ABAW5 Challenge: A Facial Affect Recognition Approach Utilizing Transformer Encoder and Audiovisual Fusion

Ziyang Zhang* Liuwei An* Zishun Cui Ao Xu Tengteng Dong Yueqi Jiang Jingyi Shi Xin Liu
Xiao Sun† Meng Wang†
Hefei University of Technology
Hefei, China
{zzzzzy, anliuwei, 2022171285, 2022111070}@mail.hfut.edu.cn
{2022180026, 2017217795, 2019213576, 2019218059}@mail.hfut.edu.cn
sunx@hfut.edu.cn eric.mengwang@gmail.com

Abstract

In this paper, we present our approach to tackling the 5th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW). The competition comprises four sub-challenges, namely Valence-Arousal (VA) Estimation, Expression (Expr) Classification, Action Unit (AU) Detection, and Emotional Reaction Intensity (ERI) Estimation. To address theµse challenges, we leverage state-of-the-art (sota) models to extract robust audio and visual features. Subsequently, these features are fused using a Transformer Encoder for the VA, Expr, and AU sub-challenges, and TEMMA for the ERI sub-challenge. To mitigate the effect of disparate feature dimensions, we introduce an Affine Module to align the features to the same dimension. Overall, our results outperform the baseline by a substantial margin across all four sub-challenges. Specifically, for the VA Estimation sub-challenge, our method attains a mean Concordance Correlation Coefficient (CCC) of 0.5342, ranking fifth overall. For the Expression Classification sub-challenge, our approach achieves an average F1 Score of 0.3337, placing fourth overall. For the AU Detection sub-challenge, our method obtains an average F1 Score of 0.4752. Lastly, for the Emotional Reaction Intensity Estimation sub-challenge, our approach yields an average Pearson's correlation coefficient of 0.3968.

1. Introduction

Sentiment analysis is a crucial research area in pattern recognition that seeks to incorporate affective dimensions into human-computer interaction. Its diverse applications span various domains, such as mental health treatment, fatigue driving detection, and consumer attitude analysis, etc [11, 17, 45].

Among the various emotional factors that influence sentiment analysis, facial information is among the most compelling and realistic. In this paper, we concentrate on facial affect recognition and extensively outline our approach to addressing the four sub-challenges of the 5th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW). These sub-challenges include the Valence-Arousal (VA) Estimation Challenge, Expression (Expr) Classification Challenge, Action Unit (AU) Detection Challenge, and Emotional Reaction Intensity (ERI) Estimation Challenge.

The Valence-Arousal (VA) Estimation Challenge involves predicting the level of valence and arousal in a time-continuous manner from audio-visual recordings. The challenge requires predicting the values of valence and arousal continuously using a database containing approximately 3 million frames annotated in terms of valence and arousal.

The Expression (Expr) Classification Challenge necessitates training a model to predict six basic expressions, the neutral state, and a category labeled 'other.' The database used for this challenge includes approximately 2.7 million frames annotated in terms of the six basic expressions (anger, disgust, fear, happiness, sadness, surprise), the neutral state, and a category labeled 'other' that denotes expressions/affective states other than the six basic ones.

In the Action Unit (AU) Detection Challenge, the facial action coding system (FACS) [6] defines a group of action units (AU) based on facial anatomy to accurately describe facial expression changes. Each facial action unit describes the apparent changes caused by a group of facial muscle movements that can express any facial expression. This challenge requires predicting whether AUs are active, and is

[*] These authors contributed equally.
[†] Corresponding Author.

a multi-label classification problem. Challenges in AU detection include insufficient labeling data, head posture interference, individual differences, and an imbalance of different AU categories. The database used for this challenge includes approximately 2.7 million frames annotated in terms of 12 action units.

The Emotional Reaction Intensity (ERI) Estimation Challenge is more challenging than facial expression classification, as emotional experience can differ significantly across individuals of different ages, genders, and cultural backgrounds. This makes ERI estimation more complex and necessitates the use of multimodal information for comprehensive emotional recognition. The challenge requires training a multimodal model to predict seven emotional experiences using the Hume-Reaction dataset, which contains subjects reacting to a wide range of various emotional video-based stimuli, including Adoration, Amusement, Anxiety, Disgust, Empathic Pain, Fear, and Surprise.

The main contribution of the proposed method can be summarized as:

1. In this paper, we aim to extract informative features from audio and visual modalities, to improve performance in affective behavior analysis. We employ the correlation toolkit to extract audio features, and use a variety of state-of-the-art models to extract facial or emotional features for the visual modality.

2. To address the issue of large dimensional differences between features, we introduce an Affine Module that aligns the features to the same dimension. Moreover, we incorporate positional encoding to capture the contextual relationship of the sequence in our model.

3. We explored all four challenges of ABAW5, and achieve excellent performance with simple models and powerful features, which outperform all the baselines with a large margin.

2. Related Work

2.1. Multimodal Features

The utilization of multimodal features, including visual, audio, and text features, has been extensively employed in previous ABAW competitions [18, 19, 21–27, 55]. We can improve the performance in affective behavior analysis tasks by extracting and analyzing these multimodal features.

In the visual modality, the facial expression is an important aspect to understand and analyze emotions. In the Facial Action Coding System (FACS) [5] proposed by Friesen and Ekman in 1978, the human face is represented by a set of specific facial muscle movements known as Action Units (AUs) and it has been widely applied in studies of facial expressions [34]. With the development of deep learning, it has been found that visual features based on convolutional [41] and transformer [10, 50] networks can achieve better results.

In the context of affective computing, audio features, which typically include energy features, time-domain features, frequency-domain features, psychoacoustic features, and perceptual features, have been extensively utilized and shown to achieve promising performance in tasks such as expression classification and VA estimation [31, 46, 59]. These features can be extracted through pyAudioAnalysis [8], which is a Python library covering a wide range of audio analysis tasks. Similar to visual features, deep learning has also been widely used in acoustic feature extraction.

The text modality has been increasingly explored to address emotion recognition tasks. To enhance the effectiveness of text modality, Word2Vec [36] and GloVe [39] have been proposed and demonstrated to achieve superior performance.In ABAW3 competition, Zhang et al. [58] utilized a word embedding extractor to extract text features, achieving promising results.

In the previous editions of the ABAW competition, many teams utilized multimodal features [14, 15, 18, 35, 57, 58]. The model proposed by Meng et al. [35] leverages both audio and visual features, ultimately achieving first place in the VA track. To fully exploit the in-the-wild emotion information, Zhang et al. [58] utilizes the multimodal information from the images, audio and text and propose a unified multimodal framework for AU detection and expression recognition. The proposed framework achieved the highest scores on both tasks. These approaches convincingly demonstrate the effectiveness of multimodal features in affective behavior analysis tasks.

2.2. Multimodal Structure

In early studies, Zadeh et al. [54] and Pérez-Rosa et al. [40] employed concatenated multimodal features to train Support Vector Machine (SVM) models, which were inadequate in effectively modeling the multimodal information. Recent research on multimodal emotion analysis has mainly used deep learning models to model both intra-modal and inter-modal information interactions. Truong et al. [48] develops a neural network model called Visual Aspect Attention Network (VistaNet), which takes visual information as the alignment source at the sentence level. This multimodal structure enables the model to pay more attention to these sentences when classifying emotions. Currently, the use of Transformer for multimodal learning has become mainstream in multimodal algorithms. In the field of image-text matching, ALBEF [28] is to some extent inspired by the CLIP [42] model, introducing the idea of multimodal contrastive learning into multimodal models, achieving the unity of multimodal contrastive learning and multimodal fusion learning. In the previous ABAW competition, [16, 47, 53, 58] utilizes transformer structures and achieves outstanding performance.

2.3. Multimodal Fusion

Multimodal research places great importance on fusion, the process of combining information extracted from multiple unimodal data sources into a unified and compact multimodal representation. Fusion methods are typically categorized according to the different stages in which fusion occurs, including early fusion, late fusion, hybrid fusion and so on [56].

During early fusion, the features are directly combined into general feature vectors for analysis by the model. Nagrani et al. [38] directly inputs a sequence of visual and auditory patches into the transformer. This early fusion model allows attention to flow freely between different spatial and temporal regions of the image, as well as across frequency and time in the audio spectrogram.

In the late fusion , each modality's features are independently analyzed, and their final outputs are then fused to acquire a better prediction. Zhang et al. [58] propose a unified late fusion framework for both Action Unit (AU) detection and expression recognition, leveraging multimodal information from images, audio, and text. The late fusion framework effectively incorporates prior multimodal knowledge, enabling effective emotion analysis from different perspectives and leading to the championship in both the AU and Expression tracks of ABAW3 competition.

As a fusion of both, Hybrid fusion combines the advantages of both early fusion and late fusion. Li et al. [29] proposes a hybrid fusion method which leads them to the winner of MuSe-Reaction competition. In [29], audio features, facial expression features and paragraph-level text embeddings are fused at the feature level and then fed into the MMA module to extract complementary information from different modalities and calculate interactions between modalities.

3. Feature Extraction

3.1. Audio Features

IS09 The INTERSPEECH 2009(IS09) feature set was introduced at the INTERSPEECH 2009 Emotion Challenge [43], and it consists of 384 features that are derived from statistical functions applied to low-level descriptor contours. To extract these features, we utilized the openSMILE toolkit [44].

VGGish VGGish [13] is a pre-trained neural network by Google for extracting speech-related features from audio signals. Its output is a 128-dimensional feature vector that can be used for speech-related tasks.

eGeMAPS The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [61] is an extension of GeMAPS. The audio feature set in eGeMAPS is designed based on expert knowledge. eGeMAPS has only 88-dimensional features compared to traditional high-dimensional feature sets, but it shows higher robustness to speech emotion modeling problems.

DeepSpectrum DeepSpectrum [1] is a method of extracting deep spectrum features from audio file spectrograms using pre-trained convolutional neural networks (CNNs). These features are obtained by forwarding spectrograms through very deep task-independent pre-trained CNNs, and extracting the activations of fully connected layers as feature vectors. The dimension of the audio feature vectors is 1024.

CNN14 To obtain the high-level deep acoustic representations, a supervised model called PANNs [7] is utilized, which has been pre-trained on the AudioSet dataset [7]. PANNs comprises of multiple systems, and for this purpose, the CNN14 system that was trained using 16 kHz audio recordings is employed to generate a feature vector consisting of 2048 dimensions.

3.2. Visual Features

EAC Erasing Attention Consistency (EAC) [60] is a novel approach for addressing noisy samples during model training. The method leverages the flip semantic consistency of facial images to create an imbalanced framework, and randomly erases input images to prevent the model from overemphasizing specific features. The EAC method, based on the ResNet50, achieves a high accuracy rate of 90.35% on the RAF-DB [30] dataset, and the dimension of the resulting visual feature vector is 2048.

ResNet18 ResNet [12] is a deep learning architecture that addresses the vanishing gradient problem in deep neural networks by introducing residual connections. These connections create shortcuts across the network, allowing the input signal to bypass multiple layers and directly propagate to the deeper layers. This helps the network learn much deeper representations. In this study, we finetune the ResNet18 on AffectNet [37], which first pretained on the MS-Celeb-1M [9], and finally obtain a 512-dimensional visual feature vector.

POSTER The two-stream Pyramid crOss-fuSion TransformER network (POSTER) [62] is proposed to address the challenges of facial expression recognition. It effectively integrates facial landmark and direct image features using a transformer-based cross-fusion paradigm and employs a pyramid structure to ensure scale invariance. Extensive experimental results demonstrate that POSTER outperforms SOTA methods on RAF-DB with 92.05%, AffectNet (7 cls) with 67.31%, and AffectNet (8 cls) with 63.34%, respectively . The dimension of the visual feature vectors is 768.

POSTER2 The proposed POSTER2 [33] aims to improve upon the complex architecture of POSTER, which achieves state-of-the-art performance in facial expression recognition (FER) by combining facial landmark and image features through a two-stream pyramid cross-fusion design. POSTER2 reduces computational cost by using a window-based cross-attention mechanism, removing the image-to-landmark branch in the two-stream design, and combining images with landmark's multi-scale features. Experimental results show that POSTER2 achieves state-of-the-art FER performance with minimum computational cost on several standard datasets. For example, POSTER2 achieves 92.21% on RAF-DB, 67.49% on AffectNet (7 cls), and 63.77% on AffectNet (8 cls) using only 8.4G FLOPs and 43.7M Params. The same visual feature dimension as POSTER is 768.

FAU Facial Action Units (FAU), originally introduced by Ekman and Friesen [5], are strongly associated with the expression of emotions. Therefore, the detection of FAU has become a popular and promising method for predicting affect-related targets. We use the OpenFace [2] open source framework to extract FAU features and get a 17-dimensional feature vector.

4. Method

The 5th Competition on ABAW included a total of 4 challenges and we participated in all of them. Although the tasks for each challenge were not the same, our model for each challenge followed a basic framework. In detail, for the former three challenges, we refer to the design of the classical Transformer model [51], and for the fourth challenge, we adopt the Transformer Encoder with Multimodal Multi-Head Attention (TEMMA) [3] model.

The overall pipeline consists of four stages. Firstly, we use existing pre-trained models or toolkits to extract the visual and audio features corresponding to each frame in the videos. Secondly, each visual or audio feature sequence is input to the Affine Module to get features with the same dimension. Thirdly, these features are concat and then input to the Transformer Encoder to model the temporal relationships. Finally, the output of the encoder is input to the Output Layer to get the corresponding output. Figure 1 shows the overall framework of our proposed method.

The main notations used in this paper are listed as follows. Given a video, we can extract the visual features or audio features corresponding to all its video frames separately. We denote all the features by $f_1, f_2, ..., f_n$, where n is the number of features.

4.1. Affine Module

In our experiments, the inputs are one or several visual features or audio features, yet their dimensions are often different, even by a large margin. The dimensions of the features are shown in Table 1. We can see that the EAC feature have 2048 dimensions while FAU has only 17 dimensions. We think that too large a dimensional difference could diminish the effect of useful features. For this purpose, we design the Affine Module. For the former three challenges, we use Linear layers to affine the features of different dimension to the same dimension. Besides, following the setup of the classical Transformer [51], we add Position Encoding(PE) to each feature sequence for conveying its contextual temporal information. It can be formulated as follows:

Feature	Modality	Dimension
IS09	A	384
CNN14	A	2048
VGGish	A	128
eGeMAPS	A	88
DeepSpectrum	A	1024
EAC	V	2048
FAU	V	17
ResNet18	V	512
POSTER	V	768
POSTER2	V	768

Table 1. The dimensions of features.

$$\hat{f_i} = (W_A f_i + b_A) + PE \tag{1}$$

where W_A and b_A are learnable parameters.

For the fourth challenge, we use one-dimensional temporal convolution network to capture temporal information for each features. For the outputs, we also add the positional encoding.

4.2. Transformer Encoder

We model the temporal feature using the classical Transformer Encoder [51]. As shown in Equation 2 and Equation 3, we first concat the outputs of the affine module and next input them to the Transformer Encoder for former three challenge. As shown in Equation 4, we use TEMMA model to obtain the temporal feature for the fourth challenge. It can be formulated as follows:

$$\hat{f} = concat(\hat{f}_1, \hat{f}_2, ..., \hat{f}_n) \tag{2}$$

$$t = TransformerEncoder(\hat{f}) \tag{3}$$

$$t = TEMMA(\hat{f}) \tag{4}$$

where t is the temporal feature.

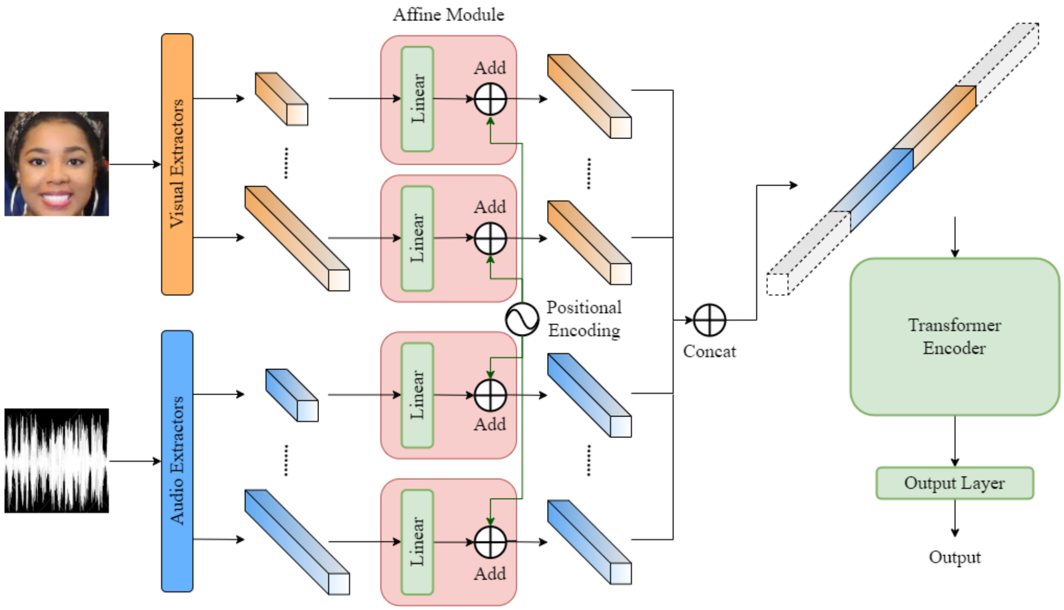

Figure 1. The overall framework of our proposed method. Visual Extractors contain EAC, ResNet18, POSTER, etc. Audio Extractors contain IS09, VGGish, eGeMAPS, etc. The design of the transformer encoder is consistent with [51].

4.3. Output Layer

After getting the temporal feature by Transformer Encoder, we input the feature t to the Output Layer. The Output Layer consists of fully-connected layer, which can be formulated as follows:

$$\hat{y} = Wt + b \tag{5}$$

which W and b are learnabel parameters. $\hat{y}$ is the predicton.

4.4. Loss Function

Since each task is different, we apply different loss function. Next, we introduce them separately.

4.4.1 Valence-Arousal(VA) Estimation Challenge

We utilize the Mean Squared Error (MSE) as the loss function for the VA challenge, which can be formulated as:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{6}$$

which y_i and $\hat{y}_i$ is the label and prediction of valence or arousal, N is the number of frames in a batch.

4.4.2 Expression (Expr) Classification Challenge

We utilize the Cross Entropy (CE) as the loss function for the Expr challenge, which can be formulated as:

$$L(y, \hat{y}) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij} \tag{7}$$

which y_{ij} and $\hat{y}_{ij}$ is the label and prediction of expression, N is the number of frames in a batch and $C = 8$ which denotes the number of expressions.

4.4.3 Action Unit (AU) Detection Challenge

We utilize the weighted asymmetric loss [32] as the loss function for the AU challenge, which can be formulated as:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} w_i[y_i \log \hat{y}_i + (1 - y_i)\hat{y}_i \log(1 - \hat{y}_i)] \tag{8}$$

which $\hat{y}_i$, y_i and w_i are the prediction (occurrence probability), ground truth and weight of the i^{th} AU. By the way, w_i is defined by the occurrence rate of the i^{th} AU in the whole training set.

4.4.4 Emotional Reaction Intensity (ERI) Estimation Challenge

We also utilize the Mean Squared Error (MSE) as the loss function for the ERI challenge, which can be formulated as:

$$L(y, \hat{y}) = \frac{1}{NC}\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij} - \hat{y}_{ij})^2 \tag{9}$$

which y_{ij} and $\hat{y}_{ij}$ is the label and prediction of expression, N is the number of frames in a batch and $C = 7$ which denotes the number of emotional reactions.

5. Experiments

5.1. Dataset

The 5th ABAW competition includes four challenges: 1) Valence-Arousal (VA) Estimation , 2) Expression (Expr) Classification, 3) Action Unit (AU) Detection, and 4) Emotional Reaction Intensity (ERI) Estimation. All challenges will accept only uni-task solutions. For the first challenge an augmented version of the Aff-Wild2 database is used. This database consists of 594 videos of around 3M frames of 584 subjects annotated in terms of valence and arousal. The second challenge involves 548 videos, while the third challenge involves 547 videos. Both challenges are based on Aff-Wild2, which is an audiovisual dataset containing approximately 2.7 million frames in total.

As for the fourth challenge, the Hume-Reaction dataset is used. The dataset is a multimodal collection of approximately 75 hours of video recordings capturing 2222 subjects from South Africa and the United States reacting to a variety of emotional video stimuli in their homes via webcam.

We use the RAF-DB and AffectNet datasets for pretraining visual feature extractors. The RAF-DB is a large-scale database which consists of approximately 30,000 facial images from thousands of individuals. Each image has been annotated independently about 40 times and then filtered using the EM algorithm to remove unreliable annotations. AffectNet dataset is a large facial expression recognition dataset containing over one million facial images collected from the internet. About half of the images have been manually annotated for seven discrete facial expressions (neutral, happy, sad, angry, fearful, surprised, and disgusted) and the intensity of valence and arousal present in the facial expression.

5.2. Experiment Setup

For the former three challenge, due to the limitation of GPU memory, we segment each video with segment length set to 256. The batch size is 128, the output dimension of the affine module is 512 or 256, the number of encoder layers is 4, and the number of attention headers is 4. The feed-forward and hidden layer dimensions are determined by the input dimension.

For the forth challenge, the number of convolutional layers is 5 and the kernel size is 3 in the input process block. The encoder blocks in the Multimodal encoder module is 4 and the number of heads in the multi-head attention layer is 4. For the inference module, the number of nodes in the last fully connected layer is 256 and the dropout is 0.2.

All the experiments are implemented with Pytorch. We adopt the Adam optimizer with the initial learning rate of 0.0001.

Features	Valence	Arousal	Mean
Baseline	0.24	0.20	0.22
EAC	0.4479	0.5878	0.5179
POSTER	0.3920	0.6317	0.5119
ResNet18	0.4762	0.5671	0.5217
POSTER2	0.5374	0.6297	0.5836
ResNet18+VGGish	0.4742	0.6220	0.5481
ResNet18+POSTER2	0.5515	0.6429	0.5972
ResNet18+POSTER2+FAU	0.4868	0.6301	0.5585
POSTER2+POSTER+VGGish	0.5003	**0.6946**	0.5975
EAC+ResNet18+POSTER2+VGGish	**0.5542**	0.6590	**0.6066**

Table 2. The results on the validation set of Valence-Arousal Estimation Challenge. Evaluation metric is Concordance Correlation Coefficient (CCC).

5.3. Experimental Results

5.3.1 Valence-Arousal(VA) Estimation Challenge

For the Valence-Arousal Estimation Challenge, Table 2 shows the results of using single feature or using multiple features at the same time on the validation set.

As can be seen in Table 2, for single features, POSTER2 performs best, and it has the best results in Valence, Arousal and Mean. It is obvious that for multiple features, the best combination of each value contains the feature POSTER2, which is a good indication of the effectiveness of the feature. In addition to this, we also find that the audio feature VGGish is useful for performance improvement.

5.3.2 Expression (Expr) Classification Challenge

For the Expression Classification Challenge, Table 3 shows the results of using single feature or using multiple features at the same time on the validation set.

For the Expression Classification Challenge, we can see that most of the features perform somewhat poorly in the classification task, both single and multiple features. The exception is the feature POSTER2, which extracts the expression information in the video quite well and achieves the best results with the F1 Score metric.

5.3.3 Action Unit (AU) Detection Challenge

For the Action Unit Detection Challenge, Table 4 shows the results of using single feature or using multiple features at the same time on the validation set.

For the Action Unit Detection Challenge, we can see that the performance of all features in the detection task is relatively close for both single and multi-features. It is obvious that FAU features can have some effect, while audio features are not effective. Our best result come from POSTER2 and FAU.

Features	F1
Baseline	0.23
EAC	0.3188
POSTER	0.3215
ResNet18	0.3176
POSTER2	**0.4055**
ResNet18+FAU	0.3287
POSTER2+POSTER	0.3630
ResNet18+POSTER2	0.3957
ResNet18+POSTER2+VGGish	0.3580
EAC+ResNet18+POSTER2+VGGish	0.3306

Table 3. The results on the validation set of Expression Classification Challenge.

Features	F1
Baseline	0.39
EAC	0.4881
POSTER	0.5046
ResNet18	0.5114
POSTER2	0.5181
ResNet18+FAU	0.5079
POSTER2+FAU	**0.5296**
POSTER2+POSTER	0.5112
ResNet18+POSTER2	0.5247
ResNet18+POSTER2+VGGish	0.5014
EAC+ResNet18+POSTER2+VGGish	0.4949

Table 4. The results on the validation set of Action Unit Detection Challenge.

5.3.4 Emotional Reaction Intensity (ERI) Estimation Challenge

For the Emotional Reaction Intensity Estimation Challenge, Table 5 shows the results of using single feature or using multiple features at the same time on the validation set.

The combination of ResNet18 features with low-level audio features such as eGeMAPS and CNN14 gives poorer results than the performance of ResNet18 features, suggesting that they are not effective. The combination of ResNet18 and DeepSpectrum yields better results. This shows that these two features can complement each other to provide more comprehensive expression information.

5.4. Test Set Performance

We use the combination of features that performed best on the validation set for each sub-challenge to test on the test set. As shown in Table 6, our methods also performs well on the test set. In total, we win 5th place in the VA sub-challenge, 4th place in the Expr sub-challenge, 9th place in the AU sub-challenge, and 5th place in the ERI sub-challenge.

Features	PCC
Baseline [4]	0.2382
ViPER [49]	0.3025
FaceRNET [20]	0.3590
Former-DFER+MLGCN [52]	0.3454
CNN14	0.1582
ResNet18	0.3893
eGeMAPS	0.0733
DeepSpectrum	0.1835
ResNet18+CNN14	0.3839
ResNet18+eGeMAPS	0.3809
ResNet18+DeepSpectrum	**0.3968**

Table 5. The results on the validation set of Emotional Reaction Intensity Estimation Challenge. Evaluation metric is Earson's Correlations Coefficient (PCC).

Challenges	Result	Evaluation metrics
VA	0.5342	CCC
Expr	0.3337	F1
AU	0.4752	F1
ERI	0.3879	PCC

Table 6. The results on the test set of the four sub-challenges.

6. Conclusion

In this paper, we present our solutions for the 5th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW), which includes four sub-challenges of Valence-Arousal (VA) Estimation Challenge, Expression (Expr) Classification Challenge, Action Unit (AU) Detection Challenge and Emotional Reaction Intensity (ERI) Estimation Challenge. Our approach leverages powerful audio and visual features, and employs an Affine Module to address potential issues resulting from differences in feature dimensions. We also integrate a Transformer encoder and TEMMA model to capture the semantic relationships between different features. Extensive experiments demonstrate that our method significantly outperforms the baseline and achieves excellent results in the four sub-challenges.

Acknowledgments

This work was supported by the National Key R&D Programme of China (2022YFC3803202), Major Project of Anhui Province under Grant 202203a05020011 and General Programmer of the National Natural Science Foundation of China (61976078).

References

[1] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy,

Alice Baird, and Björn Schuller. Snore sound classification using image-based deep spectrum features. 2017. 3

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 4

[3] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 2021. 4

[4] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, et al. The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 5–14, 2022. 7

[5] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2, 4

[6] Paul Ekman and Wallace V. Friesen. Facial action coding system. 2019. 1

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 3

[8] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12), 2015. 2

[9] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 3

[10] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2

[11] Albert Haque, Michelle Guo, Adam S. Miner, and Li Fei-Fei. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv: Computer Vision and Pattern Recognition*, 2018. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[13] 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

[14] 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. 2

[15] 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. 2

[16] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer. *arXiv preprint arXiv:2203.13472*, 2022. 3

[17] Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus pragmatics*, 2020. 1

[18] 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. 2

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

[20] Dimitrios Kollias, Andreas Psaroudakis, Anastasios Arsenos, and Paraskeui Theofilou. Facernet: a facial expression intensity estimation network. *arXiv preprint arXiv:2303.00180*, 2023. 7

[21] D Kollias, A Schulc, E Hajiyev, and S 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. 2

[22] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2

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

[24] 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. *arXiv preprint arXiv:2303.01498*, 2023. 2

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

[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. 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. 2

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi.

Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[29] Jia Li, Ziyang Zhang, Junjie Lang, Yueqi Jiang, Liuwei An, Peng Zou, Yangyang Xu, Sheng Gao, Jie Lin, Chunxiao Fan, et al. Hybrid multimodal feature extraction, mining and fusion for sentiment analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 81–88, 2022. 3

[30] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 3

[31] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10):1163, 2021. 2

[32] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 5

[33] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. Poster v2: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023. 4

[34] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017. 2

[35] 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

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 2

[37] 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

[38] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 3

[39] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[40] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, 2013. 2

[41] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[43] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. 2009. 3

[44] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013. 3

[45] Gulbadan Sikander and Shahzad Anwar. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 1

[46] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5688–5691. IEEE, 2011. 2

[47] Gauthier Tallec, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. Multi-label transformer for action unit detection. *arXiv preprint arXiv:2203.12531*, 2022. 3

[48] Quoc-Tuan Truong and Hady W Lauw. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 305–312, 2019. 2

[49] Lorenzo Vaiani, Moreno La Quatra, Luca Cagliero, and Paolo Garza. Viper: Video-based perceiver for emotion recognition. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 67–73, 2022. 7

[50] 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

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems*, 2017. 4, 5

[52] Kexin Wang, Zheng Lian, Licai Sun, Bin Liu, Jianhua Tao, and Yin Fan. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 75–80, 2022. 7

[53] Lingfeng Wang, Shisen Wang, and Jin Qi. Multi-modal multi-label facial action unit detection with transformer. *arXiv preprint arXiv:2203.13301*, 2022. 3

[54] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 2

[55] 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

[56] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020. 3

[57] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2376–2381, 2022. 2

[58] 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, 3

[59] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775. IEEE, 2018. 2

[60] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 418–434. Springer, 2022. 3

[61] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pages 65–72, 2018. 3

[62] Ce Zheng, Matias Mendieta, and Chen Chen. Poster: A pyramid cross-fusion transformer network for facial expression recognition. *arXiv preprint arXiv:2204.04083*, 2022. 3