A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition

R Gnana Praveen[1], Wheidima Carneiro de Melo[1], Nasib Ullah, Haseeb Aslam[1], Osama Zeeshan[1], Théo Denorme[1], Marco Pedersoli[1], Alessandro L. Koerich[1], Simon Bacon[2], Patrick Cardinal[1], and Eric Granger[1]
[1] LIVIA, École de technologie supérieure, Montreal, Canada
[2]Dept. of Health, Kinesiology & Applied Physiology, Concordia University, Montreal, Canada
gnanapraveen.rajasekar.1@ens.etsmtl.ca, eric.granger@etsmtl.ca

Abstract

Multimodal emotion recognition has recently gained much attention since it can leverage diverse and complementary modalities, such as audio, visual, and biosignals. However, most state-of-the- art audio-visual (A-V) fusion methods rely on recurrent networks or conventional attention mechanisms that do not effectively leverage the complementary nature of A-V modalities. This paper focuses on dimensional emotion recognition based on the fusion of facial and vocal modalities extracted from videos. We propose a joint cross-attention fusion model that can effectively exploit the complementary inter-modal relationships, allowing for an accurate prediction of valence and arousal. In particular, this model computes cross-attention weights based on the correlation between joint feature representations and individual modalities. By deploying a joint A-V feature representation into the cross-attention module, the performance of our fusion model improves significantly over the vanilla cross-attention module. Experimental results[1] on the AffWild2 dataset highlight the robustness of our proposed A-V fusion model. It has achieved a concordance correlation coefficient (CCC) of 0.374 (0.663) and 0.363 (0.584) for valence and arousal, respectively, on the test set (validation set). This represents a significant improvement over the baseline for the third challenge of Affective Behavior Analysis in-the-Wild 2022 (ABAW3) competition, with a CCC of 0.180 (0.310) and 0.170 (0.170).

1. Introduction

Emotion recognition (ER) is a challenging problem since the expressions linked to human emotions are extremely diverse across individuals and cultures. It has been extensively researched in various fields such as neuroscience, psychology, cognitive science, and computer science, leading to the advancement of a wide range of applications in, e.g., health care (e.g., assessment of anger, fatigue, depression, and pain), robotics (human-machine interaction), driver assistance (assessment of a driver's state), etc. [1]. ER problems can be formulated according to either a categorical or a dimensional model of emotions. In the categorical model, human emotions have been categorized into anger, disgust, fear, happy, sad, and surprise [2]. Subsequently, contempt has been added to these six basic emotions [3]. This categorical model of ER has been explored extensively in the field of affective computing due to its simplicity and universality. In the dimensional model, a wide range of human emotions can be analyzed on a continuous scale, where emotions can be projected onto the dimensions of valence and arousal [4]. Figure 1 illustrates the use of a two-dimensional space to represent emotional states, where valence and arousal are employed as dimen-

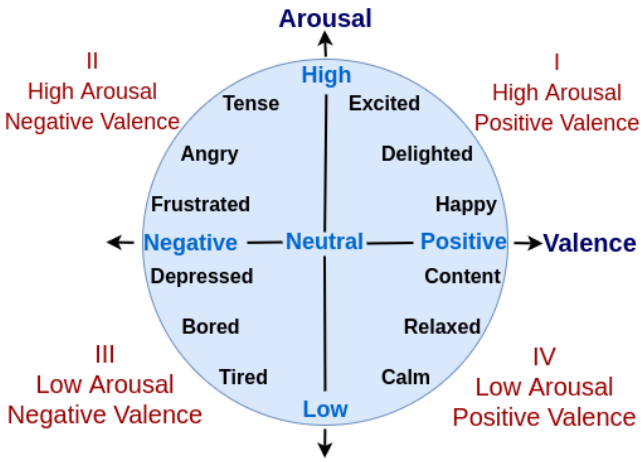

Figure 1. **The valence-arousal space.**

[1]The code is available on GitHub: https://github.com/praveena2j/JointCrossAttentional-AV-Fusion.

sional axes. Valence reflects the wide range of emotions in the dimension of pleasantness, from being negative (sad) to positive (happy). In contrast, arousal spans a range of intensities from passive (sleepiness) to active (high excitement).

Dimensional modeling of emotions is more challenging than the categorical case since it is difficult to obtain a continuous scale of annotations compared to discrete emotions. Given the continuous range of emotions, the annotations tend to be noisy and ambiguous. Several databases, such as RECOLA [5], SEWA [6], SEMAINE [7], etc., have been introduced for the task of dimensional ER. Depending on the video capture conditions, i.e., controlled or in-the-wild environments, this task can present different challenges due to poor illumination, pose variations, and background noise. Recently, Kollias et al. [8] introduced the Affwild2 dataset, which is the largest in-the-wild database for the dimensional ER task. Affwild2 is also provided with the annotations for the tasks of expression classification and action unit detection. This dataset has previously been used for challenges hosted in conjunction with CVPR 2017 [9], FG 2020 [10], and ICCV 2021 [11]. Several approaches have been proposed for previous challenges in the framework of multi-task learning [12–15]. In continuation with the previous challenges, the third competition was held in conjunction with CVPR 2022 [16] with an exclusive challenge track for valence and arousal estimation.

This paper investigates the prospect of leveraging the complementary relationship between audio (A) and video (V) modalities in videos using a joint cross attentional framework. Facial expressions are one of the most dominant channels to express human emotions. It has been shown that only one-third of human communication is conveyed through verbal components, while two-thirds of communication occurs through non-verbal components [17]. Voice also serves as a major cue in conveying human emotions as it often carries complementary information with the V modality. For instance, we can still leverage the A modality to estimate the emotional state when the facial modality is missing due to pose, blur, low illumination, etc. Similarly, when we have silent regions in the A modality, we can leverage the rich information in the V modality. In most of the existing approaches, A-V fusion is often achieved by concatenating the A and V features, which may degrade system performance [18]. Therefore, designing a fusion mechanism based on A and V features that can effectively leverage their complementary relationships is pivotal in improving the accuracy and robustness of multimodal ER systems over uni-modal approaches.

Several ER approaches have been proposed for video-based dimensional ER using convolutional neural networks (CNNs) to obtain the deep learning (DL) features, along with recurrent neural networks (RNNs) to capture the temporal dynamics [18, 19]. DL models have also been widely explored for vocal emotion recognition, typically using spectrograms with 2D-CNNs [19, 20], or raw waveforms with 1D-CNNs [18]. In most of the existing approaches for dimensional ER [18, 21], A-V fusion is performed by concatenating the deep features extracted from individual facial and vocal modalities, and then fed to a Long Short Term Memory Networks (LSTM) for predicting valence and arousal. Although LSTM-based fusion models the spatio-temporal and intra-modal relationships, and can thereby improve system performance, it does not effectively capture the inter-modal relationships across the individual modalities. Therefore, we investigate the benefits of extracting more contributive features across A and V modalities to leverage their complementary temporal relationships.

Attention mechanisms have recently gained much interest in the computer vision and machine learning communities, allowing to extract task-relevant features, and thereby improve system performance. However, most of the existing attention-based approaches for dimensional ER explore the intra-modal relationships [22]. Although a few approaches attempt to capture the cross-modal relationships using cross-attention based on transformers [21, 23], they do not effectively leverage the complementary relationship of A-V modalities. Indeed, their computation of attention weights does not consider the correlation among the A and V features. Recently, Praveen et al. [24] proposed a cross-attentional model for dimensional ER based on A-V fusion, and showed significant improvements on the RECOLA dataset [5] over state-of-the-art methods by leveraging the complementary relationships of A and V modalities. This paper introduces joint modeling of intra- and inter-modal relationships into a cross attentional framework. The cross-correlation is computed between the joint A-V feature representation, and the features of individual modalities. We show that deploying joint representation into the cross-attentional module can significantly improve the modeling of cross-modal relationships over the vanilla cross attentional model [24], while reducing the heterogeneity across modalities on the challenging in-the-wild Affwild2 dataset [8].

The main contributions of the paper are as follows. (1) A joint cross-attentional model is proposed for A-V fusion based on the joint modeling of intra- and inter-modal relationships, which effectively captures the complementary relationships across A and V modalities along with intra-modal relationships. Specifically, we use joint A-V feature representations to attend to the other modality (as well as itself) based on the attention weights computed from the cross-correlation between the individual features and joint representation. (2) The effectiveness of the proposed approach is analyzed through an extensive set of experiments and ablation studies on the Affwild2 dataset.

The rest of this paper is organized as follows. Section 2

provides a critical analysis of the relevant literature on dimensional ER and attention models for A-V fusion. Section 3 describes the proposed joint cross-attentional A-V fusion model. Sections 4 and 5 present the experimental methodology for validation and results obtained with the proposed approach, respectively.

2. Related Work

2.1. A-V Fusion Based Emotion Recognition

One of the primitive approaches using DL models for A-V fusion-based dimensional ER was proposed by Tzirakis et al. [18], where A and V features, obtained from ResNet50 and 1D-CNN, respectively, are concatenated and fed to Long short-term memory model (LSTM). Juan et al. [25] presented an empirical study of fine-tuning several layers of pretrained CNN models for V modality and used conventional A features for fusion. Nguyen et al. [26] proposed a DL model of two-stream auto-encoders and LSTM to simultaneously learn compact representative features from A and V modalities for dimensional ER. Schonevald et al. [19] explored knowledge distillation using a teacher-student model for V modality, and a CNN model for A modality using spectrograms, and combined them RNNs. Deng et al. [27] proposed an iterative self distillation method for modeling the uncertainties in the labels in a multi-task framework. Kuhnke et al. [28] proposed a two-stream A-V network, where V features are extracted from the R(2plus1)D model pretrained from an action recognition dataset, and A features are obtained from the Resnet18 model. Wang et al. [20] further improved their approach [28] by introducing a teacher-student model in a semi-supervised learning framework. The teacher model is trained on the available labels, which are further used to obtain pseudo labels for unlabeled data. The pseudo labels are finally used to train the student model, used for the final prediction. Though the approaches mentioned above have shown significant improvement for dimensional ER, they fail to effectively capture the inter-modal relationships and relevant salient features specific to the task. Therefore, we have focused on capturing the comprehensive features in a complementary fashion using attention mechanisms.

2.2. Attention Models for A-V Fusion

Attention models for A-V fusion have been widely explored in modeling intra- and inter-modal relationships between A-V modalities for various applications such as A-V event localization [29], action localization [30], emotion recognition [23], etc. Zhang et al. [31] proposed an attentive fusion mechanism, where multi-features are obtained from 3D-CNNs and 2D-CNNs for V modality, and from 2D-CNNs using spectrograms for A modality. The obtained A and V features are further re-weighted using scoring functions based on the relevant information in the individual modalities. Recently, cross-modal attention is found to be promising as effective modeling of inter-modal relationships significantly improves the system performance. Srinivas et al. [23] explored transformers with encoder layers, where cross-modal attention is deployed to integrate A and V features for dimensional ER. Tzirakis et al. [21] investigated self-attention as well as cross-attention fusion based on transformers to enable the extracted features of different modalities to attend to each other. Although these approaches have explored cross-modal attention with transformers, they fail to leverage semantic relevance among the A-V features based on cross-correlation. Zhang et al. [32] investigated the prospect of improving the fusion performance over individual modalities and proposed leader-follower attentive fusion for dimensional ER. The obtained features are encoded, and attention weights are obtained by combining the encoded A and V features. These weights are further attended to on the V features and concatenated to the original V features for final prediction.

Unlike prior approaches, we advocate for a simple yet efficient joint cross-attentional model based on joint modeling of intra- and inter-modal relationships between A and V modalities. Cross-attention has been successfully applied in several applications, such as weakly-supervised action localization [30], few-shot classification [33] and dimensional ER [34]. In most cases, cross-attention has been applied across the individual modalities. Praveen et al. [24] have shown significant improvements using cross attention based on cross-correlation across the individual features. However, we have explored joint attention between individual and combined A-V features. By deploying the joint A-V feature representation, we can effectively capture the intra- and inter-modal relationships simultaneously by allowing interactions across the modalities and oneself. Recently, joint co-attention has also been explored by Duan et al. [29] in a recursive fashion for A-V event localization and found to be promising in obtaining robust multimodal feature representations. In this paper, joint (combined) A-V features are extracted through cross-attention, where the features of each modality attend to themselves, as well as those of the other modality through cross-correlation of the concatenated A-V features and features of individual modalities. The proposed approach can significantly improve system performance by effectively leveraging the joint modeling of intra- and inter-modal relationships.

3. Proposed Approach

3.1. Visual Network

Facial expressions in videos carry rich information pertinent to both appearance and temporal dynamics, which plays a crucial role in understanding a person's emotions

[35]. Therefore, these spatial and temporal cues must be efficiently modeled to obtain robust feature representations suitable for ER. In recent years, DL models have been widely explored for analyzing facial expressions in videos. In most of these approaches [36, 37], 2D-CNN has been used in conjunction with RNNs to capture the spatial and temporal dynamics, respectively. 3D-CNNs have also been widely explored, especially for action recognition, and found to be promising in simultaneously capturing the spatial and temporal dynamics. Inspired by the performance of 3D-CNNs, authors in [38] explored R(2plus1)D networks pretrained on the Kinetics-400 action recognition dataset [20, 28]. It has outperformed conventional 2D-CNNs for dimensional ER on Affwild2 dataset. Recently, Inflated 3D-CNNs (I3Ds) [39] have provided significant improvement on action recognition data with fewer parameters than conventional 3D-CNNs, while being able to exploit the weights of several pre-trained 2D-CNN models. However, it fails to capture the long-term temporal dependencies. Temporal convolutional networks (TCN) were found to be efficient in capturing the long-term temporal dependencies [32]. Therefore, we have considered I3D with TCN to leverage both long- and short-term temporal dynamics. We have also explored other V backbones, such as the R(2plus1)D network pretrained on the Kinetics-400 dataset [20, 28], and ResNet CNNs with GRU to obtain V features and validate our fusion model (see implementation details in Section 4).

3.2. Audio Network

Several low-level descriptors such as prosodic, excitation, Mel-Frequency Cepstral Coefficients (MFCCs), and spectral descriptors have commonly been used as feature representations for the A modality in ER [25, 40]. With the advent of DL models, the performance of speech ER has been significantly improved using 1D-CNNs on raw A signals [18] or 2D-CNN models on spectrograms [19, 20]. Compared to 1D-CNNs, 2D-CNNs using spectrograms have been widely explored in the literature of speech ER, as it was found to carry significant para-lingual information about the affective state of a person [41]. Various 2D-CNN architectures such as VGGish [32] and Resnet18 [42] have been used to obtain robust feature representations of A modality for ER. Given the ubiquitous usage of spectrograms for extracting effective feature representations pertinent to the affective state of a person, we have also used spectrograms with 2D-CNNs in our framework to validate the proposed fusion model (see implementation details in Section 4).

3.3. Joint Cross-Attentional A-V-Fusion

Though A-V fusion can be achieved through unified multimodal training, it was found that simultaneous training of multimodal networks often declines over that of individual modalities [43]. This can be attributed to several factors, such as differences in learning dynamics for A and V modalities [43], different noise topologies, with some modality streams containing more or less information for the task at hand, as well as specialized input representations [44]. Therefore, we have trained DL models for the individual A and V modalities independently to extract A and V features, fed to the joint cross-attentional module for A-V fusion that outputs final valence and arousal predictions.

The V modality carries more relevant information in some video clips for a given video sequence, whereas the A modality might be more relevant for others. Since multiple modalities convey diverse information for valence and arousal, their complementary relationship needs to be effectively captured. To reliably combine these modalities, we rely on a cross-attention-based fusion mechanism to encode the inter-modal information efficiently while preserving the intra-modal characteristics. Though cross-attention has been conventionally applied across the features of individual modalities, we used cross-attention in a joint learning framework. Specifically, our joint A-V feature representation is obtained by concatenating the A and V features to attend to the individual A and V features. By using the joint representation, features of each modality attend to themself and the other modality, helping to capture the semantic inter-modal relationships across A and V. The heterogeneity among A and V modalities can also be drastically reduced by using the combined feature representation in the cross-attentional module, which further improves system performance. A block diagram of the proposed model is shown in Figure 2.

A) Training mode: Let $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$ represent two sets of deep feature vectors extracted for A and V modalities in response to a given input video sub-sequence $\boldsymbol{S}$ of fixed size, where $\boldsymbol{X_a} = \{\boldsymbol{x_a^1}, \boldsymbol{x_a^2}, ..., \boldsymbol{x_a^L}\} \in \mathbb{R}^{d_a \times L}$ and $\boldsymbol{X_v} = \{\boldsymbol{x_v^1}, \boldsymbol{x_v^2}, ..., \boldsymbol{x_v^L}\} \in \mathbb{R}^{d_v \times L}$. L denotes the number of non overlapping fixed-size clips sampled uniformly from $\boldsymbol{S}$, d_a and d_v represents the feature dimension of A and V representations, $\boldsymbol{x_a^l}$ and $\boldsymbol{x_v^l}$ denotes A and V feature vectors, respectively, for $l = 1, 2, ..., L$ clips.

As shown in Figure 2, the joint representation of A-V features, $\boldsymbol{J}$, is obtained by concatenating the A and V feature vectors: $\boldsymbol{J} = [\boldsymbol{X_a}; \boldsymbol{X_v}] \in \mathbb{R}^{d \times L}$, where $d = d_a + d_v$ denotes the feature dimension of concatenated features. This A-V feature representations ($\boldsymbol{J}$) of the given video sub-sequence ($\boldsymbol{S}$) is now used to attend to unimodal feature representations $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$. The joint correlation matrix $\boldsymbol{C_a}$ across the A features $\boldsymbol{X_a}$, and the combined A-V features $\boldsymbol{J}$ are given by:

$$\boldsymbol{C_a} = \tanh\left(\frac{\boldsymbol{X_a}^\top \boldsymbol{W_{ja}} \boldsymbol{J}}{\sqrt{d}}\right) \quad (1)$$

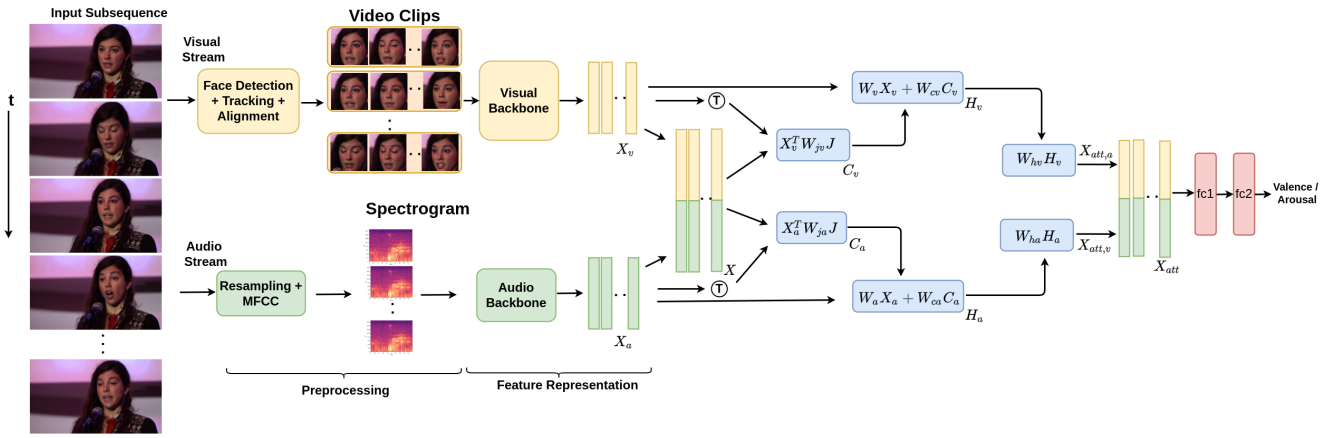

Figure 2. **An overview of the proposed joint cross-attention model for A-V fusion (training mode).**

where $\boldsymbol{W_{ja}} \in \mathbb{R}^{L \times L}$ represents learnable weight matrix across A and joint A-V features. Similarly, the joint correlation matrix for V features is given by:

$$\boldsymbol{C_v} = \tanh\left(\frac{\boldsymbol{X_v^\top W_{jv} J}}{\sqrt{d}}\right) \tag{2}$$

The joint correlation matrices $\boldsymbol{C_a}$ and $\boldsymbol{C_v}$ for A and V modalities provide a semantic measure of relevance not only across the modalities but also within the same modality. A higher correlation coefficient of the joint correlation matrices $\boldsymbol{C_a}$ and $\boldsymbol{C_v}$ shows that the corresponding samples are strongly correlated within the same modality as well as the other modality. Therefore, the proposed approach can efficiently leverage the complementary nature of A and V modalities (i.e., inter-modal relationships) and intra-modal relationships, thereby improving the system's performance. After computing the joint correlation matrices, the attention weights of A and V modalities are estimated.

Since the dimensions of joint correlation matrices ($\mathbb{R}^{d_a \times d}$) and the features of the corresponding modality ($\mathbb{R}^{L \times d_a}$) differ, we rely on different learnable weight matrices corresponding to features of the individual modalities to compute attention weights of the modalities. For the A modality, the joint correlation matrix $\boldsymbol{C_a}$ and the corresponding A features $\boldsymbol{X_a}$ are combined using the learnable weight matrices $\boldsymbol{W_{ca}}$ and $\boldsymbol{W_a}$ respectively to compute the attention weights of the A modality, which is given by:

$$\boldsymbol{H_a} = ReLu(\boldsymbol{W_a X_a} + \boldsymbol{W_{ca} C_a^\top}) \tag{3}$$

where $\boldsymbol{W_{ca}} \in \mathbb{R}^{k \times d}$, $\boldsymbol{W_a} \in \mathbb{R}^{k \times L}$ and $\boldsymbol{H_a}$ represents the attention maps of the A modality. Similarly, the attention maps ($\boldsymbol{H_v}$) of the V modality are obtained as

$$\boldsymbol{H_v} = ReLu(\boldsymbol{W_v X_v} + \boldsymbol{W_{cv} C_v^\top}) \tag{4}$$

where $\boldsymbol{W_{cv}} \in \mathbb{R}^{k \times d}$, $\boldsymbol{W_v} \in \mathbb{R}^{k \times L}$.

Finally, the attention maps are used to compute the attended features of A and V modalities. These features are obtained as:

$$\boldsymbol{X_{att,a}} = \boldsymbol{W_{ha} H_a} + \boldsymbol{X_a} \tag{5}$$

$$\boldsymbol{X_{att,v}} = \boldsymbol{W_{hv} H_v} + \boldsymbol{X_v} \tag{6}$$

where $\boldsymbol{W_{ha}} \in \mathbb{R}^{k \times L}$ and $\boldsymbol{W_{hv}} \in \mathbb{R}^{k \times L}$ denote the learnable weight matrices, respectively. The attended A and V features, $\boldsymbol{X_{att,a}}$ and $\boldsymbol{X_{att,v}}$ are further concatenated to obtain the A-V feature representation, which is given by:

$$\boldsymbol{X_{att}} = [\boldsymbol{X_{att,v}}; \boldsymbol{X_{att,a}}] \tag{7}$$

Finally, the A-V features are fed to the fully connected layers for the predictions of valence or arousal.

The concordance correlation coefficient (ρ_c) has been widely used in the literature to measure the level of agreement between the predictions (x) and ground truth (y) annotations for dimensional ER [18]. Let μ_x, and μ_y represent the mean of predictions and ground truth, respectively. Similarly, if σ_x^2 and σ_y^2 denote the variance of predictions and ground truth, respectively, then ρ_c between the predictions and ground truth is:

$$\rho_c = \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{8}$$

where σ_{xy}^2 denotes the covariance between predictions and ground truth. Although MSE has been widely used as a loss function for regression models, we use $\mathcal{L} = 1 - \rho_c$ since it is standard and common loss in the dimensional ER literature [18]. The parameters of our A-V fusion model

($\boldsymbol{W}_{\mathbf{ca}}$, $\boldsymbol{W}_{\mathbf{a}}$, $\boldsymbol{W}_{\mathbf{cv}}$, $\boldsymbol{W}_{\mathbf{v}}$, $\boldsymbol{W}_{\mathbf{ha}}$, and $\boldsymbol{W}_{\mathbf{hv}}$) are optimized according to this loss.

B) Test mode: A continuous video sequence is an input to our model during inference. Feature representations $\boldsymbol{x}_{\mathbf{a}}^{l}$ and $\boldsymbol{x}_{\mathbf{v}}^{l}$ are extracted by A and V backbones for successive input clips and spectrograms and fed to the A-V fusion model for the prediction of valence and arousal. In addition, arousal and valence predictions may be produced using multiple diverse A and V backbones that are combined through feature-level fusion or multiple A-V fusion models that are combined through decision-level fusion (see implementation details in Section 4).

4. Experimental Methodology

4.1. Dataset

Affwild2 is the most extensive database in affective computing captured from YouTube under extremely challenging environments. Though the dataset is provided with annotations for the tasks of expression classification, action unit detection, and valence-arousal, we have focused on the problem of estimating valence-arousal in this work. For the track of valence-arousal estimation challenge, there are 567 videos with the annotations of valence and arousal. Sixteen of these video clips display two subjects, both of which have been annotated. The annotations are provided by four experts using a joystick, and the final annotations are obtained as the average of the four raters. There are $2,786,201$ frames with 455 subjects, of which 277 are male and 178 female. The annotations for valence and arousal are provided continuously in the range of [-1, 1]. Some of the frames in some videos are not annotated. So we discard those frames. The dataset is split into the training, validation, and test sets. The partitioning is subject-wise so that every subject's data will be present in only one subset. The partitioning produces 341, 71, and 152 training, validation, and test videos, respectively.

4.2. Implementation Details

For the V modality, we have used the cropped, and aligned images provided by the challenge organizers [11]. For the missing frames in the V modality, we have considered black frames (i.e., zero pixels). Faces are resized to 224x224 to be fed to the I3D network. The videos are converted to sub-sequences, which are sampled uniformly to obtain non-overlapping fixed-size clips. The subsequence length and the clip length of the videos are considered to be 64 and 8 respectively, obtained by down-sampling a sequence of 256 frames by 4. Therefore, we have eight clips in each sub-sequence, resulting in $196,265$ training samples and $41,740$ validation samples, and $92,941$ test samples. The I3D model was pre-trained on the ImageNet dataset and inflated to a 3D-CNN using Affwild2 videos of facial expressions. Dropout is used with $p = 0.8$ on the linear layers to regularize the network. The initial learning rate was set to $1e-3$, and momentum of 0.8 is used for SGD. Weight decay of $5e-4$ is used. Here again, the batch size of the network is set to 8. Data augmentation is performed on the training data through random cropping, which produces a scale-invariant model. The number of epochs is set to 50, and early stopping is used to obtain the weights of the best model.

For the A modality, the vocal signal is extracted from the corresponding video and re-sampled to $44,100$Hz, which is further processed to extract short vocal segments corresponding to a clip size of 32 frames of the V network. The clips and sub-sequences of V clips are ensured to be properly synchronized with that of A clips. The spectrogram is obtained using Discrete Fourier Transform (DFT) of length 1024 for each short clip (corresponding to 32 frames), where the window length is considered to be 20 msec and the hop length to be 10 msec. Following aggregation of short-time spectra, we obtain the spectrogram of 64×107 corresponding to each sub-sequence of the V modality. Next, the spectrogram is converted to log-power-spectrum, expressed in dB. Finally, mean and variance normalization is performed on the spectrogram. Now the obtained spectrograms are fed to the Resnet18 [42] to obtain A features. Due to the availability of a large number of samples in the Affwild2 dataset, we trained the Resnet18 model from scratch. To adapt to the number of channels of the spectrogram, the first convolutional layer in the Resnet18 model is replaced by a single channel. The network is trained with an initial learning rate of 0.001, and weights are optimized using the Adam optimizer. The batch size is considered to be 64, and early stopping is used to obtain the best model for prediction.

For the A-V fusion network, the size of the concatenated A-V features J are set to be 1024. In the joint cross-attention module, the initial weights of the cross-attention matrix are initialized with the Xavier method [45], and the weights are updated using the Adam optimizer. The initial learning rate is set to be 0.001, and the batch size is fixed to 64. Also, a dropout of 0.5 is applied to the attended A-V features, and a weight decay of $5e-4$ is used for all the experiments. Finally, feature-level (decision-level) fusion is implemented by training a fully connected neural network to provide a weighted fusion of feature representations (decisions values) for arousal and valence predictions.

5. Results and Discussion

5.1. Ablation Study

Table 1 presents the results of our ablation study on the validation dataset. The performance of the proposed joint cross-attentional fusion is compared using various A and

Table 1. **Performance of our approach with different components on the development set of the Affwild2 dataset. The Resnet18 [42] is used to extract A features in all experiments.**

V Backbone	Fusion Module	Valence	Arousal
I3D	Feature Concatenation	0.531	0.468
R3D	Feature Concatenation	0.517	0.493
I3D	Cross-Attention [24]	0.541	0.517
I3D	Leader-Follower [19]	0.592	0.521
Resnet18-GRU	Joint Cross-Attention (Ours)	0.632	0.520
R3D	Joint Cross-Attention (Ours)	0.642	**0.592**
I3D	Joint Cross-Attention (Ours)	0.657	0.580
I3D-TCN	Joint Cross-Attention (Ours)	0.663	0.584
I3D-TCN + R3D	Joint Cross-Attention (Ours)	**0.670**	0.590

Table 2. **CCC of the proposed approach compared to state-of-the-art methods for A-V fusion on the Affwild2 development set.**

Method	A and V Backbones	Valence			Arousal		
		Audio	Visual	Fusion	Audio	Visual	Fusion
Kuhnke et al. [28] [FGW 2020]	A: Resnet18; V: R(2plus1)D	0.351	0.449	0.493	0.356	0.565	0.604
Zhang et al. [32] [ICCVW 2021]	A: VGGish; V: Resnet50-TCN	-	0.405	0.457	-	0.635	**0.645**
Rajasekhar et al. [24] [FG 2021]	A: Resnet18; V: I3D-TCN	0.351	0.417	0.552	0.356	0.539	0.531
Joint Cross-Attention (Ours)	A: Resnet18; V: I3D-TCN	0.351	0.417	0.663	0.356	0.539	0.584
Joint Cross-Attention (Ours)	A: Resnet18; V: I3D-TCN + R3D	0.351	-	**0.670**	0.356	-	0.590

V backbones and A-V fusion strategies. First, we have implemented I3D [39] with simple feature concatenation, where A and V features are concatenated, and fed to fully connected layers for valence and arousal prediction. Then we replaced I3D with R3D [38] and implemented a similar fusion strategy of feature concatenation. R3D was found to perform slightly better than I3D for arousal, while I3D shows superior performance for valence. We have also compared the proposed approach with other relevant attention fusion strategies in the literature. We have compared the backbones of I3D with that of leader-follower attention [32] and cross-attention [24]. Compared to the vanilla cross attention model, leader-follower attention was found to perform better.

Finally, to validate the generalization capability of the proposed fusion model, we have implemented various V backbones using I3D, R3D, Resnet18 with GRU, and I3D with TCN. Though the performance of our fusion model varies slightly with different backbones, we can observe that our proposed fusion model can outperform other attention strategies [24, 32], especially for valence. Compared to the 2D-CNN model (Resnet18 with GRU), the 3D-CNNs architectures are found to perform slightly better. Furthermore, I3D provides more improvements over valence than arousal with our fusion model when compared to R3D. By introducing TCN with I3D, the performance of the proposed fusion model is found to perform even better since it can more effectively capture long-term temporal cues than I3D alone. We have further explored the feature-level fusion of V backbones by training a full-connected network to combine I3D-TCN and R3D, which shows a slight improvement over I3D-TCN alone. Resnet18 is used as the backbone for the A modality in all the experiments conducted above.

5.2. Comparison to State-of-Art Methods

Table 2 shows our comparative results against relevant state-of-the-art A-V fusion models on the Affwild2 validation set submitted for the previous challenges [10, 11]. Most relevant approaches have been implemented with different experimental protocols and training strategies. Therefore, to have a fair comparison, we have re-implemented these approaches according to our experimental protocol and analyzed the results on the Affwild2 validation set. Similar to our A and V backbones, Kuhnke et al. [28] also used 3D-CNNs, where the R(2plus1)D model is used for the V modality, and the Resnet18 is used for the A modality. However, they use additional masks for the V modality and annotations of other tasks to refine the annotations

Table 3. **CCC of the proposed approach compared to state-of-the-art methods for A-V fusion on Affwild2 test set.**

Method	Modalities	Valence	Arousal	Mean
Situ-RUCAIM3 [46]	Audio, Visual	0.606	0.596	0.601
FlyingPigs [47]	Audio, Visual, Text	0.520	0.602	0.561
PRL [48]	Visual	0.450	0.445	0.448
HSE-NN [49]	Visual	0.417	0.454	0.436
AU-NO [50]	Audio, Visual	0.418	0.407	0.413
Joint Cross-Attention (Ours)	Audio, Visual	**0.374**	**0.363**	**0.369**
Baseline [16]	Visual	0.180	0.170	0.175

of valence and arousal. They further perform simple feature concatenation without any specialized fusion model to predict valence and arousal. Therefore, the performance with fusion was not significantly improved over the uni-modal performance. Zhang et al. [32] explored the leader-follower attention model for fusion and showed minimal improvement in fusion performance over uni-modal performances. Though they have shown significant performance for arousal than valence, it is mostly attributed to the V backbone. The proposed approach has shown significant improvement for fusion, especially for valence than arousal. Even with vanilla cross attentional fusion [24], we have shown that fusion performance for valence has been improved better than that of [32] and [28]. By deploying joint representation into the cross attentional fusion model, the fusion performance of valence has been significantly improved further. In the case of arousal, though the fusion performance is lower than that of [32] and [28], we can observe that it has improved over that of uni-modal V performance. Therefore, the proposed approach effectively captures the variations spanning a wide range of emotions (valence) than the intensities of the emotions (arousal).

We have further compared our fusion model with that of other valid submissions for the third ABAW challenge [16] on the test set as shown in Table 3. The winner of the challenge [46] also used A-V fusion and demonstrated outstanding performance for both valence and arousal. They used three external datasets to improve the generalization capability of the training model and features from multiple backbones for both V and A modalities. FlyingPigs [47] uses the text modality along with A and V modalities and achieved improvement over A-V fusion using the leader-follower attention strategy. Apart from these, AU-NO [50] is the only approach that relies on A-V fusion. They have investigated the performance of attention mechanisms such as self-attention and cross attention with that of recurrent networks. They have also used additional loss components of mean square error (MSE) and categorical cross-entropy loss along with CCC. PRL [48] and HSE-NN [49] used only visual modality, where [48] used ensemble-based strategy and [49] used external AffectNet dataset [51] for better performance. It is worth mentioning that we have not used any advanced loss components or post-processing operations on predictions using cross-validation, etc., apart from clipping the predictions to the range of [-1,1]. We did not use any external dataset or features from multiple backbones for A and V modalities. The performance of the proposed approach is solely attributed to the efficacy of our fusion model. We observed that the fusion performance has been significantly improved over the uni-modal performances, especially for valence. The proposed fusion model can be further improved using the fusion of multiple A and V backbones either through feature-level or decision-level fusion similar to that of the winner of the challenge [46].

6. Conclusion

This work introduced joint cross-attentional for A-V fusion in video-based dimensional ER, leveraging the intra- and inter-modal relationships across A and V features. In particular, the complementary relationship between A and V features is efficiently captured based on the correlation between the combined A-V features and individual A and V features. By jointly modeling the intra- and inter-modal relationships, features of each modality attend to the other modality as well as itself, resulting in robust A and V feature representations. With the proposed model, A and V backbones are first trained individually for facial (V) and vocal (A) modalities. Then, an attention mechanism based on the correlation between joint and individual features is applied to obtain the attended A and V features. Finally, the attention-weighted features are concatenated and fed to linear connected layers to predict valence and arousal values. The proposed A-V fusion model is validated experimentally on the challenging Affwild2 video dataset, using different A and V backbones. The experimental results have shown that the proposed model achieves superior multimodal performance by effectively fusing A and V modalities.

References

[1] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel. *Emotion Recognition and Its Applications*. 2014. 1

[2] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. 1

[3] D. Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16:363–368, 1992. 1

[4] H. Schlosberg. Three dimensions of emotion. *Psychological Review*, 61:81–88, 1954. 1

[5] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *FG*, 2013. 2

[6] J Kossaifi, R Walecki, Y Panagakis, J Shen, M Schmitt, F Ringeval, J Han, V Pandit, A Toisoul, B Schuller, K Star, E Hajiyev, and M Pantic. Sewa db: A rich database for a-v emotion and sentiment research in the wild. *IEEE Trans Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, 2021. 2

[7] G McKeown, M Valstar, R Cowie, M Pantic, and M Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affective Computing*, 3(1):5–17, 2012. 2

[8] D Kollias, P Tzirakis, M A Nicolaou, A Papaioannou, G Zhao, B Schuller, I Kotsia, and S Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *IJCV*, 127:907–929, 2019. 2

[9] 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

[10] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *FG*, 2020. 2, 7

[11] D Kollias and S Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *ICCVw*, 2021. 2, 6, 7

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

[13] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv:2103.15792*, 2021. 2

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

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

[16] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection and multi-task learning challenges. *arXiv:2202.10659*, 2022. 2, 8

[17] Albert Mehrabian. *Nonverbal Communication*, page 235. Routledge, 09 2017. 2

[18] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. End-to-end multimodal er using deep neural networks. *IEEE J. of Selected Topics in Signal Proc.*, 11(8):1301–1309, 2017. 2, 3, 4, 5

[19] L Schoneveld, A Othmani, and H Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Rec. Letters*, 146:1–7, 2021. 2, 3, 4, 7

[20] Lingfeng Wang, Shisen Wang, Jin Qi, and Kenji Suzuki. A multi-task mean teacher for semi-supervised facial affective behavior analysis. In *ICCV Workshop*, 2021. 2, 3, 4

[21] P Tzirakis, J Chen, S Zafeiriou, and B Schuller. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53, 2021. 2, 3

[22] Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn. Audio-visual attention networks for emotion recognition. In *Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, page 27–32, 2018. 2

[23] Srinivas Parthasarathy and Shiva Sundaram. Detecting expressions with multimodal transformers. In *STL 2021*, pages 636–643, 2021. 2, 3

[24] R. Gnana Praveen, Eric Granger, and Patrick Cardinal. Cross attentional audio-visual fusion for dimensional emotion recognition. In *FG*, 2021. 2, 3, 7, 8

[25] Juan D. S. Ortega, Patrick Cardinal, and Alessandro L. Koerich. Emotion recognition using fusion of audio and video features. In *SMC*, 2019. 3, 4

[26] Dung Nguyen, Duc Thanh Nguyen, Rui Zeng, Thanh Thi Nguyen, Son Tran, Thin Khac Nguyen, S. Sridharan, and Clinton Fookes. Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. *IEEE Trans. on Multimedia*, 2021. 3

[27] Didan Deng, Liang Wu, and Bertram E. Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *ICCV Workshop*, 2021. 3

[28] F Kuhnke, L Rumberg, and J Ostermann. Two-stream aural-visual affect analysis in the wild. In *FGW*, 2020. 3, 4, 7, 8

[29] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *WACV*, 2021. 3

[30] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *ICLR*, 2021. 3

[31] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. Multi-modal continuous valence-arousal estimation in the wild. In *IEEE FG*, 2020. 3

[32] S Zhang, Y Ding, Z Wei, and C Guan. Continuous emotion recognition with audio-visual leader-follower attentive fusion. In *ICCV Workshop*, 2021. 3, 4, 7, 8

[33] Ruibing Hou, Hong Chang, Bingpeng MA, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *NIPS*, 2019. 3

[34] R Gnana Praveen, E. Granger, and P. Cardinal. Deep weakly supervised domain adaptation for pain localization in videos. In *FG*, 2020. 3

[35] Gnana Praveen R, Eric Granger, and Patrick Cardinal. Weakly supervised learning for facial behavior analysis : A review. *arXiv:2101.09858*, 2021. 4

[36] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. on Affective Computing*, 2:92–105, 2011. 4

[37] M Wöllmer, M Kaiser, F Eyben, B Schuller, and G Rigoll. Lstm-modeling of continuous emotions in an a-v affect recognition framework. *IVC*, 31(2):153–163, 2013. 4

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 4, 7

[39] J Carreira and A Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 4, 7

[40] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *AVEC*, 2015. 4

[41] Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai. Emotion recognition from variable-length speech segments using deep learning on spectrograms. In *INTERSPEECH*, 2018. 4

[42] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6, 7

[43] W Wang, D Tran, and M Feiszli. What makes training multimodal classification networks hard? In *CVPR*, 2020. 4

[44] A Nagrani, S Yang, A Arnab, C Schmid, and C Sun. Attention bottlenecks for multimodal fusion. In *NIPS*, 2021. 4

[45] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *ICAIS*, 2010. 6

[46] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, and Qin Jin. Multi-modal emotion estimation for in-the-wild videos. *arXiv:2203.13032*, 2022. 8

[47] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. *arXiv:2203.13031*, 2022. 8

[48] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial expression analysis in video. *arXiv:2203.12891*, 2022. 8

[49] Andrey V. Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv:2203.13436*, 2022. 8

[50] 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. *arXiv:2203.13285*, 2022. 8

[51] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 8