Prior Aided Streaming Network for Multi-task Affective Analysis

Wei Zhang*, Zunhu Guo*, Keyu Chen*, Lincheng Li, Zhimeng Zhang, Yu Ding†
Runze Wu, Tangjie Lv, Changjie Fan

Virtual Human Group, Netease Fuxi AI Lab

Abstract

Automatic affective recognition has been an important research topic in the human-computer interaction (HCI) area. With the recent development of deep learning techniques and large-scale in-the-wild annotated datasets, facial emotion analysis is now aimed at challenges in real world settings. In this paper, we introduce our submission to the 2nd Affective Behavior Analysis in-the-wild (ABAW2) Competition. In dealing with different emotion representations, including Categorical Expression (EXPR), Action Units (AU), and Valence Arousal (VA), we propose a multi-task streaming network by a heuristic that the three representations are intrinsically associated with each other. Besides, we leverage an advanced facial expression embedding model as prior knowledge, which is capable of capturing identity-invariant expression features while preserving the expression similarities, to aid the down-streaming recognition tasks. In order to enhance the generalization ability of our model, we generate reliable pseudo labels for unsupervised training and adopt external datasets for fine-tuning. In the official test of ABAW2 Competition, our method ranks first in the EXPR and AU tracks and second in the VA track. The extensive quantitative evaluations, as well as ablation studies on the Aff-Wild2 dataset, prove the effectiveness of our proposed method.

1. Introduction

Recognizing and analyzing facial affective statements from human behaviors is a long-standing problem in the intersection area of the computer science and psychology community. An ideal human-computer interaction system is expected to capture the vivid human emotions, mostly conveyed by facial performances, and to react respectively. Because of the diverse environments and varying contexts where emotions occur, the perception of facial effectiveness is always natural to our human beings but never straightforward to the artificial intelligent machines. Thanks to the continuous research of psychology and rapid development of deep learning methods, especially recent published large scale in-the-wild annotated datasets e.g., *Aff-Wild* [1, 2] and *Aff-Wild2* [3], the automatic affective recognition approaches are now pushed to meet the real-world requirements.

Different from most existed facial emotion datasets [4, 5, 6] that contain only one of the three commonly used emotional representations: Categorical Expression (EXPR), Action Units (AU), and Valence Arousal (VA), the *Aff-Wild2* [3] dataset is annotated with all three kinds of emotional labels, containing extended facial behaviors in random conditions and increased subjects/frames to the former *Aff-Wild* [1, 2] dataset. Consequently, the multi-task affective recognition can benefit from it, for example, the works [7, 8, 9, 10] participated in the first Affective Behavior Analysis in-the-wild (ABAW) Competition [11].

In this work, we propose a novel multi-task affect recognition framework for the ABAW2 Competition [12]. In contrast to the previous methods which take the multiple emotion recognition problems as parallel tasks, we design our algorithm pipeline in a streaming structure to fully exploit the hierarchical relationships among the three representations including AU, EXPR and VA. Specifically, we make our single-flow network first estimates the AU vectors from input images, then the EXPR labels, and finally the VA distribution. Such arrangements are made due to a heuristic that the regressing order AU$\rightarrow$EXPR$\rightarrow$VA should match the underlying semantics of the three target representations. For instance, AU is defined by the facial action coding system (FACS) based on local patches and therefore AU-related features could provide low-level information for the global expression recognition task. Moreover, the seven-dimensional expression distributions (spanned by the categorical classes) can be compressed into 2D with the two principal components: Valence and Arousal (VA).

Another contribution of our framework is that we utilize an advanced facial expression embedding model to employ helpful prior knowledge for the downstream tasks, i.e., AU detection, Expression recognition, and VA regres-

*Equal contribution.
†Corresponding author; Email: dingyu01@corp.netease.com

sion. Despite the traditional facial expression recognition (FER) models have regressed continuous expression distributions for discrete classification, they can hardly encode the fine-grained expression features. In this work, we adopt the triplet-based expression embedding model [13] as the backbone of the entire framework. Since the expression embedding is trained to distinguish minor expression similarities between different subjects, it can provide powerful expression-related priors to the high-level emotion recognition task.

In participating in the ABAW2 Competition, we conduct extensive experiments on the *Aff-Wild2* [3] dataset. Because of the multi-task framework and streaming design, each module of our network can be fine-tuned on images with no need for all three emotion representation labels to exist. In order to improve the generalization ability of our multi-task model, we extend the training dataset with *BP4D* [5], *DFEW* [14] and *AffectNet* [15]. Besides, we produce reliable pseudo labels on the unsupervised data for augmentation, which is proved as a useful trick to enhance the performance.

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

- We propose a streaming network to handle the multi-task affect recognition problem. By heuristically designing the regression order, the streaming structure allows exploiting inner relationships across different emotional representation spaces.
- We employ an identity-invariant expression prior model as the backbone. With fine-grained expression-related features, our network can well capture the high-level information for the emotional recognition tasks.
- By using the external datasets and producing reliable pseudo labels on the unsupervised data, we manage to fine-tune our model and achieve better performance. The extensive experimental results, along with the competition scores, prove the superiority of our method.

2. Related Works

In this section, we briefly review some concepts, works, and datasets related with the affective recognition problem.

2.1. Facial Expression Representation

Representing human emotions is a fundamental research topic in the affective behavior analysis area. There are three common used facial expression representations: the seven basic emotion categories [16], the Action Units (AUs) defined by the Facial Action Coding System [17] and the two dimensional Valence and Arousal (VA) Space [18]. The seven basic emotions include Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral. AUs [17] include 32 atomic facial action descriptors based on facial muscle groups, which facilitate the physical and fine-grained understanding of human facial expressions. The detection of facial AU occurrence offers crucial information for emotion recognition [19], micro-expression detection [20], and mental health diagnosis [21]. In VA space, the valence represents the degree of emotional positiveness/negativeness and the arousal indicates whether the emotion is passive or active.

Besides, there is another branch of representation methods that tries to model the facial expressions by latent codes. FECNet [22] first proposes to learn a continuous and compact embedding space from facial images. Later, DLN [13] extends this idea by considering the identity attributes and thus proposes a disentangled framework for expression embedding learning. Apart from that, the expression embedding representation has also achieved promising results in terms of capturing fine-grained expression similarities and promoting the other emotion recognition tasks.

2.2. Automatic Affective Behavior Analysis

The ABAW2 Challenge [12] attracts a lot of research efforts into the automatic affective analysis area. Here we first review some submitted works to the competition. NISL-2021 team wins the first prize in the VA track with a model consisting of four transformer layers and a backbone of MobileFaceNet [23]. The CPIC-DIR2021 team [24] extracts multi-modal information from audio and visual signals and trains a multi-task network for the AU and EXPR recognition task, winning second place in both tracks. Maybe Next Time [25] uses a pre-trained ResNet-50 [26] as the backbone and proposes a collaboration training strategy for the AU and EXPR task, achieving third place in both tracks. Morphoboid [27] proposes a teacher-student model for the VA and EXPR task and ranks third place in the VA track.

Apart from the competition, there are many research works focusing on AU detection, expression recognition and VA regression in general scenarios. For AU detection, the previous works [28, 29, 30] first adopt facial landmarks as auxiliary information. Some recent works [31, 32, 33, 34, 35, 36, 37] also exploits the inner-dependencies among different AUs. Specifically, Li *et al*. [31] and Niu *et al*. [38] learn a constant graph for AU relation modeling. Song *et al*. [36] proposes to produce the hybrid graphs based on a random sampling method. Yang *et al*. [37] extracts the AU embeddings from textual descriptions with intra- and inter-attention mechanisms. In terms of facial expression recognition, Li *et al*. [39] and Wang *et al*. [40] propose to use region-based attention networks to alleviate occlusion problem. Considering the uncertainty that comes from inconsistent and incorrect annotations, Zeng *et al*. [41] attempts to automatically re-label the uncertain samples for robust expression recognition. For VA regression, Mehu *et* *al*. [42] observes that some AUs are sensitive to the VA value. Consequently, Chang *et al*. [43] proposes a method to filter some distinctive AU features for VA regression. However, until recently, there exist few works trying to tackle the multiple tasks simultaneously. With the collection of large-scale in-the-wild affective dataset, Kollias *et al*. [44, 45] proposes to jointly predict the three emotional modalities with one model.

2.3. Affective Recognition Dataset

The ABAW2 Competition [12] provides a benchmark dataset *Aff-Wild2* [3] for affective analysis. Extened from the previous *Aff-wild* [1], *Aff-wild2* [3] contains an increasing number of 564 annotated videos: 561 annotated for valence-arousal, 546 videos annotated for 7 basic emotion categories and 541 videos annotated for 12 AUs. *Aff-wild2* [3] is by far the largest in-the-wild dataset w.r.t all the three affective behavior tasks. Besides, there are some other facial emotion datasets. For example, *BP4D* [5] contains spontaneous expressions displayed by 41 subjects. *DFEW* [14] is composed of 16,372 movie clips annotated by 7-dimensional expression distribution vector. *AffectNet* [15] provides eight expression (the seven basic expressions plus contempt) and VA labels of 450k in-the-wild facial images.

3. Method

In this section, we introduce our method for affective behavior analysis in the ABAW2 Competition. The overall pipeline is illustrated in Fig. 1. The entire framework consists of two components: a prior model for extracting prior expression embedding knowledge, and a streaming model for exploiting the hierarchical relationships among three emotional representations.

3.1. Overview

As described in the official white paper [12], the ABAW2 Competition contains three challenges, corresponding to the three commonly used emotion representations: seven basic expression recognition, twelve action units, and two-dimensional valence and arousal space. We propose a general framework to jointly handle the three individual tasks. Despite the different psychological research backgrounds of the three emotional representations, it is widely agreed that the representations are intrinsically associated with each other [46]. One of the evidence is that similar facial muscle movements (action units) mostly indicate similar inner statements, and so do the perceived facial emotions. However, most previous research works on multi-task emotion recognition omit this fact and they just model the different tasks in parallel branches. Inspired by the observation above, we design the recognition process in a serial manner AU$\rightarrow$EXPR$\rightarrow$VA, from local action units to global emotion statements. The streaming structure is helpful to adjust the hierarchical distributions on different feature levels. For example, the optimizing energy from the most high-level VA space should be back-propagated to the low-level features and thus help the other two tasks in training.

Due to the limited subjects and unbalanced annotations of existed affective datasets, it is a challenging issue to prevent the emotion recognition model from overfitting on the disturbing factors, like image background or random noise. To tackle this problem, we adopt a prior facial expression embedding model [13], which can capture the detailed expression similarities across different people, into our framework. The expression embedding model [13] brings at least two advantages. First, by training on even larger facial image datasets with the identity invariant constraint, the embedding itself is independent of the identity attributes as well as the other low-level noisy factors, and therefore can improve the network's generalizability to unseen subjects. Second, the expression embedding model [13] is targeted for discriminating the minor expression similarities within triplet training data. It provides a nice initialization for our latter emotion recognition tasks.

Combining with the prior and the streaming model, we train our multi-task affective recognition model in an end-to-end manner. Given an image $\mathcal{I}$ with at least one of the three emotional annotations, we send it to the full network for training and compute corresponding losses on its existed labels. In the following, we will introduce the network structure and loss functions in detail.

3.2. Prior Model

We adopt the pre-trained Deviation Learning Network (DLN) from [13] as the expression prior model to our framework. In order to generate a compact and continuous expression embedding space disentangled from the identity factor, the DLN model has been trained on more than 400k annotated triplets from the FECNet dataset [22].

Following the idea from [22, 47], the DLN aims to map the similar expression image pair (*anchor* and *positive*) close to each other in the low-dimensional space, while keep the dissimilar expression image pair (*anchor* and *negative*) away from each other (See Fig. 2). To efficiently exclude the identity attributes from the extracted image features, the DLN model proposes a deviation module by subtracting the identity vectors (produced by a pre-trained face recognition model) from the facial ones.

Since the original DLN model maps the facial expression images into a 16-dimensional space, which leaves quite tight room for optimization in our problem, we only take the pre-trained deviation module from [13] that produces 512-dimensional features. Specifically, given a facial image $\mathcal{I}$ from the training dataset, the prior model is expected to generate a 512-dimensional embedding vector $\mathbf{Emb}$ that

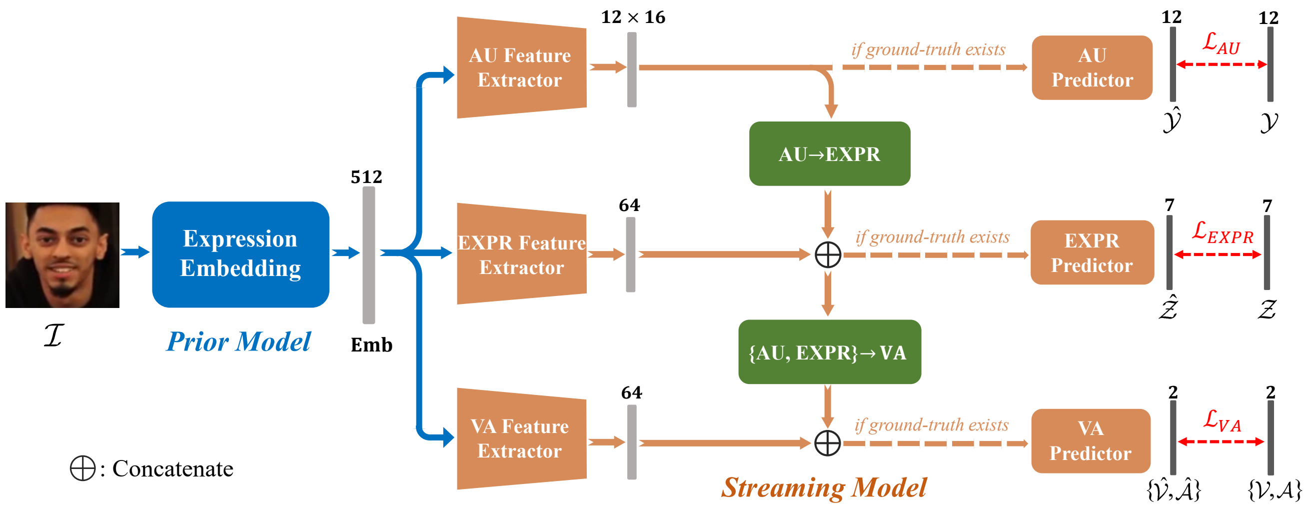

Figure 1: Pipeline of our proposed prior aided streaming network for multi-task affective recognition. Given a cropped facial image $\mathcal{I}$, we first send it to the prior model (blue) adopted from the Deviation Learning Network [13] to produce **Emb**, which is an identity-invariant and fine-grained expression embedding. Then we feed the embedding into the AU/EXPR/VA feature extractors and predictors (orange). The intermediate translation modules (green) are designed to learn the latent mapping between different emotion representations. Finally, we calculate the loss for predictions if the corresponding ground-truth label exists.

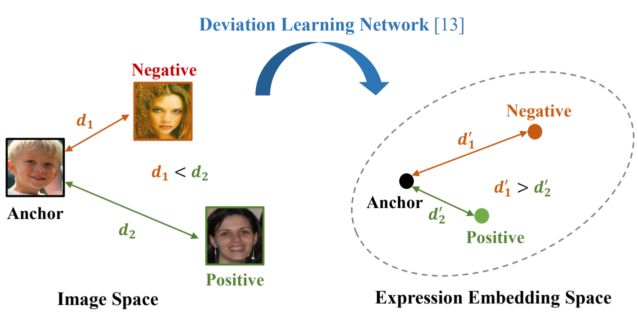

Figure 2: Illustration of the Deviation Learning Network (DLN) [13]. DLN learns a continuous and compact embedding for expression representation. By training on the FECNet dataset [22] with triplet loss, DLN tends to shrink the embedding distances between similar expressions while enlarge the distances between dissimilar expressions.

contains identity-invariant expression information. In training the entire framework, we also make the expression embedding model to be trainable and adaptively adjust the embedding vectors. The prior model serves as a useful backbone for direct expression feature extraction, which is further proved to be very helpful to boost the downstream tasks in experiments.

3.3. Streaming Model

We design the multi-task affective recognition model in a streaming structure. Specifically, following the prior generated expression embedding, we first construct three individual feature extractors to downsample the expression-related feature **Emb** from 512 to 12×16, 64, 64, respectively. After that, two streaming modules are responsible for translating the features by the pre-defined order AU→EXPR→VA. At each stage, the individually extracted feature and the translated one will be concatenated together and sent to the corresponding predicting module for loss calculation (if the corresponding ground-truth label exists). The three predictors are all made of several MLP layers along with activation units, producing the final output vectors of dimension 12, 7 and 2, respectively. In the following, we will introduce the detailed model structure as well as training losses for each task.

AU Detection. We set the AU detection task as an initial step in our streaming network. Because AU detection is target for capturing the local signals within facial movements, it actually plays a fundamental role within the affective analysis process. For AU features in $\mathbb{R}^{12\times16}$, we directly send it into a multilayer perceptron (MLP) predictor to predict the probability for each AU. In practice, 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.

We adopt two loss functions for the AU detection: multi-label circle loss [48, 49] and binary cross entropy loss. The former one is proposed for capturing the correlation between AUs, trying to simultaneously enforce all the activated AU's output value to be bigger than 0 and the non-activated AU's output value smaller than 0:

$$
\begin{aligned}
\mathcal{L}_{Circle} &= \log(1 + \sum_{i\in\Omega_0} e^{s_i}) + \log(1 + \sum_{j\in\Omega_1} e^{-s_j}), \\
\Omega_0 &= \{\, i \mid \text{if } y_i = 0 \,\}, \\
\Omega_1 &= \{\, j \mid \text{if } y_j = 1 \,\}.
\end{aligned} \tag{1}
$$

The binary cross entropy loss is used to optimize single AU classification. For each AU, we calculate the cross entropy between the prediction result $\log \hat{y}_j$ and the ground-truth y_j, which can be formulated as:

$$
\mathcal{L}_{CrossEntropy} = -\frac{1}{12}\sum_{j=1}^{12}[y_j \log \hat{y}_j + (1 - y_j)\log(1 - \hat{y}_j)]. \tag{2}
$$

The total loss of AU detection is given as:

$$
\mathcal{L}_{AU} = \mathcal{L}_{Circle} + \mathcal{L}_{CrossEntropy}. \tag{3}
$$

EXPR Recognition. Except for predicting the AU label, the intermediate AU feature is also translated to assist the expression recognition. We propose an AU$\rightarrow$EXPR module to model the latent relationship between AU and EXPR. The outputs from AU$\rightarrow$EXPR module and EXPR feature extractor are jointly sent into the EXPR predictor for expression classification. After a softmax activation function, the output vector is denoted as $\hat{\mathcal{Z}} = \{\hat{z}_1, \hat{z}_2, ..., \hat{z}_7\} \in \mathbb{R}^7$. The expression ground-truth $\mathcal{Z} = \{z_1, z_2, ..., z_7\} \in \mathbb{R}^7$ is an one-hot vector generated from the annotated expression class. To alleviate the overfitting issue, we use soft cross entropy loss for optimization as follows:

$$
\mathcal{L}_{EXPR} = \lambda \log(\hat{z}_e) + \sum_{\substack{i=1 \\ i\neq e}}^{7}(1 - \lambda)\log(\hat{z}_i). \tag{4}
$$

where e indicates the e-th expression class, i.e., $z_e = 1$, and λ is the label smoothing factor and empirically set to 0.9.

VA Regression. Finally, in order to predict the valence and arousal values, we make use of the intermediate features translated from the AU and EXPR task. The $\{$AU, EXPR$\}\rightarrow$VA module takes the joint features as input and generates another 64-dimensional feature to aid the VA regression. Specifically, the concatenated 128-dimensional feature vector is sent into VA predictor consisting of several fully-connected layers with tanh activation for generating a two-dimensional vector. In the VA track, we use the Concordance Correlation Coefficient (CCC) loss for optimization. CCC is used to evaluate the correlation between all ground truth labels and predictions. For a pair of ground-truth/regression vector $\{\mathcal{X}, \hat{\mathcal{X}}\}$, the CCC function is formulated as:

$$
CCC(\mathcal{X}, \hat{\mathcal{X}}) = \frac{2\rho_{\mathcal{X}\hat{\mathcal{X}}}\delta_{\mathcal{X}}\delta_{\hat{\mathcal{X}}}}{\delta_{\mathcal{X}}^2 + \delta_{\hat{\mathcal{X}}}^2 + (\mu_{\mathcal{X}} - \mu_{\hat{\mathcal{X}}})^2}. \tag{5}
$$

where $\delta_{\mathcal{X}}$,$\delta_{\hat{\mathcal{X}}}$ 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.

We define the batch output of VA predictions as $\hat{\mathcal{V}}$, $\hat{\mathcal{A}}$, and the annotated labels $\mathcal{V}$, $\mathcal{A}$. We compute two CCC values, $CCC(\mathcal{V}, \hat{\mathcal{V}})$ for valence and $CCC(\mathcal{A}, \hat{\mathcal{A}})$ for arousal. In general, the CCC loss for VA regression is computed as following:

$$
\mathcal{L}_{VA} = 2 - [CCC(\mathcal{V}, \hat{\mathcal{V}}) + CCC(\mathcal{A}, \hat{\mathcal{A}})]. \tag{6}
$$

In sum, the total loss of our streaming network can be formulated as:

$$
\mathcal{L}_{total} = \alpha_{AU} \cdot \mathcal{L}_{AU} + \alpha_{EXPR} \cdot \mathcal{L}_{EXPR} + \alpha_{VA} \cdot \mathcal{L}_{VA}, \tag{7}
$$

where $\alpha_{AU}, \alpha_{EXPR}$ and α_{VA} are boolean valueables indicating the existences of ground-truth labels on each track.

To conclude, our design of the streaming model comes from the idea that there exists underlying relationships between the AU, EXPR and VA representations. It is obeyed to the phenomenon that human can infer the expression categories from the AUs and approximate the VA values in 2D space from the expressions. Therefore, we propose the AU$\rightarrow$EXPR module and $\{$AU, EXPR$\}\rightarrow$VA module to mimic the above heuristics, and so that to help infer more hidden information from limited training data.

3.4. Data Augmentation

Due to the unbalanced distribution of emotional recognition datasets, we further propose two strategies to augment the training data, including adding external datasets and generating pseudo labels.

External Dataset. In addition to the original training set of *Aff-Wild2* [3], our model is further trained on the *BP4D* [5], *DFEW* [14], and *AffectNet* [15]. *BP4D* is a large-scale in-the-lab 3D video database of spontaneous facial expressions with totally 328 videos from 41 subjects. The videos are annotated with 12 AUs (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, AU24), head pose and facial landmarks. At the same time, we add the data from *DFEW* and *AffectNet* dataset for facial expression recognition. *DFEW* is a large-scale facial expression database with 16,372 video clips from movies and annotations of 7 basic expressions. *AffectNet* contains 450,000 in-the-wild images categorized into 8 basic expressions (one more category for contempt than the typical seven basic expressions) and also labelled with VA.

In the expression recognition task, we only use part of the *DFEW* dataset due to the bias between different

Track / Validation set	AU			EXPR			VA		
	F1	*TAcc*	S_{AU}	*F1*	*TAcc*	S_{EXPR}	CCC_V	CCC_A	S_{VA}
Original	**0.588**	**0.896**	**0.742**	**0.757**	0.856	**0.790**	**0.488**	**0.502**	**0.495**
Fold-1	0.602	0.903	0.753	0.753	0.843	0.783	0.574	0.581	0.578
Fold-2	**0.640**	0.903	**0.772**	**0.673**	0.830	**0.725**	0.557	**0.624**	0.591
Fold-3	0.610	0.901	0.755	0.730	**0.827**	0.762	0.455	0.609	0.532
Fold-4	0.605	0.901	0.753	0.737	0.839	0.770	**0.642**	0.600	**0.621**
Fold-5	0.609	**0.907**	0.758	0.714	**0.868**	0.765	0.591	0.621	0.606

Table 1: The validation results 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.

datasets. Specifically, we only utilize the images that achieve a high confidence on the original *Aff-Wild2*-trained model with a threshold of 0.8. In the AU detection task, the AU labels in the external datasets are not exactly the same as the *Aff-Wild2*'s. In particular, *BP4D* dataset lacks the annotations for AU25 and AU26 and adds the annotations for AU14 and AU17. So we only keep the external data with AUs that are consistent with the *Aff-Wild2* and omit the different ones. In the VA regression task, we adopt the images with valence and arousal annotations between -1.0 and -0.25 from the *AffectNet* dataset for training.

Pseudo Label. We also propose to generate reliable pseudo labels on the unannotated data to improve the network generalization ablity. We introduce two strategies for pseudo label generation: based on rules and based on teacher-student scheme.

We first exploit the underlying relationship between AU and EXPR in a manual manner. Particularly, it is observed that some AUs mostly indicate the same expression classes. For example, the AU vector $(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1)$ in the *Aff-Wild2* training set mostly occurs when conveying surprise emotion. With such knowledge, we can quickly infer the missing expression labels from explicit AU annotations. We conclude these rules from the training set and make them to generate pseudo expression labels for fine-tuning our model.

Second, we also employ the teacher-student training strategy for unsupervised domain adaptation. For the facial images without emotional annotations, we filter them with a high confidence value by model prediction results and add them for training. In this way, we produce around 500k pseudo labels for fine-tuning and the experimental results indicate obvious improvement in terms of evaluation scores.

4. Experiment

In this section, we first introduce our experimental settings. Then, we give some experimental comparison results on the validation and test set of *Aff-Wild2* [3]. We also conduct several ablation experiments to evaluate the effectiveness of each module in our framework.

4.1. Experimental Setting

We processed all videos in the *Aff-Wild2* [3], *BP4D* [5], and *DFEW* [14] datasets into frames by OpenCV and employ the OpenFace [50] detector to crop all facial images into 224×224 scale. Our training process is implemented based on PyTorch. The training procedure costs around 20 hours on an NVIDIA RTX 3090 graphics card, with a learning rate of 0.002 and batch size 80. We use a stochastic gradient (SGD) optimizer with a cosine annealing warm restart learning rate scheduler.

4.2. Metric

For AU and EXPR tasks, we calculate the F1-Score (*F1*) and total accuracy (*TAcc*) to evaluate the prediction results. For the VA regression results, we compute the Concordance Correlation Coefficient (CCC) for valence and arousal respectively (i.e., CCC_V and CCC_A).

In participating the ABAW2 Competition [51], we also report the final scores per each track. The AU and EXPR scores are defined as the weighted sum of *F1* and *TAcc*: $S_{AU} = 0.5 \times F1 + 0.5 \times TAcc$; $S_{EXPR} = 0.67 \times F1 + 0.33 \times TAcc$. While the VA score is defined as the average of CCC_V and CCC_A: $S_{VA} = (CCC_V + CCC_A)/2$.

4.3. Comparison

In practice, we first conduct 5-fold cross-validation experiments on the *Aff-Wild2* datasets (See Sec. 4.3.1). The quantitative results indicate that the splitting of the training/validation set makes a notable impact on the model precision. Therefore we propose an ensembling strategy to fuse the prediction results generated by models that are trained on different foldings. In Sec. 4.3.2, we compare our method with the baseline [12] as well as the other competitive approaches.

4.3.1 Cross validation and ensembling.

The original *Aff-Wild2* dataset [3] is split into training/validation/test set based on the video subjects. We argue

Method \ Track	AU			EXPR			VA		
	F1	*TAcc*	S_{AU}	*F1*	*TAcc*	S_{EXPR}	CCC_V	CCC_A	S_{VA}
Baseline [12]	0.367	0.193	0.280	0.260	0.460	0.326	0.200	0.190	0.195
Morphoboid [27]	–	–	–	0.351	0.668	0.455	0.505	0.474	0.489
Maybe Next Time [25]	0.461	0.876	0.669	0.604	0.728	0.645	–	–	–
CPIC-DIR2021 [24]	0.489	**0.891**	0.690	0.683	0.770	0.712	–	–	–
NISL-2021	0.450	0.846	0.652	0.431	0.653	0.504	**0.532**	0.453	**0.493**
Ours	**0.505**	0.888	**0.697**	**0.763**	**0.806**	**0.777**	0.485	**0.495**	0.490

Table 2: Comparison results between our method and the other competitive works on the ABAW2 test set. The best is indicated in bold.

that the division of training and validation set is sensitive to the model precision. To verify this point, we conduct 5-fold random cross-validation experiments and report the statics of prediction results on each fold including the original split dataset. In Tab. 1, it can be observed that our model performances are varying obviously among the different validation sets, especially on the VA task. The CCC_V metric ranges from 0.488 to 0.642 while CCC_A could be 0.502 to 0.624.

The unstable results indicate that the model performances are highly dependent on the distribution consistency between the training and validation set. To this end, we propose an ensembling strategy to improve the robustness of our prediction results. Specifically, we make the results generated by six models (five are trained on 5-fold split datasets and the other trained on the original dataset) to vote for the final predictions.

4.3.2 Test result.

Here, we report the official released comparison results on the test set of *Aff-Wild2* [3]. As shown in Tab. 2, our method wins the first prizes in the AU and EXPR tracks, and the second place in the VA track. In particular, we achieve a leading EXPR result with score of 0.777, compared to the second one 0.712 from CPIC-DIR2021 [24].

One of the most technical differences between our method and the others is the backbone/prior model. Specifically, the other works [25, 27] simply apply Resnet [26] or face recognition model [24] as backbones while our method adopts the expression priors from a pre-trained expression embedding model which can encode fine-grained and identity-invariant expression similarity information. Besides, CPIC-DIR2021 [24] and NISL-2021 take the time-sequential information into account. Maybe Next Time [25] and Morphoboid [27] implicitly exploit the correlation between different affective representations by multi-task training. Instead, we propose the streaming model to extract the interrelationships following an explicit order AU→EXPR→VA.

Method	S_{AU}	S_{EXPR}	S_{VA}
Baseline [12]	0.310	0.366	0.220
Ours w/o prior model	0.669	0.621	0.473
Ours w/o streaming model	0.677	0.664	0.447
Ours w/o data augmentation	0.742	0.790	0.495
Ours	**0.756**	**0.793**	**0.540**

Table 3: Ablation study results of the prior model, streaming structure and data augmentation module. All scores are computed based on the official validation set. The best is indicated in bold.

4.4. Evaluation

In order to evaluate the effectiveness of our proposed algorithm design, i.e., prior model, streaming network, and data augmentation, we conduct ablation studies by comparing the models trained without the corresponding components. The quantitative results shown in Tab. 3 and Fig. 3 indicate the benefits of the algorithm modules in terms of improving the affective recognition performances on each track.

Prior Model. To verify the effectiveness of the prior model, we conduct an ablation study by replacing the DLN [13] prior model with the ResNet50 [26] backbone. From the Tab. 3, it can be observed that DLN makes distinct improvements compared to the ResNet50 model. To analyze the concerned attributes of the prior model, we utilize the Grad-CAM [52] tool to visualize the feature-sensitive areas within the last layer of the prior model. Fig. 4 illustrates the several samples from *Aff-Wild2* [3] and proves that the DLN model is concentrating on the facial areas mostly conveying human emotions, such as foreheads, eyebrows, cheeks, lips, and jaws, but ignoring the less interesting areas like face boundaries and backgrounds. This phenomenon demonstrates that our prior model is capable enough of capturing the identity-invariant expression features and therefore motivating the down-streaming tasks.

Streaming Model. To prove the effectiveness of the streaming model design, we compare our method with and without the AU→EXPR and {AU, EXPR}→VA module. As shown in Tab. 3, when applying the streaming model, the

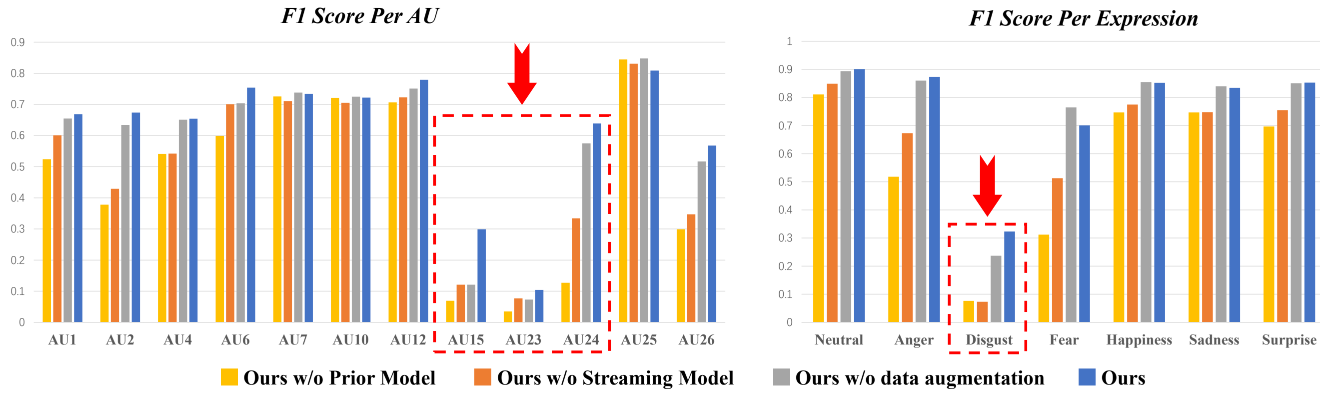

Figure 3: F1 score per each class of the AU (left) and EXPR (right). The results are calculated on the official validation set.

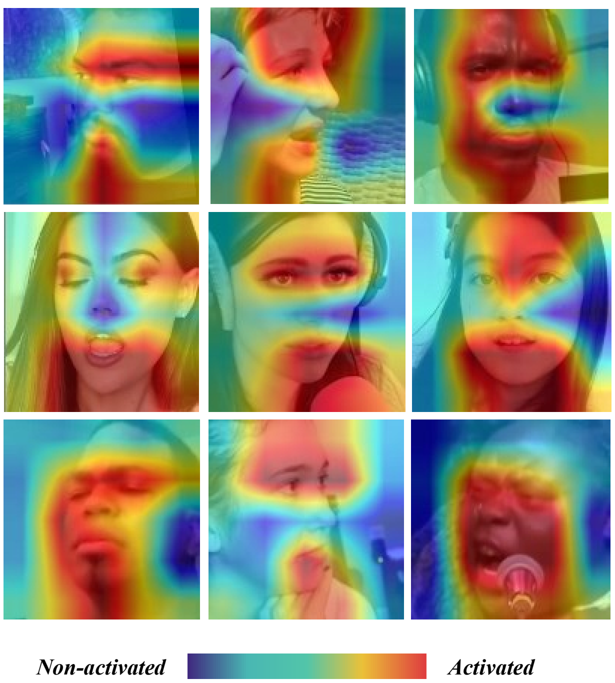

Figure 4: Grad-CAM [52] visualization of the feature activation map within the last layer of our prior model. The activated areas suggest that the prior model produces large gradients there, which means that our model is trained to focus on the expression-related face regions.

AU, EXPR, VA score is improved from 0.677, 0.664, 0.447 to 0.756, 0.793, 0.540, respectively. It suggests that the streaming model can effectively model the underlying relationships between different affective representation spaces.

Data Augmentation: Finally, we compare the experimental results with and without training on the augmented datasets. By finetuning on the external datasets and generated pseudo labels, our model reaches higher scores, especially for the AU and VA tracks (Tab. 3). As shown in Fig. 3, the F1 scores of the unbalanced categories (e.g., AU15, AU23 and AU24 in AU track and Disgust in EXPR track) improve obviously with the augmented training data. This proves that the data augmentation operation can serve as a useful strategy to promote the model performance on minority classes.

5. Limitation and Discussion

Despite we have proved the effectiveness of the streaming model for multi-task affective analysis, however, the AU→EXPR→VA order is not thoroughly evaluated yet. The current hierarchical design is simply based on our heuristics on a perception level of the three concepts. In the future, it would be meaningful to exploit and demonstrate the underlying relationships across the three representations. Besides, we did not explore our model performance on the other affective recognition dataset, due to a limitation of time and resources. In the future, we would extend the framework by considering other improvements such as aural and temporal information.

6. Conclusion

In this paper, we introduce our deep learning based framework for multi-task affective recognition in the ABAW2 Competition. We propose a streaming network by exploiting the hierarchical relationships between different emotion representations. Besides, we employ an expression prior model to help extract the identity-invariant expression features, alleviating the burden of downstream tasks. Finally, we finetune our model on the external datasets and reliable pseudo labels. In participating in the competition, we won the first prizes in the AU track and EXPR track and achieve second place in the VA track. The competition results indicate the superiority of our framework. We also conduct the ablation study to prove that each component of our method is effective to the affective recognition tasks.

References

[1] 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] 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.

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

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

[5] 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.

[6] 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.

[7] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020.

[8] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. *arXiv preprint arXiv:2009.14440*, 2020.

[9] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 632–636. IEEE, 2020.

[10] Junya Saito, Ryosuke Kawamura, Akiyoshi Uchida, Sachihiro Youoku, Yuushi Toyoda, Takahisa Yamamoto, Xiaoyu Mi, and Kentaro Murase. Action units recognition by pairwise deep architecture. *arXiv preprint arXiv:2010.00288*, 2020.

[11] 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.

[12] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.

[13] 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.

[14] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889, 2020.

[15] 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.

[16] Paul Ekman. Darwin, deception, and facial expression. *Annals of the new York Academy of sciences*, 1000(1):205–221, 2003.

[17] Paul Ekman and Wallace Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press Palo Alto*, 12, 01 1978.

[18] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[19] Maja Pantic and Léon Rothkrantz. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34:1449 – 1461, 07 2004.

[20] Guoying Zhao and Xiaobai Li. Automatic micro-expression analysis: Open challenges. *Frontiers in Psychology*, 10, 08 2019.

[21] David R Rubinow and Robert M Post. Impaired recognition of affect in facial expression in depressed patients. *Biological psychiatry*, 31(9):947–953, 1992.

[22] 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.

[23] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018.

[24] 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.

[25] 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.

[26] 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.

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

[28] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017.

[29] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eacnet: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018.

[30] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018.

[31] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019.

[32] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11917–11926, 2019.

[33] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 298–313, 2018.

[34] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017.

[35] Xianpeng Ji, Yu Ding, Lincheng Li, Yu Chen, and Changjie Fan. Multi-label relation modeling in facial action units detection. *arXiv preprint arXiv:2002.01105*, 2020.

[36] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6267–6276, 2021.

[37] 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 (CVPR)*, pages 10482–10491, June 2021.

[38] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. *arXiv preprint arXiv:1910.11012*, 2019.

[39] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[40] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[41] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.

[42] Marc Mehu and Klaus R Scherer. Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion*, 15(6):798, 2015.

[43] Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 17–25, 2017.

[44] 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.

[45] 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.

[46] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[47] Juyong Zhang, Keyu Chen, and Jianmin Zheng. Facial expression retargeting from human to avatar made easy. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[48] Huihui He and Rui Xia. Joint binary neural network for multi-label learning with applications to emotion classification. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 250–259. Springer, 2018.

[49] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020.

[50] 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.

[51] 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.

[52] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.