Transformer-based Multimodal Information Fusion for Facial Expression Analysis

Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang,
Rudong An, Bowen Ma, Yu Ding‡
Virtual Human Group, Netease Fuxi AI Lab
{zhangwei05, qiufeng, wangsuzhen, zenghao03, zhangzhimeng,
anrudong, mabowen01, dingyu01}@corp.netease.com

Abstract

Human affective behavior analysis has received much attention in human-computer interaction (HCI). In this paper, we introduce our submission to the CVPR 2022 Competition on Affective Behavior Analysis in-the-wild (ABAW). To fully exploit affective knowledge from multiple views, we utilize the multimodal features of spoken words, speech prosody, and facial expression, which are extracted from the video clips in the Aff-Wild2 dataset. Based on these features, we propose a unified transformer-based multimodal framework for Action Unit detection and also expression recognition. Specifically, the static vision feature is first encoded from the current frame image. At the same time, we clip its adjacent frames by a sliding window and extract three kinds of multimodal features from the sequence of images, audio, and text. Then, we introduce a transformer-based fusion module that integrates the static vision features and the dynamic multimodal features. The cross-attention module in the fusion module makes the output integrated features focus on the crucial parts that facilitate the downstream detection tasks. We also leverage some data balancing techniques, data augmentation techniques, and postprocessing methods to further improve the model performance. In the official test of ABAW3 Competition, our model ranks first in the EXPR and AU tracks. The extensive quantitative evaluations, as well as ablation studies on the Aff-Wild2 dataset, prove the effectiveness of our proposed method.

1. Introduction

Human affective behavior analysis plays a significant role in human-computer interaction. The basic expressions, Action Units (AU), and Valence-Arousal (VA), as three commonly-used expression representations, are usually used to infer human emotions. AU explains human ex-
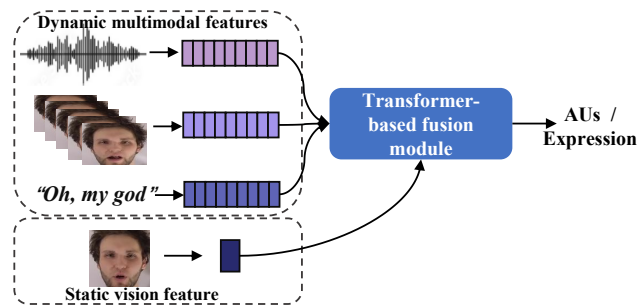
*Corresponding Author.

Figure 1. The sketch of our proposed transformer-based multimodal information fusion framework for AU detection and expression recognition. The transformer-based fusion module is used to integrate the static vision feature and dynamic multimodal features.

pression from the view of facial muscle movement based on the Facial Action Coding System (FACS) [10]. The basic expression uses discrete and explicit definitions to represent human expressions. Valence and Arousal are two continuous values used to describe human emotion states. However, most existing datasets [16, 30, 53, 54] contain only one of the three representations. Different from them, *Aff-Wild* [20,21,23,49] and *Aff-Wild2* [19,22,24–26] containing all the three representation labels have received considerable attention from both academic and industrial communities. Also, the videos in the *Aff-Wild2* [19, 22, 24–26] show human spontaneous affective behaviors in the wild, pushing the affective analysis to fit with the real-world scenarios.

Considering the fact that visual expression, spoken words, and speech prosody also imply rich emotion information, we propose a transformer-based multimodal information fusion framework for AU detection and expression recognition (See Fig. 1). First, a static vision feature extractor is used to capture the vision expression feature of the static image. Then we employ the sliding windows to obtain the sequence of adjacent frames, audio, and text. Based

on the three kinds of multimodal signals, we capture the corresponding prior features by some pretrained models or open-sourced tools. In particular, we employ the pretrained DLN [52] to extract the expression embedding feature of the cropped faces sequence. The sequence of audio features, extracted by the open-sourced librosa tool [29], consists of MFCC, pitch, and short-time energy. Also, we obtain the spoken words by ASR tools [34] and produce their word embedding based on the pretrained Bert [9].

After obtaining these features, the dynamic uni-modal features combined with the static vision feature are fed into a transformed-based fusion module. The static vision feature is injected into the multi-head cross attention module as K and V for attention coefficient computation. In this way, we can effectively integrate the static vision features and dynamic multimodal features. The integrated multimodal features are forced to pay more attention to key point parts for the current frame detection task. In addition, to alleviate the serious imbalance in the expression recognition trainset, we resample the training samples based on the number of categories and video frames. Remix [5] data augmentation is also used to relieve the imbalance. Considering that the expression labels seldom change rapidly in continuous sequence frames, we propose the filter strategy to smooth the final prediction results.

In sum, the contributions of this work are three-fold:

- To fully exploit the in-the-wild emotion information, we utilize the multimodal information from the images, audio and text and propose a unified multimodal framework for AU detection and expression recognition. With the valid prior multimodal information, our model realizes the effective affective analysis from different views.

- We introduce a transformer-based fusion module for integrating the static vision feature and dynamic multimodal features. With the multi-head cross-attention, the output integrated features capture the crucial information for affective analysis. In the ABAW3 competition, our method ranks first in both AU and expression tracks. The final test set score and quantitative experiments can prove the superiority of our method.

2. Related works

In this section, we will introduce some related works about the facial expression representations, AU detection, and expression recognition.

2.1. Facial expression representations

The main challenges in the 3rd Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW) are about the commonly-used facial expression represen-

tations: AU, basic expression categories, and Valence-Arousal. Ekman *et al.* [10] first introduces the definition of the AUs from the Facial Action Coding System (FACS), which contains 32 atomic facial action descriptors based on the movements of the facial muscles. The basic expression categories describe human affection by seven discrete semantic definitions, namely Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral. Valence-Arousal describes the emotion state of humans by two continuous values. Valence indicates how positive or negative is the human affective behavior and Arousal indicates how active or passive is the human affective behavior. The values of Valence and Arousal are between -1 and 1. Except for the above representations, recently a compact and continuous expression embedding [41] is proposed to represent fine-grained human expressions. The main idea of it is to reserve the expression similarity on a low dimensional space.

2.2. AU detection

In the area of AU detection, one of the main challenges is the bad generalization caused by the limited identities in the commonly-used AU datasets. Therefore, numerous works focus on introducing more additional information to enhance the detection task. To facilitate the model with local features, multi-task methods usually combine AU detection with landmark detection [2, 36] or landmark-based attention map prediction [11]. SEV-Net [46] proposes to utilize the textual descriptions of local details to generate a regional attention map.

In ABAW2 competition, multi-task [14, 39, 42, 51] that combines the tasks of expression recognition or VA detection is often employed to add underlying features for AU task. Zhang *et al.* [51] pretrain the backbone on the DLN [52] and propose a streaming multi-task framework, winning first place in AU and EXPR tracks. Jin *et al.* [14] with multi-task framework also introduce the multimodal information from audio and visual signals, winning second place in AU and EXPR tracks.

In ABAW3 competition, Jiang *et al.* [13] pretrain the feature extractor on Glint360K and some private commercial datasets before conducting the AU detection task in the ABAW3. They achieve second place in the AU track. Savchenko *et al.* [35] also pretrain the backbone on large facial expression datasets and finetune it for AU detection or expression recognition. They rank 5-th in the AU track and 4-th in the EXPR track. Transformer [40] is also often used. Wang *et al.* [44] proposed a two-stream transformer-based framework to model the inherent relationships of AUs. Tallec *et al.* [38] and Nguyen *et al.* [31] use the transformer-based architecture for the multi-label AU detection, taking advantage of the self-attention and cross-attention modules to learn the local features for each action unit. Multimodal information like audio is also introduced in AU task [44].
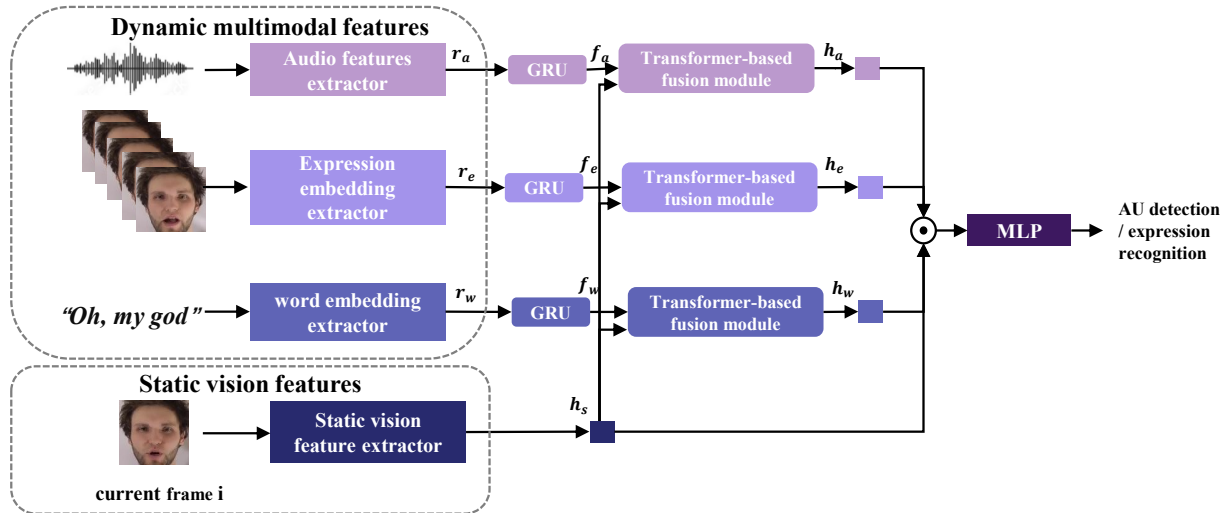
Figure 2. The pipeline of our proposed multimodal framework for AU detection and expression recognition. The static vision feature extractor captures the static vision expression features h_s from the current frame. Three dynamic multimodal feature extractors with the multimodal sequence information as input focus on providing emotion features from different views. The transformer-based fusion module is used to incorporate the static and dynamic multimodal features. The output features from three transformer-based fusion modules concatenated with h_s are sent into the multilayer perceptron (MLP) for detection tasks.

2.3. Expression recognition

Expression recognition aims to identify the basic expression categories of humans, such as anger, disgust, fear, happiness, sadness, surprise, etc. Recently, benefiting from the development of the neural network, deep-learning-based approaches are applied to end-to-end expression recognition. Some works [3, 15] employ CNN to extract the spatial and temporal features and improve expression recognition. Some other works [6, 17] utilize LSTM to model the temporal dependencies and characteristics within consecutive facial images. To focus on the most relevant features, some recent works introduce the attention mechanism into CNN [4, 27, 47, 48].

In the ABAW2 Competition, the winner in EXPR track [51] leverages an identity-invaraint expression feature and develops a multi-task streaming network to extract interaction relationships among three modal representations. Thinh *et al.* [39] takes advantage of the multi-task learning technique by combining the knowledge for two correlated tasks, AUs prediction and emotion classification. Deng *et al.* [8] is dedicated to solving the problem of emotion uncertainty and proposes a framework that predicts both emotions labels and the estimated uncertainty. Then, this work improves emotion recognition and uncertainty estimation by the self-distillation algorithm. Jin *et al.* [14] use both AU and expression annotations to train the model and apply a sequence model to further extract associations between frames in sequences. Wang *et al.* [43] aims to tackle the problem of incomplete labeled datasets in multi-task affective behavior recognition methods. This approach trains a semi-supervised model as the teacher network to predict pseudo labels for unlabeled data. Then these pseudo labels are utilized for student network training, which allows it to learn from unlabeled data.

In the ABAW3 Competition, other teams also propose effective frameworks. Jeong *et al.* [12] use the affinity loss to train the feature extractor from the images. And a multi-head attention network in an orchestrated fashion is proposed to extract diverse attention for expression recognition. To distinguish the similar expressions, Xue *et al.* [45] propose a two-stage framework, named CFC networks, to separately predict negative expressions and the other kinds of expressions. Savchenko *et al.* [35] develop a real-time framework through a lightweight EfficientNet model, which may be even implemented for video analytics on mobile devices. Phan *et al.* [32] employs the pre-trained model RegNet [33] as the backbone and then introduces Multi-Head Attention and Transformer Encoder to generate sequence representations. Kim *et al.* [18] develop a three-stream network consisting of a visual stream, a temporal stream, and an audio stream, based on Swin transformer [28] as the backbone.

3. Method

The architecture of the proposed transformer-based multimodal framework is shown in Fig. 2. The whole framework consists of a static vision feature extractor, three dynamic multimodal feature extractors and a transformer-
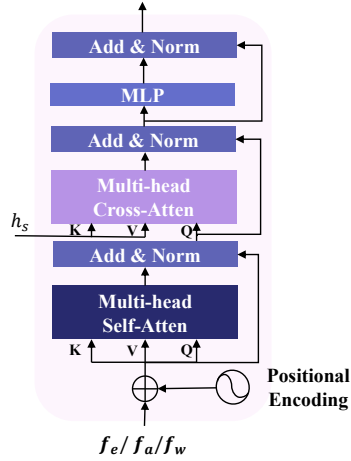
Figure 3. The structure of the transformer-based fusion module. Different from the original transformer decoder, its second multi-head cross attention uses the static vision feature h_s as K and V.

based fusion module. The static vision feature extractor captures the static vision expression features from the current frame. Also, the sequences of the cropped face frames, audio, and spoken words from the two seconds around the current frame are sent into the corresponding modal feature extractor to provide more dynamic emotion information that facilitates AU detection and expression recognition. Then, the proposed transformer-based fusion module is responsible for integrating the static vision feature and the dynamic multimodal information. The output of each modal fusion module and the static vision feature are concatenated for the final recognition task. Besides, to alleviate the imbalanced problem, we use the resample strategy and remix [5] data augmentation in the training process. A filter strategy is also utilized for smoothing the predictions.

3.1. Multimodal feature extraction

It is commonly acknowledged that human usually conveys and perceives emotion state from vision, audio and text. Based on this observation, it is a straightforward idea to combine multimodal features for full emotion information collection. The proposed framework takes advantage of the three modal information and extracts the corresponding original features from some pretrained models or open-sourced tools.

Specifically, the static vision feature extractor with the structure of Inception-Resnet [37] is used to capture the vision features h_s of the current frame for its affective analysis. To obtain an effective parameter initialization, we make our static vision feature extractor pretrained on the expression embedding task [52] by the distillation. It is proven that the expression embedding can represent the complex and fine-grained human expressions [52]. It is of great value to

expression-related tasks like AU detection and expression recognition.

Besides, we consider the dynamic multimodal information. We choose M adjacent frames before and after the current frame and compute their expression embedding features by the pretrained DLN network [52]. In our experiments, we set M to 60. The output expression embedding sequence is represented as r_e. Likely, we extract the MFCC, pitch, and short-time energy features from the audio during these frames and concatenate them as audio features sequence r_a. We choose these kinds of raw audio features due to their robustness. Also, the spoken words during these frames can be obtained by the open-source ASR tools [34] and they are encoded by pretrained Bert model [9]. r_w is the output sequence of the word embedding extractor.

3.2. Transformer-based fusion module

To effectively integrate the static and dynamic multimodal features, we propose a transformer-based fusion module that makes different modal sequence features focus on the current frame detection task. More concretely, r_e, r_a and r_w from the dynamic multimodal feature extractors are first encoded by 4 GRU [7] for capturing their individual dynamic features h_e, h_a and h_w.

Then, we choose the decoder structure of transformer [40] as our fusion module (See Fig. 3) to integrate the static vision feature h_s and other dynamic modal features. As shown in Fig. 3, different from the original transformer decoder, the K and V of the second multi-head attention are replaced by the h_s. As a result, the output of the second multi-head attention can be followed:

$$Attention(Q, K, V) = Attention(h'_{(.)}, h_s, h_s)$$
$$= softmax(\frac{h'_{(.)} h_s^T}{\sqrt{d_k}}) h_s \qquad (1)$$

Where d_k is the dimension of h_s, $h'_{(.)}$ is each modal input feature of the second multi-head cross attention module. The correlation between the current frame static vision information and the uni-modal dynamic feature can be built by the attention computation. The output integrated features can focus on the key point parts for the current frame detection task.

3.3. Training

Except for the framework structure, some network details and training tricks are introduced in this part.

Resampling. Due to the serious imbalance of the expression recognition training set, we resample the training samples from the original dataset. We observe that both category imbalance and identity imbalance exist in the training

dataset. If a video contains more than N frames of a particular class, we randomly sample N frames of that class from the video. We set N to 200 and 50 for the minority classes (Anger, Disgust, Fear, Sadness, and Surprise) and the majority classes (Neutral, Happiness, and Other) respectively. In this way, we can avoid the model overfitting to the specific identities and the majority classes.

Data augmentation. Except for the commonly-used data augmentation operations like rotation, color jitter, center crop, Flip, etc, we take advantages of remix [5] that considers the imbalanced categories based on the original mixup [50]. It controls the soft labels of mixed samples more biased towards the minority classes. The process can be followed:

$$x_{new} = \lambda x_i + (1 - \lambda)x_j$$
$$y_{new} = \lambda_y y_i + (1 - \lambda_y)y_j \tag{2}$$

In particular, the input of the model (x_{new}) is the mixed sample that is linearly combined from an arbitrary sample pair (x_i and x_j) by a mixing factor λ. λ is sampled from the beta distribution. Different from mixup [50], λ_y is related to the labels of the sample pair. The rule for λ_y is:

$$\lambda_y = \begin{cases} 0 & n_i/n_j \geq \kappa \ and \ \lambda < \tau \\ 1 & n_i/n_j \leq 1/\kappa \ and \ 1 - \lambda < \tau \\ \lambda & otherwise \end{cases} \tag{3}$$

where n_i and n_j is the number of samples with the label of y_i and y_j. κ and τ are two hyper-parameters and are usually set as 3 and 0.5. In this way, when one mixed sample is from one minority class and one majority class, the mixed label will be dominated by the minority class.

Post-process. It can be observed that expression labels seldom rapidly change in a continuous sequence of frames. Therefore, we use the filter operation to smooth the model prediction results, removing some glitches in the temporal sequences. In particular, for each hopping point in the sequences of predictions, we first calculate the number of consecutive frames with the same prediction labels. If the number of consecutive frames is less than the threshold, we consider these points as outliers and count the most common values in adjacent frames to correct the hopping points. In addition, due to the data imbalance, we use a lower threshold and smaller adjacent window length for the minority classes in Aff-Wild2 [19] dataset.

Loss function. For AU detection, we utilize the cross-entropy loss function $\mathcal{L}_{AU\_CE}$ and circle loss function $\mathcal{L}_{AU\_Circle}$ like the work [51] for training. For expression recognition, we use the cross-entropy loss $\mathcal{L}_{EXPR\_CE}$ for training. In AU detection task, the direct output of the MLP is $\mathcal{S} = \{s_1, s_2, ..., s_{12}\} \in \mathbb{R}^{12}$ without scaling. The AU probability $\hat{\mathcal{Y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_{12}\}$ can be computed by sigmoid activation function for the output $\mathcal{S}$, and the ground-truth binary AU label is $\mathcal{Y} = \{y_1, y_2, ..., y_{12}\} \in \mathbb{B}^{12}, \mathbb{B} =$

$\{0, 1\}$, where 1 denotes the corresponding action unit is activated and vice versa. The loss functions for AU task can be formulated as:

$$\mathcal{L}_{AU\_CE} = -\frac{1}{12} \sum_{j=1}^{12} [y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)]. \tag{4}$$

$$\mathcal{L}_{AU\_Circle} = \log(1 + \sum_{i \in \Omega_0} e^{s_i}) + \log(1 + \sum_{j \in \Omega_1} e^{-s_j}),$$
$$\Omega_0 = \{ i \mid if \ y_i = 0 \},$$
$$\Omega_1 = \{ j \mid if \ y_j = 1 \}. \tag{5}$$

In EXPR task, the softmax output of the MLP is $\hat{\mathcal{Z}} = \{\hat{z}_1, \hat{z}_2, ..., \hat{z}_8\}$, and the one-hot ground-truth EXPR label is $\mathcal{Z} = \{z_1, z_2, ..., z_8\} \in \mathbb{C}^8, \mathbb{C} = \{0, 1\}$. The loss function for EXPR task can be formulated as:

$$\mathcal{L}_{EXPR\_CE} = -\frac{1}{8} \sum_{j=1}^{8} z_j \log \hat{z}_j. \tag{6}$$

4. Experiments

In this section, we will first introduce our used datasets and the implementation details. Then we evaluate our model on the ABAW3 competition metrics. To prove the effectiveness of each module in our framework, We also present several ablation studies.

4.1. Datasets

The 3rd Workshop and Competition on Affective Behavior Analysis in-the-wild provides the Aff-wild2 datasets as the official datasets. It contains 564 in-the-wild videos and around 2.8M frames from Youtube. Most of them show human spontaneous affective behaviors. In detail, there are 548 videos in Aff-wild2 [19] annotated with 12 AUs, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, and AU26. The ABAW3 competition provides 295 of them as the trainset, 105 of them as the validation set, and 141 of them as the test set. Aff-wild2 [19] contains 548 videos annotated with discrete expression categories, namely neutral, happiness, anger, disgust, fear, sadness and surprise, and other. Different from the ABAW2 [26], the new added other class means the expression state is different from the other seven basic expressions. The ABAW3 competition provides 248 of them as the trainset, 70 of them as the validation set, and 228 of them as the test set. Aff-wild2 [19] contains 567 videos annotated with valence and arousal. The ABAW3 competition provides 341 of them as the trainset, 71 of them as the validation set, and 152 of them as the test set. ABAW3 creates

Val Set	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU23	AU24	AU25	AU26	Avg.
Official	55.3	**48.9**	56.7	**62.8**	74.4	75.5	73.6	28.1	10.5	20.8	83.9	39.1	52.5
fold-1	60.2	55.1	58.6	65.3	74.8	**77.8**	71.8	24.2	**13.8**	26.3	**87.7**	**39.6**	**54.6**
fold-2	**66.9**	54.3	**66.1**	62.8	78.9	**71.8**	73.6	**15.2**	**7.80**	31.6	85.4	38.4	54.4
fold-3	59.1	51.9	59.3	67.8	**79.3**	73.6	**74.5**	27.1	12.9	**16.2**	84.9	35.4	53.5
fold-4	**54.7**	49.9	55.2	**69.6**	73.1	74.4	72.9	**30.2**	10.8	21.5	**82.0**	**30.9**	**52.1**
fold-5	62.3	**57.4**	**54.2**	64.5	76.2	76.1	**72.3**	28.7	11.8	19.4	86.8	33.7	53.6

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). The highest and lowest scores are both indicated in bold.

Val Set	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Other	Avg.
Official	57.6	34.7	14.0	20.0	53.3	**43.0**	37.9	**54.7**	39.4
fold-1	54.6	**39.1**	10.0	23.3	**50.6**	34.7	35.3	55.9	37.9
fold-2	61.4	**8.10**	31.7	38.7	60.2	33.4	40.2	55.1	**41.1**
fold-3	**65.8**	29.9	19.8	22.0	51.7	23.6	**27.8**	**61.9**	37.8
fold-4	**54.2**	33.9	6.0	14.1	**60.3**	21.9	48.5	59.6	37.3
fold-5	56.4	25.7	**5.0**	**8.1**	60.0	29.7	41.8	62.3	**36.1**

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). The highest and lowest scores are both indicated in bold.

a static version of Aff-wild2 [19] s-Aff-wild2 that is used in the Multi-task learning challenge. The ABAW3 competition provides 145,273 frames as the trainset, 27,087 frames as the validation set, and 51,245 frames as the test set. The official use F1 score as the metric for AU detection and expression recognition and Concordance Correlation Coefficient as the metric for VA detection. We use the official trainset for training and evaluate our model on the official validation set in the AU and EXPR tracks. Final scores of ABAW3 competition are computed on the unseen official test set.

4.2. Implementation details

We use OpenCV to process all videos in the *Aff-Wild2* dataset [24] into frames and crop all facial images into 224×224 scale by OpenFace [1] detector. The static vision feature extractor is initialized by the pretrained expression embedding model distilled from DLN [52]. The audio feature extraction relies on the open-sourced tool librosa [29]. We obtain the spoken words by the open-sourced ASR tool [34] and extract their word embedding by the pretrained Bert [9] model. Our training process is implemented based on PyTorch. The training procedure runs about 20 epochs on an NVIDIA RTX 3090 graphics card with a learning rate of 0.002 and batch size of 80. We use a stochastic gradient (SGD) optimizer with a cosine annealing warm restart learning rate scheduler.

4.3. Results on the validation set

The official ABAW3 provides the official training/validation/test set based on the videos. To relieve the disturbance caused by different dataset divisions, we also perform the 5-fold random cross-validation. Tab. 1 and Tab. 2 presents the AU detection results of the official validation set and our 5-fold validation sets. From the above table, it can be observed that the different dataset division results in the fluctuating average F1 score, especially for the expression recognition task. The AU average F1 score ranges from 52.1% to 54.6% while the expression average F1 score ranges from 36.1% to 41.1%. The reason is that the trainset contains limited identities, especially for the minority classes. To alleviate the bad influence, we make the final decisions for the test set prediction by voting on these models using different dataset divisions.

4.4. Results on the test set

We present the final results on the official test set for AU detection and expression recognition task in the Tab. 5 and Tab. 6. In particular, our method achieves 49.9% F1 score in AU track and 35.9% F1 score in EXPR track, winning the first prizes in the AU and EXPR tracks.

Observing these methods, Transformer structure is usually used. Tallec *et al.* [38] and Nguyen *et al.* [31] use it on the vision feature for AU detection. Phan *et al.* [32] use it for expression recognition. Wang *et al.* [44] use the transformer encoder structure to model the relationships between different AUs. Also, it is practical to use extra datasets for

Condition	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU23	AU24	AU25	AU26	Avg.
only static	54.0	44.9	53.1	62.9	73.1	74.1	74.3	26.7	13.0	16.8	82.2	35.5	50.9
w/o exp_emb	53.3	42.8	55.3	62.3	75.0	75.2	**74.8**	28.2	11.1	19.8	83.8	32.7	51.2
w/o audio	55.2	49.6	**57.4**	**63.7**	75.1	75.1	74.1	15.0	4.20	21.6	83.7	**40.0**	51.2
w/o word	**56.9**	**49.7**	55.0	62.8	**75.5**	72.3	73.1	15.3	7.70	**28.6**	74.6	37.9	50.8
w/o trans	53.5	42.2	54.6	62.0	73.8	74.0	73.3	**29.9**	**16.5**	24.0	82.9	33.1	51.7
Ours	55.3	48.9	56.7	62.8	74.4	**75.5**	73.6	28.1	10.5	20.8	**83.9**	39.1	**52.5**

Table 3. Ablation study results of different modality features and transformer-based fusion module for AU detection task. All scores are computed based on the official validation set.The highest is indicated in bold.

Condition	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Other	Avg.
only static	62.1	27.7	2.32	3.98	49.1	15.3	30.1	45.4	29.5
w/o exp_emb	62.7	**45.6**	10.3	11.5	51.7	17.9	30.3	**57.4**	35.9
w/o audio	**65.3**	41.9	5.06	10.7	54.4	32.0	30.2	50.6	36.3
w/o word	60.8	45.1	2.20	18.8	48.3	23.3	25.8	36.4	32.6
w/o trans	64.2	33.6	3.00	**23.3**	**55.0**	28.4	31.0	52.5	36.4
Ours	57.6	34.7	**14.0**	20.0	53.3	**43.0**	**37.9**	54.7	**39.4**

Table 4. Ablation study results of different modality features and transformer-based fusion module for Expression recognition task. All scores are computed based on the official validation set. The highest is indicated in bold.

Method	AU Avg F1 score
Tallec *et al.* [38]	44.3
Savchenko *et al.* [35]	47.3
Wang *et al.* [44]	48.8
Nguyen *et al.* [31]	49.0
Jiang *et al.* [13]	49.8
our method	**49.9**

Table 5. The AU F1 scores (in %) of different models on the official Aff-wild2 test set. The highest is indicated in bold.

Method	EXPR Avg F1 score
Kim *et al.* [18]	27.2
Phan *et al.* [32]	28.6
Savchenko *et al.* [35]	30.3
Xue *et al.* [45]	32.2
Jeong *et al.* [12]	33.8
our method	**35.9**

Table 6. The expression F1 scores (in %) of different models on the official Aff-wild2 test set. The highest is indicated in bold.

pretraining [13, 35]. Xue *et al.* [45] propose a Coarse-to-Fine network to improve the prediction for the hard negative expression classes.

Different from them, we use more valid multimodal features. Also, our transformer decoder structure is for multimodal feature fusion.

4.5. Ablation study

To prove the effectiveness of the multimodal features and our transformer-based fusion module, we conduct ablation studies by comparing the models trained without the corresponding features or modules. We present the quantitative results for AU detection and expression recognition in Tab. 3 and Tab. 4. As shown in the tables, each modality and our fusion module can facilitate AU detection and expression recognition.

Multimodal features. To indicate the effectiveness of the dynamic multimodal features, we compare our complete model with the model that only uses static vision features (*only static*). From the Tab. 3 and Tab. 4, it can be observed that multimodal features bring about a distinctive improvement for both AU and expression tasks. Next, we will give a detailed analysis of each modality's features.

To verify the benefits of the expression embedding features, we conduct an ablation study by removing the expression embedding extractor (*w/o exp_emb*). From the Tab. 3 and Tab. 4, it can be observed that the expression embedding feature plays an important role in AU and expression tasks. Without it, AU average F1 score falls from 52.5% to 51.2% and the expression average F1 score falls from 39.4% to 35.9%.

To prove the effectiveness of the word embedding features, we compare our complete model with the model without the word embedding extractor (*w/o word*). As shown in Tab. 3 and Tab. 4, without word feature, AU average F1 score falls from 52.5% to 50.8% and expression average F1

score falls from 39.4% to 32.6%. It can be observed that word embedding features play a more important role in expression recognition than AU detection.

Likely, we conduct the ablation study by removing the audio features extractor (*w/o audio*) to prove the effectiveness of the MFCC, pitch, and energy features. As shown in Tab. 3 and Tab. 4, without audio features, AU average F1 score falls from 52.5% to 51.2% and expression average F1 score falls from 39.4% to 36.3%. It can be indicated that audio information can facilitate the AU and expression tasks.

Transformer-based fusion module. We perform an ablation study by replacing the transformer-based fusion module with concatenation to evaluate its significance. In the Tab. 3 and Tab. 4, without the fusion module, AU average F1 score falls from 52.5% to 51.7% and expression average F1 score falls from 39.4% to 36.4%. This proves that the transformer-based fusion module promotes the model performance effectively. In addition, the rising degree of F1 score brought by the fusion module is slightly less than multimodal features overall. This indicates that multimodal features are more important.

5. Conclusion

In this paper, we introduce our transformer-based multimodal information fusion framework for AU detection and expression recognition in the ABAW3 Competition. We propose to exploit the static vision feature and three kinds of dynamic multimodal features to fully collect the human emotion clues. Besides, to integrate the static features and dynamic multimodal features, we utilize the transformer decoder structure. In participating in the ABAW3 competition, we won the first prizes in the AU track and EXPR track. The competition results prove the superiority of our method. Also, the quantities ablation studies indicate that each multimodal feature and fusion module can improve the model performance for affective tasks.

References

[1] 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. 6

[2] Carlos Fabian Benitez-Quiroz, Yan Wang, and Aleix M Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, pages 3990–3999, 2017. 2

[3] Ran Breuer and Ron Kimmel. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*, 2017. 3

[4] Junyan Cheng, Iordanis Fostiropoulos, Barry Boehm, and Mohammad Soleymani. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2447–2458, 2021. 3

[5] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *European Conference on Computer Vision*, pages 95–110. Springer, 2020. 2, 4, 5

[6] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017. 3

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 4

[8] Didan Deng, Liang Wu, and Bertram E Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3557–3566, 2021. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 4, 6

[10] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1, 2

[11] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

[12] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Yuchul Jung, and Jin-Woo Jeong. Facial expression recognition based on multi-head cross attention network, 2022. 3, 7

[13] Wenqiang Jiang, Yannan Wu, Fengsheng Qiao, Liyu Meng, Yuanyuan Deng, and Chuanhe Liu. Facial action unit recognition with multi-models ensembling, 2022. 2, 7

[14] 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, 3

[15] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015. 3

[16] Miyuki Kamachi, Michael Lyons, and Jiro Gyoba. The japanese female facial expression (jaffe) database. *Availble: http://www. kasrl. org/jaffe. html*, 01 1997. 1

[17] Dae Hoe Kim, Wissam J Baddar, Jinhyeok Jang, and Yong Man Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10(2):223–236, 2017. 3

[18] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer, 2022. 3, 7

[19] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1, 5, 6

[20] 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. 1

[21] 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. 1

[22] 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. 1

[23] 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. 1

[24] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1, 6

[25] 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

[26] 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, 5

[27] Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In *Conference on Empirical Methods in Natural Language Processing*, pages 150–161, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 3

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[29] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015. 2, 6

[30] 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. 1

[31] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial expression analysis in video, 2022. 2, 6, 7

[32] Kim Ngan Phan, Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. Expression classification using concatenation of deep neural network for the 3rd abaw3 competition, 2022. 3, 6, 7

[33] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 3

[34] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624. 2, 4, 6

[35] Andrey V. Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices, 2022. 2, 3, 7

[36] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, 2021. 2

[37] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 4

[38] Gauthier Tallec, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. Multi-label transformer for action unit detection, 2022. 2, 6, 7

[39] Phan Tran Dac Thinh, Hoang Manh Hung, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. Emotion recognition with incomplete labels using modified multi-task learning technique. *arXiv preprint arXiv:2107.04192*, 2021. 2, 3

[40] 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, 4

[41] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5683–5692, 2019. 2

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

[43] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis. *arXiv preprint arXiv:2107.04225*, 2021. 3

[44] Lingfeng Wang, Shisen Wang, and Jin Qi. Multi-modal multi-label facial action unit detection with transformer, 2022. 2, 6, 7

[45] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. 3, 7

[46] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2021. 2

[47] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[48] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3

[49] 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

[50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[51] 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*, pages 3539–3549, 2021. 2, 3, 5

[52] 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, 4, 6

[53] 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. 1

[54] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016. 1