Netease Fuxi Youling Crowdsourcing[1] platform for data cleaning and management. For the AU detection track, we use the extra dataset BP4D [91] to supplement some of the limited AU categories in Aff-wild2. For the expression recognition track, we use the extra dataset RAF-DB [40] and AffectNet [51] to supplement the Anger, Disgust, and Fear data.

4.3. Implementatal Setting

We utilize retinaface [8] to detect faces for each frame and normalize them to a size of 224 × 224 pixels. We pre-train an MAE on a large facial images dataset that consists of several open-source face images datasets (i.e., AffectNet [51], CASIA-WebFace [74], CelebA [46] and IMDB-WIKI [59], Webface260M [96]). We use this MAE as the basic feature extractor to capture the visual information for facial images in each track. The pre-training process is trained for 800 epochs with a batch size of 4096 on 8 NVIDIA A30 GPUs, using the AdamW optimizer [47]. For the tasks of AU detection, expression recognition, and VA estimation, we incorporate the temporal, audio, and other information to further improve the performance. At this stage, the training data consists of continuous video clips of 100 frames. The learning rate is set as 0.0001 using the AdamW optimizer. To reduce the gap caused by data division, we conduct five-fold cross-validation for all the tracks.

4.4. Evaluation on the Validation Dataset

Results for AU Detection. Tab. 1 presents the results of our method on the official validation set and the five-fold cross-validation in the AU detection track. As the average F1 scores suggested, the results of the five-fold cross-validation

[1] https://fuxi.163.com/solution/data

Table 1. The AU F1 scores (in %) of models that are trained and tested on different folds (including the original training/validation set of *Aff-Wild2* dataset).

Val Set	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU23	AU24	AU25	AU26	Avg.
Official	55.29	51.40	65.81	68.61	76.08	75.00	75.24	37.65	18.89	30.89	83.41	44.98	56.94
fold-1	62.61	46.20	71.22	77.71	67.44	69.69	74.62	36.32	29.43	21.75	81.56	40.73	56.61
fold-2	64.23	54.35	73.85	77.33	77.49	76.70	80.74	29.05	28.96	18.47	87.71	43.63	59.37
fold-3	58.55	48.37	60.05	71.22	72.43	74.29	75.43	29.81	19.52	32.86	83.37	47.63	56.13
fold-4	53.34	39.34	66.26	70.67	66.51	69.39	71.76	39.49	25.17	32.40	82.27	40.05	54.72
fold-5	53.50	44.68	63.45	72.02	69.72	74.00	78.24	38.81	23.67	7.56	81.24	43.67	54.22

Table 2. The expression F1 scores (in %) of models that are trained and tested on different folds (including the original training/validation set of *Aff-Wild2* dataset).

Val Set	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Other	Avg.
Official	70.21	73.93	50.34	21.83	59.05	66.41	36.51	66.11	55.55
fold-1	70.06	37.21	32.12	22.71	61.77	77.61	45.62	51.58	49.83
fold-2	67.36	44.45	21.21	42.50	62.22	78.24	36.67	70.00	52.83
fold-3	73.64	71.60	45.01	23.25	47.67	77.05	46.81	65.56	56.32
fold-4	65.41	71.00	53.70	23.27	61.62	61.79	27.76	72.68	54.65
fold-5	64.03	31.23	35.66	67.64	67.97	69.75	52.12	55.64	55.51

Table 3. The VA CCC scores of models that are trained and tested on different folds (including the original training/validation set of *Aff-Wild2* dataset).

Val Set	Valence	Arousal	Avg.
Official	0.5523	0.6531	0.6027
fold-1	0.6408	0.6195	0.6302
fold-2	0.6033	0.6758	0.6395
fold-3	0.6773	0.6961	0.6867
fold-4	0.6752	0.6486	0.6619
fold-5	0.6591	0.7019	0.6801

are consistent with the results obtained on the official validation set (56.94% on the F1 score). This demonstrates the superior generalization capability of our method.

Based on the detection results from each category, our method achieves comparatively high F1 scores on AU1, AU2, AU4, AU6, AU7, AU10, AU12, and AU25, all of which exceed 50% on the official validation dataset. Particularly, our method attains 83.41% on AU25. Conversely, the F1 score for AU15 (*lip tightening*), AU23 (*lip puckering*), and AU24 (*lip pressing*) are relatively low, especially with AU23 scoring only 18.89%. We speculate that the challenges in detecting the three categories may stem from the nuanced nature of the human face when displaying these three expressions. The subtle variations in these expressions pose challenges in capturing distinct features compared to other AUs, thus diminishing the detection accuracy.

Table 4. Final result comparisons with other participating teams. Our team (*i.e.* **NetEase Fuxi AI Lab**) attains the highest results on the three competitive tracks. We color code the best results. Average CCC indicates the average CCC scores of Valence and Arousal.

Tracks	Teams	Average CCC	CCC-V	CCC-A	F1 (%)
VA	Baseline [34]	0.2010	0.2110	0.1910	-
	SUN CE [11]	0.5608	0.5355	0.5861	-
	CtyunAI [95]	0.5640	0.564	0.6057	-
	DeepAVER [53]	0.5807	0.5418	0.6196	-
	Ours [90]	0.6721	0.6873	0.6569	-
EXPR	Baseline [34]	-	-	-	22.50
	HSEmotion [60]	-	-	-	34.14
	USTC-IAT-United [76]	-	-	-	35.34
	CtyunAI [95]	-	-	-	36.25
	Ours [90]	-	-	-	50.05
AU	Baseline [34]	-	-	-	36.50
	USTC-IAT-United [77]	-	-	-	48.40
	HSEmotion [60]	-	-	-	48.78
	CtyunAI [95]	-	-	-	49.41
	Ours [90]	-	-	-	56.01

Results for Expression Recognition. The results on the official validation dataset and the five-fold datasets are shown in Tab. 2. Our method achieves similar results to the official validation set in four of the folds, with the results in fold-1 being 5.72% lower compared to the official validation set. This indicates that our method performs well in the majority of data distributions and demonstrates a certain level of generalization capability. Additionally, the "Fear" and "Surprise" categories show relatively low results in the official validation set, with F1 scores of 21.83% and 36.51%, respectively, while achieving relatively higher results in certain folds. This implies that our method may not fully learn the features of the two categories during the training process, resulting in lower performance for these categories.

Results for VA Estimation. Tab. 3 shows the VA CCC scores on the official validation dataset and our five-fold validation sets. As the average scores indicated, the results of the five-fold cross-validation experiments are consistently higher than those on the official validation dataset (with an average VA CCC score of 0.6027). This demonstrates that our method performs more robustly and reliably across the entire dataset. Moreover, the CCC score of our method for predicting Arousal is higher than that for predicting Valence, *e.g.,* the Arousal CCC score is about 0.1 higher than the Valence score on the official validation set. This proves that our model performs better at predicting emotional arousal (Arousal) than predicting emotional valence (Valence) on the Aff-Wild2 dataset.

4.5. Evaluation on the Test Dataset

We display the competition final evaluation results of competitive teams and ours on the three tracks in Tab. 4. Note that, the final evaluation is conducted on an unseen test set. Our method demonstrates significant superiority over all competing teams, achieving first place in all three tracks. Specifically, in the VA estimation task, our method outperforms the second-place team by 15.7% in terms of average CCC. While DeepAVER [53] focuses more on leveraging complementary information across different modalities, our approach extensively explores the information between various modalities and emphasizes enhancing the robustness of the model to complex shooting environments.

In the expression recognition task, CtyunAI [95] attains second place with the F1 score of 0.3625, while our method attains 0.5005. CtyunAI and our method adopt a similar feature extraction process, *i.e.,* employing the MAE as the visual feature extractor and leveraging a Transformer-based structure for multimodal feature fusion. The difference is that our method divides the dataset based on the characteristics of backgrounds and trains the model in an ensemble learning strategy. In the AU detection task, the F1 score of CtyunAI [95] in the second place is 0.4941, which is 0.066 lower than our score. We still categorize the main factor influencing the final results of the two teams is whether they focus on the impact of complex background on the detection accuracy.

Table 5. The ablation experiments on the official validation set to evaluate the effectiveness of our new MAE, Transformer-based Multi-modal Fusion (TMF) and the scene data division (SDD).

Tracks	MAE ABAW5 [89]	MAE ABAW6	TMF	SDD	avg_CCC	F1 (%)
VA	✓	×	×	×	0.5483	-
	×	✓	×	×	0.5647	-
	✓	×	✓	×	0.5525	-
	×	✓	✓	×	0.5786	-
	×	✓	✓	✓	**0.6027**	-
EXPR	✓	×	×	×	-	46.79
	×	✓	×	×	-	49.28
	✓	×	✓	×	-	48.93
	×	✓	✓	×	-	52.59
	×	✓	✓	✓	-	**55.55**
AU	✓	×	×	×	-	54.83
	×	✓	×	×	-	55.69
	✓	×	✓	×	-	55.86
	×	✓	✓	×	-	**56.94**
	×	✓	✓	✓	-	56.63

4.6. Ablation Study

To demonstrate the effectiveness of our designs, we present the results of the ablation studies in Tab. 5 and Tab. 6.

Enhaned MAE pre-training. From Tab. 5, it can be observed that the use of new MAE in ABAW6 enhances the performance of the three tasks, not only in static image training but also when TMF is added. The difference between MAE ABAW5 and MAE ABAW6 is the amount of the self-supervised training data. MAE ABAW6 expands the data by almost double that of MAE ABAW5. The experiment results prove the effectiveness of the new MAE and facial data augmentation.

Transformer-based Multi-modal Fusion (TMF). In the three tasks, the use of temporal multi-modal information further improves the model performance. Specifically, the average CCC of VA increases from 0.5647 to 0.5786 by adding the TMF. This improvement also occurs in the AU (0.5569→0.5694) and EXPR tracks (0.4928→0.5259). This illustrates that multi-modal information plays an important role in analyzing human facial expressions.

Scene Data Division (SDD). In the aff-wild2 dataset, a common scenario involves "reaction videos," where individuals record their reactions to another video or content, often providing real-time commentary, facial expressions, and opinions. This type of data constitutes a large portion of the dataset. And a large portion of them exhibit similar emotional expressions. Therefore, we specifically train this subset of reaction video data separately and then merge the prediction results with those of the entire dataset, which we refer to as SDD. From the Tab. 5, SDD effectively improve the CCC in VA track and F1 score in EXPR track. However, SDD does not show a significant effect in the AU track. This may be because these reaction videos exhibit similar expres-

Table 6. The ablation experiments on the official validation set to evaluate the impact of different visual and audio features combination. The metrics for VA, EXPR, and AU are average CCC, F1 score (%), and F1 score (%), separately. Each individual track experiment has already defaulted to using the corresponding uni-task visual features pre-trained on the aff-wild2 dataset. The pre-trained model defaults to using the MAE architecture.

Visual				Audio			Tracks		
$affn_{VA}$	$affn_{EXPR}$	raf_{EXPR}	$BP4D_{AU}$	Hubert	wav2vec	vggish	VA	EXPR	AU
✓	✓	×	×	✓	×	×	0.5698	50.04	55.82
✓	✓	✓	×	✓	×	×	0.5721	50.96	56.37
✓	✓	✓	×	×	×	×	0.5685	49.71	**56.94**
✓	✓	✓	✓	✓	×	×	0.5751	51.15	54.73
✓	✓	✓	✓	✓	✓	×	0.5741	51.98	54.89
✓	✓	✓	✓	✓	×	✓	**0.5786**	**52.59**	55.71

sions in emotion categories and VA intensity, but there are significant differences in the distribution of AU.

Feature selection in TMF. In the stage of TMF, we also use several combinations of features from different pre-trained models. In the visual modality, we utilize the uni-task model of the MAE fine-tuned on Aff-wild2 ($aff2_{AU}$,$aff2_{VA}$,$aff2_{EXPR}$). In addition, we also fine-tuned the MAE model on multiple tasks, including AU, EXPR, and VA tasks in different datasets such as BP4D ($BP4D_{AU}$), AffectNet ($affn_{EXPR}$ and $affn_{VA}$), and RAF-DB (RAF_{EXPR}). The features from these models are also combined in the TMF training. In the audio modality, we try the features of Hubert [67], wav2vec2 [1] and vggish [19] from the open-source pre-trained models. Tab. 6 shows the results of different feature combinations in the three tracks. In TMF, each single track defaults to using a uni-task model as one kind of visual feature. This is not separately listed in Tab. 6. It can be observed that the visual features from extra datasets facilitate EXPR and VA tracks. However, adding the visual features based on BP4D and audio features harms the performance of the AU metrics. This may be due to the differences in the scenes and annotation rules of the BP4D dataset, leading to a gap between the datasets. Aff-wild2, AffectNet, and RAF-DB are all in-the-wild datasets. But BP4D is an in-the-lab dataset. Also, the combination of Hubert and Vggish features is most beneficial for the EXPR and VA tasks.

5. Conclusion

In summary, our study contributes to advancing Affective Behavior Analysis, aiming to make technology emotionally intelligent. Through a comprehensive evaluation of the ABAW competition, we address five competitive tracks. Our method designs integrate emotional cues from multi-modal data, ensuring robust expression features. We achieve significant performance across all tracks, indicating the effectiveness of our approach. These results highlight the potential of our method in enhancing human-machine interactions and technological advancements toward devices understanding and responding to human emotions.

6. Acknowledgments

This work is supported by the 2022 Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054). We express our gratitude to all those who have contributed to this endeavor.

References

[1] 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. 3, 8

[2] Soufiane Belharbi, Marco Pedersoli, Alessandro Lameiras Koerich, Simon Bacon, and Eric Granger. Guided interpretable facial expression recognition via spatial action unit cues. *arXiv preprint arXiv:2402.00281*, 2024. 1, 2

[3] Patricia J Bota, Chen Wang, Ana LN Fred, and Hugo Plácido Da Silva. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE access*, 7:140990–141020, 2019. 2

[4] Paul Buitelaar, Ian D Wood, Sapna Negi, Mihael Arcan, John P McCrae, Andrejs Abele, Cecile Robin, Vladimir Andryushechkin, Housam Ziad, Hesam Sagha, et al. Mixedemotions: An open-source toolbox for multimodal emotion analysis. *IEEE Transactions on Multimedia*, 20(9): 2454–2465, 2018. 2

[5] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2020. 2

[6] Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 2

[7] Zijun Cui, Chenyi Kuang, Tian Gao, Kartik Talamadupula, and Qiang Ji. Biomechanics-guided facial action unit detection through force modeling. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8694–8703, 2023. 2

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 5

[9] Rongkang Dong and Kin-Man Lam. Bi-center loss for compound facial expression recognition. *IEEE Signal Processing Letters*, 2024. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Denis Dresvyanskiy, Maxim Markitantov, Jiawei Yu, Peitong Li, Heysem Kaya, and Alexey Karpov. Sun team's contribution to abaw 2024 competition: Audio-visual valence-arousal estimation and expression recognition, 2024. 6

[12] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020. 3

[13] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021. 1

[14] Chiara Filippini, David Perpetuini, Daniela Cardone, Antonio Maria Chiarelli, and Arcangelo Merla. Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review. *Applied Sciences*, 10(8):2924, 2020. 1

[15] Matthias Franz, Marc A Nordmann, Claudius Rehagel, Ralf Schäfer, Tobias Müller, and Daniel Lundqvist. It is in your face—alexithymia impairs facial mimicry. *Emotion*, 21(7): 1537, 2021. 1

[16] Riccardo Gervasi, Federico Barravecchia, Luca Mastrogiacomo, and Fiorenzo Franceschini. Applications of affective computing in human-robot interaction: State-of-art and challenges for manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 237(6-7):815–832, 2023. 1

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[18] Shuangjiang He, Huijuan Zhao, Li Yu, Jinqiao Xiang, Congju Du, and Juan Jing. Compound facial expression recognition with multi-domain fusion expression based on adversarial learning. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 688–693. IEEE, 2022. 1

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 3, 8

[20] Alison C Holland, Garret O'Connell, and Isabel Dziobek. Facial mimicry, empathy, and emotion recognition: a meta-analysis of correlations. *Cognition and Emotion*, 35(1):150–168, 2021. 1

[21] Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 512–520, 2020. 2

[22] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1, 2

[23] 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. 1

[24] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 1

[25] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[26] 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. 1, 2

[27] 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. 1

[28] 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.

[29] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[30] 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)*, pages 637–643. IEEE, 2020.

[31] 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.

[32] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emo-

tional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023. 1, 2

[33] 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.

[34] 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. 1, 6

[35] Jure Kranjec, S Beguš, G Geršak, and J Drnovšek. Non-contact heart rate and heart rate variability measurements: A review. *Biomedical signal processing and control*, 13:102–112, 2014. 2

[36] Beibei Kuang, Xueting Li, Xintong Li, Mingxiao Lin, Shanrou Liu, and Ping Hu. The effect of eye gaze direction on emotional mimicry: A multimodal study with electromyography and electroencephalography. *NeuroImage*, 226:117604, 2021. 1

[37] Bharat Lal, Raffaele Gravina, Fanny Spagnolo, and Pasquale Corsonello. Compressed sensing approach for physiological signals: A review. *IEEE Sensors Journal*, 2023. 2

[38] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022. 2

[39] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020. 1, 2

[40] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. 5

[41] Yante Li, Xiaohua Huang, and Guoying Zhao. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing*, 436:221–231, 2021. 1, 2

[42] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *arXiv preprint arXiv:2308.10175*, 2023. 2

[43] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 2

[44] Xiaolong Liu, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Yuanyuan Deng, Zhaopei Huang, Liyu Meng, Yuchen Liu, and Chuanhe Liu. Evaef: Ensemble valence-arousal estimation framework in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5862–5870, 2023. 1

[45] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–4. IEEE, 2019. 2

[46] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3, 5

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[48] Danielle Lottridge, Mark Chignell, and Aleksandra Jovicic. Affective interaction: Understanding, evaluating, and designing for human emotion. *Reviews of Human Factors and Ergonomics*, 7(1):197–217, 2011. 3

[49] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022. 2

[50] Javier Marín-Morales, Carmen Llinares, Jaime Guixeres, and Mariano Alcañiz. Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*, 20(18):5163, 2020. 2

[51] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 3, 5

[52] 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. 1

[53] R. Gnana Praveen and Jahangir Alam. Recursive cross-modal attention for multimodal fusion in dimensional emotion recognition, 2024. 6, 7

[54] R Gnana Praveen, Patrick Cardinal, and Eric Granger. Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023. 1

[55] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *arXiv preprint arXiv:2305.18891*, 2023. 2

[56] Minglun Ren, Nengying Chen, and Hui Qiu. Human-machine collaborative decision-making: An evolutionary roadmap based on cognitive intelligence. *International Journal of Social Robotics*, 15(7):1101–1114, 2023. 1

[57] I Michael Revina and WR Sam Emmanuel. A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, 33(6):619–628, 2021. 1, 2

[58] Albert D Ritzhaupt, Rui Huang, Max Sommer, Jiawen Zhu, Anita Stephen, Natercia Valle, John Hampton, and Jingwei Li. A meta-analysis on the influence of gamification in formal educational settings on affective and behavioral outcomes. *Educational Technology Research and Development*, 69(5):2493–2522, 2021. 1

[59] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018. 3, 5

[60] Andrey V. Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction, 2024. 6

[61] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6248–6257, 2021. 1

[62] Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:290–300, 2022. 2

[63] Rukshani Somarathna, Tomasz Bednarz, and Gelareh Mohammadi. Virtual reality for emotion elicitation–a review. *IEEE Transactions on Affective Computing*, 2022. 3

[64] Boštjan Šumak, Saša Brdnik, and Maja Pušnik. Sensors and artificial intelligence methods and algorithms for human–computer intelligent interaction: A systematic mapping study. *Sensors*, 22(1):20, 2021. 1

[65] Martina Szabóová, Martin Sarnovskỳ, Viera Maslej Krešňáková, and Kristína Machová. Emotion analysis in human–robot interaction. *Electronics*, 9(11): 1761, 2020. 1

[66] Gauthier Tallec, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. Multi-label transformer for action unit detection. *arXiv preprint arXiv:2203.12531*, 2022. 1, 2

[67] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP*, 2022. 3, 8

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[69] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020. 1, 2

[70] Tanja SH Wingenbach, Mark Brosnan, Monique C Pfaltz, Peter Peyk, and Chris Ashwin. Perception of discrete emotions in others: Evidence for distinct facial mimicry patterns. *Scientific reports*, 10(1):4692, 2020. 1

[71] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[72] Xinyu Yang, Yizhuo Dong, and Juan Li. Review of data features-based music emotion recognition methods. *Multimedia systems*, 24:365–389, 2018. 2

[73] Dongjie Ye, Zhangkai Ni, Hanli Wang, Jian Zhang, Shiqi Wang, and Sam Kwong. Csformer: Bridging convolution and transformer for compressive sensing. *IEEE Transactions on Image Processing*, 2023. 2

[74] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3, 5

[75] 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. 1

[76] 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

[77] 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. 6

[78] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *European conference on computer vision*, pages 318–333. Springer, 2016. 2

[79] Xin Yu and Fatih Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[80] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3760–3768, 2017.

[81] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018.

[82] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 908–917, 2018.

[83] Xin Yu, Fatemeh Shiri, Bernard Ghanem, and Fatih Porikli. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2148–2164, 2019. 2

[84] 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. 1

[85] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, Yu Ding, Runze Wu, Tangjie Lv, and Changjie Fan. Prior aided streaming network for multi-task affective analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3539–3549, 2021. 2

[86] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6759–6768, 2021. 2

[87] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 2428–2437, 2022. 2

[88] Wei Zhang, Lincheng Li, Yu Ding, Wei Chen, Zhigang Deng, and Xin Yu. Detecting facial action units from global-local fine-grained expressions. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[89] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multi-modal facial affective analysis based on masked autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2023. 1, 2, 3, 7

[90] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge, 2024. 6

[91] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 5

[92] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Classifier learning with prior probabilities for facial action unit recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5108–5116, 2018. 2

[93] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 2

[94] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 1

[95] 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, 2024. 6, 7

[96] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 3, 5