Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

Jointly Learning from Unimodal and Multimodal-Rated Labels in Audio-Visual Emotion Recognition

Lucas Goncalves[1], Huang-Cheng Chou[2], Ali N. Salman[1], Student Member, IEEE, Chi-Chun Lee[2], Senior Member, IEEE, and Carlos Busso[1,3], Fellow, IEEE

[1]The University of Texas at Dallas, Richardson, TX 75080 USA
[2]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[3]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Corresponding author: Carlos Busso (email: busso@cmu.edu).

This study was funded by the National Science Foundation (NSF) under grant CNS-2016719.

ABSTRACT *Audio-visual emotion recognition* (AVER) has been an important research area in *human-computer interaction* (HCI). Traditionally, audio-visual emotional datasets and corresponding models derive their ground truths from annotations obtained by raters after watching the audio-visual stimuli. This conventional method, however, neglects the nuanced human perception of emotional states, which varies when annotations are made under different emotional stimuli conditions—whether through unimodal or multimodal stimuli. This study investigates the potential for enhanced AVER system performance by integrating diverse levels of annotation stimuli, reflective of varying perceptual evaluations. We propose a two-stage training method to train models with the labels elicited by audio-only, face-only, and audio-visual stimuli. Our approach utilizes different levels of annotation stimuli according to which modality is present within different layers of the model, effectively modeling annotation at the unimodal and multimodal levels to capture the full scope of emotion perception across unimodal and multimodal contexts. We conduct the experiments and evaluate the models on the CREMA-D emotion database. The proposed methods achieved the best performances in macro-/weighted-F1 scores. Additionally, we measure the model calibration, performance bias, and fairness metrics considering the age, gender, and race of the AVER systems.

INDEX TERMS multimodal learning, emotion recognition, audio-visual sentiment analysis, affective computing, emotion analysis, multi-label classification

I. INTRODUCTION

AUDIO-VISUAL Emotion Recognition (AVER) is integral to *human-computer interaction* (HCI). AVER is a technology that identifies human emotions by analyzing expressive signals within audio and visual channels [1]. AVER relies on audiovisual data annotated by humans to provide perceptual evaluations that describe perceived emotions. Video clips comprise audio and visual components, and it is standard to obtain from the video separate audio and visual streams, then employ a unified ground truth label for training the model based on the annotators' combined audio-visual perception [2]–[4]. However, human perception of emotions can vary when the modality of the emotional stimulus changes [5]—for example, when data contains only voice signals without corresponding visual cues, or vice versa. Furthermore, the emotional information conveyed in speech and facial features is different, so it is not clear that the audio-visual label is appropriate to describe the information individually conveyed in these modalities. This study seeks to explore whether a combination of labels derived from voice-only, face-only, and audio-visual stimuli can enhance the performances of the AVER systems.

In the field of AVER, studies have adopted methods that leverage audio-visual signals through distinct processing layers. These models typically process audio-only and visual-only inputs in isolation, followed by their integration using

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

some different fusion techniques such as feature-level [6], [7], decision-level [8], [9], or model-level integration [10], [11]. We hypothesize that models may exhibit improved learning when the audio-only and visual-only layers are trained with a combination of labels derived from audio-only and audio-visual stimuli for acoustic layers and video-only and audio-visual stimuli for visual layers. This hypothesis aligns with the premise that multimodal systems should incorporate distinct cues from heterogeneous dimensions and shared spaces [12], acknowledging that while modality fusion can amplify signals, it should not overlook the unique characteristics inherent to each modality [13].

We propose a two-stage training strategy to incorporate audio-only, video-only and audio-visual labels. In the initial stage, the model is trained with separate audio and visual inputs using labels generated from the corresponding unimodal stimuli, along with separate unimodal layers that incorporate audio-visual stimuli. In the second stage, we fix the weights of the models from the first stage and introduce additional layers to process the outputs from these models. Training continues with labels derived from audio-visual stimuli to enable our model to integrate information from both audio and visual channels within our shared layers.

In this study, alongside examining the proposed methodology, we conduct an ablation study by training models with either audio-only or video-only inputs, using labels derived from audio-only, video-only, or audio-visual stimuli. For the audio input encoding, we employ the WavLM Large model, which has demonstrated superior performance in emotion recognition tasks according to the SUPERB leaderboard [14]. For the visual input encoding, we select the Mobilenetv2 [15] as the primary backbone. Our experiments are carried out on the CREMA-D emotion dataset [16], which is the only corpus that includes emotional annotations based on voice-only, face-only, and audio-visual stimulus modalities. The CREMA-D dataset also provides demographic characteristics of the speakers, such as gender, age, and race, enabling us to assess performance bias and fairness within AVER systems. Additionally, we examine the model's calibration level. The contributions and findings of this work are listed as follows:

- A training strategy for AVER systems with a mixture of unimodal and audio-visual-rated labels, which achieves better recognition rates and model calibration.
- The paper extensively evaluate AVER model calibration, performance bias, and fairness across demographics
- The proposed approach has higher recognition rates, when we evaluate the performance using a multi-label formulation to accommodate the co-occurrence of emotions.

II. Related Work and Background

A. Impact of Auditory and Visual Modalities

Emotional cues can be conveyed via various modalities, such as voice and facial expressions. The information is not conveyed the same across modalities, since each modality provides a unique perspective in the expression of emotions. Therefore, it is expected that the perceived emotion will vary depending on the available modality. Paulmann et al. [5] revealed that humans have different emotional perceptions depending on stimulus (e.g., audio, video). Humans have higher recognition of decoding emotions from multi-modal emotional stimuli than the uni-modal stimuli (e.g., voice-only or face-only). Rigoulot et al. [17] conducted experiments to analyze the effects of vocal cues on facial expression. Focusing on four basic emotions (fear, anger, happiness, and a neutral state), the authors found that the presence of both acoustic and visual cues offers supplementary guidance for decoding emotional cues in facial expressions. Among the four emotions considered in this study, this phenomenon was notably more pronounced in the processing of fear, indicating that acoustic channels significantly enhance the interpretation of emotional cues in facial expressions. Yu et al. [18] worked on exploring different levels of modality cues by auto-generating unimodal labels and using a multi-task approach to train their model. The above studies indicate differences in the perception of emotion depending on the particular modality that is available. Using multimodal data can also provide richer training information, which can help mitigate challenges related to data limitations [19]. Hence, we aim to utilize the labels perceived by humans from unimodal and multimodal-based stimuli for training an AVER system. We expect that AVER systems using these labels can achieve better performance in recognizing emotions from audio-visual data.

B. Multi-label Emotion Recognition

Emotion perception is naturally subjective because of different emotional experiences, cultures, and gender. It is common to see disagreement among raters when they are elicited by the same emotional stimulus. Most previous computational studies about emotion recognition regard disagreement as noise and define the emotion recognition task as a single-label task. However, the emerging semantic space theory [20] concluded that emotion perception is high-dimensional, and blended after collecting self-reports to describe emotion perception elicited by emotional images, face videos, audio, music, and speech. Therefore, we summarize prior works that formulate emotion recognition as a multi-label task, highlighting the varied methodologies adopted to address discrepancies in emotional annotations.

1) Facial Expression Recognition

Previous studies [21], [22] have demonstrated that a single facial expression can simultaneously convey multiple emo-

tions. To accommodate this complex phenomenon, previous study transformed traditional categorical labels into distributional labels. By employing distributional label learning techniques [23], they trained facial expression recognition systems to interpret these nuanced labels. This approach allows for the representation of multiple emotions within a single dataset entry, capturing both predominant and minor emotions through the distributional labels. Unlike conventional methods that rely on single labels, this distributional labeling strategy effectively captures the complexity of mixed emotions.

2) Speech Emotion Recognition

Most prior research on *speech emotion recognition* (SER) treats this problem as a single-label task, often excluding data samples lacking consensus labels. The approaches use techniques such as plurality [24] or majority voting [25] to define an emotional state. However, studies [26]–[28] have demonstrated that utilizing distributional labels (also known as soft labels) during the model training stage can improve the performance of SER systems on single-label test sets. Riera et al. [29] argued against removing any data from the test set to accurately assess the real-world performance of SER systems. Furthermore, Chou et al. [30] proposed considering all emotional ratings to compute distributional labels based on the frequency of each emotion. Assuming that multiple raters annotated a file, they introduce a threshold to convert the distributional predictions of the system into binary decisions.

3) Audio-Visual Emotion Recognition

In the field of AVER, the most common approach in prior studies [31], [32] has been to treat disagreements among raters as noise, opting for aggregation rules to establish a consensus emotion as the ground truth. In this work, we adopt a broader definition of emotional states and investigate the presence of multiple emotional states within a single data sample. To model this complexity, we draw inspiration from Chou et al. [30] for determining the learning target. We then apply a thresholding method to obtain multi-hot labels, which serve as the basis for evaluating our model's performance.

III. Task Definition

Previous studies typically frame AVER as a single-label recognition task, discarding data points lacking annotator consensus. This strategy simplifies the test set but does not reflect practical scenarios where predictions are required for all samples. Notably, these discarded data points often exhibit ambiguous cues that evoke multiple coexisting traits. In this work, we consider all perceived annotations and formulate the task as a multi-label recognition problem. We calculate the proportion of the evaluations assigned to each emotional class by the annotators, forming a distribution. We select all the emotional classes with proportions above a given threshold. We use the threshold $1/C$ to binarize the distribution probabilities, which is the approach followed in previous studies [29], [33]. C is the number of emotional classes. This step removes emotions that are not consistently provided by annotators, reducing label noise. As a result, we create a multi-hot vector which allows each sample to contain multiple emotions.

In this work we convert the raw annotations into the training/testing labels obtained using unimodal and multimodal labels. We consider all labels elicited by voice-only, facial-only, and audio-visual stimuli. The labels are distributional labels for the training stage, and are converted into binary vectors when the values are higher than the defined threshold. We allow the samples to have more than one emotion to reflect the nature of emotion perception that could involve mixed emotions from a psychology perspective [20].

IV. Methodology

This section presents our proposed two-stage training strategy to consider unimodal and multimodal labels. We implement a framework based on a previous audio-visual emotion recognition research, the *versatile audio-visual learning* (VAVL) model [11], which presents a basic structure that perfectly fits our explored methodology, facilitating the implementation of a two-stage training system. In the first stage, the model is trained on a blend of labels derived from both multimodal and unimodal perceptions of emotion. Subsequently, in the second stage, its shared layers are further trained on perceived labels generated from multimodal stimuli. This strategy aims to recognize modality-specific cues that are specifically found in either speech or facial expressions, and general trends that only appear when multimodal stimuli are available.

A. Versatile Audio-Visual Learning Model

First, we briefly describe the VAVL methods, highlighting the reasons why this audio-visual model fits our training strategy. VAVL employs a versatile combination of acoustic-only and visual-only layers that independently process audio-visual content before merging into shared layers for joint learning. This framework fits well our two-stage approach, allowing us to independently train the acoustic-only and visual-only layers in the first stage. In stage 2, we integrate these now-frozen layers with a series of shared layers to perform audio-visual prediction using audio-visual labels.

B. Using Mixture of Emotion Perception

This section explains the two-stage training system with a mixture of emotion-perceived labels. An overview of our proposed approach is presented in Figure 1. We provide additional details on the framework configuration in Section V.D.

During stage 1, we concentrate on training the unimodal layers of our model, specifically focusing on the acoustic

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2025.3530274

layers (depicted with green blocks in Figure 1a) and the visual layers (illustrated with orange blocks in Figure 1a). This initial stage involves separate training for the acoustic and visual components. The acoustic layers are divided into two equal sets: one set is trained using voice-only labels, while the other is adjusted using labels from audio-visual stimuli. Similarly, the visual layers are also split into two equal sets, with one set trained on facial-only labels and the other refined with audio-visual labels. The motivation for splitting the layers in this manner is to harness the distinct and complementary information captured by each modality when exposed to both unimodal and multimodal stimuli labels. We aim to preserve the unique characteristics of each sensory input, ensuring that no valuable information is disregarded. As listed in Section V.C, we include baselines that ablate the importance of using this strategy in our result in Section VI to demonstrate the efficacy of our approach.

After completing stage 1, the pre-trained blocks from the acoustic module (*learnable weighted sum* (LW-Sum) and transformer layers) and the visual module (1D *convolutional neural network* (CNN) and transformer layers) are frozen. Then, we integrate the shared layers (illustrated with purple blocks in Figure 1b) into the model. The second stage involves training solely the shared layers, utilizing the audio-visual perceived labels as the learning target. Inputs from both the acoustic and visual layers are sequentially fed through the shared layers, incorporating residual connections during this process. For the final audio-visual prediction, we employ *attentive statistics pooling* (AS-Pool) on all audio and visual outputs from the shared layers and merge these outputs to feed into the *fully-connected* (FC) layers for making the audio-visual predictions

V. Experimental Settings

A. Dataset

This study employs the CREMA-D corpus [16] to train and evaluate the proposed strategy. CREMA-D is an audiovisual dataset featuring high-quality recordings from 91 actors (48 male, 43 female) of diverse racial and ethnic backgrounds, performing sentences with specific emotional intent. The dataset includes 7,442 clips, each evaluated by an average of 7 raters, totaling 5.26 hours of data. Sentence durations range from 0.51 to 5.01 seconds, with a mean of 2.54 seconds.

CREMA-D's unique labeling strategy involves perceptual evaluations under three conditions—audio-only, video-only, and audio-visual—making it particularly suited for our study. To the best of the authors' knowledge, this is the only multi-modal emotional dataset with this labeling approach, making it ideal for our proposed methodology. We approach the AVER problem in CREMA-D as a six-class multi-label classification task with speaker-independent data splits, covering anger, disgust, fear, happiness, sadness, and neutral states. The dataset's demographic details allow us to conduct a comprehensive fairness evaluation of the trained AVER systems.

B. Evaluation Metrics

1) Emotion Recognition

In our evaluation framework, we utilize the macro-F1 score and weighted-F1 score, which are the most suitable metrics for our task since it simultaneously assess recall and precision rates, to provide a balanced measure of the AVER systems' performance. Our evaluation process adopts a threshold-based approach [29] for scenarios involving multi-label classifications to accurately identify the target classes from the ground truth data. Specifically, a prediction for a particular class is deemed correct if its proportional representation among all predictions exceeds the threshold of $(1/C)$, where C is the total number of emotional classes under consideration, as shown in [33]. Utilizing this approach allows for a nuanced and precise calculation of F1 scores, effectively capturing the performance of our emotion recognition systems in recognizing a wide range of emotional states even in imbalanced scenarios.

2) Model Calibration

We assessed the calibration of AVER system predictions using the *Brier Score* (BS) [34], calculated for each emotion and averaged. BS values range from 0 to 1, with lower scores indicating better calibration. The averaged BS is given by:

$$BS = \frac{1}{C}\sum_{e=1}^{C}\left(\frac{1}{N^e}\sum_{i=1}^{N}(P_i^e - T_i^e)^2\right) \tag{1}$$

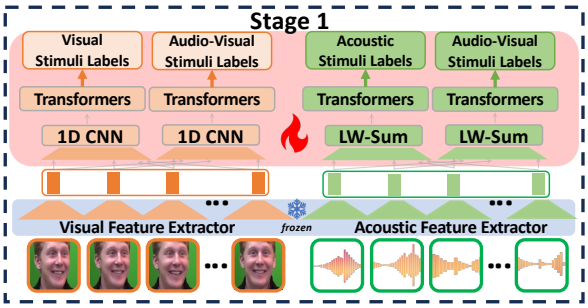
where C is the number of emotion classes, e denotes an emotion, N^e is the sample count for emotion e, P_i^e is the predicted probability, and T_i^e is the ground truth (0 or 1). The Brier Score was chosen for its ability to jointly capture accuracy and calibration, offering a continuous metric that avoids the potential biases introduced by the binning process required in alternatives such as the *expected calibration error* (ECE). Additionally, its straightforward calculation and adaptability to multi-label settings make it a robust choice for assessing probabilistic quality across multiple emotion classes.

3) Performance Bias

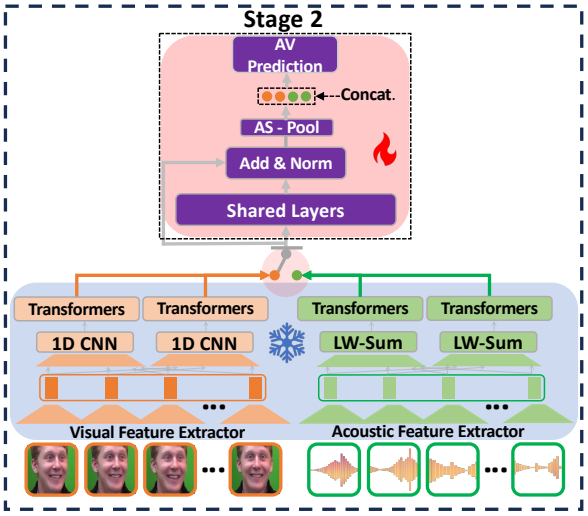
We define performance bias as the difference in macro-F1 scores across demographic groups (gender, age, race). The Macro-F1 Bias is calculated as:

$$Bias = \frac{1}{2} \times \text{mean}\left(\sum_{g=1}^{G}\sum_{i=1}^{G}|A_g - A_i|_{,i\neq g}\right) \tag{2}$$

where g and i represent groups, G is the total number of groups, and A_g and A_i are the macro-F1 scores for groups g and i. Lower bias values indicate less performance disparity across groups.



(a) Stage 1 training. At this stage the visual and acoustic model layers are trained separately.



(b) Stage 2 training. At this stage the acoustic and visual model layers trained from stage 1 are frozen, we then implement the shared layers and audio-visual prediction layers to be trained.

- Trainable Parameters | - Frozen Parameters

FIGURE 1: Overview of our two-stage training method.

4) Fairness

We evaluate fairness using *demographic parity difference* (DPD) and *equalized odds difference* (EOD) across gender, race, and age. Lower DPD values indicate more uniform selection rates across groups, while lower EOD values suggest more equitable rates of true/false positives/negatives.

C. Baseline Models

We implement a series of experiments to validate the proposed method. We investigate the impact of the diverse levels of annotation stimuli and the integration of these perceptual evaluations on the performances of the AVER models. All the following models are trained with a class-balanced objective function, originally proposed by Cui et al. [35], to address the imbalanced annotation distributions and ensure the accuracy and calibration of the emotion recognition systems. We evaluate several models, which We list by an index. We include the model index in the tables to refer to the models.

1) Baselines with Unimodal or Multimodal Labels

- Models indexed from 1 to 3 only process the acoustic inputs and are trained utilizing perceptual evaluations elicited by audio-only, voice-only, or audio-visual (AV) modalities, respectively. These models are employed using one of the sets of the green acoustic layers shown in Figure 2a.
- Models indexed from 4 to 6 only process the visual inputs and are trained utilizing perceptual evaluations elicited by audio-only, voice-only, or audio-visual (AV) modalities, respectively. These models are employed by using one of the sets of the orange visual layers shown in Figure 2b.
- Models indexed from 7 to 9 incorporate both acoustic and visual inputs and are trained utilizing perceptual evaluations elicited by audio-only, voice-only, or audio-visual (AV) modalities, respectively. These models follow the overall structure presented in Figure 2c; where employ the use of a single set of layers for each modality before the shared layers and we train all layers in one single step.

2) Two-Stage Approach

Figure 1 presents our proposed two-stage approach. In this section, we implement alternative versions of our two-stage models to understand the contributions of our model selection. We also index the models, including these indexes in the result tables.

- Model indexed as 10 represents a baseline version of our two-stage approach, which employs the exclusive use of the perceptual evaluations elicited by the audio-visual modality for training in both stages. This model, indexed with the number 10, serves as the base benchmark for our study. It does not incorporate acoustic layers trained with acoustic stimuli labels or visual layers trained with visual stimuli labels. We refer to this approach as the **AV** elicited method.
- Model indexed with the number 11 contrasts with model 10 by exclusively utilizing acoustic layers trained with acoustic stimuli labels and visual layers trained with visual stimuli labels in the first stage. The shared layers in the second stage are then trained using the perceptual evaluations elicited by the audio-visual modality. We refer to this approach as the **Mixed** elicited method.
- Model indexed with the number 12 corresponds to our final proposed methodology, which we refer to as the **Proposed** method.

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2025.3530274

3) State-Of-The-Art Baselines

Additionally, we benchmarked five AVER frameworks to evaluate our proposed model, utilizing code from their respective repositories or specifications from their associated papers.

VAVL: Goncalves et al. [11] proposed the VAVL framework, which combines acoustic-only and visual-only layers for processing, followed by shared layers for joint feature learning.

MulT: Tsai et al. [38] introduced the MulT method, a cross-modal transformer originally for human language time series. We adapted it for bimodal representations, focusing on visual and acoustic features.

SFAV: Chumachenko et al. [36] developed the SFAV approach, which uses late and intermediate transformer fusion techniques to handle incomplete audiovisual data.

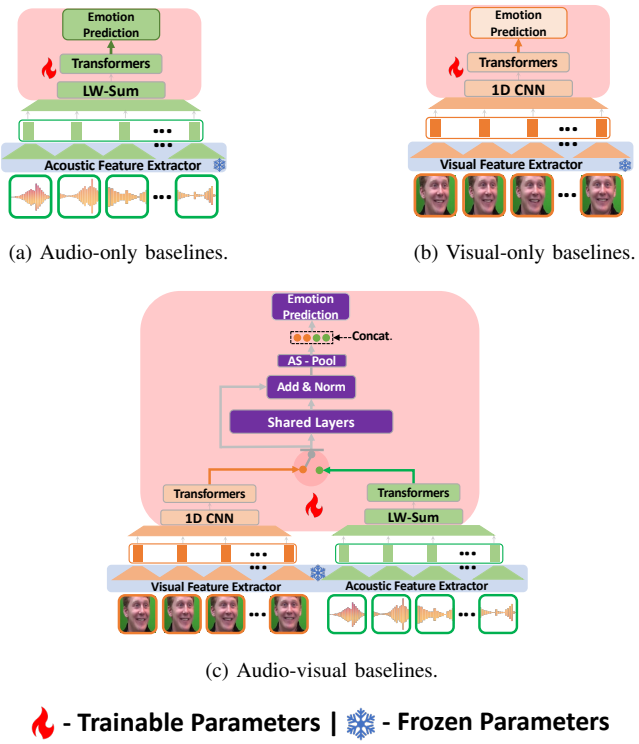
AuxFormer: Goncalves and Busso [31] presented the AuxFormer architecture, a model using transformer layers, auxiliary networks, and modality dropout for robust cross-modal representations.

TSLTM: Huang et al. [37] proposed the TSLTM method, integrating transformers and *long short-term memory* (LSTM) networks to fuse audio and visual data, modeling long-term emotional dynamics.

D. Implementation Details

1) Acoustic and Visual Features

Within the CREMA-D corpus, we have access to raw video and audio recordings, enabling the extraction of features from both audio and visual modalities. Our acoustic feature extraction leverages the pre-trained WavLM-large architecture [39], sourced from Hugging Face at "microsoft/wavlm-large." This model comprises 24 transformer layers and approximately 317M parameters. For feature extraction, we follow the approach used in previous studies [14], utilizing the frozen pre-trained WavLM model to extract features from the output hidden states of all 24 transformer layers, as well as from the hidden state output of the WavLM model's 1D CNN encoder. Consequently, each audio sequence processed by our WavLM feature extractor yields a feature set denoted as $x_a \in \mathbb{R}^{N_a \times 25 \times 1,024}$, where N_a represents the sequence length of the acoustic feature, 1,024 is the dimensionality of the hidden states, and 25 corresponds to the total number of hidden states from which features are extracted.



(a) Audio-only baselines. (b) Visual-only baselines.

(c) Audio-visual baselines.

FIGURE 2: Figure illustrates the baseline models indexed as 1 to 9. Where models indexed 1 to 3 are audio-only, models indexed 4 to 6 are visual-only, and models indexed 7 to 9 are audio-visual.

TABLE 1: Overview of models performances. We use similar structure to previous tables. The columns, **Macro-F1** and **Weighted-F1**, show the average macro-F1 and weighted-F1 scores, the lower, and upper bound of the confidence interval between 2.75% and 97.5%. Bold results assert statistical significance

Index	Audio	Video	Elicited	Macro F1 ↑	Weighted F1 ↑
1			Voice	0.645 (0.638, 0.652)	0.650 (0.643, 0.656)
2	✓	-	Face	0.654 (0.647, 0.661)	0.661 (0.655, 0.668)
3			AV	0.695 (0.688, 0.701)	0.699 (0.694, 0.705)
4			Voice	0.568 (0.560, 0.574)	0.570 (0.564, 0.576)
5	-	✓	Face	0.632 (0.626, 0.639)	0.634 (0.627, 0.641)
6			AV	0.658 (0.652, 0.665)	0.657 (0.650, 0.663)
7			Voice	0.674 (0.668, 0.681)	0.672 (0.666, 0.678)
8	✓	✓	Face	0.719 (0.712, 0.725)	0.719 (0.713, 0.725)
9			AV	0.762 (0.757, 0.768)	0.761 (0.756, 0.766)
10			AV	0.763 (0.757, 0.768)	0.758 (0.753, 0.763)
11	✓	✓	Mixed	0.769 (0.764, 0.774)	0.764 (0.759, 0.769)
12			Proposed	**0.772 (0.766, 0.777)**	**0.767 (0.762, 0.772)**

TABLE 2: Overview of our proposed approach and other SOTA AVER. Results are reported with **Macro-F1** and **Weighted-F1**, show the and confidence interval between 2.75% and 97.5% for each result. Bold results assert statistical significance

Audio	Video	Model	Macro F1 ↑	Weighted F1 ↑
		Ours	**0.772 (0.766, 0.777)**	**0.767 (0.762, 0.772)**
		VAVL [11]	0.762 (0.757, 0.768)	0.761 (0.756, 0.766)
		AuxFormer [31]	0.742 (0.737, 0.748)	0.741 (0.734, 0.747)
✓	✓	**SFAV** [36]	0.731 (0.725, 0.737)	0.728 (0.723, 0.734)
		TLSTM [37]	0.710 (0.704, 0.716)	0.705 (0.699, 0.711)
		MulT [38]	0.743 (0.738, 0.750)	0.741 (0.736, 0.748)

TABLE 3: Overview of audio-visual emotion recognition performances for each emotion in F1 scores and the measure of model calibration as the average Brier Score (BS). The mark "✓" in the columns , **Audio** and **Video**, represents the models use the modality as input. The column, "Elicited", means the training labels collected by giving which modalities as emotional stimulus. All results contain the average, lower and upper bound of the confidence interval between 2.75% and 97.5% for each result (lower bound, upper bound) in F1 scores. We also use the "grey" background color and bold font to show the best performances in the table.

Index	Audio	Video	Elicited	Anger	Sadness	Disgust	Fear	Neutral	Happiness	BS ↓
1			**Voice**	0.713 (0.695, 0.730)	0.603 (0.583, 0.622)	0.607 (0.588, 0.624)	0.607 (0.589, 0.623)	0.709 (0.699, 0.720)	0.631 (0.609, 0.654)	0.158
2	✓	-	**Face**	0.683 (0.665, 0.700)	0.546 (0.527, 0.566)	0.641 (0.625, 0.658)	0.572 (0.553, 0.590)	0.760 (0.750, 0.772)	0.720 (0.704, 0.739)	0.155
3			**AV**	0.735 (0.718, 0.753)	0.614 (0.594, 0.632)	0.664 (0.649, 0.680)	0.668 (0.652, 0.684)	0.761 (0.751, 0.774)	0.725 (0.708, 0.742)	0.146
4			**Voice**	0.507 (0.489, 0.525)	0.528 (0.510, 0.548)	0.571 (0.556, 0.588)	0.464 (0.445, 0.483)	0.646 (0.636, 0.657)	0.688 (0.666, 0.712)	0.184
5	-	✓	**Face**	0.565 (0.546, 0.585)	0.551 (0.534, 0.571)	0.594 (0.577, 0.611)	0.553 (0.534, 0.570)	0.724 (0.712, 0.736)	0.807 (0.793, 0.823)	0.160
6			**AV**	0.585 (0.565, 0.604)	0.578 (0.560, 0.594)	0.622 (0.607, 0.638)	0.580 (0.564, 0.597)	0.731 (0.720, 0.742)	0.854 (0.840, 0.868)	0.157
7			**Voice**	0.722 (0.704, 0.738)	0.640 (0.619, 0.659)	0.627 (0.610, 0.644)	0.626 (0.609, 0.643)	0.705 (0.694, 0.715)	0.727 (0.708, 0.747)	0.157
8	✓	✓	**Face**	0.718 (0.700, 0.736)	0.612 (0.594, 0.631)	0.676 (0.661, 0.691)	0.651 (0.632, 0.667)	0.793 (0.782, 0.804)	0.862 (0.848, 0.875)	0.142
9			**AV**	0.784 (0.768, 0.801)	0.656 (0.636, 0.673	**0.726 (0.711, 0.740)**	0.718 (0.704, 0.734	**0.811 (0.802, 0.821)**	0.879 (0.866, 0.892)	0.131
10			**AV**	0.770 (0.753, 0.786)	0.668 (0.650, 0.686)	0.717 (0.704, 0.731)	0.713 (0.698, 0.727)	0.797 (0.787, 0.807	**0.911 (0.900, 0.922)**	0.135
11	✓	✓	**Mixed**	**0.793 (0.777, 0.808**	**0.685 (0.667, 0.702)**	0.712 (0.698, 0.726)	0.712 (0.696, 0.726)	0.805 (0.795, 0.815)	0.909 (0.897, 0.919)	0.130
12			**Proposed**	0.792 (0.776, 0.808)	0.682 (0.665, 0.699)	0.722 (0.708, 0.736	**0.723 (0.707, 0.736)**	0.804 (0.793, 0.815)	0.909 (0.898, 0.920)	**0.129**

To extract emotional visual features from facial images, we employ a CNN-based model, specifically the pre-trained MobileNetV2 [15] model. We fine-tuned the MobileNetV2 model for facial expression recognition on the AffectNet [40] for 60 epochs, saving the weights of the model that achieves the best validation accuracy score. Once the model has been trained, we proceed to freeze the weights, enabling us to extract facial features from images extracted from videos within the dataset used in this study. We crop the image using the bounding box provided by a face detection algorithm and rotate the image to ensure that the line connecting the two eyes is parallel to the x-axis, thereby normalizing the image for consistent feature extraction. Finally, we feed the normalized image through our trained MobileNetV2 model. We use the 1,280-feature vector, extracted post-global pooling layer, as the feature representation for each image. Consequently, the feature set for each visual sequence extracted by our MobileNetV2 feature extractor is denoted as $x_v \in \mathbb{R}^{N_v \times 1,280}$, where N_v denotes the number of frames in the sequence, and 1,280 represents the dimensionality of the features from the MobileNetV2 model's global pooling layer.

2) Model Configuration and Training Settings

Figure 1b illustrates the entire framework of the model. The visual, acoustic, and shared layers are constructed using transformer blocks, each with an encoder hidden layer of 512 dimensions and eight attention heads. The acoustic, visual, and shared transformer layers consist of three, three, and two layers, respectively. The transformer block follows the standard structure of *multi-head attention* (MHA) layers, as introduced in [41]. Within the transformer block, the input is used to generate the Q, K, and V matrices. The attention mechanism is computed simultaneously for a set of queries, which are aggregated into a single matrix Q. Similarly, the keys and values are aggregated into matrices K and V, respectively, as described in Equation 3.

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

where d_k is the dimensionality of the key vectors. The scaling factor $\sqrt{d_k}$ prevents the dot product between Q and K from becoming too large, ensuring stable gradients in the softmax function. MHA allows the model to attend to multiple representation subspaces simultaneously, overcoming the limitations of single-head attention. It is computed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{4}$$

where each attention head is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{5}$$

here, $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ are learnable parameter matrices.

The features input to each specific layer, as shown in Figure 1, are processed as follows: the visual feature extraction output is denoted as $x_v \in \mathbb{R}^{N_v \times 1,280}$, while the acoustic feature extraction output is $x_a \in \mathbb{R}^{N_a \times 25 \times 1,024}$. To align these feature dimensions with the corresponding transformer layers, we apply distinct operations to the audio and visual inputs:

1) For the audio features, a *Learnable Weighted Sum* (LW-Sum) layer is used to aggregate the 25 extracted layers into a single feature matrix. Specifically, the weighted sum is computed as:

$$z_a = \sum_{i=1}^{25} w_i \cdot x_a^{(i)}, \tag{6}$$

where $x_a^{(i)} \in \mathbb{R}^{N_a \times 1,024}$ represents the i-th extracted layer, and w_i are learnable weights constrained by $\sum_{i=1}^{25} w_i = 1$ (enforced using a softmax layer). This operation results in the matrix $z_a \in \mathbb{R}^{N_a \times 1,024}$ for each batch sequence.

2) For the visual features, a 1D *convolutional neural network* (1D-CNN) layer is employed to adjust the feature dimension from 1,280 to 1,024. The operation is defined as:

$$z_v = \text{ReLU}(\text{Conv1D}(x_v)), \tag{7}$$

where Conv1D applies a convolution with kernel size k, stride s, and output channels $1,024$, transforming $x_v \in \mathbb{R}^{N_v \times 1,280}$ into $z_v \in \mathbb{R}^{N_v \times 1,024}$.

By applying these operations, both audio and visual features are aligned to $\mathbb{R}^{N \times 1,024}$, ensuring compatibility with the transformer layers. At the model's head, *attention-statistic pooling* (AS-Pool) layers [42] are employed to aggregate frame-level features into a fixed-length representation. AS-Pool layers compute a weighted mean and variance across the input sequence, where the weights are determined by an attention mechanism. This mechanism ensures that the model can dynamically focus on the most informative parts of the input sequence, adapting to varying temporal dynamics in the data. Then, we follow with *fully connected* (FC) layers, that are utilized for prediction. These pooling layers use an attention mechanism to assign different weights to various frames from the transformer layers' outputs. The model is optimized using the AdamW optimizer, configured with a learning rate (`lr`) of 1×10^{-5}, a weight decay of 5×10^{-7}, and beta parameters of 0.95 and 0.999. Training is conducted over 50 epochs with a batch size of 32. All experiments were done in a NVIDIA A100 48Gb, and the total of GPU hours are around 500 hours within a single NVIDIA A100 GPU.

VI. Results and Analyses

In this section, we discuss the outcomes derived from implementing our proposed methodology. We train the models indexed from 1 to 9 to reveal the impact of perceptual evaluation elicited by the varying modalities on the audio-visual condition. Results in Table 1 show that the models trained with the audio-visual labels achieve the best values in macro-F1 and weighted-F1 scores. Overall, audio-visual label usage leads to a relative improvement of 12.08% and 5.49% compared to the voice-only and face-only labels in the macro-F1 score, respectively. Also, the models that can take both acoustic and visual inputs lead to better performances than those with unimodal input, achieving a relative improvement of 8.07% and 15.95% compared to the acoustic-input and visual-input models in the macro-F1 score, respectively. The results are aligned with the findings in Paulmann and Pell [5], which showed that humans perform better in processing emotional cues on audio-visual stimuli than on unimodal stimuli.

Regarding the two-stage models, notably, the performances of models indexed as 11 (mixed) and 12 (proposed), which integrate the mixture of the pre-trained models trained with the perceptual evaluation elicited by varying modalities, surpass that of the model solely focused on audio-visual labels (model indexed as 10). These findings indicate that perceptual evaluations elicited by diverse stimuli modalities contains supplementary emotional cues, augmenting the effectiveness of AVER systems.

Table 2 presents additional results, directly comparing our proposed method against SOTA AVER baselines. We observe that our method consistently outperforms the SOTA AVER methods with statistical significance in all cases.

A. Model Calibration and Per-Emotion Results

Table 3 summarizes model calibration and macro-F1 scores per emotion. Models 1–9 show improved performance with audio-visual labels. Our proposed method (12) achieved the best calibration and excelled in recognizing fear. Models 11 and 12, which both utilize of our proposed approach of combining unimodal and multimodally rated labels for training, outperformed others in anger and sadness recognition, highlighting the effectiveness of our two-stage approach with perceptual evaluations across diverse stimuli.

B. Measure of Bias and Fairness Across Groups

Table 5 shows the performance, performance bias, and fairness metrics for all models concerning male and female speakers. Despite a larger proportion of male speakers in the database, AVER systems consistently demonstrate better emotion recognition performance for female speakers. This performance difference can be attributed to an imbalance in the emotion distribution between genders. Specifically, male samples are heavily skewed toward the "Neutral" emotion, which accounts for over 51% (2,006 out of 3,928 samples) of the male data, compared to about 39% (1,366 out of 3,510 samples) in female data. This dominance of the "Neutral" class in male samples leads the model to perform well on this prevalent emotion but struggle with less represented emotions such as "Angry," "Sad," and "Disgust." Since the macro-F1 score equally weights each emotion regardless of its frequency, the model's difficulties with these less frequent emotions in male data lower the overall performance metric. In contrast, the female data exhibits a more balanced distribution of emotions, allowing the model to perform more consistently across all emotion classes, resulting in higher macro-F1 scores for female speakers. The imbalance in the male dataset may also cause the model to overfit to the "Neutral" class, reducing its ability to generalize across other emotions within male samples. This effect is mitigated in the female dataset due to the more even distribution of emotions, enhancing the model's overall performance on female data. Although we observe this difference in performance accross gender, our proposed method appears effective in minimizing this performance

TABLE 4: Overview of recognition performances, performance bias, and fairness of the audio-visual emotion recognition systems on race groups. We use similar structure to previous tables. The column, "Macro-F1 Bias", means the differences between the macro-F1 scores of different race groups.

Index	Audio	Video	Elicited	African American	Asian	Caucasian	Unknown	MacroF1 Bias	DPD	EOD
1			Voice	0.646 (0.631, 0.660)	0.632 (0.605, 0.659)	0.645 (0.636, 0.655)	0.599 (0.528, 0.662)	**2.60 (5.84, 0.39)**	0.052	**0.098**
2	✓	-	Face	0.662 (0.648, 0.675)	0.677 (0.654, 0.699)	0.648 (0.640, 0.657)	0.616 (0.542, 0.678)	3.27 (5.70, 2.18)	0.154	0.348
3			AV	0.697 (0.683, 0.709)	0.711 (0.687, 0.733)	0.694 (0.686, 0.702)	0.572 (0.503, 0.631)	7.02 (9.29, 5.25)	0.052	0.205
4			Voice	0.536 (0.520, 0.551)	0.533 (0.507, 0.559)	0.582 (0.574, 0.590)	0.438 (0.356, 0.503)	7.27 (11.11, 4.50)	0.622	0.696
5	-	✓	Face	0.638 (0.625, 0.651)	0.641 (0.615, 0.666)	0.630 (0.622, 0.639)	0.468 (0.402, 0.523)	8.79 (11.27, 7.39)	0.519	0.573
6			AV	0.662 (0.648, 0.674)	0.653 (0.626, 0.677)	0.660 (0.653, 0.668)	0.464 (0.411, 0.511)	9.99 (12.44, 8.39)	0.535	0.619
7			Voice	0.676 (0.662, 0.689)	0.647 (0.620, 0.672)	0.677 (0.669, 0.685)	0.630 (0.544, 0.700)	2.80 (6.97, 1.42)	0.064	0.127
8	✓	✓	Face	0.716 (0.704, 0.727)	0.720 (0.697, 0.741)	0.721 (0.714, 0.728)	0.566 (0.515, 0.607)	7.80 (10.05, 6.72)	0.115	0.122
9			AV	0.759 (0.747, 0.771)	**0.780 (0.758, 0.799)**	0.762 (0.756, 0.770)	0.639 (0.591, 0.679)	7.15 (8.53, 6.05)	**0.035**	0.121
10			AV	0.765 (0.755, 0.776)	0.752 (0.729, 0.772)	0.764 (0.758, 0.771)	0.627 (0.560, 0.685)	7.12 (10.32, 4.57)	0.125	0.139
11	✓	✓	Mixed	**0.771 (0.761, 0.782)**	0.765 (0.745, 0.784)	0.770 (0.763, 0.776)	0.619 (0.563, 0.664)	7.71 (10.30, 6.11)	0.157	0.111
12			Proposed	0.770 (0.759, 0.780)	0.769 (0.747, 0.789)	**0.774 (0.768, 0.781)**	**0.668 (0.604, 0.724)**	5.31 (8.39, 3.25)	0.113	0.101

TABLE 5: Overview of recognition performances, performance bias, and fairness of the audio-visual emotion recognition systems on male and female groups. The column, "Macro-F1 Bias", quantifies the differences between the macro-F1 scores of male and female speakers.

Idx	Male	Female	Macro-F1 Bias ↓	EOD ↓
1	0.634 (0.624, 0.645)	0.653 (0.642, 0.663)	**1.86 (1.80, 1.80)**	0.019
2	0.634 (0.623, 0.643)	0.672 (0.662, 0.681)	3.81 (3.84, 3.83)	0.039
3	0.680 (0.671, 0.689)	0.708 (0.699, 0.718)	2.80 (2.78, 2.89)	0.032
4	0.544 (0.534, 0.555)	0.587 (0.578, 0.598)	4.32 (4.41, 4.31)	0.031
5	0.617 (0.608, 0.626)	0.646 (0.636, 0.655)	2.90 (2.88, 2.90)	0.075
6	0.643 (0.633, 0.653)	0.672 (0.663, 0.682)	2.86 (2.99, 2.90)	0.019
7	0.662 (0.653, 0.672)	0.683 (0.674, 0.692)	2.08 (2.13, 2.06)	0.026
8	0.697 (0.689, 0.706)	0.739 (0.730, 0.747)	4.12 (4.15, 4.11)	0.047
9	0.747 (0.739, 0.755)	0.776 (0.768, 0.784)	2.89 (2.87, 2.89)	0.027
10	0.750 (0.743, 0.758)	0.774 (0.767, 0.781)	2.37 (2.44, 2.35)	0.032
11	0.757 (0.750, 0.764)	0.780 (0.773, 0.788)	2.35 (2.34, 2.35)	**0.011**
12	**0.762 (0.754, 0.770)**	**0.781 (0.774, 0.788)**	1.86 (1.98, 1.84)	0.018

bias between genders. Specifically, the models indexed as 11 and 12 utilizing our proposed approach rank as the top two in performance. Furthermore, the performance bias for these models approaches the lowest observed values among baseline models. Although the bias in models indexed 11 and 12 is slightly above the minimum seen in baseline models, these models exhibit a large relative improvement in performance – 20.19% and 19.6%, respectively – over the model indexed as 1, which has the lowest bias. Regarding fairness, the EOD value for model indexed as 11 is the lowest among all models, highlighting its superiority in ensuring equitable treatment across gender groups within AVER systems.

Table 4 presents the performance bias, results, and fairness of all models across different racial groups. Models indexed as 11 and 12 generally outperform others across all racial categories, with the exception of the Asian group. The results of performance bias and fairness in Table 4 underscore the necessity of employing strategies specifically aimed at reducing bias and enhancing fairness within AVER systems.

VII. Discussion and Limitations

While we investigate the performance bias and fairness of models across demographic groups, including gender, age, and race, it is crucial to acknowledge that gender identity encompasses a spectrum beyond the binary male/female [43]. For a more comprehensive analysis, a dataset with annotations reflecting this broader spectrum of gender identities is required. Moreover, our findings indicate that to effectively reduce performance bias across racial groups, the models require further methodologies aimed at bias mitigation. Consequently, additional measures are important to enhance the fairness of all models.

VIII. Conclusion and Future Work

This work addresses a critical gap in traditional AVER systems that rely on uniform annotations and overlook the complexity of human emotional perception across different stimuli. We implement models that consider a full spectrum of emotions by recognizing co-occurrence in samples during training and evaluation. By integrating a two-stage approach using labels from audio-only, face-only, and audio-visual stimuli, our study better reflects how humans perceive emotions through both unimodal and multimodal channels. This method improves our model's accuracy in predicting emotional states, enhancing performance on the CREMA-D emotion database with superior macro-/weighted-F1 scores. We extensively evaluate model calibration, performance bias, and fairness across demographics, essential for equitable AVER systems. Our findings indicate that AVER systems perform better on female speakers despite having more male data, and our method reduces this bias. In future work, we plan to implement additional methods during training AVER systems to mitigate the performances bias and increase fairness.

REFERENCES

[1] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021.

[2] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using gaussian mixture models for face and voice," in *2008 Tenth IEEE International Symposium on Multimedia*, 2008, pp. 250–257.

[3] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2156–2170, October-December 2022.

[4] ——, "Learning cross-modal audiovisual representations with ladder networks for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023),*, Rhodes island, Greece, 2023.

[5] S. Paulmann and M. D. Pell, "Is there an advantage for recognizing multi-modal emotional stimuli?" *Motivation and Emotion*, vol. 35, pp. 192–201, 2011.

[6] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *International conference on Multimodal interaction (ICMI 2015)*, Seattle, WA, USA, November 2015, pp. 467–474.

[7] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *International conference on multimodal interaction (ICMI 2014)*, Istanbul, Turkey, November 2014, pp. 494–501.

[8] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, June 2011.

[9] B. Sun *et al.*, "Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 125–137, June 2016.

[10] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *ICMI 2020*, Utrecht, The Netherlands, October 2020, pp. 400–404.

[11] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for emotion recognition," *IEEE Transactions on Affective Computing*, vol. to appear, 2024.

[12] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and trends in multimodal machine learning: Principles, challenges, and open questions," 2023.

[13] S. Partan and P. Marler, "Communication goes multimodal," *Science*, vol. 283, no. 5406, pp. 1272–1273, 1999.

[14] H. Wu *et al.*, "Emo-superb: An in-depth look at speech emotion recognition," 2024.

[15] M. Sandler *et al.*, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] H. Cao *et al.*, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[17] S. Rigoulot and M. D. Pell, "Emotion in the voice influences the way we scan emotional faces," *Speech Comm.*, vol. 65, pp. 36–49, 2014.

[18] W. Yu *et al.*, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *AAAI Conference*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[19] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, p. Article 46, Apr. 2023, open access. [Online]. Available: https://doi.org/10.1186/s40537-023-00727-2

[20] A. S. Cowen and D. Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021.

[21] Y. Zhou, H. Xue, and X. Geng, "Emotion Distribution Recognition from Facial Expressions," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1247–1250.

[22] Y. Li *et al.*, "Self-Supervised Exclusive-Inclusive Interactive Learning for Multi-Label Facial Expression Recognition in the Wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3190–3202, 2022.

[23] Xin Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[24] L. Goncalves *et al.*, "Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 247–254.

[25] M. Gokilavani, H. Katakam, S. A. Basheer, and P. Srinivas, "Ravdness, crema-d, tess based algorithm for emotion recognition using speech," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022, pp. 1625–1631.

[26] H. M. Fayek *et al.*, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *International Joint Conference on Neural Networks (IJCNN 2016)*, Vancouver, BC, Canada, July 2016, pp. 566–570.

[27] H.-C. Chou and C.-C. Lee, "Every Rating Matters: Joint Learning of Subjective Labels and Individual Annotators for Speech Emotion Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 5886–5890.

[28] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, USA, Oct 2017, pp. 415–420.

[29] P. Riera *et al.*, "No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, Graz, Austria, September 2019, pp. 11–15.

[30] H.-C. Chou *et al.*, "Exploiting Annotators' Typed Description of Emotion Perception to Maximize Utilization of Ratings for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*. Singapore: IEEE, May 2022, pp. 7717–7721.

[31] L. Goncalves and C. Busso, "AuxFormer: Robust Approach to Audiovisual Emotion Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7357–7361.

[32] Y. Lei and H. Cao, "Audio-Visual Emotion Recognition With Preference Learning Based on Intended and Multi-Modal Perceived Labels," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2954–2969, 2023.

[33] H. Chou *et al.*, "Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–15, jun 5555.

[34] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[35] Y. Cui *et al.*, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, California, USA, June 2019.

[36] K. Chumachenko *et al.*, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *2022 ICPR*, Los Alamitos, CA, USA, aug 2022, pp. 2822–2828.

[37] J. Huang *et al.*, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020*, 2020, pp. 3507–3511.

[38] Y.-H. Tsai *et al.*, "Multimodal transformer for unaligned multimodal language sequences," in *Association for Computational Linguistics (ACL 2019)*, vol. 1, Florence, Italy, July 2019, pp. 6558–6569.

[39] S. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, October 2022.

[40] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, January-March 2019.

[41] A. Vaswani *et al.*, "Attention is all you need," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 5998–6008.

[42] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[43] M. Bass, L. J. Gonzalez, L. Colip, N. Sharon, and J. Conklin, "Rethinking gender: The nonbinary approach," *American Journal of Health-System Pharmacy*, vol. 75, no. 22, pp. 1821–1823, 11 2018.