Multi-modal Facial Affective Analysis based on Masked Autoencoder

Wei Zhang, Bowen Ma, Feng Qiu, Yu Ding[*]
Virtual Human Group, Netease Fuxi AI Lab
{zhangwei05,mabowen01,qiufeng,dingyu01}@corp.netease.com

Abstract

Human affective behavior analysis focuses on analyzing human expressions or other behaviors to enhance the understanding of human psychology. The CVPR 2023 Competition on Affective Behavior Analysis in-the-wild (ABAW) is dedicated to providing high-quality and large-scale Aff-wild2 for the recognition of commonly used emotion representations, such as Action Units (AU), basic expression categories (EXPR), and Valence-Arousal (VA). The competition is committed to making significant strides in improving the accuracy and practicality of affective analysis research in real-world scenarios. In this paper, we introduce our submission to the CVPR 2023: ABAW5. Our approach involves several key components. First, we utilize the visual information from a Masked Autoencoder (MAE) model that has been pre-trained on a large-scale face image dataset in a self-supervised manner. Next, we finetune the MAE encoder on the image frames from the Aff-wild2 for AU, EXPR and VA tasks, which can be regarded as a static and uni-modal training. Additionally, we leverage the multi-modal and temporal information from the videos and implement a transformer-based framework to fuse the multi-modal features. Our approach achieves impressive results in the ABAW5 competition, with an average F1 score of 55.49% and 41.21% in the AU and EXPR tracks, respectively, and an average CCC of 0.6372 in the VA track. Our approach ranks first in the EXPR and AU tracks, and second in the VA track. Extensive quantitative experiments and ablation studies demonstrate the effectiveness of our proposed method.

1. Introduction

In recent years, there has been a growing interest in the research of human affective behavior analysis due to its potential to provide a more accurate understanding of human emotions, which can be applied to design more friendly human-computer interaction. The commonly used human

---

[*]Corresponding Author.

expression representations include Action Unit (AU), basic expression categories (EXPR), and Valence-Arousal (VA). Specifically, AU is first proposed by Paul Ekman and Wallace Friesen in the 1970s [12]. It depicts the local regional movement of faces which can be used as the smallest unit to describe the expression. Basic expression categories divide expressions into a limited number of groups according to the emotion categories, e.g., happiness, sadness, etc. VA contains two continuous values Valence (V) and Arousal (A), which are ranged between -1 and 1. They can be used to describe the human emotional state. V represents the degree of positivity or negativity of emotion; A describes the level of intensity or activation of emotion.

The fifth Competition on Affective Behavior Analysis in-the-wild (ABAW5) [26] is organized to focus on handling the obstacles in the process of human affective behavior analysis. It makes great efforts to construct large-scale multi-modal video datasets Aff-wild [24, 27, 65] and Aff-wild2 [22, 23, 25, 28–30]. The proposal of these datasets has greatly promoted the development of facial expression analysis in the wild and accelerated the practical implementation of related industries. Aff-wild2 contains 598 videos and most of them have the three kinds of frame-wise annotated labels: AU, basic expression categories and VA. ABAW5 proposes three challenges of detecting these three kinds of expression representations.

In this paper, we introduce our submission to the ABAW5. First of all, we pre-train a Masked Autoencoder (MAE) [13, 36] on a large-scale facial dataset in a self-supervised manner. Then, we choose the MAE encoder as our visual feature extractor to capture the visual features of the faces. Due to the extensive quantity of faces included in the dataset, the features extracted with the MAE encoder have strong generalization capabilities. We also finetune the MAE encoder for the specific tasks of AU detection, EXPR classification, and VA estimation. This training process only use the static and vision modal data. To further exploit temporal and multi-modal information, we design a Temporal and Multi-modal Fusion (TMF) that divides the videos into several short clips and performed clip-wise training on the downstream tasks mentioned above. Dur-

ing this process, we utilized the finetuned MAE encoder to extract visual features from each frame, while also incorporating pre-trained audio models such as Hubert [53], Wav2vec2 [3], and Vggish [14] to capture acoustic features. The visual and acoustic features are concatenated and fed into a Transformer structure to capture temporal information for the downstream tasks. Additionally, we proposed several effective post-processing policies aimed at smoothing predictions and further enhancing model performance.

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

- Our approach employs a highly effective MAE feature extractor to capture visual features. The MAE model is pre-trained on a large-scale facial image dataset and exhibits remarkable generalization ability for diverse downstream face-related tasks.

- Our approach employs a Temporal and Multi-modal Fusion (TMF) to leverage the benefits of temporal and multi-modal information. With a strategic selection of optimal vision and audio features, our approach further improves the performance of the model.

- In the ABAW5 competition, our approach ranks first in both AU and EXPR tracks and ranks second in VA track. The final test set score and quantitative experiments can prove the superiority of our method.

2. Related Works

In this section, we introduce some recent works on the relevant tasks in the CVPR2023: ABAW5 competition, including AU detection, expression recognition and VA estimation in the wild. We also briefly introduce the self-supervised learning method in facial affective analysis.

2.1. AU Detection

For AU detection in the wild, there exist some challenges regarding the limited identity information and interference of diversity poses, illumination or occlusions. These disturbances restrain the model generation and cause the overfitting to the noise information. Several studies propose the use of a multi-task framework to incorporate additional auxiliary information as regularization, which introduces extra label constraints. Specifically, Zhang et al. [68] proposed a streaming model that simultaneously performs AU detection, expression recognition, and VA regression. Similarly, Jin et al. [18] and Thinh et al. [52] combine the tasks of AU detection with expression recognition. JAA-Net [49] performs landmarks detection and AU detection at the same time.

Another effective approach to enhance the model generalization is to utilize the related tasks' pre-trained backbones. Jiang et al. [17] use IResnet100 [11] that pre-trains on Glint360K [1] and some private commercial datasets

before conducting the AU detection task in the ABAW3. Savchenko et al. [45] utilize the EfficientNet [51] that pretrained on the face recognition task as the backbone. Zhang et al. [68, 71] introduce the pre-trained expression embedding model as the backbone and win the first prizes in ABAW2 and ABAW3.

Multi-modal information is also involved in ABAW competitions. Zhang et al. [71] capture three modalities of information - vision, acoustic, and text - and fused them using a transformer decoder structure. Jin et al. [19] extract the vision features from IResNet100 and the audio features from Mel Spectrogram. They also use the transformer structure for the fusion of multi-modal features.

2.2. Expression Recognition

The goal of expression recognition is to classify an input image into one of the basic emotion classes, such as happiness or sadness. Similar approaches to exploit the extra information regularization are mentioned in Sec. 2.1. Zhang et al. [68] utilize the prior expression embedding model and propose a multi-task framework. Phan et al. [42] employ the pre-trained model RegNet [43] as the backbone and add the Transformer [54] structure to extract the temporal information. Kim *et al.* [21] use Swin transformer [33] as the backbone and exploit the extra auxiliary from the audio modal. Wang et al. [57] propose a semi-supervised framework to predict pseudo-labels for unlabeled data, which helps improve the model's generalization to some extent. Xue *et al.* [62] develop a CFC network that uses different branches to train the easy-distinguished and hard-distinguished emotion categories.

2.3. VA Estimation

For VA estimation, several studies [7, 55, 58, 68] leverage the correlation between VA and AU or VA and EXP, proposing multi-task frameworks. This enables these methods to extract supplementary information from other tasks, particularly for data without VA labels but possessing AU or EXPR labels. Moreover, many works [20, 39, 48, 64, 66, 70] propose multi-modal frameworks, which leverage hidden features from vision, audio, or text. The Transformer structure is also commonly used in VA tasks for feature fusion. Several works [32, 39, 70, 74] utilize it for this purpose.

2.4. Self-supervised Learning in Affective Analysis

It has been pointed out [65] that annotating the corresponding emotion/AU/VA labels from real-world facial images costs a large amount of time/labor, which limits the development of the affective analysis. It is a potential solution to make use of the self-supervised learning (SSL) method to exploit more knowledge from the existing large-scale unlabelled data. There have been several works to develop SSL methods to enhance the accuracy in the area

of affective analysis. Shu et al. [50] explore different strategies in the choice of positives and negatives to enforce the expression-related features and reduce the interference of other facial attributes. They improve the accuracy of expression recognition based on the contrastive SSL methods (e.g. SimCLR [6]). Ma et al. [36] pre-train the Masked Autoencoder (MAE) structure on the large-scale face images and finetune it on the AU detection, which achieves the state-of-the-art performance on the BP4D [72] and DISFA [38]. In the ABAW5 competition, Zhang et al. [70], Liu et al. [32], and Wang et al. [60] all employ the pre-trained MAE to extract vision features and secure a position among the top few performers.

3. Method

This section will present our method proposed for AU detection, expression recognition and VA estimation in ABAW5 competition. The pipeline can be seen in Fig. 1. It consists of two steps: MAE pre-train and Temporal and Multi-modal Fusion (TMF). We first introduce the MAE structure with good generalization to extract the vision features. Subsequently, we employ the TMF to integrate the temporal acoustic features with the temporal visual features to extract more effective information for completing the task. Finally, we adopt some smoothing policies to generate the final predictions.

3.1. MAE Pre-training

Different from the traditional MAE, our MAE is pre-trained on the facial image dataset to focus on learning the facial vision features. We construct a large-scale facial image dataset that contains images from the existing facial image datasets, e.g., AffectNet [40], CASIA-WebFace [63], CelebA [34] and IMDB-WIKI [44]. The total number of images is about 2.17M. Then we pre-train the MAE model on the dataset in a self-supervised manner. Specifically, our MAE consists of a ViT-Base encoder and a ViT decoder based on the structure of Vision Transformer (ViT) [10]. The MAE pre-training procedure follows a masking-then-reconstruct method, whereby images are first divided into a series of patches (16x16) and 75% of them are randomly masked. These masked images are sent to the MAE encoder and the complete images should be reconstructed by the MAE decoder (See Fig. 1 left Step1). The loss function of MAE pre-training is the pixel-wise L2 loss to make the reconstructed images close to the target images.

Once the self-supervised learning is complete, we remove the MAE decoder and replace it with a fully connected layer attached to the MAE encoder (See Fig. 1 left Step2). This allows us to fine-tune downstream tasks: AU detection, expression recognition, and VA estimation on the Aff-wild2 dataset. It is important to note that this process is based on frame-wise training, without taking into account

temporal or other modal information. The corresponding loss functions for these three tasks are as follows:

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

(1)

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

(2)

$$\mathcal{L}_{VA\_CCC} = 1 - CCC(\hat{v}_{batch_i}, v_{batch_i}) \\ +1 - CCC(\hat{a}_{batch_i}, a_{batch_i})$$

(3)

$$CCC(\mathcal{X}, \hat{\mathcal{X}}) = \frac{2\rho_{\mathcal{X}\hat{\mathcal{X}}} \delta_{\mathcal{X}} \delta_{\hat{\mathcal{X}}}}{\delta_{\mathcal{X}}^2 + \delta_{\hat{\mathcal{X}}}^2 + (\mu_{\mathcal{X}} - \mu_{\hat{\mathcal{X}}})^2}.$$

(4)

where $\hat{y}$, $\hat{z}$, $\hat{v}$ and $\hat{a}$ denote the model's predictions for AU, expression category, Valence, and Arousal, respectively. The symbols without hats refer to the ground truth. $\delta_{\mathcal{X}}, \delta_{\hat{\mathcal{X}}}$ indicate the standard deviations of $\mathcal{X}$ and $\hat{\mathcal{X}}$, respectively. $\mu_{\mathcal{X}}$ and $\mu_{\hat{\mathcal{X}}}$ are the corresponding means and $\rho_{\mathcal{X}\hat{\mathcal{X}}}$ is the correlation coefficient. For the AU and EXPR tasks, we utilize weighted cross-entropy as the loss function. The weights for different categories, represented by W_{au_j} and W_{exp_j}, are inversely proportional to the class number in the training set.

3.2. Temporal and Multi-modal Fusion

To further exploit the temporal and multi-modal features for AU, EXPR and VA tasks, we design the sequence-based model which combines the audio features. To concretely, we first divide the videos into several short clips, each having an equivalent frame number of K. We construct the Temporal and Multi-modal Fusion (TMF) to perform the sequence-wise training which can be seen in Fig. 1 right.

Given a video clip C_i and the corresponding audio clips A_i, we use the finetuned MAE encoder and some existing pre-trained audio feature extractor models (e.g. Hubert [53], Wav2vec2 [3], vggish [14].) to extract the vision and acoustic features F_{vis}^i and F_{aud}^i for each frame separately. Then we concatenate F_{vis}^i and F_{aud}^i and sent them into a Transformer [54] encoder structure to exploit the temporal correlations between them. The Transformer encoder comprises of four encoder layers with a dropout ratio of 0.3. The output of the Transformer encoder is then directed towards a fully connected layer to resize the final output size, which is tailored to fit various tasks. In the training process, the parameters of MAE encoder and audio feature extractor are fixed without training. During the computation of loss, we flatten the sequence result of a clip and use the same loss function as Equ. 1, 2, 3.
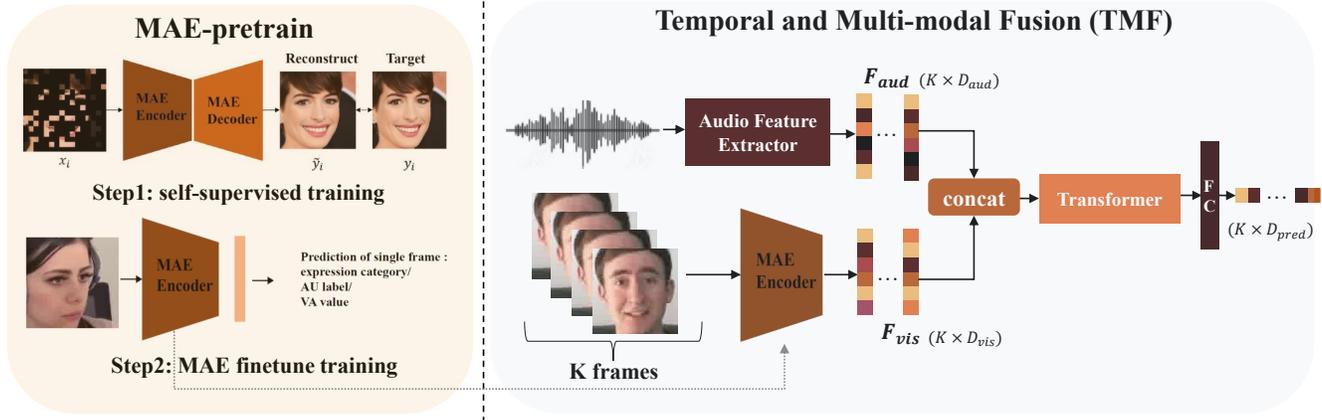
Figure 1. Illustration of our pipeline. In the first stage, we pre-train a MAE in SSL manner. The input face image is randomly masked 75% and MAE learns to reconstruct the missing pixels. Then, the MAE encoder is fine-tuned on the Aff-wild2 for AU detection, EXPR classification, and VA estimation. In the second stage, we propose a transformer-based structure TMF that utilizes temporal vision and audio information to further facilitate the challenge tasks.

Since the final prediction is made frame-by-frame, applying a smoothing policy can enhance the stability of the predictions. We find that the cropped face images from videos occasionally miss some frames potentially due to the frames not being able to detect faces. Therefore, to maintain continuity, we use the nearest frame to replace the lost frame and produce the prediction for each frame. Then, we use the Gaussian filter to smooth the probability for AU or EXPR and VA values. The sigma of Gaussian filter is set according to different tasks. The detailed setting can be seen in the experiment section.

4. Experiment

4.1. Experimental Setting

We pre-processed all videos in the Aff-Wild2 datasets into frames by OpenCV and employ the OpenFace [4] detector to crop all facial images into 224×224 scale. We find that OpenFace incorrectly detects some images and there are no faces in them. We use the NetEase Fuxi Youling crowdsourcing platform to check and remove incorrect images. We pre-train MAE on large-scale facial image datasets for 800 epochs using the AdamW [35] optimizer. We set the batch size as 4096 and the learning rate as 0.0024. Our training process is implemented based on PyTorch and trained on 8 NVIDIA A30 GPUs. In the MAE fine-tuning phase, we set the batch size to 512 and the learning rate to 0.0001, while using AdamW as the optimizer. During TMF training, we set the video clip length to 100, with a batch size of 32 and a learning rate of 0.0001. The TMF training process was completed in approximately 20 epochs using the AdamW optimizer. All the experiments are carried out with the Fuxi Youling Platform that is based Agent-Oriented Programming (AOP) in order to facilitate

task modeling.

Metrics. For AU detection and expression classification, we calculate the F1-Score (F1) for each class to evaluate the prediction results. For VA estimation, we calculate the Concordance Correlation Coefficient (CCC) for valence and arousal respectively. The definition of CCC can be seen equ. 4. The specific definitions for each challenge are as follows:

$$S_{AU} = \frac{1}{N_{au}} \sum F1_{au_i} \qquad (5)$$

$$S_{EXP} = \frac{1}{N_{exp}} \sum F1_{exp_i} \qquad (6)$$

$$S_{VA} = 0.5 * (CCC(\hat{v}, v) + CCC(\hat{a}, a)) \qquad (7)$$

4.2. Results on Validation set

Action Unit Detection Challenge. For AU detection, We evaluate the performance of our model using the F1 scores in equ. 5. To improve the generalization capability of our model, we also perform 5-fold cross-validation based on randomly split videos from the existing labeled data. We present the F1 scores of each AU and their average F1 in Tab. 1. From the table, we find that the performances of AU4, AU15, and AU24 in different data splits have a large variance (AU4: $53.40\% \sim 71.84\%$, AU15: $27.78\% \sim 44.44\%$, AU24: $17.34\% \sim 28.51\%$). The reason for this could be that the number of these AUs is relatively low, which means the learning for these AUs is not sufficient and heavily relies on the data splits.

Expression Classification Challenge. For the EXPR challenge, we also perform 5-fold cross-validation based on randomly selected video clips from existing labeled data.

Val Set	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU23	AU24	AU25	AU26	Avg.
Official	55.26	51.35	56.70	67.25	75.75	75.11	75.82	31.21	17.34	**33.77**	83.90	42.26	55.86
fold-1	**64.94**	52.31	**71.84**	74.59	69.10	69.92	74.05	35.38	**28.51**	21.56	77.97	42.14	56.86
fold-2	63.90	**53.86**	71.59	**75.72**	**77.15**	**76.26**	**79.39**	27.78	28.28	23.86	**86.20**	40.56	**58.71**
fold-3	61.80	51.10	60.85	70.70	73.08	75.15	74.84	29.79	23.71	27.35	79.22	**48.33**	56.33
fold-4	51.59	39.27	53.40	64.69	66.73	71.83	72.80	**44.44**	23.87	27.60	76.70	39.12	52.67
fold-5	52.69	44.59	60.90	70.67	70.11	73.25	76.04	43.84	23.86	10.30	76.96	39.96	53.60

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	65.11	44.61	48.91	18.85	56.46	60.95	32.20	64.32	48.93
fold-1	61.92	34.22	46.96	16.96	53.65	77.77	30.08	50.12	46.46
fold-2	74.60	16.22	41.49	**70.22**	**65.26**	57.42	**40.29**	60.31	53.23
fold-3	62.16	36.00	21.99	19.40	60.12	65.68	32.38	**72.16**	46.24
fold-4	63.67	**57.74**	**56.09**	15.09	64.97	**88.98**	25.18	72.01	**55.47**
fold-5	**77.54**	25.86	13.12	19.50	57.36	42.75	29.51	50.57	39.53

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.

Val Set	Valence	Arousal	Avg.
Official	0.4643	0.6407	0.5525
fold-1	0.5927	0.6542	0.6234
fold-2	0.5647	0.6267	0.5957
fold-3	0.5679	0.6959	0.6319
fold-4	0.5567	0.6456	0.6011
fold-5	**0.6478**	**0.7056**	**0.6767**

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

Team	Test Set	
	Rank	F1-score
PRL [56]	#5	51.01
SZFaceU [60]	#4	51.28
USTC-IAT-United [64]	#3	51.44
SituTech [32]	#2	54.22
Ours	#1	**55.49**

Table 4. Final competition results (average F1 score in %) on the AU test set of ABAW5.

We show our experimental results on the official and 5-fold validation sets in Tab. 2. We evaluate the model by the average F1 metric in equ. **??**. From the table, we also find the model performances for Anger, Disgust, and Fear are relatively unstable (Anger: 16.22% ∼ 57.74%, Disgust: 13.12% ∼ 56.09%, Fear: 15.09% ∼ 70.22%). At the same time, these three categories have the lowest frequency of occurrence in the dataset, which leads to their poor performances.

Valence-Arousal Estimation Challenge. For VA estimation, We evaluate the model by the CCC of Valence and Arousal in equ. 4. To enhance the model generalization, we also perform the 5-fold cross-validation according to random video split in the existing labeled data. The experimental results can be seen in Tab. 3.

4.3. Results on Test Set

Action Unit Detection Challenge In ABAW5 competition, we need to predict the labels of the official test set. Our proposed method achieves an average F1 score of 55.49% for AU detection, thereby winning first place in this track (See Tab. 4). The second-place team SituTech and the fourth-place team SZFaceU in this track also leverage MAE features. Similar to our approach, they use a large-scale facial image dataset for MAE pre-training and use it as the feature extractor. Apart from this, SituTech incorporates additional visual features from DenseNet [16], IRes-Net100 [11], and MobileNet [15], pre-trained on the expression recognition task based on AffectNet [40], FER+ [5], and RAF-DB [31]. In contrast, we fine-tune MAE on Aff-Wild2 and incorporate it into our multi-modal information training framework. SZFaceU proposes a Spatial-Temporal Graph Learning module to construct a graph representation for spatial-Temporal features. The third-place team pro-

Team	Test Set	
	Rank	F1-score
HSE-NN-SberAI [48]	#5	32.92
HFUT-MAC [73]	#4	33.37
CtyunAI [74]	#3	35.32
SituTech [32]	#2	40.72
Ours	#1	**41.21**

Table 5. Final competition results (average F1 score in %) on the EXPR test set of ABAW5.

Team	Test Set	
	Rank	CCC
HFUT-MAC [73]	#5	0.5342
CtyunAI [74]	#4	0.5666
CBCR [67]	#3	0.5913
SituTech [32]	#1	**0.6414**
Ours	#2	0.6372

Table 6. Final competition results (average CCC) on the VA test set of ABAW5.

poses to use LANet [59] to extract the local features for AU. They model the correlations between AUs through Graph Neural Networks. PRL utilizes the Video Vision Transformer [2] to capture the temporal expression movements in the video.

Expression Classification Challenge In Tab. 5, our approach achieves an average F1 score of 41.21% in EXPR track, winning first place in this track. Our approach, as well as the second-place team (SituTech), the third-place team (CtyunAI) incorporate multi-modal information from both vision and acoustic modalities. Besides the above-mentioned vision features, SituTech uses the audio features of Hubert [53], wav2vec [3] and ECAPA-TDNN [9]. CtyunAI use three kinds of vision features: arcface [8], EfficientNet-b2 [47] and DAN [61] and two kinds of audio features: wav2vec2 [3] and wav2vec2-emotion [41]. HFUT-MAC uses POSTER2 [37] as the feature extractor and employs an affine module to align features of varying dimensions to a uniform dimension. Subsequently, a transformer is utilized to integrate the temporal features. HSE-NN-SberAI applies EfficientNet CNN from HSEmotion library [46] to extract effective prior expression knowledge and use MLP for classification.

VA Estimation Challenge In Table 6, our approach achieves an average CCC score of 0.6372 in the VA track, securing second place. Notably, our results are highly competitive with the first-place team's score of 0.6414. After extracting three modalities of information: vision, acoustic and text, CBCR use TCN to capture the temporal features and channel attention network (CAN) for features fusion. The methods of other teams are roughly similar to the analysis of other challenges.

Method	AU	EXPR	VA
w/o. MAE finetune	51.04	41.34	0.5169
w/o TMF	55.27	46.79	0.5483
Ours	**55.86**	**48.93**	**0.5525**

Table 7. Ablation studies that discuss the significance of MAE fine-tune and Temporal and Multi-modal Fusion. The evaluation metrics used for AU, EXPR, and VA are average F1 (%), average F1 (%), and CCC, respectively.

4.4. Ablation Studies

In this section, we conduct several experiments to discuss the significance of each module of our approach, including MAE finetuning training, Temporal and Multi-modal Fusion (TMF), the selection of features and the policies of post-processing. All the experiments are perform on the official training and validation set.

MAE finetuning training. To prove the effectiveness of our procedure of finetuning MAE on Aff-wild2 (See Fig. 1 (left: step 2)), we conduct the experiment that directly uses the initial MAE features to operate the TMF training. The result can be seen in Tab. 7 (w/o. MAE finetune). It can be found that there is a significant drop in the performance metrics across all three tracks. The average F1 score for AU decreases from 55.86% to 51.04%, and for EXPR it decreased from 48.93% to 41.34%. The average CCC of VA also decreases from 0.5525 to 0.5169. This illustrates that MAE finetuning can effectively exploit the static vision features based on a single image, which provides valuable prior knowledge for learning the temporal vision feature.

Temporal and Multi-modal Fusion. To illustrate the significance of TMF, we conduct an experiment that removes the TMF and uses the model trained on the single frame for evaluation. The results are presented in Tab. 7 (w/o TMF). It is evident that the inclusion of TMF leads to an improvement in AU F1, EXPR F1, and VA CCC scores by 0.59%, 2.14% and 0.0042, respectively. This proves that the temporal and multi-modal features can further exploit more hidden clues for these three tasks. Notably, TMF is particularly effective for EXPR compared to other challenges.

The choice of multi-modal features. In this section, we analyze the impact of using different multi-modal features, as presented in Tab. 8. Specifically, we experiment with two types of vision features: expression embedding from DLN [69] and the aforementioned MAE features. Our results demonstrated that using MAE features alone has a more significant advantage in all three challenges. Hence, for multi-modal information fusion, we only utilized the features extracted by MAE. In terms of audio features, we try three kinds of features: Hubert [53], wav2vec2 [3] and Vggish [14]. From Tab. 8, we observe that the utilization of audio features does not enhance the performance of the

Vis_fea	Audio_fea	AU	EXPR	VA
MAE	-	**55.86**	47.73	0.5483
DLN	-	50.49	39.12	0.4789
MAE + DLN	-	54.76	44.69	0.5169
MAE	Hubert	54.89	48.03	0.5405
MAE	Wav2vec2	53.14	47.91	0.5364
MAE	Vggish	51.68	47.71	0.5157
MAE	Hubert+Vggish	52.28	**48.93**	**0.5525**
MAE	Hubert+wav2vec2	52.55	48.18	0.5471
MAE	Hubert+wav2vec2+Vggish	53.54	47.89	0.5521

Table 8. Ablation studies that discuss the significance of different selection of vision and audio features. The evaluation metrics used for AU, EXPR, and VA are average F1 (%), average F1 (%), and CCC, respectively.

Policy	AU	EXP	VA
w/o. smooth	55.86	48.93	0.5525
gaussian_filter	**56.01**	**49.16**	**0.5640**
median_filter	55.91	48.98	0.5537
average_filter	55.95	49.08	0.5595

Table 9. Ablation studies that discuss the influence of different smooth policies. The evaluation metrics used for AU, EXPR, and VA are average F1 (%), average F1 (%), and CCC, respectively.

AU task. However, the incorporation of audio features enhances the performance of the EXPR and VA challenges. For EXPR and VA tracks, our experiments show that the best performance is achieved by combining Hubert and Vggish features.

Smooth Policy Due to the frame-by-frame prediction of the challenges, smooth policies can effectively eliminate some noisy predictions and enhance prediction stability. In this part, we discuss the different policies we used for smoothing in Tab. 9. From the table, we can find that the metrics have varying degrees of improvement by using smooth policies. We use Scipy library to realize these smooth approaches. In different challenges, there exist slight differences in hyper-parameter settings. For instance, we set the sliding window size of median_filter and average_filter in AU, EXP and VA tracks as 10, 25 and 50, respectively. We set the sigma of gaussian_filter in AU, EXP and VA tracks as 5, 25 and 25, respectively. This is because human facial physical movements of expressions tend to change frequently over a short period of time, whereas emotional states often exhibit a small range of change. Therefore, AU changes are more sensitive compared to EXP and VA. Our final prediction for the official test set combines some of the mentioned smooth policies.

5. Conclusion

This paper introduces our submission to the ABAW5 competition for the tasks of AU detection, expression recognition, and VA estimation. Our approach begins by pre-training a MAE in a self-supervised manner, using a large-scale facial image dataset. This enables the MAE to learn a variety of general features associated with human faces. Subsequently, we finetune the MAE using static images from Aff-wild2 dataset. Then we propose Temporal and Multi-modal Fusion (TMF) to exploit the multi-modal information from the vision and audio temporal features. In participating in the ABAW5 competition, we won the first prizes in the AU track and EXPR track and second prize in the VA track. The quantities ablation studies indicate that each module and procedure of our approach can improve the model performance for affective tasks.

Acknowledgments

References

[1] Xiang An, Jiangkang Deng, Jia Guo, Ziyong Feng, Xuhan Zhu, Yang Jing, and Liu Tongliang. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 6

[3] 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. 2, 3, 6

[4] 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

[5] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016. 5

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning

of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Didan Deng, Liang Wu, and Bertram E. Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition, 2021. 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[9] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*. ISCA, oct 2020. 6

[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. 3

[11] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE, 2021. 2, 5

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

[13] 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. 1

[14] 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. 2, 3, 6

[15] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 5

[16] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 5

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

[18] 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

[19] 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

[20] Vincent Karas, Mani Kumar Tellamekala, Adria Mallol-Ragolta, Michel Valstar, and Björn W. Schuller. Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. 2022. 2

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

[23] 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

[24] 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

[25] 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

[26] 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. 1

[27] 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

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

[29] 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

[30] 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

[31] 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. 5

[32] Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, and Wenqiang Jiang. Multi-modal expression recognition with ensemble method, 2023. 2, 3, 5, 6

[33] 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. 2

[34] 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

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[36] Bowen Ma, Rudong An, Wei Zhang, Yu Ding, Zeng Zhao, Rongsheng Zhang, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Facial action unit detection and intensity estimation from self-supervised representation. *arXiv preprint arXiv:2210.15878*, 2022. 1, 3

[37] 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. 6

[38] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 3

[39] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, and Qin Jin. Multimodal emotion estimation for in-the-wild videos, 2022. 2

[40] 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

[41] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021. 6

[42] 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. 2

[43] 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. 2

[44] 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

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

[46] Andrey V Savchenko. Hsemotion: High-speed emotion recognition library. *Software Impacts*, 14:100433, 2022. 6

[47] Andrey V Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2359–2366, 2022. 6

[48] Andrey V. Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition, 2023. 2, 6

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

[50] Yuxuan Shu, Xiao Gu, Guang-Zhong Yang, and Benny Lo. Revisiting self-supervised contrastive learning for facial expression recognition. *arXiv preprint arXiv:2210.03853*, 2022. 3

[51] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[52] 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

[53] 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. 2, 3, 6

[54] 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, 3

[55] Manh Tu Vu and Marie Beurton-Aimar. Multitask multi-database emotion recognition, 2021. 2

[56] Tu Vu, Van Thong Huynh, and Soo Hyung Kim. Vision transformer for action units detection, 2023. 5

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

[58] Lingfeng Wang, Shisen Wang, Jin Qi, and Kenji Suzuki. A multi-task mean teacher for semi-supervised facial affective behavior analysis, 2021. 2

[59] Qiangchang Wang and Guodong Guo. Ls-cnn: Characterizing local patches at multiple scales for face recognition. *IEEE Transactions on Information Forensics and Security*, 15:1640–1653, 2019. 6

[60] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, shiling Wu, Weicheng Xie, and Linlin Shen. Spatio-temporal au relational graph representation learning for facial action units detection, 2023. 3, 5

[61] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021. 6

[62] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. 2

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

[64] Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Local region perception and relationship learning combined with feature fusion for facial action unit detection, 2023. 2, 5

[65] 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, 2

[66] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3, 2022. 2

[67] Su Zhang, Ziyuan Zhao, and Cuntai Guan. Multimodal continuous emotion recognition: A technical report for abaw5, 2023. 6

[68] 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

[69] 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. 6

[70] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition, 2023. 2, 3

[71] 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

[72] 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. 3

[73] Ziyang Zhang, Liuwei An, Zishun Cui, Ao xu, Tengteng Dong, Yueqi Jiang, Jingyi Shi, Xin Liu, Xiao Sun, and Meng Wang. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge, 2023. 6

[74] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer, 2023. 2, 6