IEEE Open Journal of Signal Processing

Received 24 June 2025; revised 26 October 2025; accepted 22 December 2025. Date of publication 26 December 2025; date of current version 27 January 2026. The review of this article was arranged by Associate Editor Z. Duan.

Digital Object Identifier 10.1109/OJSP.2025.3648710

Contextual Attention for Robust Audio-Visual Emotion Recognition

LUCAS GONCALVES [1] (Member, IEEE), HUANG-CHENG CHOU [2,3] (Member, IEEE), ALI N. SALMAN [1], CHI-CHUN LEE [2] (Senior Member, IEEE), AND CARLOS BUSSO [4] (Fellow, IEEE)

[1]The University of Texas at Dallas, Richardson, TX 75080 USA
[2]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan
[3]Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90007 USA
[4]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA

CORRESPONDING AUTHOR: Carlos Busso (email: busso@cmu.edu).

This work was supported by the National Science Foundation under Grant CNS-2016719.

ABSTRACT *Audio-visual emotion recognition* (AVER) often performs well under ideal conditions but faces significant challenges in scenarios with missing modalities (e.g., missing frames of audio and/or video). Addressing these challenges is crucial for the effective deployment of AVER systems in *human-computer interaction* (HCI) applications, where robustness can significantly impact user experience. This study introduces a novel approach that enhances AVER robustness by leveraging a decoder-like summarizer structure. This structure processes audio and visual content and generates contextual summaries that effectively capture emotional cues even when modalities are degraded. To enhance system resilience against missing modalities, we integrate modality dropout during training, enabling the summarizer to adaptively handle these scenarios. We define the context summary length as the number of learnable query tokens used in the summarizer, a fixed hyperparameter in our model. We analyze how varying context summary lengths affect performance, identifying an optimal balance between compression and expressiveness. In addition to improving robustness, we systematically evaluate model calibration across emotions in current *state-of-the-art* (SOTA) AVER methods. Our experiments on the MSP-IMPROV and CREMA-D databases demonstrate that our model achieves superior performance across macro-, micro-, and weighted-F1 scores, both under ideal conditions and in scenarios with modality losses. Additionally, we conduct ablation studies to assess the impact of different context lengths on our summarizer structure in terms of overall AVER performance.

INDEX TERMS Multimodal learning, missing modality, emotion recognition, audio-visual sentiment analysis, affective computing, emotion analysis, multi-label classification, model calibration.

I. INTRODUCTION

Human communication benefits significantly from recognizing emotional signals expressed through speech and facial expressions, which are crucial for enhancing the message conveyed during interactions. Therefore, to optimize emotion recognition systems, it is often important to involve the integration of cues from both speech [1], [2], [3], [4] and facial expressions [5], [6]. *Audio-visual emotion recognition* (AVER) relies on acoustic and visual signals to recognize emotions. However, the input signals could be corrupted or missing when AVER systems are integrated into realistic applications, such as *human-computer interaction* (HCI) systems. Addressing robustness to missing modalities plays a vital role in making these systems deployable in actual applications. In practice, visual information may be unavailable or unreliable due to occlusion, head pose (non-frontal views), motion blur, or cropping. Likewise, acoustic information may also be missing (e.g., target speakers listening to their interlocutors). Such segments should be treated as missing frames. The multimodal model should attend only to reliable frames and down-weight or ignore unreliable ones before fusion, unifying missing-modality and missing-frame cases under a single framework. The problem of missing frames or modality is also an important challenge in other multimodal tasks, including audio-visual question answering, representation learning, diarization, and multimodal sentiment prediction [7], [8], [9],

[10]. In addition to high performance, another desirable property of AVER systems is to have a calibrated system, where the model's predicted probabilities reflect the true likelihood of the outcomes. A few studies have evaluated in-depth performance biases related to model calibration. An uncalibrated system can perpetuate existing biases or underperform in non-ideal conditions [11]. Our vision is that multimodal systems should handle missing modalities, providing calibrated predictions.

Multimodal emotion recognition leverages diverse information from various modalities to enhance model performance. However, these systems face challenges when modalities are missing, often necessitating the presence of all relevant modalities to properly function, particularly in real-world applications [12]. Issues such as packet losses in speech signals or missing acoustic and visual frames can degrade the performance of applications such as *automatic speech recognition* (ASR) and multimodal systems [11], [13]. To compensate, techniques such as data cancellation and generation [14], [15], cascaded enhancements [16], and joint representation learning [17] are used. Research has also explored handling missing data through strategies such as feature interpolation [18] and ensemble fusion [19]. Studies have explored training modifications such as modality dropout, which involves zeroing out portions of the data to improve the robustness and performance in deep learning models [20], [21].

This study proposes a *robust in audio-visual emotion recognition* (RAVER) framework. The key innovation of the proposed method is a context summarizer implemented with an attention mechanism. The strategy consists of a decoder-like network that dynamically focuses on the most relevant segments of the input features, learning how to address cases where input information is missing. This approach can extract relevant information even in extreme conditions where most of the frames are missing. The effectiveness of our proposed method is validated through experiments conducted on the MSP-IMPROV [22] and CREMA-D [23] datasets. Our proposed method enhances the robustness of AVER systems under conditions of missing modalities. Compared to five established *state-of-the-art* (SOTA) AVER baselines, our system demonstrates superior resilience and reliability, ensuring more consistent performance even in extreme scenarios where 95% or more of acoustic or visual inputs are missing. The evaluation also examines the overall model calibration of RAVER in comparison to the baselines. RAVER not only achieves SOTA performance but also has lower bias than other existing AVER systems. In this study, lower performance bias refers to reduced disparity in performance across emotions. RAVER shows smaller gaps between emotional classes and a lower Brier score, indicating better calibration and more equitable predictions across all categories.

To the best of our knowledge, RAVER is the first AVER model to utilize a decoder-style contextual summarizer with learnable query tokens that selectively retrieve salient frames and explicitly compress variable-length sequences, improving robustness when audio and/or visual inputs are missing. We pair this strategy with a modality dropout approach during training. We evaluate our solution under extreme random masking and a two-state Markov packet-loss setting, and analyze calibration (Brier) in addition to standard metrics. Our contributions are threefold:

- We introduce a decoder-style contextual summarizer with learnable queries that enhances robustness to missing inputs in AVER.
- We present consistent gains under extreme masking (up to 100%) and bursty Markov packet loss, including cases where one modality is partially or entirely absent.
- We report the Brier score and per-emotion metrics toward transparency and better calibration: on CREMA-D, RAVER attains the lowest Brier (0.119) and shows reduced emotion-specific variability.

II. RELATED WORK AND BACKGROUND

This paper focuses on enhancing robustness in AVER systems. This section reviews previous studies, focusing on emotion recognition formulations, strategies to mitigate performance loss due to missing modalities, and existing SOTA AVER studies.

A. MULTI-LABEL EMOTION RECOGNITION

Most studies working on AVER systems rely on labels derived from human-rated perceptual evaluations. Given the differences in the perception of emotion, it is common to observe disagreement in the labels among public emotional databases [22]. To deal with the disagreement among raters, most studies [24], [25] utilized majority vote or plurality rule to aggregate a single emotional class for each sample, removing samples that have no consensus emotional class [26]. This common practice in the field of AVER [20], [27] regards disagreements among raters as noise. However, emotion perception is subjective and people have different interpretations of the expressed emotional behaviors, influenced by their gender, unique emotional experiences, or different culture [26]. Also, studies in psychology reveal that emotion perception is high-dimensional, overlapped, and blended [28]. In this context, it is appealing to consider emotion recognition as a multi-label recognition task.

Several studies have formulated emotion recognition as a multi-label classification task. For example, studies have revealed that human facial expressions can simultaneously convey multiple emotions instead of a single emotion [29], [30]. Chou et al. [31] computed distributional labels based on the frequency that each emotion was selected by the annotators as ground truth for training *speech emotion recognition* (SER) systems. They used a threshold to convert the distribution of the selected emotions by the annotators into binary decisions, defining the threshold as $1/C$, where

C represents the number of emotional classes. We follow this approach in this study.

B. MISSING MODALITY

During human-human interactions, modalities used for emotion recognition may be intermittently missing or unreliable due to factors such as facial occlusion or acoustic noise [32], [33]. Traditional multimodal methods often assume ideal conditions and do not account for these challenges. However, recent studies have proposed various strategies to handle missing modalities [11], [20], [21]. For instance, Mittal et al. [12] proposed using *canonical correlation analysis* (CCA) to generate proxy features for missing modalities. Chen et al. [34] introduced a heterogeneous graph-based hypernode framework for multimodal fusion of incomplete data. Liu et al. [35] and Zuo et al. [13] utilized contrastive learning and mapping representation techniques, respectively, to enhance modality-invariant features to capture and compensate for missing inputs. In *audio-visual speech recognition* (AVSR), Dai et al. [36] showed that applying video-dropout improves robustness to missing frames but can introduce modality bias. They proposed a knowledge-distillation framework to mitigate the bias. In speaker diarization, Cheng and Li [37] proposed a multi-input multi-output *target speaker voice activity detection* (TSVAD) model capable of operating in audio-only, video-only, or audio-visual modes by dynamically adapting to missing inputs. Other approaches have used generative models to reconstruct missing modalities [13], [15], [17], [38], [39], [40], [41], [42]. For example, Du et al. [15] developed a semi-supervised multiview deep generative framework that treats missing modalities as latent variables to be integrated during inference. Pham et al. [17] treated missing modalities as a translation problem, learning to map between source and target modalities. A cycle consistency loss is used to ensure that translating from one modality to another and back retains the original information, promoting accurate reconstruction and preserving essential features across modalities.

Ma et al. [43] utilized the *Hirschfeld-Gebelein-Rényi* (HGR) maximal correlation to extract common information between audio-visual modalities. Parthasarathy and Sundaram [21] focused on visual input ablations during training, where visual frames were randomly removed. Unfortunately, this approach does not address the problem when acoustic features are missing. More recently, Goncalves and Busso [20] investigated the simultaneous dropout of both audio and visual modalities during training to enhance robustness to missing modalities. Unlike previous methods that rely on dropout as a form of implicit regularization or use generative methods to reconstruct missing modalities, our approach explicitly learns to reweight and summarize representations based on the available data. By integrating modality dropout with our context summarizer, the model not only generalizes better to missing modalities but also minimizes performance loss across varying missing data conditions.

C. AUDIO-VISUAL EMOTION RECOGNITION

In the domain of AVER, various frameworks have been developed to address the challenges of multimodal data integration and emotion recognition under diverse conditions. Tsai et al. [44] developed *MulT*, a multimodal transformer that facilitates cross-modal interactions between modalities. The approach generates bimodal representations that are concatenated and further processed for prediction tasks. Goncalves and Busso [20], [45] introduced the *AuxFormer* architecture, a transformer-based framework that enables cross-modal representations through transformer layers. This model used cross-modal layers to transfer representations from queries in one modality to the keys and values of another. This strategy improves robustness through modality dropout and auxiliary networks, enabling effective operation when one or both modalities are available.

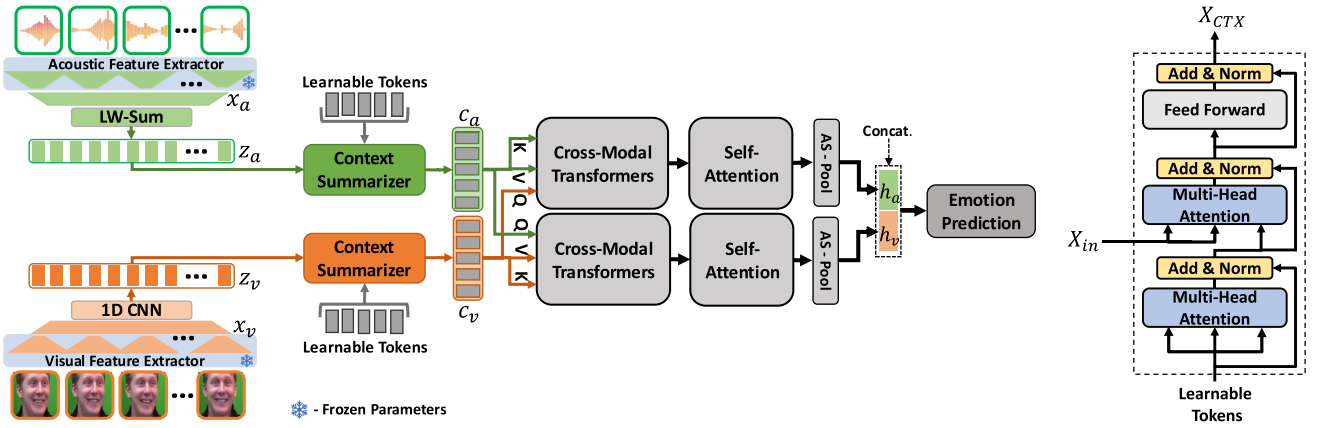
Chumachenko et al. [46] presented a method using *self-attention fusion for audio-visual* (SFAV) emotion recognition, designed to perform optimally in scenarios where data from either audio or visual modalities might be incomplete. The architecture utilizes advanced fusion techniques, such as late and intermediate transformer fusion, to robustly integrate features across modalities, highlighting the importance of adaptable fusion mechanisms in AVER. Additionally, Huang et al. [47] proposed the *transformer and long short-term memory* (TSLTM) approach, which combines the transformer architecture's ability to capture long-term dependencies with *long short-term memory* (LSTM) networks, enhancing emotional recognition across temporal sequences. This model employs multi-head attention to merge modalities into a coherent semantic space, demonstrating enhanced performance over traditional methods. More recently, Goncalves et al. [48] introduced the *versatile audio-visual learning* (VAVL), utilizing a blend of acoustic-only and visual-only layers that process audio-visual content independently before merging into shared layers for joint learning.

Our model incorporates robust training techniques, such as modality dropout, designed to enhance performance in environments where modalities may be partially or completely absent. In contrast to previous works, we introduce a context summarizer to learn how to handle varying modality input conditions and handle sequential data summarization using attention.

D. RELATION TO PREVIOUS WORK

While cross-modal attention and modality dropout are widely used in multimodal learning, RAVER introduces a distinct contribution through its *contextual summarizer*. Unlike prior frameworks such as AuxFormer [20] and VAVL [48], which rely on standard attention or pooling for temporal aggregation, RAVER employs decoder-style learnable query tokens to dynamically summarize relevant segments from each modality.

Additionally, while modality dropout was utilized in AuxFormer as a training strategy to improve performance under missing modalities, RAVER integrates it more tightly with



(a) Proposed framework overview. (b) Context summarizer.

FIGURE 1. **The figures show an overview of our proposed RAVER framework. (a) Architecture with feature extractor, context summarizer, cross-modal transformers, self-attention, pooling layers, and emotion prediction head; (b) detailed decoder-based structure of the context summarizer.**

the summarization process. The contextual summarizer learns to reweight and extract informative features even under high levels of modality degradation. This fusion of dropout and summarization enables RAVER to remain effective in extreme scenarios (e.g., 95% frame loss), where earlier models tend to degrade sharply.

This summarization mechanism enables robust performance under missing modality conditions, as shown in Section V-A. To our knowledge, this is the first use of attention-based summarization for enhancing robustness in audio-visual emotion recognition.

III. METHODOLOGY

This section describes our proposed approach for enhancing the robustness of an AVER system in scenarios with missing modalities. Fig. 1(a) depicts the overall framework. The proposed framework has three main blocks: (1) separate encoders for acoustic and visual modalities to extract features, (2) transformer-based context summarizers to extract relevant information from the available frames, and (3) cross-modal attention mechanisms to perform AVER.

A. FEATURE EXTRACTION

The first block in RAVER extracts modality-specific representations. We employ pre-trained feature encoders with frozen parameters to process audio and visual inputs separately. For acoustic features, we use WavLM [49], extracting hidden states from all 24 transformer layers and its 1D *CNN* encoder. This process yields a feature set $x_a \in \mathbb{R}^{N_a \times 25 \times 1,024}$, where N_a is the sequence length, 25 is the number of extracted layers, and 1,024 is the hidden state dimensionality. For a 4-second clip, N_a is approximately 200, based on 20 ms frame hops.

For the visual feature extraction, we use MobileNetV2 [50]. The faces are detected using the MTCNN toolkit [51] and aligned to normalize eye orientation before cropping. The processed images are then passed through the frozen model, extracting a 1,280-dimensional feature vector from the post-global pooling layer. This process results in a feature set $x_v \in \mathbb{R}^{N_v \times 1,280}$, where N_v is the sequence length (considering all frames in 30 fps videos) and 1,280 is the feature dimensionality. For a 4-second video, N_v is approximately 120 frames at 30 fps.

Once the features for the modalities are extracted, the feature representations are $x_a \in \mathbb{R}^{N_a \times 25 \times 1,024}$ for acoustic features, and $x_v \in \mathbb{R}^{N_v \times 1,280}$ for visual features. To ensure consistency for feature dimensions across modalities, we apply distinct operations to the audio and visual inputs. For the audio features, we use a *learnable weighted sum* (LW-Sum) layer, as used in Wu et al. [49], which aggregates the 25 extracted layers into a feature matrix $z_a \in \mathbb{R}^{N_a \times 1,024}$ per audio sequence. For the visual features, a 1D convolution (Conv1D, $k = 1$) reduces the feature dimension from 1,280 to 1,024, resulting in $z_v \in \mathbb{R}^{N_v \times 1,024}$. With $k = 1$, Conv1D acts as a time-distributed linear projection, equivalent to applying the same fully connected layer to each frame, preserving temporal order without adding extra nonlinearity before the contextual summarizer. These features are then processed by two independent context summarizers, which refine the resulting acoustic and visual features with learnable tokens that build a summarized context from the most useful cues present in each modality.

B. CONTEXT SUMMARIZER

The structure of the context summarizer is based on the decoder layout of the Transformers [52], as shown in Fig. 1(b). We draw inspiration from Iashin et al. [53]. In our work, we define N_{ctx} as the number of learnable tokens per modality, which is set to 128 based on the experiments we highlighted in Section IV-F. These tokens, $\mathbf{Q}_a \in \mathbb{R}^{N_{\text{ctx}} \times d}$ and $\mathbf{Q}_v \in \mathbb{R}^{N_{\text{ctx}} \times d}$, serve as trainable queries in a cross-attention mechanism that

"pulls" relevant information from the acoustic and visual feature sequences:

$$\mathbf{c}_{\mathrm{a}} = \mathrm{Decoder}\big(\mathbf{Q}_{\mathrm{a}}, \mathbf{A}\big) \in \mathbb{R}^{N_{\mathrm{ctx}} \times 1,024}, \quad (1)$$

$$\mathbf{c}_{\mathrm{v}} = \mathrm{Decoder}\big(\mathbf{Q}_{\mathrm{v}}, \mathbf{V}\big) \in \mathbb{R}^{N_{\mathrm{ctx}} \times 1,024}, \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{N_{\mathrm{a}} \times d}$ and $\mathbf{V} \in \mathbb{R}^{N_{\mathrm{v}} \times d}$ are the flattened audio and video features, respectively. Because $N_{\mathrm{ctx}} \ll N_{\mathrm{a}}, N_{\mathrm{v}}$, the summarizer condenses the original sequences into a more compact set of representations, effectively ignoring empty or irrelevant frames (e.g., from missing data). This architecture enables the summarizer to learn compressed representations from the most informative frames, which is critical in non-ideal scenarios with missing or noisy data. Unlike standard pooling, the learnable tokens dynamically attend to useful segments within each modality stream.

C. CROSS-MODAL TRANSFORMER LAYERS

The learned output contexts from each modality are processed through *cross-modal transformers* (CMT). The embedded context features for each modality are used to generate the Q, V, and K matrices. To compute attention scores from one modality to another, we use cross-modal attention as presented in Tsai et al. [44]. Using CMT layers is ideal for our purpose. When one modality is entirely missing, the cross-modal sublayer creates a near-zero update and, via the residual connections inside the transformer layers, it preserves the representation of the available modality. Then, the layers to follow operate in a unimodal manner.

Subsequently, the cross-modal outputs are input into the self-attention layers that have a structure similar to that of the cross-modal attention layers. However, there is no exchange of information from one feature modality to another. The Q, V, and K matrices are contained within their respective self-attention layers, computing separate feature representations. The use of self-attention modules after the cross-modal layers helps our model to emphasize areas of the cross-modal contexts that are more useful for our discriminative task. At this stage, the outputs of the self-attention layers are still shaped as $c_a \in \mathbb{R}^{N_{\mathrm{ctx}} \times 1,024}$ and $c_v \in \mathbb{R}^{N_{\mathrm{ctx}} \times 1,024}$.

Lastly, these representations are passed through two separate pooling layers to obtain flattened representations for each modality sequence, shaped as $h_a \in \mathbb{R}^{1 \times 1,024}$ and $h_v \in \mathbb{R}^{1 \times 1,024}$. These pooling layers employ *attentive statistics pooling* (AS-Pool) [54], which utilizes an attention mechanism to allocate varying weights to different frames. This technique enables the calculation of both weighted means and standard deviations for the frames sourced from each self-attention layer. The processed outputs from the pooling layers are subsequently concatenated and fed into a series of *fully connected* (FC) layers for emotion prediction.

IV. EXPERIMENTAL SETTINGS

A. DATASETS

This study employs the CREMA-D corpus [23] and the MSP-IMPROV corpus [22] for training and evaluation. We selected these corpora because their labels are based on perceptual evaluations from several evaluators, and the sessions include clear audio recordings and frontal facial views of the participants. These conditions are not present in some of the other multimodal databases (e.g., the IEMOCAP corpus [55] has only three annotators per sentence and does not provide frontal face videos of the participants).

The CREMA-D is an audio-visual database featuring high-quality video clips of 91 actors from diverse gender, racial, ethnic, and age groups. The dataset comprises 7,442 video clips (we exclude four corrupted files). There are 2,443 unique raters. Each speech sample was annotated on average by 9.84 annotators (the maximum number of annotators per sample was 12 and the minimum was 6). The lengths of the clips vary, with an average duration of 2.54 seconds (the maximal duration is 5.01 seconds and the minimal duration is 0.51 seconds). We defined the task within the CREMA-D dataset using "speaker-independent" splits for six-category multi-label classification task, involving the emotions anger, disgust, fear, happiness, sadness, and neutral state similar to the setup presented in Chou et al. [56].

The MSP-IMPROV corpus [22] serves as the second audio-visual resource that we consider. This corpus was devised to probe the nuances of emotional perception in conversational scenarios. Rather than simply instructing actors to repeat sentences with varied emotions, a protocol was employed to create authentic emotional interaction turns by designing hypothetical two-person scenarios that would prompt a participant to utter the target sentence with the desired emotion. With 20 target sentences spanning four emotional states (happiness, sadness, anger, and neutrality), 80 unique scenarios were produced. This segment of the corpus includes 652 speech instances. Additionally, the corpus encompasses all other interactions leading up to the target utterance (4,381 spontaneous speech turns) and natural interactions between the actors during the breaks of the improvisations (2,785 natural speech turns). It further consists of read renditions of the target sentences in the four emotions (620 rehearsed speech instances). In total, the MSP-IMPROV corpus comprises 7,818 spontaneous speech instances and 620 read phrases. The annotation was conducted through a crowd-sourcing protocol, with a minimum of five workers annotating each sentence. Similar to the CREMA-D setup, we defined the task within the MSP-IMPROV dataset using "speaker-independent" splits. The only difference is that this dataset provides four emotions: anger, sadness, happiness, and neutral state, leading us to set up a four-category multi-label classification task.

B. TASK DEFINITION

This work formulates the emotion recognition task as a multi-label problem, which differs from the common single-label classification approach used in previous AVER studies. For this purpose, we first consider all emotional ratings to compute the distribution of the labels using the counts of the selected emotions. Then, we follow the work of Riera et al. [57] and Chen et al. [58] to use a threshold, $1/C$ (where

TABLE 1. Example of an Annotation From the CREMA-D Corpus in Which Six Raters Selected Anger (A) and Two Raters Selected Sadness (S). Other Emotions (Disgust (D), Fear (F), Happiness (H), and Neutral State (N)) Were Not Selected.

Label Processing	Raw Annotation	A*6, S*2
	Label for Training Stage	(0.75,0.25,0.0,0.0,0.0,0.0,0.0)
	Label for Testing Stage	(1,1,0,0,0,0)

C is the number of emotions) to binarize the presence of each emotion in the speech signal, creating hard-decisions for each emotion. Taking as example one of the datasets utilized in this study, for CREMA-D the number of target emotions is six, so the value of the threshold is set to 1/6. This "ad-hoc" design can mitigate the potential noise problem in the labels. Table 1 shows one example of the label processing approach used in this study. The labels are distributional labels for the training stage and are converted into binary vectors when the values are higher than the defined threshold.

C. EVALUATION METRICS

1) PERFORMANCE METRICS

Our evaluation framework employs macro-F1, micro-F1, and sample-precision scores to provide a clear assessment of the AVER systems' performance. These F1 metrics effectively measure both recall and precision rates. Macro-F1 computes the F1 score independently for each emotional class and averages the results, treating all classes equally. Micro-F1 aggregates counts across all classes before computing the score, giving more weight to frequent classes. Including both metrics provides complementary insights to evaluate the overall performance of the models. During evaluation, a prediction is considered correct if the emotion(s) present in the sentence according to the thresholding approach described in Section IV-B. This approach allows precise F1 score calculations, capturing the systems' ability to recognize a diverse range of emotions in imbalanced class scenarios.

2) CALIBRATION

We use probabilistic calibration, which is defined as the agreement between predicted probabilities and empirical frequencies in the predictions. Among predictions made with confidence p for an emotion e, approximately a fraction p should be correct. We quantify calibration with the *Brier score* (BS) [59], which is defined as the mean squared error between predicted probabilities and binary targets. For a multi-label setting with N samples and C emotions, we compute a macro-averaged BS as

$$\text{BS} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{e=1}^{C}\left(p_i^{(e)} - y_i^{(e)}\right)^2,$$

where $p_i^{(e)} \in [0, 1]$ is the probability for emotion e on sample i, and $y_i^{(e)} \in \{0, 1\}$ is the corresponding target. Lower BS indicates better calibration [60]. Note that calibration is complementary to thresholded recognition metrics (e.g., F1); a model can be well-calibrated but show different per-emotion F1, and vice versa.

D. IMPLEMENTATION DETAILS

The section describes how we implement the models in detail, including feature extraction, model configuration, training strategy, and baseline methods.

1) FEATURE EXTRACTORS DETAILS

For the visual feature extraction, we employ the pre-trained MobileNetV2 model [50], which has been fine-tuned on the AffectNet corpus [61]. AffectNet is a large-scale facial expression dataset containing more than one million facial images collected from the internet and annotated for eight discrete emotions (e.g., happiness, sadness, anger) as well as continuous valence and arousal values. We fine-tune the MobileNetV2 architecture on the 8-class emotion classification task to be used for feature extraction. The images are processed through MobileNetV2 to extract a 1,280-dimensional feature vector from the global pooling layer for each image, denoted as $x_v \in \mathbb{R}^{N_v \times 1,280}$, where N_v represents the total frames per sequence. For the acoustic features, we use the pre-trained WavLM-large model [58], available at "microsoft/wavlm-large.".[1] This process yields an acoustic feature set $x_a \in \mathbb{R}^{N_a \times 25 \times 1,024}$, with N_a as the sequence length, 1,024 as the dimensionality of the hidden states, and 25 representing the number of layers of hidden states used.

2) MODEL CONFIGURATION AND TRAINING SETTINGS

Fig. 1(a) illustrates the framework of the model. The context summarizer is configured with 128 learnable tokens, 6 layers with multi-head attention blocks (8 heads each), a dropout of 0.1, and an embedding dimension of 1,024. Both the cross-modal and self-attention layers have the same structure, featuring 3 layers, 8 heads, an embedding dimension of 1,024, an embedding dropout of 0.25, and dropouts for attention and residuals at 0.1. The prediction head includes three FC layers with 2,048, 1,024, and 256 dimensions. Adopting the approach of modality dropout strategy outlined in previous AVER studies [20], we implemented modality dropout during training to ensure our approach can properly handle non-ideal environments where modalities might be partially or completely absent. Specifically, for each training batch, we randomly zero out acoustic features 20% of the time, visual features 20% of the time, and leave features unchanged 60% of the time. This training strategy prepares the model to handle scenarios not encountered under optimal conditions. We train on a 6-class or 4-class emotion classification task, using the class balanced objective function, detailed in subsection IV-D3, for 60 epochs at a learning rate (`lr`) of 1×10^{-5}. We utilize the AdamW optimizer with a weight

[1] https://huggingface.co/microsoft/wavlm-large

TABLE 2. Overview of Models' Performances on the CREMA-D and MSP-IMPROV Datasets. The Columns Show the Average macro-F1 (Macro-F1), micro-F1 (Micro-F1), and Sample-Precision (Sample-Precision). In Brackets, We Present the Lower and Upper Bounds of the Confidence Interval Between 2.75% and 97.5% for Each Result.

Model	CREMA-D			MSP-IMPROV		
	Macro-F1 Score ↑	Micro-F1 Score ↑	Sample-Precision ↑	Macro-F1 Score ↑	Micro-F1 Score ↑	Sample-Precision ↑
TLSTM [47]	0.710 (.704, .716)	0.705 (.699, .711)	0.705 (.699, .711)	0.648 (.639, .657)	0.710 (.703, .716)	0.733 (.725, .741)
SFAV [46]	0.731 (.725, .737)	0.731 (.725, .736)	0.728 (.723, .734)	0.642 (.633, .650)	0.708 (.703, .715)	0.716 (.709, .724)
MulT [44]	0.743 (.738, .750)	0.743 (.738, .749)	0.741 (.736, .748)	0.651 (.643, .659)	0.710 (.703, .716)	0.728 (.720, .735)
AuxFormer [20]	0.742 (.737, .748)	0.742 (.737, .748)	0.741 (.734, .747)	0.672 (.663, .680)	0.728 (.722, .734)	0.740 (.733, .748)
VAVL [48]	0.772 (.767, .778)	0.770 (.765, .775)	0.770 (.765, .775)	0.680 (.673, .689)	0.749 (.741, .754)	0.771 (.766, .776)
RAVER	**0.777 (.771, .782)**	**0.772 (.766, .777)**	**0.772 (.766, .777)**	**0.692 (.683, .700)**	**0.752 (.746, .758)**	**0.779 (.772, .786)**

TABLE 3. Comparison of macro-F1 Scores Between Single-Modality Methods and Multimodal Models Under 100% Missing Audio or Video Conditions

Model	100% Missing Audio (FER)	100% Missing Video (SER)
FER [65]	0.658 (0.652, 0.665)	–
SER [66]	–	0.709 (0.703, 0.714)
AuxFormer [20]	0.710 (0.708, 0.711)	0.704 (0.703, 0.706)
SFAV [46]	0.673 (0.671, 0.674)	0.644 (0.642, 0.645)
TLSTM [47]	0.699 (0.697, 0.700)	0.678 (0.676, 0.680)
MulT [44]	0.705 (0.704, 0.707)	0.678 (0.676, 0.680)
VAVL [48]	0.659 (0.657, 0.661)	0.706 (0.704, 0.707)
RAVER (Ours)	**0.737 (0.736, 0.739)**	**0.751 (0.750, 0.753)**

decay of 5×10^{-7} and beta parameters of 0.95 and 0.999, and utilizing a batch size of 32. Experiments were conducted on an NVIDIA A100 GPU.

3) CLASS-BALANCED OBJECTIVE FUNCTION

This study not only takes model accuracy into account but also cares about the reliability and robustness of AVER systems. Miscalibrated systems mainly result from the imbalanced annotation distributions, so we follow Chou et al. [62] to introduce the *class-balanced cross-entropy* (CBCE) loss to mitigate the miscalibration of models' predictions. The CBCE was originally proposed by Cui et al. [63], and the main operation of the CBCE is using a weighting factor to control the loss values of each class during the training phase. The weighting factor is defined as $\frac{1-\beta}{1-\beta^{n_i}}$, where n_i is the number of positive samples in the i^{th} emotional class in the training set, the C is the number of emotional classes and $\beta \in (0, 1]$ is a hyperparameter. The number of weighting factors equals the number of target emotions. The CBCE value can be calculated using (3):

$$\mathcal{L}_{CBCE} = \sum_{i=1}^{C} \left(\frac{1-\beta}{1-\beta^{n_i}} \cdot \mathcal{L}_{CE}^{(i)} \right), \tag{3}$$

where $\mathcal{L}_{CE}^{(i)}$ is the value of the cross-entropy loss [64] for the i^{th} emotion.

E. BASELINES

We benchmarked our proposed model with five AVER frameworks, utilizing code from their respective repositories or specifications from associated papers.

VAVL: Proposed by Goncalves et al. [48], the VAVL model employs a versatile combination of acoustic-only and visual-only layers that process audio-visual content before merging into shared layers for joint learning discriminative features for AVER.

MulT: Proposed by Tsai et al. [44], the MulT model uses a cross-modal transformer framework designed for human language time series. We adapt this architecture to generate bi-modal representations by focusing only on visual and acoustic features, excluding the textual branch.

SFAV: Chumachenko et al. [46] presented a model for audio-visual emotion recognition that adapts to incomplete data by using robust fusion techniques such as late and intermediate transformer fusion for feature integration.

AuxFormer: Presented by Goncalves and Busso [20], the AuxFormer model uses transformer layers to create cross-modal representations, incorporating unimodal auxiliary networks and modality dropout to enhance robustness. This approach achieves strong performance in both multimodal and unimodal scenarios (e.g., audio-only or video-only).

TSLTM: Proposed by Huang et al. [47], the TSLTM model combines transformers and LSTM networks in their architecture for continuous emotion recognition, leveraging multi-head attention to fuse audio and visual data into a shared semantic space to model long-term emotional dynamics.

F. CONTEXT LENGTH OF RAVER SUMMARIZER

Our method involves one key parameter: the context size of the summarizer. Initial experiments on the development set explored various context sizes, including 16, 32, 64, 128, and 256. Fig. 2 presents the average of performances across different context lengths, simultaneously considering both audio and visual missing frames. To generate this plot, we averaged the results on CREMA-D from acoustic ablations and visual ablations at different percentage levels of missing frames. These findings suggest that while the performance across various context lengths is generally similar, a context size of 128 consistently achieved the highest

TABLE 4. Overview of Recognition Performances and Model Calibration of the AVER Systems for Each Emotion on the CREMA-D Dataset

Model	Brier Score ↓	Anger ↑	Sadness ↑	Disgust ↑	Fear ↑	Neutral ↑	Happiness ↑
TLSTM [47]	0.143	0.742 (.729, .753)	0.610 (.596, .624)	0.632 (.620, .643)	0.673 (.663, .684)	0.782 (.774, .790)	0.884 (.874, .892)
SFAV [46]	0.134	0.742 (.727, .757)	0.617 (.597, .637)	0.648 (.631, .664)	0.708 (.695, .723)	0.788 (.777, .800)	0.880 (.867, .892)
AuxFormer [20]	0.134	0.752 (.735, .768)	0.654 (.636, .673)	0.668 (.653, .684)	0.697 (.682, .712)	0.810 (.801, .821)	0.873 (.860, .886)
MulT [44]	0.128	0.765 (.749, .782)	0.637 (.617, .656)	0.666 (.649, .682)	0.713 (.699, .729)	0.803 (.793, .814)	0.875 (.861, .888)
VAVL [48]	0.128	0.789 (.774, .804)	0.669 (.651, .687)	**0.735 (.720, .748)**	0.726 (.712, .740)	**0.815 (.804, .825)**	0.900 (.888, .911)
RAVER	**0.119**	**0.796 (.782, .812)**	**0.669 (.650, .688)**	0.733 (.717, .747)	**0.736 (.721, .751)**	0.805 (.795, .816)	**0.922 (.910, .932)**

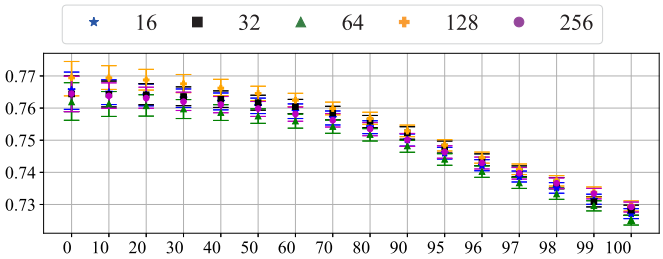
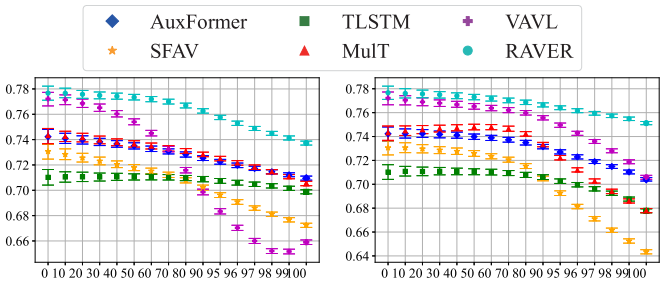


FIGURE 2. Preliminary experiments exploring the optimal context length. Y-axis represents macro-f1 scores on the CREMA-D dataset.



(a) Missing % of acoustic modality (b) Missing % of visual modality

FIGURE 3. Analysis of our system's performance when modalities are masked from the CREMA-D dataset. The macro-F1 results (y-axis) are obtained from masking acoustic or video frames from 0% to 100% (x-axis). Note that the x-axis uses a step size of 10% from 0 to 90 and 1% from 90 to 100 to provide finer granularity in extreme cases. Results are plotted with the mean, lower, and upper bound values.

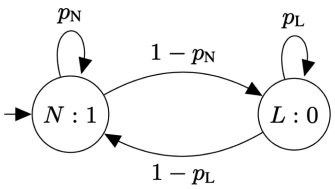


FIGURE 4. *Markov Chain* packet-loss model to simulate the loss of frames. State N:1 indicates that the frame is received, and state L:0 indicates that the frame is lost.

macro-F1 score under various conditions of missing audio and video on the development set. Therefore, we proceed with the experimental results and analysis of RAVER with these settings.

V. RESULTS AND ANALYSES

We aim to evaluate performance scores and bias of AVER systems across emotions. To fairly compare the effectiveness of the proposed method, we also show the results of the five different existing SOTA AVER systems mentioned in Section IV-E. Table 2 summarizes overall results on both the CREMA-D and MSP-IMPROV emotion datasets. As observed, RAVER consistently outperforms the other models across both datasets, demonstrating significant improvements over the SOTA AVER methods. In the following sections, we further explore the effectiveness of the proposed model under non-ideal conditions, such as missing modalities, and conduct emotional state bias analyses against SOTA AVER methods. Due to space limitations, we present detailed results from the CREMA-D database, although similar performance trends were observed with the MSP-IMPROV dataset.

A. MISSING MODALITY CONDITIONS

This section evaluates the performance of the AVER models in scenarios where modalities are missing. We use two methods to simulate missing data conditions: random masking and packet loss.

1) SIMULATION BY RANDOM MASKING

We simulate missing data conditions by randomly masking either visual or acoustic information at the frame level. As depicted in Fig. 3, we progressively mask input frames, starting from 0% and increasing to 90% in 10% increments, following previous studies [11], [20]. To further examine performance under extreme conditions, we apply masking from 95% to 100% in 1% increments. This analysis provides insight into how model performance varies with reduced audio-visual data and evaluates the model's ability to handle severe data absence.

Fig. 3 demonstrates that our proposed approach achieves the best AVER performance across all baselines and maintains strong performance across varying missing content percentages. Unlike VAVL, SFAV, or MulT, RAVER does not experience sharp performance drops, even in extreme cases (e.g., 95% or more missing content).

While Table 2 shows only a slight improvement of RAVER over the VAVL framework under ideal conditions, noticeably, our missing modality analysis reveals a larger gap under degraded scenarios. VAVL's performance

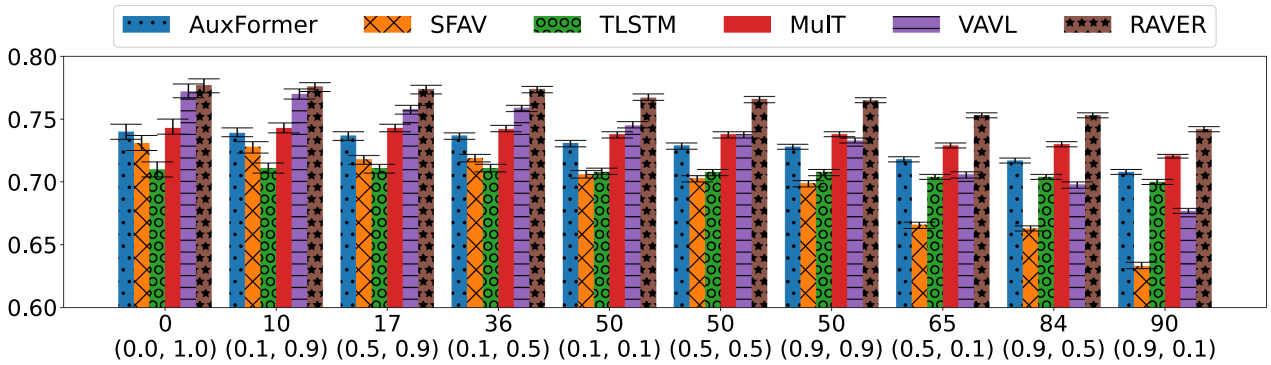


FIGURE 5. Overview of macro-F1 scores (Y-axis) on the CREMA-D dataset using the Markov chain packet-loss simulation of missing audio and video frames. The figure reports the *approximated percentage of dropped frames* (ADF) in the X-axis. The X-axis also reports the parameter values of the Markov chain packet-loss model used for each scenario (P_L, P_N).

significantly drops as missing frames increase, whereas RAVER maintains superior robustness, making it more effective for real-world applications with modality degradation. The results underscore the effectiveness of our proposed RAVER method, which leverages context summarization and modality dropout to ensure robustness even under extreme conditions.

2) SIMULATION BY A MARKOV CHAIN PACKET-LOSS

Simulating random missing frames is not always a proper strategy representing practical applications. For example, when transmitting frames over the Internet, the probability of missing frames increases if the previous frame was missed. To simulate audio-visual testing scenarios with missing data, we utilize the two-state *Markov Chain* packet-loss model, commonly used to simulate packet loss on the Internet. Following the strategy employed in other studies for emotion recognition tasks [11], [67], we establish frame-level probabilities for packet loss (p_L) and transmission (p_N), generating a randomly sampled binary sequence mask for the test sentence. Fig. 4 shows this model. The initial frame always begins in the non-loss state. This simulation masks audio and visual content simultaneously within the same regions.

Fig. 5 illustrates the comprehensive test results across varying values of p_L and p_N. The figure also includes the *approximated percentage of dropped frames* (ADF). As observed, the proposed RAVER achieves the best performance in all conditions compared to the baselines, demonstrating the effectiveness and resilience of the proposed method to audio-visual missing scenarios.

An interesting possible scenario is asynchronous cases of missing modalities. In practice, audio and video may be missing at different times. Our architecture also supports this setting. We maintain independent frame-level masks for audio and video (m_a, m_v); each modality's contextual summarizer attends only to unmasked frames, and cross-modal fusion operates after summarization. Consequently, intermittent and unsynchronized drops in either modality are handled without modifying the model architecture or training strategy.

3) COMPARISON WITH SINGLE-MODALITY EMOTION RECOGNITION

To contextualize the robustness of our framework under extreme missing modality conditions, we compare RAVER to recent single-modality models for *speech emotion recognition* (SER) [66] and *face emotion recognition* (FER) [65]. Table 3 summarizes the macro-F1 scores under the 100% modality loss scenarios on the CREMA-D dataset.

Table 3 shows that RAVER consistently outperforms the strongest unimodal baselines, particularly under the 100% missing visual or acoustic input. The SER model in Chou et al. [66] uses WavLM as the speech representation. Notably, RAVER exceeds this SER model despite both relying on the same acoustic features. This evaluation demonstrates that our architecture's summarization and fusion design provides robustness and generalization even in single-modality scenarios. During multimodal training, the models can learn robust representations that bring advantages even when only a single modality is available. For example, we have observed in our previous work that multimodal models evaluated with a single modality outperform the performance of their corresponding unimodal models [20], [68].

We attribute RAVER's resilience under extreme missing modality conditions to its contextual summarizer. The learnable query tokens within the summarizer enable the model to focus on the remaining informative frames even when 95% or more of the input is missing. This strategy stands in contrast to traditional models that rely on uniform pooling or full-sequence attention.

B. MODEL CALIBRATION AND PER-EMOTION RESULTS

In this work, we allow samples to have more than one emotion (multi-label formulation). This section aims to verify the calibration of the model's predictions and the recognition performance for each emotion. Table 4 presents measures of model calibration using the Brier Score and performance evaluations using the macro-score. RAVER outperforms all baseline models on four of the six emotion classes, indicating a stronger ability to generalize across emotional categories.

While the overall performance of RAVER achieves a Macro-F1 score of 0.77 for CREMA-D, as detailed in Fig. 2, we observe performance disparities between emotions, such as happiness (with a sample-F1 score of 0.922) and sadness (with a sample-F1 score of 0.669). Despite these performance disparities, Table 4 shows that RAVER achieves the lowest Brier Score (0.119) among the tested models. Therefore, RAVER not only achieves the best performance but also is the most well-calibrated model across emotions. This result highlights that RAVER not only improves overall robustness but also reduces emotion-specific variability more effectively than prior methods. While RAVER makes improvements, these results highlight the need for further effort to achieve equal performance across different emotions during training.

VI. CONCLUSIONS AND FUTURE WORK

This paper aims to enhance the robustness of an audio-visual emotion system under missing modality. Our approach shows robustness in handling missing modalities. It outperforms five current SOTA AVER systems, particularly in scenarios where over 95% of the input modality signals are absent.

In future research, we plan to address the fairness of AVER systems, motivated by our findings of bias in current leading AVER models. Our goal is to intentionally modify our architecture and training strategies to create models that are not only robust but also fair. Additionally, we intend to integrate the textual modality into the systems, as suggested by the study of Schmitz et al. [69], which indicates that the text modality exhibits less bias compared to audio and visual modalities. The CREMA-D corpus is particularly valuable as it includes demographic information about the speakers. Future advancements can be achieved by incorporating this feature in new data collections.

REFERENCES

[1] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5084–5088.

[2] A. Reddy Naini et al., "The Interspeech 2025 challenge on speech emotion recognition in naturalistic conditions," in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 4668–4672.

[3] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, San Antonio, TX, USA, Oct. 2017, pp. 415–420.

[4] L. Goncalves et al., "Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *Proc. Speaker Lang. Recognit. Workshop*, 2024, pp. 247–254.

[5] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[6] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *Proc. 7th Int. Seminar Speech Prod.*, Ubatuba-SP, Brazil, Dec. 2006, pp. 549–556.

[7] K. R. Park, H. J. Lee, and J. U. Kim, "Learning trimodal relation for audio-visual question answering with missing modality," in *Proc. Comput. Vis.*, in ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham, Switzerland, Springer, 2025, vol. 15073, pp. 42–59.

[8] G. Chochlakis, C. Lavania, P. Mathur, and K. J. Han, "Tackling missing modalities in audio-visual representation learning using masked autoencoders," in *Proc. Interspeech*, 2024, pp. 4678–4682.

[9] A. Wuerkaixi, K. Yan, Y. Zhang, Z. Duan, and C. Zhang, "Dyvise: Dynamic vision-guided speaker embedding for audio-visual speaker diarization," in *Proc. IEEE 24th Int. Workshop Multimedia Signal Process.*, Shanghai, China, 2022, pp. 1–6.

[10] R. Lin and H. Hu, "MissModal: Increasing robustness to missing modality in multimodal sentiment analysis," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 1686–1702, 2023, doi: 10.1162/tacl_a_00628.

[11] W.-C. Lin, L. Goncalves, and C. Busso, "Enhancing resilience to missing data in audio-text emotion recognition with multi-scale chunk regularization," in *Proc. ACM Int. Conf. Multimodal Interaction*, Paris, France, Oct. 2023, pp. 207–215.

[12] T. Mittal et al., "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, Feb. 2020, vol. 34, pp. 1359–1367.

[13] H. Zuo et al., "Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[14] J. Lin et al., "A time-domain convolutional recurrent network for packet loss concealment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Toronto, ON, Canada, Jun. 2021, pp. 7148–7152.

[15] C. Du et al., "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *Proc. ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, Oct. 2018, pp. 108–116.

[16] A. Triantafyllopoulos et al., "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1691–1695.

[17] H. Pham et al., "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, Jan./Feb. 2019, vol. 33, no. 1, pp. 6892–6899.

[18] C. Setz et al., "Using ensemble classifier systems for handling missing data in emotion recognition from physiology: One step towards a practical system," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction Workshops*, Amsterdam, Netherlands, Sep. 2009, pp. 1–8.

[19] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 206–218, Oct.–Dec. 2011.

[20] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2156–2170, Oct.–Dec. 2022.

[21] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Proc. Int. Conf. Multimodal Interaction*, Utrecht, The Netherlands, Oct. 2020, pp. 400–404.

[22] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan.–Mar. 2017.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014.

[24] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C. Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 22–38, Nov. 2021.

[25] C.-C. Lee, T. Chaspari, E. M. Provost, and S. S. Narayanan, "An engineering view on emotions and speech: From analysis and predictive models to responsible human-centered applications," *Proc. IEEE*, vol. 111, no. 10, pp. 1142–1158, Oct. 2023.

[26] A. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proc. Nat. Acad. Sci.*, vol. 114, no. 38, pp. E7900–E7909, Sep. 2017.

[27] Y. Lei and H. Cao, "Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2954–2969, Oct.–Dec. 2023.

[28] A. S. Cowen and D. Keltner, "Semantic space theory: A computational approach to emotion," *Trends Cogn. Sci.*, vol. 25, no. 2, pp. 124–136, 2021.

[29] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, no. 6–7, pp. 884–906, Jun. 2019.

[30] Y. Li, Y. Gao, B. Chen, Z. Zhang, G. Lu, and D. Zhang, "Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3190–3202, May 2022.
[31] H.-C. Chou et al., "Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, May 2022, pp. 7717–7721.
[32] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image Vis. Comput.*, vol. 26, no. 7, pp. 1052–1067, Jul. 2008.
[33] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, May 2022, pp. 6447–6451.
[34] J. Chen and A. Zhang, "HGMF: Heterogeneous graph-based fusion for multimodal data with incompleteness," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Jul. 2020, pp. 1295–1305.
[35] M. Liu et al., "Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interaction*, Istanbul, Turkey, Nov. 2014, pp. 494–501.
[36] Y. Dai et al., "A study of dropout-induced modality bias on robustness to missing video frames for audio–visual speech recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 27445–27455.
[37] M. Cheng and M. Li, "Multiinput multioutput targetspeaker voice activity detection for unified, flexible, and robust audiovisual speaker diarization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 3522–3536, 2025.
[38] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, London, U.K., Aug. 2024, pp. 1158–1166.
[39] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang, "Metric learning on healthcare data with incomplete modalities," in *Proc. Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 3534–3540.
[40] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 4971–4980.
[41] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4400–4407.
[42] N. Wang, H. Cao, J. Zhao, R. Chen, D. Yan, and J. Zhang, "M2R2: Missing-modality robust emotion recognition framework with iterative data augmentation," *IEEE Trans. Artif. Intell.*, vol. 4, no. 5, pp. 1305–1316, Oct. 2023.
[43] F. Ma, S. L. Huang, and L. Zhang, "An efficient approach for audio-visual emotion recognition with missing labels and missing modalities," in *Proc. IEEE Int. Conf. Multimedia Expo*, Shenzhen, China, Jul. 2021, pp. 1–6.
[44] Y.-H. Tsai, S. Bai, P. Liang, J. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, vol. 1, pp. 6558–6569.
[45] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audio-visual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7357–7361.
[46] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *Proc. 26th Int. Conf. Pattern Recognit.*, Los Alamitos, CA, USA, Aug. 2022, pp. 2822–2828.
[47] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3507–3511.
[48] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 1, pp. 306–318, Jan.–Mar. 2025.
[49] H. Wu et al., "EMO-SUPERB: An in-depth look at speech emotion recognition," Mar. 2024, *arXiv:2402.13018v4*.
[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[51] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with MTCNN," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng.*, 2017, pp. 424–427.
[52] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5998–6008.
[53] V. Iashin, W. Xie, E. Rahtu, and A. Zisserman, "Sparse in space and time: Audio-visual synchronisation with trainable selectors," in *Proc. Brit. Mach. Vis. Conf.*, 2022, pp. 1–15.
[54] K. Okabe et al., "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.
[55] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *J. Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.
[56] H.-C. Chou, L. Goncalves, S.-G. Leem, A. N. Salman, C.-C. Lee, and C. Busso, "Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule," *IEEE Trans. Affect. Comput.*, vol. 16, no. 1, pp. 41–55, Jan.–Mar. 2025.
[57] P. Riera, L. Ferrer, A. Gravano, and L. Gauder, "No sample left behind: Towards a comprehensive evaluation of speech emotion recognition systems," in *Proc. Workshop Speech, Music Mind*, Graz, Austria, Sep. 2019, pp. 11–15.
[58] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
[59] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.
[60] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007, doi: 10.1198/016214506000001437.
[61] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.
[62] H.-C. Chou, L. Goncalves, S.-G. Leem, C.-C. Lee, and C. Busso, "The importance of calibration: Rethinking confidence and performance of speech multi-label emotion classifiers," in *Proc. INTERSPEECH*, 2023, pp. 641–645.
[63] Y. Cui et al., "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 9268–9277.
[64] I. J. Good, "Rational decisions," *J. Roy. Stat. Soc., Ser. B (Methodological)*, vol. 14, no. 1, pp. 107–114, 1952.
[65] L. Goncalves, H.-C. Chou, A. N. Salman, C.-C. Lee, and C. Busso, "Jointly learning from unimodal and multimodal-rated labels in audio-visual emotion recognition," *IEEE Open J. Signal Process.*, vol. 6, pp. 165–174, 2025.
[66] H.-C. Chou, H.-T. Wu, and C.-C. Lee, "Stimulus modality matters: Impact of perceptual evaluations from different modalities on speech emotion recognition system performance," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Hyderabad, India, 2025, pp. 1–5.
[67] M. Mohamed and B. Schuller, "ConcealNet: An end-to-end neural network for packet loss concealment in deep speech emotion recognition," May 2020, *arXiv:2005.07777*.
[68] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Proc. Interspeech*, Incheon, South Korea, Sep. 2022, pp. 1168–1172.
[69] M. Schmitz, R. Ahmed, and J. Cao, "Bias and fairness on multimodal emotion detection algorithms," 2022, *arXiv:2205.08383*.

LUCAS GONCALVES (Member, IEEE) received the Ph.D. degree in electrical engineering from The University of Texas at Dallas (UTD), Richardson, TX, USA, in 2024. From 2022 to 2024, he was the recipient of Erik Jonsson School Excellence in Education Doctoral Fellowship. He is currently an Applied Scientist with Amazon, USA. His research interests include multimodal signal processing and deep learning, with emphasis on audio–visual learning, speech and language technologies, and vision–language models.

HUANG-CHENG CHOU (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2016 and 2024, respectively. From 2021 to 2022, he was a Visiting Scholar with the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, TX, USA. He is currently a Postdoctoral Scholar - NSTC Fellow with the University of Southern California (USC), Los Angeles, CA, USA. His research interests include affective computing, trustworthy AI, speech emotion recognition (SER), audio-visual emotion recognition (AVER), automatic deception detection (ADD), and the development of fair, calibrated, and robust large-scale speech and audio models. He was the recipient of the ACLCLP Doctoral Dissertation Award (Honorable Mention) in 2024 and APSIPA ASC Best Regular Paper Award in 2019. He is a member of IEEE Signal Processing Society, International Speech Communication Association (ISCA), and Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

ALI N. SALMAN received the B.S. and M.S. degrees in computer science from Indiana State University, Terre Haute, IN, USA, in 2015 and 2017, respectively, and the Ph.D. degree in electrical engineering from the University of Texas at Dallas, Richardson, TX, USA, in 2024. He is currently a Research Scientist with ARRAY Innovation. His research interests include affective computing, retrieval-augmented generation (RAG) systems, and facial analysis.

CHI-CHUN LEE (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2007 and 2012, respectively. He is currently a Professor with the Department of Electrical Engineering, National Tsing Hua University (NTHU), Hsinchu, Taiwan. He is the coauthor on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. His research interests include speech and language, affective computing, health analytics, and behavioral signal processing. He has been an Associate Editor-in-Chief of IEEE TRANSACTION ON AFFECTIVE COMPUTING (since 2025), IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE (2025-2027), IEEE TRANSACTION ON MULTIMEDIA (2019-2020), *Journal of Computer Speech and Language* (since 2021), *APSIPA Transactions on Signal and Information Processing* (2022-2024) and TPC member of IEEE SLTC, APSIPA IVM and MLDA committee. He is the General Chair of ASRU 2023 and Area Chair of Interspeech 2016, 2018, and 2019, respectively. He was the recipient of the NSTC Outstanding Research Award (2024), CIEE Outstanding Electrical Engineering Professor Award (2025), IICM K. T. Li Cornerstone Award (2024), and NTHU-Novatek Distinguished Talent Chair (2025). He led a team to the 1st place in Emotion Challenge in Interspeech 2009 and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is also an ACM and ISCA member.

CARLOS BUSSO (Fellow, IEEE) received the B.S. and M.S. (with highest Hons.) degrees in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2008. He is currently a Professor with Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, where he is also the Director of Multimodal Speech Processing (MSP) Laboratory. His research interests include human-centered multimodal machine intelligence and applications, focusing on the broad areas of speech processing, affective computing, multimodal behavior generative models, and foundational models for multimodal processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He was the recipient of NSF CAREER Award, ICMI Ten-Year Technical Impact Award, in 2014, Hewlett Packard Best Paper Award at IEEE ICME 2011 (with J. Jain), Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie), Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie), Distinguished Alumni Award in the Mid-Career/Academia category by the Signal and Image Processing Institute (SIPI) at the University of Southern California, in 2023, and 2023 ACM ICMI Community Service Award. His students were awarded the third prize IEEE ITSS Best Dissertation Award (N. Li) in 2015 and AAAC Student Dissertation Award (W.-C. Lin) in 2024. He is the Senior Area Editor of IEEE/ACM SPEECH AND LANGUAGE PROCESSING. He is a member of AAAC and senior member of ACM. He is an ISCA fellow.