MAVEN: Multi-modal Attention for Valence-Arousal Emotion Network

Vrushank Ahire† 2022csb1002@iitrpr.ac.in
Kunal Shah† 2023aim1003@iitrpr.ac.in
Mudasir Nazir Khan† 2024aim1006@iitrpr.ac.in
Nikhil Pakhale† 2024dss1013@iitrpr.ac.in
Lownish Rai Sookha† lownish.23csz0010@iitrpr.ac.in
M. A. Ganaie† mudasir@iitrpr.ac.in
Abhinav Dhall* abhinav.dhall@monash.edu

†Department of Computer Science and Engineering, Indian Institute of Technology Ropar, Punjab, India
*Department of Data Science and AI, Monash University, Melbourne, Australia

Abstract

Dynamic emotion recognition in the wild remains challenging due to the transient nature of emotional expressions and temporal misalignment of multi-modal cues. Traditional approaches predict valence and arousal and often overlook the inherent correlation between these two dimensions. The proposed Multi-modal Attention for Valence-Arousal Emotion Network (MAVEN) integrates visual, audio, and textual modalities through a bi-directional cross-modal attention mechanism. MAVEN uses modality-specific encoders to extract features from synchronized video frames, audio segments, and transcripts, predicting emotions in polar coordinates following Russell's circumplex model. The evaluation of the Aff-Wild2 dataset using MAVEN achieved a concordance correlation coefficient (CCC) of 0.3061, surpassing the ResNet-50 baseline model with a CCC of 0.22. The multistage architecture captures the subtle and transient nature of emotional expressions in conversational videos and improves emotion recognition in real-world situations.

1. Introduction

Emotion recognition is a critical challenge in affective computing with applications spanning human-computer interaction, healthcare, education, and entertainment. While traditional approaches have focused on categorical emotion classification (e.g., happiness, sadness, anger), dimensional models representing emotions along continuous valence (pleasure-displeasure) and arousal (activation-deactivation) scales have gained prominence for their ability to represent subtle emotional nuances [21, 46]. These dimensional models support the idea that emotions exist on a continuous spectrum, enriching the understanding of the complexity of human emotional experiences.

The detection and analysis of human emotions in natural settings pose significant challenges due to variations in expressions, individual differences, cultural influences, and the often subtle nature of emotional cues [18, 52]. This complexity requires a multi-modal approach, as people naturally express and perceive emotions through various channels, including facial expressions, voice intonation, language, and gestures [5, 53]. Integrating complementary information from each modality enhances the robustness and accuracy of emotion recognition systems. While facial expressions might reveal visible emotional cues, speech prosody can uncover subtle emotional undertones, and linguistic content can provide contextual information essential for accurate interpretation [12, 32].

The field has witnessed significant advancements in emotion recognition through deep learning approaches in recent years. Initial research emphasizes unimodal systems in facial expression analysis using efficient deep learning models like EfficientNet and extracting frame-level features through EmotiEffNet [39, 40]. Approaches such as EmoFAN-VR achieved notable success in controlled environments but often struggled with in-the-wild scenarios characterized by varying illumination, occlusions, and pose variations [8, 21]. Researchers explored other modalities separately, creating specialized models for emotion recognition in speech using spectral features and recurrent architectures such as Vesper [4] and text-based systems using natural language processing techniques.

The shift towards multi-modal systems reflects how humans perceive and express emotions through various channels simultaneously [53]. Initial multi-modal approaches employed simple fusion strategies, such as feature concatenation or decision-level integration, which failed to capture the inter-dependencies between modalities [5]. Advanced techniques were developed, including tensor-based fusion, specifically Tucker Tensor Regression and Tensor Regression Networks [34], bilinear pooling, and various ensemble techniques like EVAEF [29] . While these approaches have enhanced performance, they often treat different modalities as independent sources of information. To address this, Meng et al. [33] uses temporal encoders, specifically transformer-based and Long Short-Term Memory (LSTM)-based networks, to better understand how emotions evolve over time in a video.

Attention mechanisms have emerged as a promising approach to address these limitations by enabling models to focus on relevant features across modalities. Early attention-based methods primarily implemented self-attention within individual modalities or simple cross-modal attention between pairs of modalities. Praveen et al. [37] implemented a joint cross-attention fusion model, and Zhang et al. [54] introduced TEMMA, a multi-head attention module. These approaches demonstrated improved performance but created limited pathways for information exchange, failing to utilize the complementary nature of multi-modal data completely. Most existing works have concentrated on categorical emotion recognition [16], with comparatively less exploration of attention mechanisms for continuous valence-arousal prediction.

The Affective Behavior Analysis in-the-wild (ABAW) competitions have significantly advanced research in this domain by providing standardized benchmarks and challenging in-the-wild datasets [13, 24, 26, 27]. Despite advancements, several significant challenges still remain unaddressed. First, existing fusion approaches often struggle to combine information across modalities while maintaining temporal coherence effectively [6]. Second, attention mechanisms have not been fully exploited to capture the complex inter-relationships between different emotional cues [51]. Third, the direct regression of valence-arousal values in Cartesian coordinates may not optimally align with psychological models of emotion [46]. We propose a novel Multi-modal Attention for Valence-Arousal Emotion Network (MAVEN) with several key innovations to address these limitations. The key contributions are:

- We employ state-of-the-art (SOTA) modality-specific encoders: Swin Transformer for visual data, HuBERT for audio, and RoBERTa for text, to extract robust feature representations from each data stream.
- The proposed model employs a cross-modal attention mechanism that uses six distinct attention pathways and enables interactions between all modality pairs through weighted attention from other modalities. This design allows each modality to both inform and be informed by others, creating a rich information exchange network.

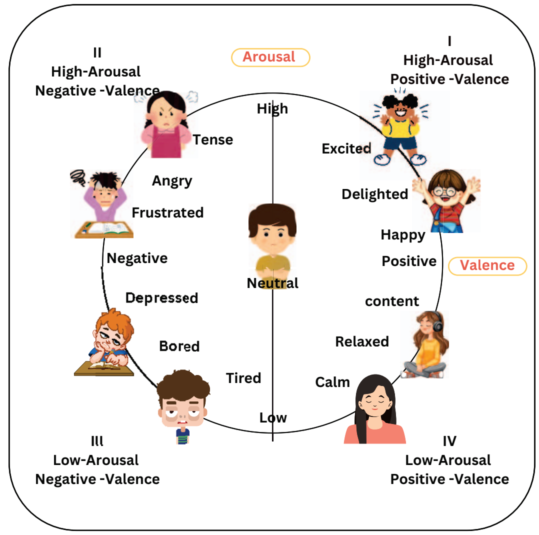


Figure 1. Valence-Arousal Emotion Circumplex

- This is followed by self-attention within its modality-specific encoder, utilizing a multi-headed attention module akin to that used in the Bidirectional Encoder Representations from Transformers (BEiT) model.
- Additionally, we exploit polar coordinate form, representing emotions as angle (θ) and intensity (I) rather than directly predicting valence-arousal values. This representation aligns naturally with psychological models of the emotion circumplex [38].
- Extensive experiments on the Aff-Wild2 dataset [18] demonstrate that the proposed approach significantly outperforms existing methods as measured by the Concordance Correlation Coefficient (CCC) [19].

The results validate the effectiveness of our bidirectional multi-modal attention architecture and polar coordinate prediction framework for capturing complex emotional expressions in conversational videos.

This paper is organized as follows: Section 2 reviews related work on emotion recognition, multi-modal fusion, and attention mechanisms. Section 3 presents our proposed methodology, which includes the multi-modal attention architecture and the polar coordinate prediction framework. Section 4 details the experimental setup, results based on the Aff-Wild2 dataset, and the ablation studies and analyzes the contributions of individual components. Finally, Section 5 concludes the paper and discusses future directions.

2. Related Work

This section reviews the evolution of emotion recognition approaches, starting with traditional unimodal methods, progressing to multi-modal fusion techniques, and examining the role of attention mechanisms. It also explores the shift from categorical to dimensional models of emotion representation.

2.1. Unimodal Emotion Recognition

Early research in emotion recognition often focused on single modalities such as visual cues, audio signals, or textual content.

2.1.1. Visual-Based Approaches

Visual emotion recognition primarily focuses on analyzing facial expressions and body gestures. Early methods relied on handcrafted features to capture spatio-temporal information, but these often required significant prior knowledge and demonstrated a disconnect between the extracted features and actual emotional expressions. Recently, deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have become the preferred approach [12, 54, 55]. For facial expressions, models like SISTCM (Super Image-based Spatio-Temporal Convolutional Model) [44] and two-stream LSTM models have been developed to capture local spatio-temporal features and global temporal cues related to emotional changes, with SISTCM utilizing 2D convolution for efficiency. The advancements in body gesture recognition have emerged through representation methods based on body joint movements, including the Attention-based Channel-wise Convolutional Model (ACCM), which employs channel-wise convolution and attention mechanisms to learn key joint features. Beyond facial expressions and body gestures, other visual cues like eye movement and body posture also play a significant role in emotion recognition [7].

2.1.2. Audio-Based Approaches

Audio-based emotion recognition, also known as Speech Emotion Recognition (SER), involves analyzing speech signals to identify emotional states. Traditional methods focused on extracting low-level features such as Mel-frequency cepstral coefficients (MFCCs), pitch, and energy [4]. Statistical analysis was then frequently applied to these handcrafted features. However, with the advancement of deep learning, techniques such as Deep Neural Networks (DNNs), CNNs using spectrograms or MFCCs as inputs, and RNNs like LSTM networks and Gated Recurrent Units (GRUs) have demonstrated significant improvements in SER performance. Additionally, attention mechanisms have gained popularity in this field, allowing models to focus on the most important parts of the speech signal.

2.1.3. Text-Based Approaches

Text-based emotion recognition, often referred to as sentiment analysis, examines written text to identify the emotions or sentiments expressed within it. Early methods primarily relied on lexicon-based approaches and traditional machine learning classifiers, such as Support Vector Machines, which were trained using features like bag-of-words or Term Frequency - Inverse Document Frequency (TF-IDF). Deep learning techniques have revolutionized text-based emotion recognition. CNNs can extract local patterns, while RNNs, LSTMs, and GRUs- effectively capture sequential dependencies in text. Pre-trained word embeddings like Word2Vec and GloVe are used to represent words semantically. Transformer networks have also achieved SOTA results in text-based emotion recognition by effectively modeling long-range dependencies and contextual -modal approaches, which can be affected by noise, occlusions, or the masking of emotions [21], researchers have increasingly turned their attention to multi-modal emotion recognition (MER) [5]. MER utilizes information from various modalities, including audio, visual (such as facial expressions and body gestures), and text, to achieve more accurate emotion recognition. The fundamental principle behind this approach is that different modalities can provide complementary insights into human emotions.

Various multi-modal fusion techniques have been explored:

- Early Fusion (Feature Fusion): This approach involves concatenating features extracted from different modalities at an early stage and then feeding the combined feature vector into a classifier. For example, audio and visual features might be extracted using CNNs and then combined before being processed by an RNN.
- Late Fusion (Decision Fusion)**:** In late fusion, separate classifiers are trained for each modality, and their individual predictions (e.g., probability scores) are then combined using methods like averaging, weighted averaging, or voting to make a final emotion prediction.
- Hybrid Approaches: These methods combine aspects of early and late fusion. For instance, features from different modalities might be processed separately by subnetworks, and the resulting intermediate representations are then fused before the final classification layer.

Deep learning models have been instrumental in developing advanced multi-modal fusion strategies. Techniques like concatenation, element-wise addition or multiplication, and more sophisticated methods using attention mechanisms or bilinear pooling are commonly employed to integrate information from multiple modalities.

2.2. Attention Mechanisms for Emotion Recognition

Attention mechanisms have become a crucial component in modern emotion recognition systems, enabling models to selectively focus on the most relevant parts of the input data across different modalities and time steps [26, 29]. In visual emotion recognition, attention can help the model focus on emotion-relevant facial regions or specific body joints. In

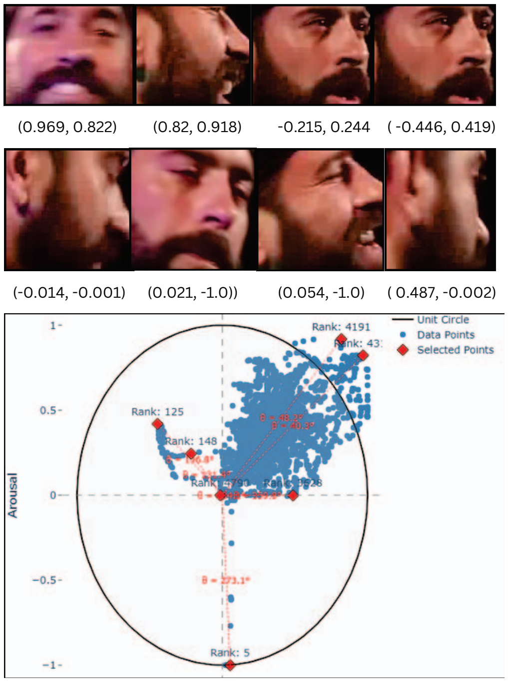


Figure 2. Valence-Arousal Distribution for a video sample from the Aff-Wild2 dataset

audio processing, attention mechanisms can weigh different parts of the speech signal based on their importance for emotion recognition. For text analysis, attention allows the model to identify the words that are most indicative of the expressed emotion.

In multi-modal emotion recognition, attention mechanisms play a vital role in learning the inter-modal relationships and determining the contribution of each modality to the final prediction. Cross-attention mechanisms allow the model to attend to relevant information in one modality based on the content of another modality, facilitating efficient information fusion. Self-attention mechanisms enable the model to capture intra-modal dependencies by attending to different parts of the same modality. Multi-head attention, as used in Transformer networks [6], allows the model to attend to different aspects of the input simultaneously.

2.3. Dimensional Models of Emotion

Early research in emotion recognition primarily focused on identifying a specific set of basic emotions, such as anger, disgust, fear, happiness, sadness, surprise, and sometimes neutrality [54]. However, researchers have shifted towards dimensional models of emotion [34, 43]. These models depict emotions as points within a continuous space characterized by dimensions like valence (pleasantness) and arousal (intensity) [52]. Additionally, some models incorporate a third dimension, which is dominance (the level of control).

The Valence-Arousal (VA) space is the most commonly used dimensional representation of emotions. In this model, valence ranges from negative to positive, while arousal varies from passive to active. This two-dimensional framework allows for a circumplex representation of emotions, where different specific emotions can be mapped to distinct regions within the space. The Affective Behavior Analysis in-the-wild (ABAW) competitions have played a significant role in advancing valence-arousal estimation in real-world scenarios by providing large-scale annotated datasets, such as Aff-Wild and Aff-Wild2 [13, 24, 26–28].

Building on previous advancements, we propose MAVEN, a novel multi-modal attention framework for valence-arousal estimation. The next section describes our proposed model.

3. Proposed Model: MAVEN

This section details MAVEN (Multi-modal Attention for Valence-arousal Emotion Network), our novel approach to continuous emotion recognition in conversational videos. Figure 3 illustrates the overall architecture of our proposed model.

3.1. Overview

MAVEN integrates information from three modalities: visual, audio, and text, employing an attention mechanism that enhances the exchange of information between these modalities. The system consists of five key components:

- Modality-specific feature extractors for visual, audio, and textual data
- A comprehensive cross-modal attention mechanism with bidirectional information flow
- A bidirectional multi-headed self-attention module for each modality
- A BEiT-based encoder refinement layer
- A polar coordinate-based emotion prediction framework

Modality-Specific Feature Extraction

3.2. Visual Features

We employ the Swin Transformer [31] to extract visual features from facial regions in each video frame. The Swin Transformer combines local attention with shifted windows, enabling efficient modeling of hierarchical visual features while maintaining linear computational complexity relative to image size.

3.2.1. Patch Embedding

Given a video sequence $\{V_1, V_2, ..., V_T\}$ with T frames, each frame $V_i \in \mathbb{R}^{H \times W \times 3}$ is split into non-overlapping patches of size $P \times P$:

$$X_p = \text{PatchPartition}(V_i) \in \mathbb{R}^{\frac{HW}{P^2} \times (P^2 \cdot 3)} \tag{1}$$

These patches are then projected into a d_v-dimensional embedding space:

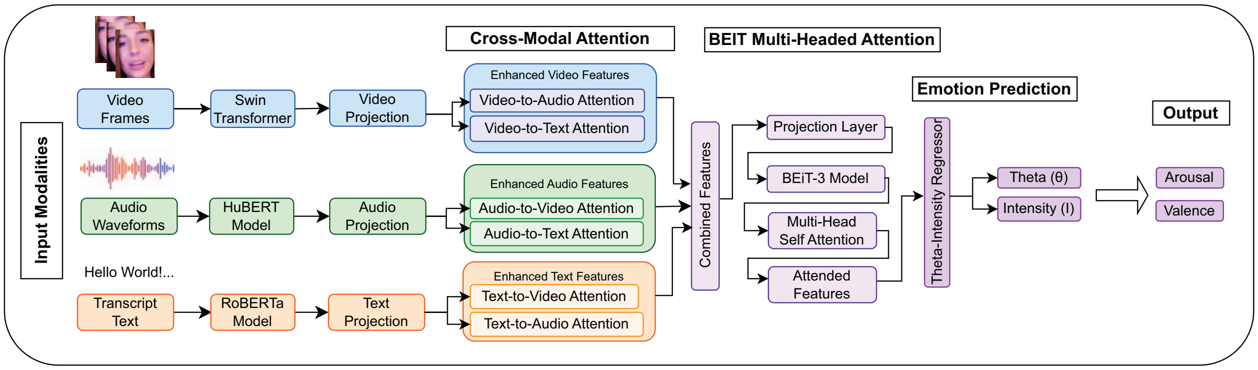


Figure 3. Multi-modal Attention for Valence-Arousal Emotion Network (MAVEN) Architecture. The model processes visual, audio, and text inputs through specialized feature extractors, fuses information via cross-modal attention pathways, refines representations with bidirectional self-attention, and predicts emotions using a polar coordinate system.

$$Z_0 = X_p W_p + b_p, \quad Z_0 \in \mathbb{R}^{\frac{HW}{P^2} \times d_v} \tag{2}$$

where $W_p \in \mathbb{R}^{(P^2 \cdot 3) \times d_v}$ and $b_p \in \mathbb{R}^{d_v}$ are learnable parameters.

3.2.2. Shifted Window Multi-Head Self-Attention (SW-MSA)

To model long-range dependencies efficiently, attention is applied within local windows that shift across layers to capture cross-window interactions:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{3}$$

where $Q, K, V = Z_{l-1}W_Q, Z_{l-1}W_K, Z_{l-1}W_V$, with $d_k = d_v/h$ and h as the number of attention heads.

3.2.3. Visual Feature Output

After processing through L transformer layers with hierarchical merging, we obtain the final visual feature sequence:

$$F_V = \text{SwinT}(V_1, V_2, ..., V_T) \in \mathbb{R}^{T \times d_v} \tag{4}$$

where F_V represents the sequence of visual features with dimension $d_v = 1024$.

3.3. Audio Features

For audio processing, we utilize HuBERT [10], a self-supervised speech representation learning model with superior performance in capturing acoustic characteristics relevant to emotion recognition.

3.3.1. Feature Extraction

Given the audio signal $A(t)$ corresponding to the video frames, we first compute log-mel spectrogram features:

$$X_A = \text{MelFilterBank}(\text{STFT}(A(t))) \in \mathbb{R}^{T' \times F}, \tag{5}$$

where T' is the number of time frames in the spectrogram and F is the number of mel frequency bands.

where T' is the number of time frames in the spectrogram and F is the number of mel frequency bands.

3.3.2. Audio Encoder

A CNN maps X_A to initial hidden representations:

$$Z_A = E_a(X_A) \in \mathbb{R}^{T'' \times d_a} \tag{6}$$

where T'' is the reduced temporal dimension and d_a is the feature dimension.

3.3.3. Audio Feature Output

After processing through the HuBERT model, we obtain the final audio feature sequence:

$$F_A = \text{HuBERT}(A_1, A_2, ..., A_T) \in \mathbb{R}^{T \times d_a} \tag{7}$$

where F_A represents the sequence of audio features with dimension $d_a = 768$.

3.4. Text Features

We employ RoBERTa [30], a robust variant of BERT, to extract semantic features from transcribed speech, capturing emotional markers from linguistic content.

3.4.1. Token Embedding

For a sequence of tokens $\{t_1, t_2, ..., t_S\}$ derived from the transcribed speech, we compute embeddings as:

$$E_t = E_{\text{token}}(t_i) + E_{\text{pos}}(i), \quad E_t \in \mathbb{R}^{S \times d_t} \tag{8}$$

where E_{token} is the token embedding function, E_{pos} is the positional embedding function, and d_t is the embedding dimension.

3.4.2. Multi-Head Self-Attention

The token embeddings are processed through multiple layers of self-attention:

$$\mathrm{MHA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)W_O, \quad (9)$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (10)$$

where each attention head computes the above for $i = 1,\ldots,h$.

3.4.3. Text Feature Output

The processed textual features are obtained as:

$$F_T = \mathrm{RoBERTa}(t_1, t_2, ..., t_S) \in \mathbb{R}^{S\times d_t} \quad (11)$$

where F_T represents the sequence of text features with dimension $d_t = 768$.

To align text features with visual and audio features that have temporal dimension T, we apply temporal interpolation:

$$F_T' = \mathrm{TemporalInterpolate}(F_T) \in \mathbb{R}^{T\times d_t} \quad (12)$$

3.5. Cross-Modal Attention

The core contribution of our approach is the bidirectional cross-modal attention mechanism, which enables information exchange between all pairs of modalities. This mechanism generates six distinct attention pathways: (visual $\rightarrow$ audio, visual $\rightarrow$ text, audio $\rightarrow$ visual, audio $\rightarrow$ text, text $\rightarrow$ visual, text $\rightarrow$ audio). This allows each modality to both provide information to and receive information from the others.

For each modality pair (m,n) where $m,n \in \{\mathrm{visual},\mathrm{audio},\mathrm{text}\}$ and $m \neq n$, we compute cross-modal attention as follows:

$$\text{Cross-Attention}(Q_m, K_n, V_n) = \mathrm{softmax}\left(\frac{Q_m K_n^T}{\sqrt{d_k}}\right) V_n \quad (13)$$

where:

- $Q_m = F_m W_Q^m$ is the query matrix derived from modality m
- $K_n = F_n W_K^n$ is the key matrix derived from modality n
- $V_n = F_n W_V^n$ is the value matrix derived from modality n
- $W_Q^m \in \mathbb{R}^{d_m\times d_k}$, $W_K^n \in \mathbb{R}^{d_n\times d_k}$, and $W_V^n \in \mathbb{R}^{d_n\times d_v}$ are learnable parameter matrices
- d_k is the dimension of the keys and queries

For example, the attention from audio to visual features is computed as:

$$A_{A\rightarrow V} = \text{Cross-Attention}(W_Q^A F_A, W_K^V F_V, W_V^V F_V) \quad (14)$$

Similarly, we compute attention for the other five directional pathways, where $A_{V\rightarrow A}$, $A_{V\rightarrow T}$, $A_{A\rightarrow T}$, $A_{T\rightarrow V}$, and $A_{T\rightarrow A}$ are obtained via cross-attention mechanisms using modality-specific query, key, and value projections of F_V, F_A, and F_T', with dimensions $A_{A\rightarrow V}, A_{T\rightarrow V} \in \mathbb{R}^{T\times d_v}$, $A_{V\rightarrow A}, A_{T\rightarrow A} \in \mathbb{R}^{T\times d_a}$, and $A_{V\rightarrow T}, A_{A\rightarrow T} \in \mathbb{R}^{T\times d_t}$.

3.6. BEiT-based Encoder Refinement

Following cross-modal attention, we apply a two-step refinement process: (1) Bidirectional Self-Attention Refinement, where intra-modal dependencies are strengthened, and (2) BEiT-based Encoder Refinement, which further contextualizes and integrates multi-modal features. To enhance intra-modal relationships, we apply self-attention within each modality:

$$F_V^{\text{refined}} = \mathrm{SelfAtt}(F_V^{\text{enhanced}}) \in \mathbb{R}^{T\times d_v}, \quad (15)$$

$$F_A^{\text{refined}} = \mathrm{SelfAtt}(F_A^{\text{enhanced}}) \in \mathbb{R}^{T\times d_a}, \quad (16)$$

$$F_T^{\text{refined}} = \mathrm{SelfAtt}(F_T^{\text{enhanced}}) \in \mathbb{R}^{T\times d_t}. \quad (17)$$

The self-attention module employs a multi-head attention mechanism followed by residual connections and layer normalization:

$$\mathrm{SelfAtt}(F) = \mathrm{LayerNorm}\left(\mathrm{MHA}(F,F,F) + F\right) \quad (18)$$

where:

- $\mathrm{MHA}(\cdot)$ denotes the multi-head self-attention operation,
- F represents the input feature matrix for a given modality (F_V, F_A, F_T),
- $\mathrm{LayerNorm}(\cdot)$ ensures stable training by normalizing the updated features.

After intra-modal refinement, the features are concatenated and projected into a unified multi-modal representation:

$$F_{concat} = \mathrm{Concat}(F_V^{refined}, F_A^{refined}, F_T^{refined}) \quad (19)$$

$$F_{projected} = F_{concat} W_{proj} + b_{proj} \in \mathbb{R}^{T\times d_{proj}} \quad (20)$$

where $F_{concat} \in \mathbb{R}^{T\times(d_v+d_a+d_t)}$, $W_{proj} \in \mathbb{R}^{(d_v+d_a+d_t)\times d_{proj}}$, and $b_{proj} \in \mathbb{R}^{d_{proj}}$ are learnable parameters.

To capture long-range dependencies and contextualized representations, we employ two BEiT-based transformer encoders [2]. The attention mechanism is formulated as:

$$\mathbf{d}_{t,t'} = \tanh(\mathbf{W}_d \mathbf{h}_t + \mathbf{W}'_d \mathbf{h}_{t'} + \mathbf{b}_d), \tag{21}$$

$$\alpha_{t,t'} = \mathbf{v}_d^T \mathbf{d}_{t,t'}, \tag{22}$$

$$a_{t,t'} = \text{softmax}(\alpha_{t,t'}), \tag{23}$$

$$\mathbf{l}_t = \sum_{t'=1}^{T} a_{t,t'} \mathbf{h}_{t'}. \tag{24}$$

where:

- $\mathbf{h}_t$ and $\mathbf{h}_{t'}$ are hidden states at positions t and t',
- $\mathbf{W}_d$, $\mathbf{W}'_d$, $\mathbf{b}_d$, and $\mathbf{v}_d$ are learnable parameters,
- $\mathbf{l}_t$ is the refined representation at position t.

The final multi-modal feature representation is obtained after passing through two consecutive BEiT-based encoders:

$$F_{refined} = \text{BEiT}_2(\text{BEiT}_1(F_{projected})) \in \mathbb{R}^{T \times d_{beit}} \tag{25}$$

This process ensures that the final multi-modal embedding captures context-aware emotional cues, allowing for robust downstream emotion prediction.

3.7. Polar Coordinate Emotion Prediction

The final component of our proposed architecture is a specialized emotion prediction framework that operates in polar coordinates, aligning with the valence-arousal emotion circumplex model widely used in affective computing.

3.7.1. Feature Pooling

First, we apply temporal average pooling to obtain a fixed-dimensional representation:

$$F_{pooled} = \frac{1}{T} \sum_{t=1}^{T} F_{refined}[t] \in \mathbb{R}^{d_{beit}} \tag{26}$$

3.7.2. Emotion Predictor

A feedforward neural network with three fully connected layers and ReLU activations processes the pooled features:

$$h_1 = \text{ReLU}(F_{\text{pooled}} W_1 + b_1) \in \mathbb{R}^{d_1}, \tag{27}$$

$$h_2 = \text{ReLU}(h_1 W_2 + b_2) \in \mathbb{R}^{d_2}, \tag{28}$$

$$[I, \theta] = h_2 W_3 + b_3 \in \mathbb{R}^2. \tag{29}$$

where $W_1 \in \mathbb{R}^{d_{beit} \times d_1}$, $W_2 \in \mathbb{R}^{d_1 \times d_2}$, $W_3 \in \mathbb{R}^{d_2 \times 2}$, $b_1 \in \mathbb{R}^{d_1}$, $b_2 \in \mathbb{R}^{d_2}$, and $b_3 \in \mathbb{R}^2$ are learnable parameters.

3.7.3. Polar Conversion

The network predicts two values: intensity I and angle θ. These are converted to valence and arousal coordinates:

$$\text{valence} = I \cdot \cos(\theta) \tag{30}$$

$$\text{arousal} = I \cdot \sin(\theta) \tag{31}$$

This parametrization aligns with psychological models of emotion, where similar emotions are adjacent in the emotion circumplex. It also enforces the constraint that emotions with similar valence-arousal values have similar representations in the model's internal space.

3.8. Training and Optimization

The model is trained to minimize the Mean Squared Error (MSE) between predicted and ground-truth valence-arousal values:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} [(v_i - \hat{v}_i)^2 + (a_i - \hat{a}_i)^2] \tag{32}$$

where v_i and a_i are the ground-truth valence and arousal values, $\hat{v}_i$ and $\hat{a}_i$ are the predicted values, and N is the number of samples.

We employ the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-2}. To prevent overfitting, we apply dropout with a rate of 0.2 after each fully connected layer in the emotion predictor.

4. Experiments and Results

In this section, we present the experimental results to evaluate the performance of the proposed model. The analysis focuses on understanding the effectiveness of the model across CCC metric.

4.1. Dataset

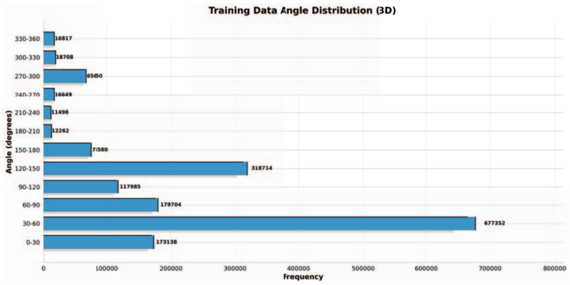
We evaluate our approach on the Aff-Wild2 dataset , the largest in-the-wild audiovisual database for valence-arousal estimation, containing 545 videos with 2.8M frames, each annotated with continuous valence and arousal values in [-1, 1] [28]. Following the ABAW competition's train/validation/test split, we extract facial regions using a face detector, aligning them to 112×112 pixels. Audio is sampled at 48kHz, and text transcriptions are generated via automatic speech recognition.

4.2. Implementation Details

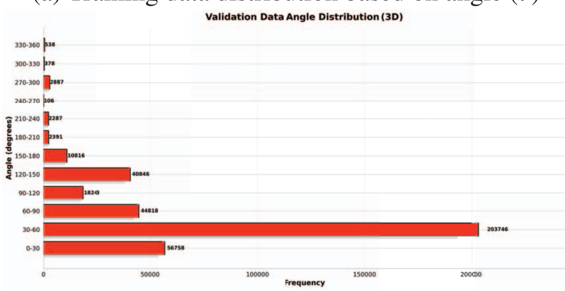
We train the model for 200 epochs using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 16, and a weight decay of 1×10^{-3}. The training is conducted on NVIDIA A100 GPUs.

4.3. Evaluation Metrics

Following the standard protocol in the ABAW competition, we use the Concordance Correlation Coefficient (CCC) as the primary evaluation metric:



(a) Training data distribution based on angle (θ)



(b) Validation data distribution based on angle (θ)

Figure 4. Example frames from the Aff-Wild2 dataset showing facial region extraction and alignment.

$$\text{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where ρ is the Pearson correlation coefficient, σ_x and σ_y are the standard deviations, and μ_x and μ_y are the means of the predicted and ground truth values, respectively.

The overall performance is measured by the average CCC of valence and arousal:

$$\text{CCC}_{avg} = \frac{\text{CCC}_v + \text{CCC}_a}{2}$$

4.4. Results and Comparison

Table 1 compares MAVEN's performance against the ABAW ResNet-50 baseline.

Table 1. Results and Comparison of Different Models

Model	Valence CCC	Arousal CCC	CCC Avg
Baseline [28]	0.2400	0.2000	0.2200
MAVEN (*proposed*)	**0.3068**	**0.3054**	**0.3061**

MAVEN achieves a CCC Avg of 0.3061, surpassing the baseline model's 0.22, demonstrating the effectiveness of the multi-modal attention framework.

4.5. Ablation Study

To analyze MAVEN's design choices, we conduct ablation studies by removing individual modalities and evaluating their impact. Table 2 presents the performance of uni-modal variants on the Aff-Wild2 dataset, measured using the Concordance Correlation Coefficient (CCC).

Table 2. Results and Comparison of Different Models

Model Variant	Valence CCC	Arousal CCC	CCC Avg
Visual only	0.1048	0.2459	0.1754
Audio only	0.1283	0.0683	0.0299
Text only	0.0019	0.0006	0.0013

The contributions of each modality are as follows:

- **Visual Modality**: Achieves the highest average CCC (0.1754), excelling in arousal detection (0.2459) by capturing subtle facial cues like widened eyes or furrowed brows, critical for emotion recognition in conversational videos.
- **Audio Modality**: Yields a lower average CCC (0.0299), with better valence performance (0.1283) through prosody and pitch, but struggles in noisy in-the-wild settings, limiting its reliability.
- **Text Modality**: Provides minimal impact (CCC Avg: 0.0013), offering contextual cues to disambiguate emotions like sarcasm, though sparse emotional markers in transcripts constrain its contribution.

These results highlight the necessity of multi-modal integration, as MAVEN's full model achieves a significantly higher CCC (0.3061), leveraging the synergistic strengths of all modalities.

5. Conclusion

This paper introduces MAVEN, a multi-modal attention architecture for valence and arousal recognition. It features a bidirectional cross-modal attention mechanism for effective information exchange, a self-attention refinement module for improved modality specificity, and a polar coordinate prediction framework for intuitive emotional representation. Achieving a SOTA CCC of **0.3061**, the model demonstrates that both the attention mechanism and polar coordinate framework significantly enhance emotion recognition by effectively utilizing complementary multi-modal cues. Future work for our proposed methodology lies in enhancing the temporal modeling capabilities of the framework to better capture the dynamics of emotion evolution.

References

[1] Abhinav Anthiyur Aravindan, Sriram Kalyan Chappidi, Anirudh Thumma, and Rohini Palanisamy. Prediction of arousal and valence state from electrodermal activity using wavelet based resnet50 model. *Current Directions in Biomedical Engineering*, 9(1):555–558, 2023.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 7

[3] JX Chen, PW Zhang, ZJ Mao, YF Huang, DM Jiang, and YN Zhang. Accurate eeg-based emotion recognition on combined features using deep convolutional neural networks. *Ieee Access*, 7:44317–44328, 2019.

[4] Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2024. 1, 3

[5] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, 8:168865–168878, 2020. 1, 2, 3

[6] Denis Dresvyanskiy, Maxim Markitantov, Jiawei Yu, Heysem Kaya, and Alexey Karpov. Multi-modal arousal and valence estimation under noisy conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4773–4783, 2024. 2, 4

[7] Cunhang Fan, Heng Xie, Jianhua Tao, Yongwei Li, Guanxiong Pei, Taihao Li, and Zhao Lv. Icaps-reslstm: Improved capsule network and residual lstm for eeg emotion recognition. *Biomedical Signal Processing and Control*, 87:105422, 2024. 3

[8] Tom Gotsman, Neophytos Polydorou, and Abbas Edalat. Valence/arousal estimation of occluded faces from vr headsets. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pages 96–105. IEEE, 2021. 1

[9] Sebastian Handrich, Laslo Dinges, Frerk Saxen, Ayoub Al-Hamadi, and Sven Wachmuth. Simultaneous prediction of valence/arousal and emotion categories in real-time. In *2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 176–180. IEEE, 2019.

[10] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460, 2021. 5

[11] Stephen Khor Wen Hwooi, Alice Othmani, and Aznul Qalid Md Sabri. Deep learning-based approach for continuous affect prediction from facial expression images in valence-arousal space. *IEEE Access*, 10:96053–96065, 2022.

[12] Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Saddik, and Soonil Kwon. Mser: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, 245:122946, 2024. 1, 3

[13] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 2, 4

[14] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023.

[15] Dimitrios Kollias and Stefanos Zafeiriou. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv preprint arXiv:1811.07771*, 2018.

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

[17] Dimitrios Kollias and Stefanos Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 12(3):595–606, 2020.

[18] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1, 2

[19] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 2

[20] 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.

[21] 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*, 127(6):907–929, 2019. 1, 3

[22] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020.

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

[24] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5889–5898, 2023. 2, 4

[25] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2813–2821, 2024.

[26] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Alice Baird, Chris Gagne, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4587–4598, 2024. 2, 3, 4

[27] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Abhinav Dhall, Shreya Ghosh, Chunchang Shao, and Guanyu Hu. 7th abaw competition: Multi-task learning and compound expression recognition. *arXiv preprint arXiv:2407.03835*, 2024. 2

[28] Dimitrios Kollias, Panagiotis Tzirakis, AS Cowen, S Zafeiriou, I Kotsia, Eric Granger, Marco Pedersoli, SL Bacon, Alice Baird, C Gagne, et al. Advancements in affective and behavior analysis: The 8th abaw workshop and competition. 2025. 4, 7, 8

[29] Xiaolong Liu, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Yuanyuan Deng, Zhaopei Huang, Liyu Meng, Yuchen Liu, and Chuanhe Liu. Evaef: Ensemble valence-arousal estimation framework in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5863–5871, 2023. 2, 3

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[32] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, and Qin Jin. Multimodal emotion estimation for in-the-wild videos. *arXiv preprint arXiv:2203.13032*, 2022. 1

[33] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Chuanhe Liu, and Qin Jin. Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2345–2352, 2022. 2

[34] Anna Mitenkova, Jean Kossaifi, Yannis Panagakis, and Maja Pantic. Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 2, 4

[35] Wenxuan Mou, Oya Celiktutan, and Hatice Gunes. Group-level arousal and valence recognition in static images: Face, body and context. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2015.

[36] João Ribeiro Pinto, Tiago Gonçalves, Carolina Pinto, Luís Sanhudo, Joaquim Fonseca, Filipe Gonçalves, Pedro Carvalho, and Jaime S Cardoso. Audiovisual classification of group emotion valence using activity recognition networks. In *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 114–119. IEEE, 2020.

[37] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, Patrick Cardinal, et al. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2486–2495, 2022. 2

[38] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 2

[39] Andrey V Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv preprint arXiv:2203.13436*, 2022. 1

[40] Andrey V Savchenko. Emotieffnets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5716–5724, 2023. 1

[41] Andrey V Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, 2024.

[42] Andrey V Savchenko. Hsemotion team at abaw-8 competition: Audiovisual ambivalence/hesitancy, emotional mimicry intensity and facial expression recognition. *arXiv preprint arXiv:2503.10399*, 2025.

[43] Arman Savran, Ruben Gur, and Ragini Verma. Automatic detection of emotion valence on faces using consumer depth cameras. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 75–82, 2013. 4

[44] Huihui Song, Qingshan Liu, Guojie Wang, Renlong Hang, and Bo Huang. Spatiotemporal satellite image fusion using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):821–829, 2018. 3

[45] Rajkumar Theagarajan, Bir Bhanu, and Albert Cruz. Deepdriver: Automated system for measuring valence and arousal in car driver videos. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2546–2551. IEEE, 2018.

[46] Gyanendra K Verma and Uma Shanker Tiwary. Affect representation and recognition in 3d continuous valence–arousal–dominance space. *Multimedia Tools and Applications*, 76: 2159–2183, 2017. 1, 2

[47] Jie Wei, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, and Yizhuo Dong. Learning facial expression and body gesture visual information for video emotion recognition. *Expert Systems with Applications*, 237:121419, 2024.

[48] Hong-Xia Xie, I Li, Ling Lo, Hong-Han Shuai, Wen-Huang Cheng, et al. Technical report for valence-arousal estimation in abaw2 challenge. *arXiv preprint arXiv:2107.03891*, 2021.

[49] Miku Yanagimoto and Chika Sugimoto. Recognition of persisting emotional valence from eeg using convolutional neural networks. In *2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 27–32. IEEE, 2016.

[50] Hao-Chun Yang and Chi-Chun Lee. An attribute-invariant variational learning for emotion recognition using physiology. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1184–1188. IEEE, 2019.

[51] Jun Yu, Gongpeng Zhao, Yongqi Wang, Zhihong Wei, Zerui Zhang, Zhongpeng Cai, Guochen Xie, Jichao Zhu, Wangyuan Zhu, Shuoping Yang, et al. Improving valence-arousal estimation with spatiotemporal relationship learning and multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7878–7885, 2024. 2

[52] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 1, 4

[53] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237: 121692, 2024. 1

[54] Ziyang Zhang, Liuwei An, Zishun Cui, Ao Xu, Tengteng Dong, Yueqi Jiang, Jingyi Shi, Xin Liu, Xiao Sun, and Meng Wang. Abaw5 challenge: A facial affect recognition approach utilizing transformer encoder and audiovisual fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2023. 2, 3, 4

[55] Weiwei Zhou, Chenkun Ling, and Zefeng Cai. Emotion recognition with clip and sequential learning. *arXiv preprint arXiv:2503.09929*, 2025. 3