Robust Audiovisual Emotion Recognition: Aligning Modalities, Capturing Temporal Information, and Handling Missing Features

Lucas Goncalves, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE*

Abstract—Emotion recognition using audiovisual features is a challenging task for human-machine interaction systems. Under ideal conditions (perfect illumination, clean speech signals, and non-occluded visual data) many systems are able to achieve reliable results. However, few studies have considered developing multimodal systems and training strategies to build systems that can perform well under non ideal conditions. Audiovisual models still face challenging problems such as misalignment of modalities, lack of temporal modeling, and missing features due to noise or occlusions. In this paper, we implement a model that combines auxiliary networks, a transformer architecture, and an optimized training mechanism to achieve a robust system for audiovisual emotion recognition that addresses, in a principled way, these challenges. Our evaluation analyzes how well this model performs in ideal conditions and when modalities are missing. We contrast this method with other multimodal fusion methods for emotion recognition. Our experimental results based on two audiovisual databases demonstrate that the proposed framework achieves: 1) improvements in emotion recognition accuracy, 2) better alignment and fusion of audiovisual features at the model level, 3) awareness of temporal information, and 4) robustness to non-ideal scenarios.

Index Terms—Multimodal emotion recognition, audiovisual fusion and alignment, attention model, auxiliary networks.

1 Introduction

Emotion recognition plays an important role in human interactions. Emotional states carry important information about an individual, including her/his intention, attitude, and decision making process. During an interaction, we mainly use verbal and visual cues to recognize emotional states. Speech carries expressive information such as intonation, rhythm, and resonance. Facial expressions are used to emphasize and/or give more context to an utterance. These modalities are essential for humans to effectively convey and perceive a message. *Human computer interaction* (HCI) aims to achieve the same emotional perception as humans. Therefore, an effective model for multimodal emotion recognition should be robust to non ideal conditions, mimicking how humans interact in real-world scenarios.

Previous studies have considered emotion recognition using mostly textual, vocal, or visual features [1], [2]. These models can be successful in areas such as movie rating sentiment analysis [3], speech emotional recognition for virtual assistants [4], and facial expression recognition based on a single picture [5]. However, early studies conducted by De Silva et al. [6] and Chen et al. [7] have shown that bimodal methods generally perform better than unimodal emotion recognition systems. While these models have achieved good performance, they do not necessarily process the data the way humans recognize emotions during an interaction. In daily interactions, verbal and visual cues are aligned and simultaneously perceived as the message is transmitted during a conversation. Therefore, a model should combine audiovisual cues similarly to how humans perceive these cues. One of the issues with many current audiovisual models is the assumption that audiovisual features are timely synchronized. However, studies have shown that audiovisual features are not necessarily synchronized, showing a time-shift difference longer than three video frames (i.e., hundred of milliseconds) [8], [9], [10]. Bregler and Konig [11] reported that the best alignment shift for audiovisual modalities is 120 milliseconds, assuming that lip movements precede speech. Furthermore, this time-variant phase between the modalities is not consistent time-wise, where either of the modalities can precede the other [8], [12], [13]. Another important aspect for multimodal emotional modeling is to consider the dynamic nature of emotions, which are externalized over-time. Knowing an emotional state from previous time-steps/frames helps in predicting the current emotion [14], [15]. It is essential to capture temporal information so that the model can track emotional changes within the speaking turn. Furthermore, during naturalistic human-to-human interactions, sometimes a modality will be missing for certain periods of time. For example, someone's face might not be visible during parts of an interaction or speech may not be audible for some period of time. Therefore, it is important to have a model that can handle these scenarios as well. Recent advances in deep learning approaches have worked on models that fuse [16] and align modalities [17] or successfully handle missing modalities [18]. However, to the best of our knowledge, we have not encountered a method that is able to address all these issues at once.

Our main contribution is the proposal of a novel audiovisual emotion recognition model, which combines three ma-

• *L. Goncalves and C. Busso are with the Erik Jonsson School of Engineering & Commputer Science, The University of Texas at Dallas, Richardson TX 75080.*
E-mail: goncalves@utdallas.edu, busso@utdallas.edu

Manuscript received May xx, 2022; revised XX XX, 20xx.

jor components: auxiliary networks, a transformer architecture, and an optimized training mechanism. The proposed model achieves robustness for multimodal emotion recognition, ensuring alignment between modalities, capturing temporal information, and handling missing features. The approach uses visual features extracted using a pretrained VGG-Face model at the frame level from the speaker's face and *low-level descriptors* (LLDs) extracted from speech. A key feature of our approach is the use of the dot-product attention mechanism and positional encodings, which are used to combine and temporally align the audiovisual modalities. We train our model using an optimized training mechanism to add robustness to non-ideal scenarios where modalities might be missing. This goal is also achieved by simultaneously training two auxiliary networks, which separately embed each modality into new vector spaces. The representations obtained from the auxiliary networks are used to provide an extra layer of information to our framework. Since each auxiliary network only contains features from a single modality, they add robustness to the system in scenarios when only one modality is available, leading to good performance even when evaluated in a unimodal setting.

The proposed method achieves the highest performance, with a micro F1-score of 76.1% and macro F1-score of 70.3% on the MSP-IMPROV corpus, and a micro F1-score of 77.3% and macro F1-score of 77.2% on the CREMA-D corpus. In our experiments, we show that our model can perform well under conditions where only visual or acoustic cues are available, demonstrating the model's ability to obtain strong performance with low access to visual and acoustic data. Our architecture is able to shift focus between what modalities are available at specific times, guaranteeing that its performance is not drastically affected by non-ideal scenarios. We compare this attention based model with three strong attention based models that use similar architectures for modality fusion. Further contributions include the experiments with modality ablations of the model to understand how our model performs under non-ideal conditions and the implementation of architectural ablations to understand how certain blocks in the model's architecture affects its performance.

This paper is organized as follows: Section 2 reviews previous studies that are relevant to this work. Section 3 describes the model architecture introduced in this study. Section 4 presents the experimental settings of the model's architecture, databases used to train and test the model, and feature extraction methods used in our evaluation. Section 5 presents our experimental evaluations, contrasting the results of the proposed model with the results of strong baselines. Lastly, Section 6 summarizes the contributions of this work while analyzing the advantages and disadvantages of this approach, as well as discussing possible future research directions.

2 Related Work

This study explores the use of the attention mechanism to combine modalities and generate alignment between audiovisual features for emotion recognition. Previous studies have focused on feature level integration [19], [20], [21], [22], [23], [24], decision level integration [25], [26], [27], [28], and model level integration [17], [18], [29], [30], [31], [32], [33], [34], [35]. Over recent years, important contributions have been made on model level integration using deep learning formulations. With the introduction of the transformer architecture [36], multimodal models have been proposed that perform textual, visual, and acoustic feature fusion using this framework. We focus our bibliography review on the most recent studies, which have considered similar problems as our study.

2.1 Audiovisual Temporal Modeling and Alignment

Studies in the past have shown that audiovisual features are not synchronized, meaning there is a time-shift difference between the audio and visual streams [8], [9], [10]. Therefore, several studies have worked on dealing with alignment and temporal modeling of audiovisual features [37]. Some studies have proposed aligning the modalities using *canonical correlation analysis* (CCA) [38], [39]. Halperin et al. [40] proposed a method that uses *dynamic time warping* (DTW) to find an optimal temporal audiovisual alignment. Another method was proposed by Chung and Zisserman [41], where they made the assumption that in television broadcasts the audio and video are usually synced. The authors then used this assumption as a standard to generate a model that determines the lip-synchronization error in videos using a contrastive loss between audio and visual features to achieve synchronization. However, such an assumption does not necessarily offer proper alignment [12], since audio and video misalignment is often non-uniform and irregular.

In the area of multimodal emotion recognition, authors have also explored methods to achieve proper multimodal alignment. Recently, attention mechanisms have been widely used for audiovisual temporal modeling and alignment. Chao et al. [42] proposed a method that considers sub-sequences of a whole sequence and uses a soft attention mechanism to align audio and visual streams. Wang et al. [43] presented the AlignNet, a model for audiovisual alignment which uses attention, pyramidal processing, warping, and affinity to generate audiovisual alignment through implicit mappings between the modalities. Tsai et al. [17] proposed a multimodal transformer architecture for unaligned language sequences. This model uses 1D temporal convolutions to retrieve temporally aware input textual, acoustic, and visual feature vectors and build a cross-modal transformer model that injects and aligns different modalities using an attention mechanism. Due to the unified approach to build cross-modal relationships between the modalities, it is difficult for these models to retain modality-specific temporal modeling information. Additionally, classical approaches mentioned earlier, often fail to capture long-range cross-modal relationships and requires manual pre-processing steps to define specific word or utterance times to perform proper alignment.

2.2 Handling Missing Modalities

During human-human interactions, modalities used for emotion recognition may be missing or unreliable for unforeseeable periods of time (e.g., facial occlusion, acoustic

noise in the environment). Traditional multimodal methods do not consider these scenarios, assuming ideal-scenarios. A few recent studies have presented strategies to handle missing modalities [18], [44], [45], [46], [47], [48]. Mittal et al. [44] proposed using CCA to create a check step, which generates proxy features to replace missing modalities. Du et al. [48] used a semi-supervised multiview deep generative framework to process missing modalities and treat them as latent variables to be integrated out during inference. Chen et al. [45] proposed a heterogeneous graph-based hypernode framework to enable multimodal fusion of incomplete data within a graph structure.

More recently, Ma et al. [46] proposed the use of a correlation loss based on the *Hirschfeld-Gebelein-Rényi* (HGR) maximal correlation to extract common information between the audiovisual modalities to handle missing modalities. Parthasarathy and Sundaram [18] proposed training strategies to train neural networks with missing modalities by generating ablations to the visual inputs during training. This approach disregards introducing ablations to the audio modality, which leads to a system that is highly dependable on acoustic features and it is not able to handle scenarios where audio is missing or noisy. Although our proposed method uses a similar strategy to the one presented by Parthasarathy and Sundaram [18], our method considers both audio and visual modalities to ensure robustness against any missing modality.

2.3 Relation to Prior Work

Our proposed model expands and presents important contributions with respect to previous studies. It builds on our preliminary study [49], extending the analysis and experimental evaluations to understand the benefits of each component in our framework. The attention mechanism present in the transformer architecture offers the possibility of easy integration of multimodal features. The closest study to our paper is the work of Tsai et al. [17], which proposed a multimodal transformer architecture for unaligned language sequences. Our proposed method implements pieces of their idea by using cross-modal alignment with transformers. Our model adapts this formulation to audiovisual tasks, enhancing it with auxiliary networks and a combination of loss functions, which have been successful strategies in tasks involving images or videos [50], [51]. Our method implements a training method which strategically ablates both the visual and acoustic features to enhance our model's robustness to missing modalities. Additionally, we incorporated the use of learnable scalar weights to later self-attention layers present in our model to scale and emphasize the most relevant features for our task. The resulting architecture provides a strong solution for aligning the modalities, capturing temporal information, and handling missing features.

3 Methodology

The proposed framework consists of three networks: the main audiovisual network $\mathcal{F}_{av}(\bullet)$, the auxiliary acoustic network $\mathcal{F}_{a}(\bullet)$, and the auxiliary visual network $\mathcal{F}_{v}(\bullet)$. The three networks rely on the attention mechanism to generate strong representations, which have proven to work well not only for *natural language processing* (NLP) tasks, but also for multimodal tasks. The dot-product attention mechanism allows the network to compute attention scores between the modalities to best align their features, while combining their representations at the model level. The resulting features from the attention fusion approach are then passed through fully connected layers to perform the final classification.

Figure 1 depicts an overview of the model. Before providing a detailed description of each of its blocks, we summarize the main components of our framework, and their roles. The proposed framework has five major components: 1) *Projection of Features and Positional Encodings* (Sec. 3.1). At this step, the network receives the acoustic and visual features. Since the dimensions of the visual and acoustic features do not match, this component is used to project the visual features to a lower dimensional space to match the dimension of the acoustic features. Subsequently, we add positional encodings to both modalities to preserve temporal information. 2) *Fusion Attention Layers* (Sec. 3.2). This block aligns and generates visual-audio and audio-visual cross-modal representations with multi-head attention layers. 3) *Self-Attention Layers with Learnable Scaling Parameters* (Sec. 3.3). This block computes attention scores within the representations obtained from the fusion attention layers to give more emphasis to areas that are more useful for our discriminative task. 4) *Multilayer Perceptron and Classification* (Sec. 3.4). This block processes the concatenated representations generated by the self-attention layers through fully-connected layers to perform emotion recognition. 5) *Audio and Visual Auxiliary Networks* (Sec. 3.5). These auxiliary networks are identical unimodal architectures composed of self-attention layers and fully-connected layers to individually extract additional representations from the acoustic and visual features. These auxiliary networks help the model handle missing modalities. Another important part of the model is the optimized training strategy designed to increase robustness against missing modalities (Sec. 3.6). This Section describes these blocks in detail.

3.1 Projection of Features and Positional Encodings

The first step in our model is to ensure that the features extracted from the visual (Sec. 4.2.1) and acoustic (Sec. 4.2.2) data have the same dimensions. The visual frame-level feature vector obtained from the fine-tuned VGG-face model [52] is $x_v \in \mathbb{R}^{N_v \times 4096}$, where N_v is the dimension of the visual feature sequence, and 4,096 is the feature vector dimension. The acoustic feature vector extracted from the LLDs is $x_a \in \mathbb{R}^{N_a \times 130}$, where N_a is the dimension of the acoustic feature sequence, and 130 is the feature vector dimension. Our fusion attention layers require that the feature embedding size between the modalities are equal. Our strategy is to keep the extracted feature vectors as intact as possible. Therefore, we do not perform representation projection for both modalities. Instead, we reduce the dimension of the visual modality to match the dimension of the audio modality, mapping the feature vectors from 4,096 to 130 ($\bar{x}_v \in \mathbb{R}^{N_v \times 130}$). We implement this projection with a 1D-convolutional layer.

Once the visual and acoustic modalities have matching dimensions, we add 1D-positional embeddings to the feature vectors, following the method used by Vaswani et al.

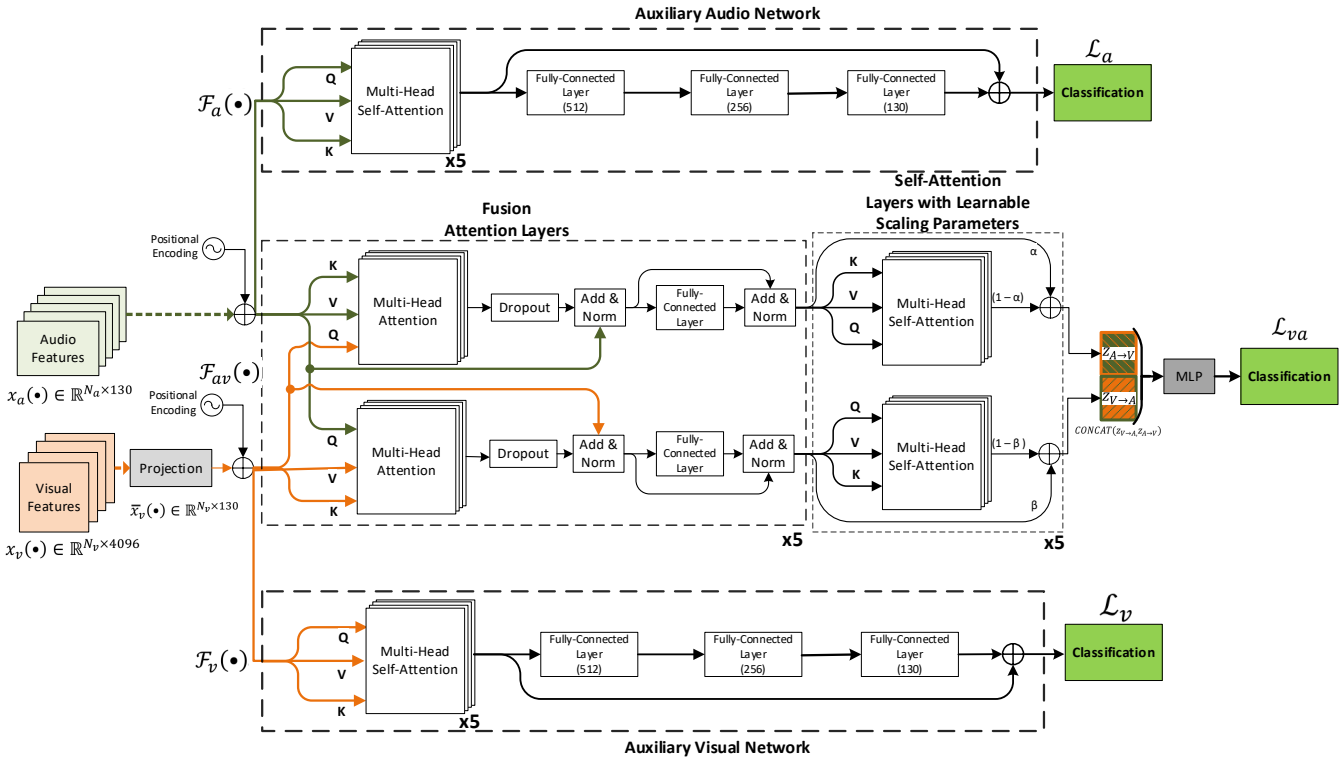


Fig. 1. Model Overview. Three networks are used to construct our framework: the main audiovisual network $\mathcal{F}_{av}(\bullet)$, the auxiliary acoustic network $\mathcal{F}_a(\bullet)$, and the auxiliary visual network $\mathcal{F}_v(\bullet)$. The proposed model receives acoustic and visual features, fusing them with cross-modal attention layers. We further encode the cross-modal features with self-attention layers equipped with residual connections. The auxiliary networks are used to generate separate unimodal representations for each modality to increase robustness against missing features.

[36]. Equation 1 shows the process, where x_m^i represents the input features for the modality m, with $i \in \{1, \ldots, N\}$ corresponding to the ith frame in a sequence of length N. $\mathbf{E}_{pos}$ represents the positional embedding added to the input features. The positional embeddings are important for our model. Since our model is accepting input sequences in parallel, it needs the positional embeddings to retain the temporal information. Enabling the input sequences to carry temporal information helps the network with creating the alignment in the multi-head attention layers (Sec. 3.2).

$$\mathbf{z}_0 = [x_m^1; x_m^2; \cdots; x_m^N] + [\mathbf{E}_{pos} \in \mathbb{R}^{N \times 130}] \tag{1}$$

3.2 Fusion Attention Layers

The fusion attention layers aim to align and combine the feature representations of the modalities. The block uses two separate *multi-head attention* (MHA) layers that concurrently fuse the audiovisual modalities. The embedded feature vectors are used to generate the Q, V, and K matrices. To inject information from one modality to the other, we share the Q vectors from one modality (m_1) at the MHA layer to the other modality (m_2) and vice-versa as shown in Equation 2 and depicted in Figure 1 under the block for the fusion attention layers.

$$\mathbf{z}_{m_1 \to m_2} = softmax(\frac{Q_{m_2} K_{m_1}^\top}{\sqrt{N_{m_1}}}) V_{m_1} \tag{2}$$

The attention heads are each independently initialized. Therefore, each head computes a different attention score from one modality to another. With this approach, the model is able to better determine how much attention each frame of one modality has to pay to certain frames of the second modality. This setting consequently enables the model to use attention scores to not only fuse the two modalities, but also align the audiovisual features. Furthermore, this block also includes two *layer normalization* (LN) and residual connections, referred to as *Add & Norm* in Figure 1. The normalized layers at this step are then added to earlier feature matrices connected via residual connections [53], [54] (Eqs. 3-4). These earlier layers are represented by $\mathbf{z}_{l-1}$ (feature matrix obtained before the MHA layer) and $\mathbf{z}'_l$ (feature matrix obtained before the FC layer). The residual connections contain representations from the same modality that was present in the K and V matrices of the preceding MHA layer. Residual connections are included so the model does not suffer from vanishing gradients, due to the model depth, or suffer from losing positional embedding information through the layers. The fusion attention layers utilize an FC layer to embed the features and generate $\mathbf{z}_l$ before passing the features forward into the self-attention layers with learnable scaling parameters. The activation function is the *rectified linear unit* (ReLU) function (Eq. 4).

$$\mathbf{z}'_l = MHA(LN(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \tag{3}$$

$$\mathbf{z}_l = LN(FC(LN(\mathbf{z}'_l)) + \mathbf{z}'_l) \tag{4}$$

3.3 Self-Attention Layers with Learnable Scaling Parameters

After the inter-modality injections and alignment through the cross-attention mechanism is completed, the two audiovisual feature vectors are passed through the intra-modality self-attention layers equipped with residual connections. The self-attention layers have a structure similar to that of our fusion attention layers. However, there is no information flowing from one feature vector pipeline to the other. The Q, V, and K matrices are contained within their own self-attention layer, computing separate feature representations. The use of the self-attention module after the fusion attention layers helps our model emphasizes areas of the representations extracted from the fusion attention layers that are more useful for our discriminative task. The self-attention layers have the layer-independent learnable scaling weights α and β. The learnable scaling weights α and β are added to increase our model's capability to accordingly scale representations in situations where acoustic or visual features are missing. Learnable scaling weights increase the robustness of our model against missing modalities.

3.4 Multilayer Perceptron and Classification

The representations are obtained from the self-attention layers with learnable scaling parameters by extracting classification tokens from the last time step of each sequential combination of audiovisual features. The last step of the encoded sequences have a receptive field of the entire sequence length, which gives us a global representation of the input sequence. Retrieving tokens from the last time step helps us minimize the effect of having embeddings from acoustic and visual features that could be in different ranges, and ensures that we cover the entire temporal range of each stream. Two token vectors are obtained from the self-attention layers, which are concatenated, as indicated in Equation 5,

$$\mathbf{z}_{class} = CONCAT(\mathbf{z}_{V\rightarrow A}, \mathbf{z}_{A\rightarrow V}) \tag{5}$$

where $\mathbf{z}_{class}$ is the resulting vector from the concatenation of the tokens obtained from the audio to visual fusion $\mathbf{z}_{A\rightarrow V}$ layers and the tokens obtained from the visual to audio fusion $\mathbf{z}_{V\rightarrow A}$ layers. The classification vector ($\mathbf{z}_{class}$) is then passed through a *multilayer perceptron* (MLP). The output layer ultimately receives the embeddings from the last hidden layer and outputs the classification probabilities.

3.5 Audio and Visual Auxiliary Networks

In addition to the main audiovisual fusion and alignment architecture, we simultaneously train two auxiliary networks, one for the visual features and the other for the acoustic features. These auxiliary networks separately embed each modality into new latent spaces. The two identical auxiliary networks are incorporated in our model to provide extra layers of meaningful representations for our classification task. We were inspired by the study of Szegedy et al. [51], which showed the benefits of auxiliary networks for very deep models. Differently from the work by Szegedy et al. [51], in our framework the auxiliary networks are not introduced in middle layers of the backbone network. We made the architecture choice of placing the auxiliary networks at the input because we understood that it is important for the auxiliary networks to process and learn from the unimodal input features independently from the layers present in our fusion networks. Additionally, right at the beginning of our fusion network, we utilize cross-attention to combine the acoustic and visual features. Therefore, if we had the auxiliary networks placed at the middle layer of the backbone network, we would not be able to generate unimodal representations. In our model, the auxiliary networks are especially important because they help the model handle missing modalities. Each auxiliary network is unimodal. Therefore, having the auxiliary networks ensures that our model still contains layers which carry full information even though the other modality is missing. The auxiliary networks also have residual connections across their fully-connected layers with shared weights to ensure that important information is not lost during training and inference.

Our complete framework is composed by three networks, and each network has its own cross-entropy loss $\mathcal{L}_m$, where $m \in \{a, v, av\}$, a represents the auxiliary audio network, v represents the visual auxiliary network, and av represents the audiovisual network. This formulation with the auxiliary networks is inspired by the study of Piergiovanni et al. [50], which explored a mechanism that learns multimodal representations by distilling shared information through losses from multiple streams of networks. Our model combines the cross-entropy losses generated by the three networks to get the total loss $\mathcal{L}_{total}$,

$$\mathcal{L}_{total} = \lambda_{va}\mathcal{L}_{va} + \lambda_a\mathcal{L}_a + \lambda_v\mathcal{L}_v, \tag{6}$$

where $\lambda_{va}, \lambda_a, \lambda_v$ are the weights for the specific losses of the audiovisual, audio, and visual networks, respectively.

3.6 Optimized Training Procedure

Our goal with the proposed model is to achieve robust performance in non-ideal environments with missing modalities. We consider cases with partial or total missing information for the audio/video feature sequence. We accomplish this goal with an *optimized training procedure*. We randomly split each batch in three groups. We replace the acoustic features with zeros in the first group (20%), and the visual features with zeros in the second group (20%). The audio and visual features in the third group (60%) are not altered. The optimized training procedure aims to simulate the presence of non-ideal scenarios that our model would not have seen if it was only trained under ideal condition settings. We compare this optimized training procedure with a standard training procedure consisting of using the entire training data available without discarding any audio or visual feature.

4 Experimental Settings

4.1 Corpora

This study uses the CREMA-D corpus [55] and the MSP-IMPROV corpus [56]. The CREMA-D corpus is an audiovisual dataset containing high quality recordings collected from a racially and ethnically diverse group of 91 actors (48 male and 43 female). The actors were given a set

of sentences and asked to say every sentence targeting a specific emotional state. Videos were recorded with green screen in the background. Two directors supported this data collection effort, where 51 actors worked with one director, and 40 actors worked with the second director. The emotional labels were rated by at least seven annotators under different conditions: audio only, video only, and audiovisual recordings. In total, 7,442 clips were collected and rated by 2,443 raters. The distributions of the classes for the audiovisual modality consensus labels are: 1,067 anger clips (10,054 ratings), 1,222 disgust clips (11,429 ratings), 1,180 fear clips (11,153 ratings), 1,230 happy clips (11,730 ratings), 672 sad clips (6347 ratings), and 2,071 neutral clips (19450 ratings). In this study, we used the perceived emotions from the audiovisual modality provided in the CREMA-D corpus. We use six classes for this corpus: anger, disgust, fear, happiness, sadness, and neutral state.

The MSP-IMPROV corpus [56] is the second audiovisual database used in this study. The corpus was collected to explore the perception of emotion [57]. One of the requirements was to have sentences with the same lexical content, but with different emotions. Instead of asking actors to read sentences expressing different emotions, the corpus relied on a more sophisticated protocol to elicit spontaneous renditions of the target sentences. The protocol created hypothetical dyadic scenarios that led one of the participants to say a target sentence in a given emotion. The corpus includes 20 target sentences using four different emotional states (happiness, sadness, anger, and neutrality), generating 80 scenarios. This portion of the corpus consists of 652 speaking turns. The corpus also includes the rest of the interactions that led one of the actors to utter the target sentence (4,381 spontaneous speaking turns), and natural interactions collected between the dyadic scenarios (2,785 natural speaking turns). It also includes read recordings of the target sentences, expressing the target emotional classes (620 read speaking turns). Overall, the MSP-IMPROV corpus contains 7,818 non-read speaking turns, and 620 read sentences. The corpus was annotated with a crowd-sourcing protocol that tracked in real time the quality of the workers, stopping the evaluation when their quality dropped below an acceptable threshold [58]. Each sentence was annotated by at least five workers, where the plurality rule was used to define the consensus labels. This study uses four emotional states from this corpus: anger, sadness, happiness, and neutral state.

For both datasets, we use the emotional labels provided by raters after watching the audiovisual stimuli. These labels are used to train and assess all the networks present in our framework, including the unimodal auxiliary networks.

4.2 Features and Preprocessing

4.2.1 Visual Features

The visual features are obtained by extracting frames from every clip present in the corpora used in this study. We use the OpenFace toolkit [59] to detect and extract faces from every image frame using bounding boxes. The bounding boxes of the faces are extracted using the *Multi-Task Cascaded Convolutional Neural Network* (MTCNN) face detection algorithm [60]. After extracting the bounding boxes, we normalize the pixel intensities to a range between -1 and 1, and resize the images to a predetermined dimension of $224 \times 224 \times 3$. Following the face extraction, we use the VGG-face model [52], which we fine-tune for a multi-class emotion recognition task using the AffectNet corpus [5]. We implement a 7-class problem with AffecNet: neutral, happiness, sadness, surprise, fear, disgust, and anger. We attach three *fully-connected* (FC) layers to the original VGG-face model and fine-tune this model for 50 epochs using *adaptive moment estimation* (ADAM) as the optimizer and a learning rate set to 0.000075. The fine-tuned VGG-face model obtains facial feature representations at the frame level from the datasets. The representation obtained from the fine-tuned VGG-face model is retrieved from the first fully connected layer of the model with an array dimension of 4,096, which is then concatenated row-wise with all the other frames within each clip from the datasets to be used as input to the video branch of our audiovisual model. We use zero-padding to ensure the feature vector sequence length is the same for all the videos.

4.2.2 Acoustic Features

The acoustic features are obtained by extracting the audio from each video clip in the datasets. We use the OpenSmile toolkit [61] to extract the *low level descriptors* (LLDs) included in the feature set for the paralinguistic challenge in Interspeech 2013 [62]. LLDs include spectral, prosodic, and energy-based acoustic features extracted at the frame level, such as the energy, fundamental frequency (f0), and *Mel-frequency cepstral coefficients* (MFCCs). These features have been shown to provide relevant information for emotion recognition tasks [63]. The features were extracted using window lengths of 32ms with a step size of 16ms over the entire audio for each sentence provided in the datasets. The resulting LLDs consist of 130 frame-based acoustic features, which are Z-normalized and then concatenated row-wise creating a $N_a \times 130$ input matrix, where N_a is the length of the audio sequence.

After extracting the audio and visual features, we use zero-padding to ensure the feature vector sequences have the same length across all sequence vectors. Otherwise, the dot-product operations cannot be performed on vectors that have unmatched dimensions.

4.3 Implementation Details

The multi-head attention layers and self-attention layers are all five layers deep, and each layer has 10 attention heads. We set a dropout rate of $p = 0.25$ on the output embeddings obtained from the attention layers, along with a $p = 0.1$ dropout rate for the residual connections. The model uses ADAM as the optimizer with an initial learning rate set to 0.000725. During training, we have a learning decay set to five epochs. We set the gradient clipping threshold to 0.8, batch size to 32, and use ReLU as the activation function. λ_{va}, λ_a, and λ_v are set to have equal contribution (i.e., $\frac{1}{3}$). We explore other combinations in Section 5.3.4 as part of our evaluation. The representations learned from the fusion layers and self-attention layers are passed to a MLP containing three fully-connected layers with dropout and ReLU activation function to make predictions. The two

intermediate hidden layer dimensions of the MLP are set to 260 hidden units. We set our output dropout rate to $p = 0.2$. This model is trained to optimize the cross-entropy loss between the predictions and ground truth labels. The model was trained for 25 epochs with an early stopping criterion based on the development loss. The models were implemented in PyTorch and trained using an Nvidia QUADRO RTX 8000.

For each database, the recordings are randomly split such that around 70% of the data is in the train set, 15% of the data is in the development, and 15% of the data is in the test set. The random splits are implemented in a speaker-independent manner, where there is no speaker overlap in the data included in the train, development, and test sets. We trained each model 20 times using random seeds with different splits every time. The 20 seeds are saved and re-used in every single model presented in the study to ensure a fair comparison of the results across models (i.e., all models are trained and tested with the same partitions). All the models are trained using the train set, and the model achieving the best results on the development set is saved and used to make predictions on the test set.

4.4 Evaluation Metrics

An emotional state prediction is obtained for each video sequence available in our test set. We report the final results on the test set using the 'micro' and 'macro' F1-scores. The 'micro' F1-score is calculated using the global number of true positives, false negatives, and false positives. This metric is sensitive to imbalance on the number of samples per emotional class. The 'macro' F1-score separately calculates the F1-scores for each class and aggregates these scores equally weighting each class. This metric provides a bigger penalization when a model does not perform well with the minority classes. Every experiment in this work is run 20 times with different random seeds, reporting the average metrics. We also perform statistical analyses to evaluate our model with baseline models using a one-tailed matched pair t-test with significance level at p-value = 0.05, comparing the macro F1-scores and micro F1-scores obtained by each model for the 20 trials. All the models are trained using the optimized and standard training strategies mentioned in Section 3.6.

4.5 Baselines

We compare our proposed model with three strong audiovisual architectures and one unimodal architecture.

4.5.1 Baseline 1: Audiovisual Framework without Auxiliary Networks

The first baseline is a framework derived from our model. This model uses our proposed model without the auxiliary networks and shared loss mechanism. The baseline corresponds to the main audiovisual fusion network $\mathcal{F}_{av}(\bullet)$. This baseline helps us quantify the performance gain obtained by adding the two adjacent auxiliary networks to our framework. The implementation settings of this baseline are the same as the proposed model. We refer to this model as *baseline 1*.

4.5.2 Baseline 2: Multimodal Transformer (MulT)

Tsai et al. [17] proposed a multimodal transformer architecture for human language time-series data. This model uses 1D temporal convolutions to retrieve temporally aware textual, acoustic, and visual feature vectors. They built a cross-modal transformer model that injects and aligns different modalities using an attention mechanism. This model builds pairs of bimodal representations that are formed by having the Keys and Values of one modality interact with the Queries of a target modality. Six bimodal vectors are generated. Vectors with similar target modalities are concatenated and passed to another transformer layer that generates representations that are then used for classification.

We implement the model following the description of the architecture presented in their study. However, we removed the textual branch from their original architecture, reducing the number of cross-modal transformers from 6 to 2 from their original architecture to adapt their method from three modalities (textual, visual, and acoustic) to two modalities (audiovisual). We refer to this model as *baseline 2*.

4.5.3 Baseline 3: Cross-modal Transformer Architecture

Parthasarathy et al. [18] used a cross-modal transformer architecture for audiovisual recognition of emotional attributes. The same architecture was used for *automatic speech recognition* (ASR) tasks achieving competitive results [64]. The architecture consists of several encoding layers that independently encode the representations from each modality and project these representations into Q, K, and V matrices to be the inputs of a transformer block. The Keys and Values of one modality are concurrently passed into one transformer block with the Queries of a different modality to fuse their representations. A residual connection is then used to add representations from the dot-product attention layers to its corresponding encoder representations using two learnable scalar weights. The resulting two final components are summed together and passed through a dense layer to get predictions.

We reproduce the implementation of this architecture following the model descriptions presented in Paraskevopoulos et al. [64] and Parthasarathy et al. [18]. We implement the model so that it is trained and evaluated using the F1-score metric instead of the *concordance correlation coefficient* (CCC) metric used in the original paper, adjusting the model's task for multi-class emotion recognition. We refer to this model as *baseline 3*.

4.5.4 Baseline 4: Unimodal Emotion Recognition Framework

We build unimodal baselines to compare the robustness of our proposed model for scenarios where video-only of acoustic-only features are available for emotion recognition. The architecture used here has similar components to our proposed model. Figure 2 depicts an overview of the unimodal baselines. This framework is composed of an auxiliary network and a network containing self-attention layers with learnable scaling parameters. Since this model only receives one modality as input, the cross-modality attention mechanism is not used. The difference here is that

This article has been accepted for publication in IEEE Transactions on Affective Computing. 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/TAFFC.2022.3216993

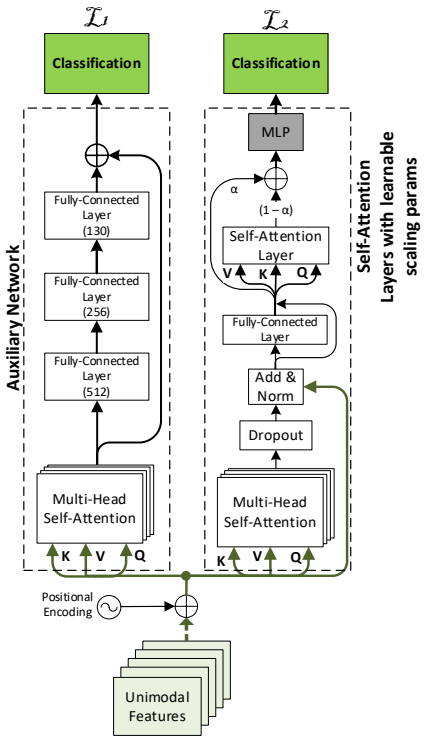


Fig. 2. Overview of the unimodal framework used to compare the performance of our approach in visual-only and acoustic-only scenarios.

we have two training losses instead of three present in our proposed model (one main loss and one auxiliary loss). During the training of our unimodal network the losses are equally weighted. Likewise, the networks' outputs are equally averaged during inference. The training settings used here are the same settings employed for our proposed model. We train two different unimodal models using this framework: a video-only emotion recognition model and an audio-only emotion recognition model.

5 Experimental Results

5.1 Comparison with Multimodal Baselines

This section compares our proposed model with the multimodal baselines. All the models are trained using both the standard and optimized training methods discussed in Section 3.6. Table 1 reports the average macro F1-scores and average micro F1-scores of the models obtained after running 20 trials. The results show that our proposed model has a strong performance that is significantly better than the baseline models for both datasets. Figure 3 provides further comparison between the models trained with the optimized training mechanism. For the CREMA-D corpus, we observe in Figure 3(a) that our model shows more consistent results compared to the performance spread of the baselines. For the MSP-IMPROV corpus, Figure 3(b) again demonstrates the consistent performance of our proposed model across all the experiments. Our proposed model had lower performance spread than baseline 1 and baseline 2. Baseline 3 shows a slightly smaller spread than our proposed model on the MSP-IMPROV corpus, but the performance is significantly lower than our method. Overall, our model trained with both training regimes is able to outperform the baselines' results with consistent F1-scores for both databases.

TABLE 1
Comparison between our model with audiovisual baselines, reporting the average F1-score values across 20 trials. The symbol $*$ indicates that our method is significantly better than the other three baselines.

Standard Training Procedure				
	CREMA-D		**MSP-IMPROV**	
Architecture	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Our Model	**0.769***	**0.768***	**0.762***	**0.706***
Baseline 1	0.705	0.702	0.751	0.687
Baseline 2 [17]	0.707	0.704	0.745	0.677
Baseline 3 [18]	0.624	0.608	0.733	0.663
Optimized Training Procedure				
	CREMA-D		**MSP-IMPROV**	
Architecture	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Our Model	**0.773***	**0.772***	**0.761***	**0.703***
Baseline 1	0.736	0.734	0.742	0.681
Baseline 2 [17]	0.724	0.722	0.744	0.673
Baseline 3 [18]	0.629	0.615	0.732	0.667

5.2 Robustness to Missing Modalities

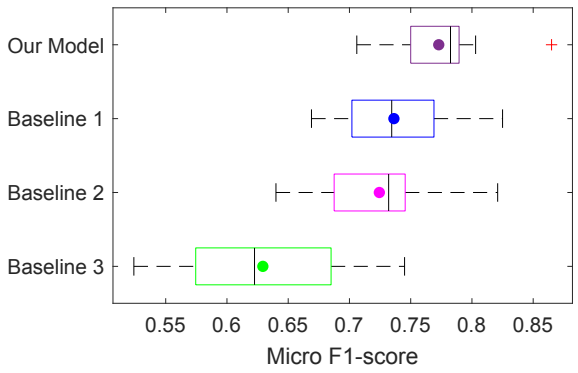
This section investigates how changes to the input data affect the model's performance. We investigate the effect of the optimized training mechanism on our model's performance in non-ideal scenarios. We conducted modality ablation studies by separately clipping either the entire audio or the video sequences (Sec. 5.2.1). We also randomly replace some frames of the feature vectors with zeros with different probabilities (Sec. 5.2.2). We present the results obtained from baseline 4, which was developed as a version of our proposed model trained in a unimodal fashion with either visual or acoustic features. Because baseline 4 is tested without dropping any feature, it is expected that this baseline will competitively perform in comparison to the multimodal models tested with missing information.

5.2.1 Modality Clipping

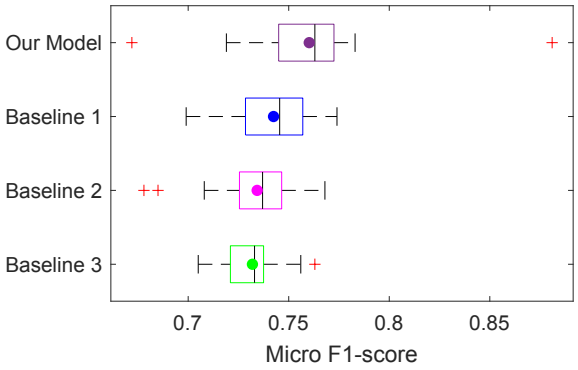
We start our modality ablation analysis by clipping either the entire visual or acoustic feature sequences (i.e., zeroing the entire feature sequence). Table 2 reports the mean F1-score value of the unimodal baseline 4 and our model trained with the optimized training approach, when a specific modality is clipped while the other modality is kept unchanged. It also shows the results of the baselines implemented with the optimized training method. The first part of the table contains the results with only visual features. The second part of the table contains the results with only the acoustic features.

Table 2 shows a clear strong performance from our proposed model in visual-only and acoustic-only scenarios, achieving significant improvements over baselines 1, 2, and 3. Table 2 shows the importance of the optimized training strategy combined with the proposed framework to deal with missing modalities. Our proposed model is the only model to achieve results that approach or exceed the performance of baseline 4 in the single-modality scenarios explored in this section.

Figure 4 compares side-by-side the performance of every multimodal model explored in this study trained with the



(a) CREMA-D corpus



(b) MSP-IMPROV corpus

Fig. 3. Evaluation of the optimized training procedure used to train the proposed model and the baselines conducted on the 20 trials on the (a) CREMA-D corpus, and (b) MSP-IMPROV corpus. The middle black bar represents the median micro F1-score values for each model. The outside bars represent the first and third quartiles. The middle colored dots represent the mean F1-score values for each model. The dash lines represent the minimum and maximum values and the red crosses represent outliers.

standard and optimized training methods for all the scenarios. The dashed orange line in the bar graphs represents baseline 4's performance in visual-only emotion recognition, and the solid green line represents baseline 4's performance in acoustic-only emotion recognition. Across both databases, all the multimodal models gain robustness to visual-only and acoustic-only scenarios when trained with the optimized training mechanism. However, the optimized training mechanism on the baselines is not enough to achieve the strong performance in non-ideal scenarios that our proposed model achieves. Our proposed model is the only architecture able to overcome baseline 4's performance by a 1.1% micro F1-score increase on visual-only scenarios and by a 1.7% micro F1-score increase on acoustic-only scenarios for the CREMA-D dataset. On the MSP-IMPROV dataset, our model's performance only has a 0.5% lower micro F1-score on the visual-only scenarios, and only a 0.4% lower micro F1-score on the acoustic-only scenarios compared to baseline 4.

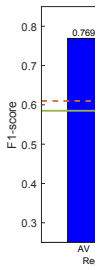
5.2.2 *Random Masking of Features*

This section analysis the performance of the system when part of the features are missing. We randomly zero-out either visual or acoustic data at the frame-level. We conduct this analysis by increasing the number of available frames from 10% to 90% in increments of 10%. For example, the 70% condition indicates that 30% of the frames selected at random from the modality that we are masking are replaced by zeros. This evaluation aims to analyze the robustness of the model with partial information for one of the modalities.

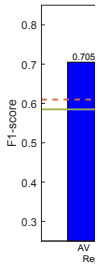
TABLE 2
Performance analysis with missing modality. The performances of the audiovisual models using the optimized training approach and tested with only acoustic or only visual features are compared with the unimodal baseline 4. The table reports the average F1-score values across 20 trials. The symbol $*$ indicates that our model is significantly better than all other baselines, including baseline 4.

Visual Features Only				
	CREMA-D		**MSP-IMPROV**	
Architecture	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Our Model	**0.622***	**0.615***	0.721	0.608
Baseline 4	0.611	0.604	**0.726**	**0.635**
Baseline 1	0.561	0.551	0.709	0.596
Baseline 2 [17]	0.572	0.566	0.692	0.526
Baseline 3 [18]	0.552	0.541	0.714	0.622
Acoustic Features Only				
	CREMA-D		**MSP-IMPROV**	
Architecture	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Our Model	**0.602***	**0.592***	0.625	**0.532**
Baseline 4	0.585	0.576	**0.629**	**0.532**
Baseline 1	0.541	0.528	0.563	0.453
Baseline 2 [17]	0.524	0.513	0.545	0.366
Baseline 3 [18]	0.327	0.244	0.511	0.378

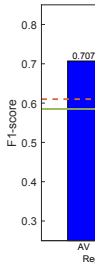
Figure 5 reports the mean micro F1-scores of our model and the baselines as we increase the number of available frames for the modality that we are masking. The upper solid blue horizontal line in Figure 5 represents the best performance obtained by our model when full-audiovisual features are available. The lower dashed horizontal red line represents the performance of the unimodal baseline 4 for the modality that it is not affected by the random masking process. We see that even with only 10% of the frames for the modality that we are masking, our model's performance is already higher than baseline 4 in all scenarios. When only 10% of visual information is available, in addition to the acoustic features, our model's performance is 15.4% higher than baseline 4 for the CREMA-D dataset (Fig. 5(b)), and 9.2% higher than baseline 4 for the MSP-IMPROV dataset (Fig. 5(d)). Baselines 1, 2 and 3 have lower performance than our proposed approach in all scenarios. When acoustic features are masked on the MSP-IMPROV dataset, the performance for baselines 1, 2, and 3 are 1.3%, 0.8%, and 0.2% lower than baseline 4 when only 10% of the acoustic features are available, respectively. Under the same setting, our model outperforms baseline 4 by 0.8% (Fig. 5(c)). Overall, the analysis shows that our proposed approach is robust against missing frames. It consistently increases its performance as more acoustic and/or visual features become available. When the number of missing frames is between 40% and 50% of either acoustic or visual features across conditions, our model approaches the performance of the full audiovisual scenario with a performance gap that is less than 1% for both databases.



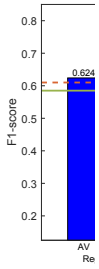
(a) O



(c) B

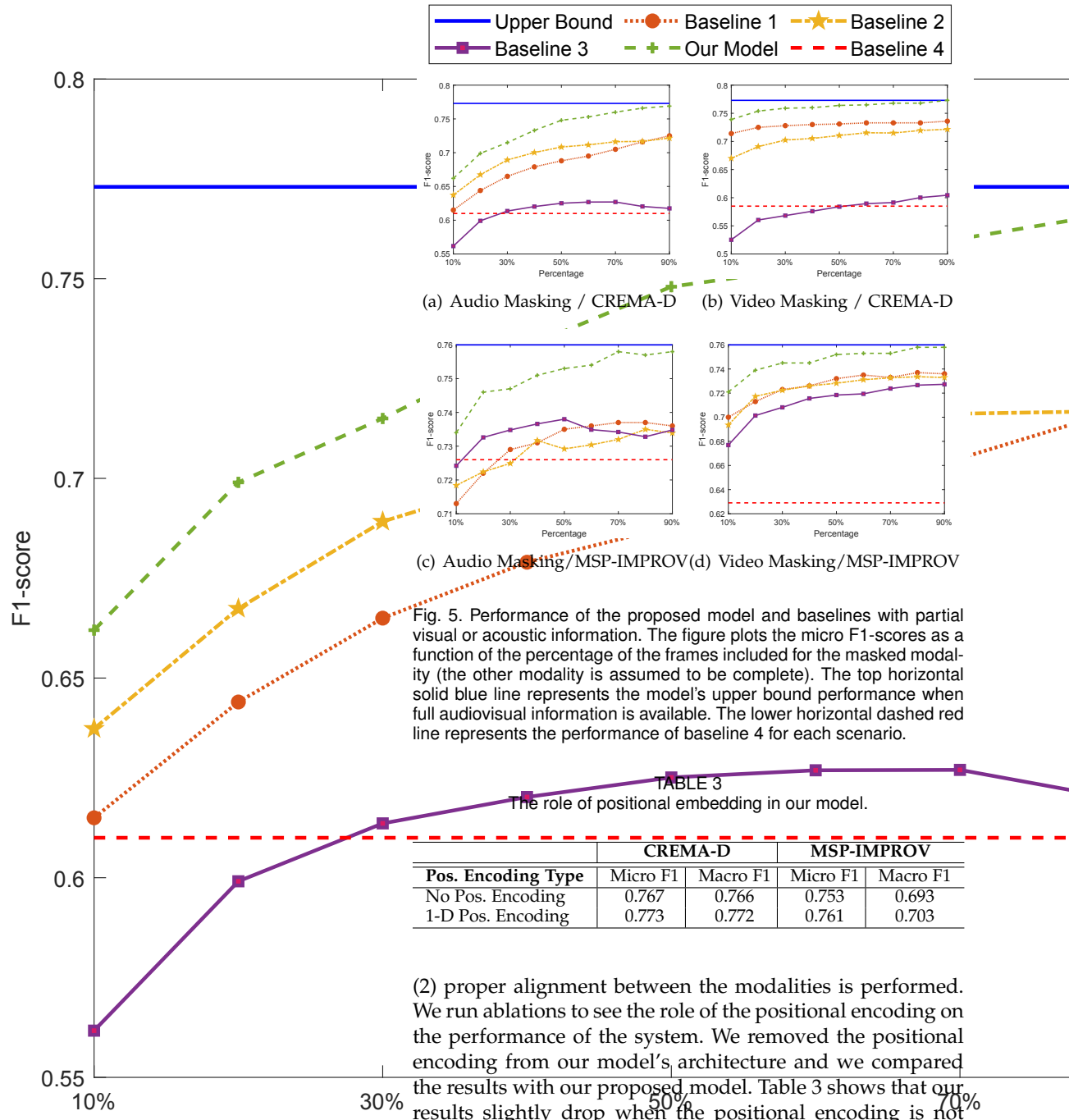


(e) Bas



(g) Bas

Fig. 4. C
dard and
increase
AV = Au
lines rep
Orange



(a) Audio Masking / CREMA-D (b) Video Masking / CREMA-D

(c) Audio Masking/MSP-IMPROV (d) Video Masking/MSP-IMPROV

Fig. 5. Performance of the proposed model and baselines with partial visual or acoustic information. The figure plots the micro F1-scores as a function of the percentage of the frames included for the masked modality (the other modality is assumed to be complete). The top horizontal solid blue line represents the model's upper bound performance when full audiovisual information is available. The lower horizontal dashed red line represents the performance of baseline 4 for each scenario.

TABLE 3
The role of positional embedding in our model.

	CREMA-D		MSP-IMPROV	
Pos. Encoding Type	Micro F1	Macro F1	Micro F1	Macro F1
No Pos. Encoding	0.767	0.766	0.753	0.693
1-D Pos. Encoding	0.773	0.772	0.761	0.703

5.3 Architecture Ablations

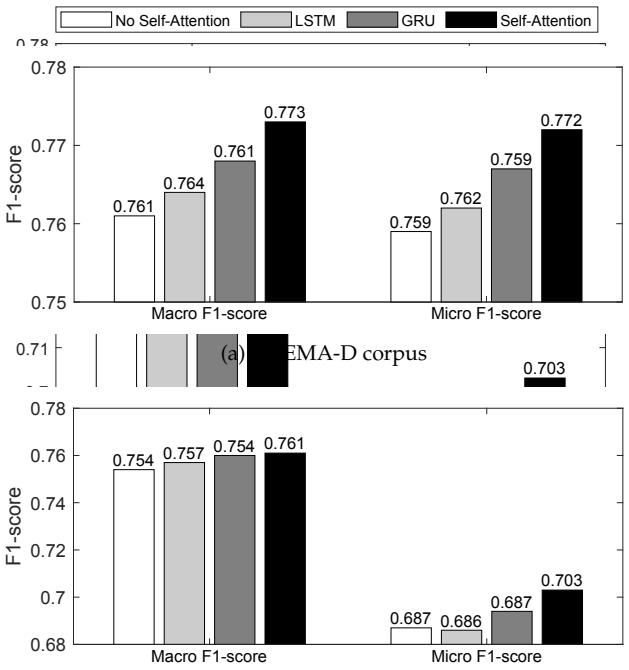
We perform a controlled architecture ablation study by evaluating the effect of various components of our model on the performance of our system. We run ablations on the main pieces of our architecture: positional encoding, self-attention layers, ways to retrieve the classification token, and different values for lambda to balance the shared loss mechanism.

5.3.1 *Positional Encoding*

We use positional encoding in our model to ensure that (1) temporal information is properly retained by our model and (2) proper alignment between the modalities is performed. We run ablations to see the role of the positional encoding on the performance of the system. We removed the positional encoding from our model's architecture and we compared the results with our proposed model. Table 3 shows that our results slightly drop when the positional encoding is not present. The drop in performance, although small, is larger for the MSP-IMPROV dataset. One potential reason for this result is the durations of the input sequences (up to 31 seconds) on the MSP-IMPROV corpus, which are generally longer than the durations of the input sequences on the CREMA-D dataset (up to 5 seconds). The results obtained in this experiment are consistent with other studies ([65], [66]) that explored positional encoding ablations and observed a small decrease in performance when no positional encoding is used.

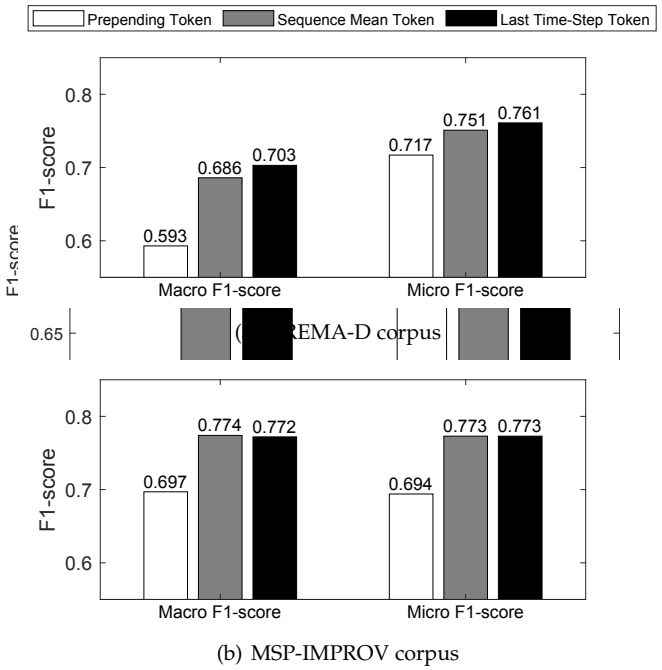
5.3.2 *Self-Attention Layers*

After processing our data through the modality fusion attention layers, the features are passed through the self-attention layers with learnable scaling parameters (Fig. 1). The self-attention layers compute representations within the same



(b) MSP-IMPROV corpus

Fig. 6. Effects of removing the self-attention layers with learnable scaling parameters, or replacing them with a LSTM or GRU layer. The figure reports the average F1-scores obtained across the 20 trials.



(b) MSP-IMPROV corpus

Fig. 7. Comparison of results obtained with different methods for classification token retrieval. The figure reports the average F1-score values across 20 trials. Our framework uses the *last time-step token* method.

sequence to emphasize regions that are more useful for our task. Figure 6 ablates the self-attention layers by removing them, or replacing them with either a *long short-term memory* (LSTM) or *gated recurrent unit* (GRU) layer.

Complete removal of the self-attention layers with learnable scaling parameters results in a 0.7% decrease in macro F1-score and a 1.6% in micro F1-score for the MSP-IMPROV dataset. For the CREMA-D database, it decreases the macro F1-score by 1.2% and the micro F1-score by 1.3%. Figure 6 shows that having the self-attention layers with learnable scaling parameters yields better results than replacing them with an LSTM or GRU layer. However, adding an LSTM or GRU layer is often preferred over just removing the self-attention layer.

5.3.3 Classification Token Generation

In our model, representations of the sequential data obtained from the self-attention layers are extracted from the last time-step of every sequence to be used as a *classification* (CLS) token. The two token vectors obtained from the top and bottom self-attention layers are concatenated (Eq.5), and used as the final hidden representation for our classification task (Section 3.4). This section explores this approach, referred to as *last time-step token*, with two alternative methods.

The first approach that we compare our model to is with a prepending token approach. This method requires a CLS token, which is prepended to the beginning of every input sequence fed into the model. After this step, everything gets processed through the transformer layers. Then, this prepended token is extracted at the final self-attention layer and used to perform the classification task. This is the same method used by Devlin et al. [67] when introducing the BERT model. We refer to this approach as the *prepending token*. Figure 7 shows that this method decreases the performance in both databases. The second approach explores using the mean over the final hidden layer representations of the processed sequential data, instead of adding a classification token to our sequential data. The resulting mean value is then used as a classification token. We refer to this approach as the *sequence Mean Token*. Using the sequence mean as a token leads to improvements over the results obtained by the *prepending token* method. Figure 7 shows that the *last time-step token* has almost the same performance as using the *sequence mean token* approach for the CREMA-D corpus. However, the results on the MSP-IMPROV corpus show stronger performance when using the *last time-step token*. We hypothesize that the differences across corpora are due to the differences in duration of the speaking turns. The *last time-step token* approach works well across both datasets. The last step of the representations is obtained from the self-attention layers with learnable scaling parameters, which have a receptive field of the entire sequence, giving a global representation.

5.3.4 Hyper-parameters for Shared Losses

Each network ($\mathcal{F}_{av}(\bullet)$, $\mathcal{F}_{a}(\bullet)$, $\mathcal{F}_{v}(\bullet)$) in our model has its own loss $\mathcal{L}_m$, where $m \in \{a, v, av\}$. The model combines the training cross-entropy losses from these networks during training to get the total loss. Table 4 presents the results when using different values for the weights $\lambda_{va}, \lambda_a, \lambda_v$ (Eq. 6). We only include results on the CREMA-D corpus due to computational constraints.

In our experiments, we make changes to the distributions of the weight values for λ_a, λ_v, and λ_{va}. The weights are distributed such that they sum up to 1. Table 4 shows that the best performance is achieved around values that

TABLE 4
F1-scores of the proposed system on the CREMA-D dataset with different weights to combine the losses (Eq. 6).

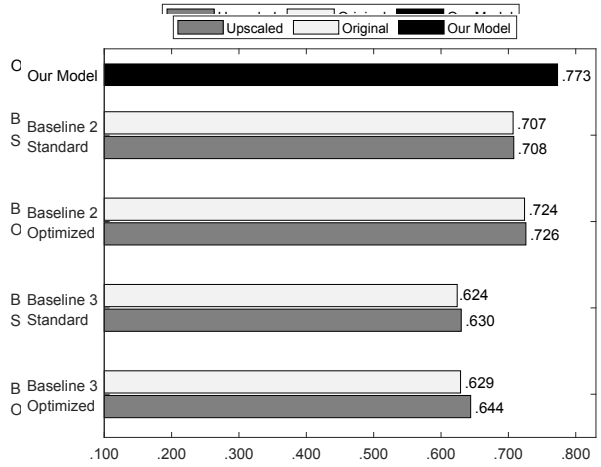
$(\lambda_{va}, \lambda_a, \lambda_v)$	Micro F1-score	Macro F1-score
(0.20, 0.40, 0.40)	0.773	0.772
(0.30, 0.35, 0.35)	0.774	0.773
(0.33, 0.33, 0.33)	0.773	0.772
(0.40, 0.30, 0.30)	0.773	0.773
(0.50, 0.25, 0.25)	0.769	0.768
(0.60, 0.20, 0.20)	0.772	0.771
(0.70, 0.15, 0.15)	0.769	0.768
(0.80, 0.10, 0.10)	0.770	0.769
(0.00, 0.50, 0.50)	0.738	0.738
(0.90, 0.00, 0.10)	0.738	0.736
(0.90, 0.10, 0.00)	0.742	0.743

are more evenly distributed amongst all three loss weights. We see a small performance decrease when the distribution shifts to a more uneven area and more weight is given to the audiovisual fusion network ($\mathcal{F}_{av}(\bullet)$). A larger drop in performance is observed when one of the loss' weights is set to zero for any of the networks. The experiments show that as long as the weights of the losses are evenly distributed and none are set to zero, the model is not very sensitive to these hyper-parameters.

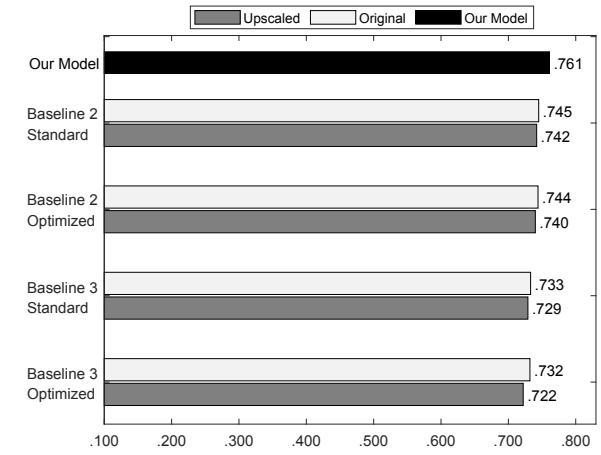
5.4 Trainable Parameters and Training Time

This section discusses the model complexity, and the time required to train the models. Baseline 1 presented in this study consists of a model which shares mostly the same architecture as our main proposed framework. The difference is that baseline 1 does not include the auxiliary networks. We have seen in our experiments that the auxiliary networks help our model to greatly improve performance over baseline 1. Table 5 shows that this performance boost comes with an addition of 1.48 times more parameters. Our proposed approach also sees a training time increase of around 35 seconds per epoch over baseline 1. When compared with baselines 2 and 3, we see that our model is substantially more complex than these baselines. Our model has 19.7 times more parameters that baseline 2 and 3.2 times more parameters than baseline 3. Additionally, we also see that our model takes 2.8 times longer and 2.1 times longer to train than baselines 2 and 3, respectively. Lastly, our unimodal baseline 4 also contains a large amount of parameters, since it was directly derived from our proposed audiovisual framework.

To ensure that the performance gains obtained by our model over baselines 2 and 3 are not solely achieved due to a higher number of parameters of our model compared to the baselines, we conduct an additional set of experiments where we upscale baselines 2 and 3 by increasing its number of parameters to values comparable to our proposed framework (∼7M). To upscale these models, we have updated the size of the transformer layer reception and the number of hidden layers in their fully-connected layers. Figure 8 shows the results. Increasing the number of parameters for the baselines does not result in a consistent increase in performances. The scaled-up version of baseline 2 has approximately 20 times more parameters than the original model (from 358,384 to 7,070,406 parameters). The scaled-up version results in minimal improvements, in some cases,



(a) CREMA-D corpus



(b) MSP-IMPROV corpus

Fig. 8. Experimental results reporting micro F1-scores from our proposed frameworks' best performance and upscaled versions of baselines 2 and 3. The bar graphs contain side-by-side comparisons with the original versions of baselines 2 and 3 reported in Table 1.

TABLE 5
Number of trainable parameters and training times per epoch for our proposed model and the baselines.

	# Params.	Train. Time
Our Model	7,054,748	83 sec/epoch
Baseline 1	4,776,994	48 sec/epoch
Baseline 2 [17]	358,384	30 sec/epoch
Baseline 3 [18]	2,211,628	39 sec/epoch
Baseline 4	4,364,354	46 sec/epoch

over the original baseline model. The scaled-up version of baseline 3 has approximately 3.2 times more parameters than its original model (from 2,211,628 to 7,125,278 parameters). Figure 8 shows that the scaled-up version of baseline 3 results in consistent performance increase for the CREMA-D dataset over the non-scaled method. However, we do not see improvements on the MSP-IMPROV dataset. Importantly, the performances of the upscaled baselines are lower than the results obtained by our proposed method in all cases.

6 Conclusions

This study presented an effective approach for robust audiovisual emotion recognition using a combination of the transformer architecture with auxiliary networks. The transformer framework and temporal information added by the positional encoding enable the model to perform audiovisual fusion and alignment. The auxiliary networks with a shared loss mechanism and the optimized training method increase the robustness of the model against missing modalities, preventing our framework from losing important information during training. The performance in emotion recognition of our model achieves a micro F1-score of 77.3% and a macro F1-score of 77.2% on the CREMA-D database, and a micro F1-score of 76.1% and a macro F1-score of 70.3% on the MSP-IMPROV database under ideal conditions. Statistical analyses revealed that our framework not only achieves superior performance when compared to competitive baselines, but also achieves proper alignment between the modalities while retaining temporal information, robustness to missing modalities, and better fusion of audiovisual features at the model level. The experimental evaluations demonstrated the benefits of auxiliary networks, temporal alignment, and model level fusion of audiovisual features. The proposed model can also be easily adapted to predict on emotional attributes (e.g. arousal, valence, dominance). Our architecture can be adapted to perform this task by simply changing our loss function to CCC and updating the model's output to the desired number of dimensions to be explored. Results from the ablation studies show how architecture changes can affect our proposed model's performance. Even though the model performed well without positional embeddings, as seen in Table 3, the additional temporal information provided by the positional embeddings allows the model to better align the audiovisual data within the attention layers. Furthermore, our results show that the optimized training method adds robustness to the model and improves the overall performance in ideal scenarios.

There are several research directions opened by this study. Given the availability of large amounts of unlabeled data across many datasets with rich emotional contents, we can explore self-supervised methods [68] for pre-training our framework aiming to strengthen the model's performance and robustness in all possible scenarios. A limitation of this study is the use of datasets containing acted or elicited emotions by actors. We would like to expand the evaluation of the proposed architecture in other databases containing more naturalistic data. Expanding on the idea of robustness against missing modalities, we can also explore the robustness of our proposed approach to noisy environments for emotion recognition systems [69], [70]. We also aim to explore situations where human voice is not present, so the system may still receive ambient sound with noisy acoustic features that are not zeros. To handle this situation, we want to explore the use of voice activity detectors, signal to noise ratio filters, and music detection filters to identify inputs which can be regarded as noise, silence, or music. We could analyze if better results can be achieved by removing these problematic inputs from the model's pipeline at inference, and replacing these segments with zeros. We can also expand this architecture to tasks that require more modalities to be combined (e.g., text, speech, and facial features) and explore the utilization of other features as input to the network (e.g. raw data). Furthermore, a promising future direction is to incorporate this framework to solve other multimodal tasks, such as audiovisual automatic speech recognition [71]. One limitation of our approach is the high complexity of the framework. We leave as future work the exploration of strategies that could reduce its complexity without affecting its performance.

Acknowledgments

This work was supported by NSF under Grant IIS-1718944

References

[1] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, January 2009.

[2] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *AAAI Conference on Artificial Intelligence (AAAI 2021)*, vol. 35, Virtual Conference, February 2021, pp. 10 790–10 797.

[3] M. Yasen and S. Tedmori, "Movies reviews sentiment analysis and classification," in *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT 2019)*, Amman, Jordan, April 2019, pp. 860–865.

[4] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, February 2021.

[5] 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.

[6] L. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *International Conference on Information, Communications and Signal Processing (ICICS)*, vol. I, Singapore, 1997, pp. 397–401.

[7] L. Chen, T. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion / expression recognition," in *Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1999, pp. 366–371.

[8] T. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082–1089, May 2006.

[9] C. Benoît, "The intrinsic bimodality of speech communication and the synthesis of talking faces," in *The Structure of Multimodal Dialogue II*, M. Taylor, F. Néel, and D. Bouwhuis, Eds. John Benjamins Publishing Company, March 2000, pp. 485–502.

[10] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[11] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1994)*, vol. 2, Adelaide, Australia, April 1994, pp. 669–672.

[12] V. V. Wassenhove, K. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, no. 3, pp. 598–607, 2007.

[13] J. Maier, M. Di Luca, and U. Noppeney, "Audiovisual asynchrony detection in human speech," *Journal of Experimental Psychology Human Perception and Performance*, vol. 37, no. 1, pp. 245–256, February 2011.

[14] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April-June 2013.

[15] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1983–1986.

[16] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategie," *APSIPA Transactions on Signal and Information Processing*, vol. 3, no. 1, p. e12, November 2014.

[17] Y.-H. Tsai, S. Bai, P. Liang, J. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Association for Computational Linguistics (ACL 2019)*, vol. 1, Florence, Italy, July 2019, pp. 6558–6569.

[18] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *International Conference on Multimodal Interaction (ICMI 2020)*, Utrecht, The Netherlands, October 2020, pp. 400–404.

[19] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, February 2013.

[20] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using Gaussian mixture models for face and voice," in *IEEE International Symposium on Multimedia (ISM 2008)*, Berkeley, CA, USA, December 2008, pp. 250–257.

[21] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 33–48, March 2010.

[22] 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.

[23] S. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, and Y. Bengio, "Combining modality specific deep neural networks for emotion recognition in video," in *ACM on International Conference on Multimodal Interaction (ICMI 2013)*, Sydney, Australia, December 2013, pp. 543–550.

[24] 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.

[25] 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.

[26] L. Todorovski and S. Džeroski, "Combining classifiers with meta decision trees," *Machine Learning*, vol. 50, no. 3, pp. 223–249, March 2003.

[27] D. Morrison, R. Wang, and L. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, February 2007.

[28] B. Sun, L. Li, X. Wu, T. Zuo, Y. Chen, G. Zhou, J. He, and X. Zhu, "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.

[29] Z. Zeng, J. Tu, B. Pianfetti, and T. Huang, "Audio-visual affective expression recognition through multistream fused HMM," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, June 2008.

[30] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *International Conference on Multimodal Interaction (ICMI 2014)*, Istanbul, Turkey, November 2014, pp. 508–513.

[31] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, January 2004.

[32] A. Zadeh, T. B. Y. C. Lim, and L. Morency, "Convolutional experts constrained local model for 3D facial landmark detection," in *IEEE International Conference on Computer Vision Workshops (ICCVW 2017)*, Venice, Italy, October 2017, pp. 2519–2528.

[33] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Systems*, vol. 161, no. 1, pp. 124–133, December 2018.

[34] W. Rahman, M. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Association for Computational Linguistics (ACL 2020)*, Online, July 2020, pp. 2359–2369.

[35] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2803–2807.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 5998–6008.

[37] F. Tao and C. Busso, "Aligning audiovisual features for audiovisual speech recognition," in *IEEE International Conference on Multimedia and Expo (ICME 2018)*, San Diego, CA, USA, July 2018, pp. 1–6.

[38] H. Bredin and G. Chollet, "Audiovisual speech synchrony measure: Application to biometrics," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 70186, pp. 1–11, December 2007.

[39] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, November 2007.

[40] T. Halperin, A. Ephrat, and S. Peleg, "Dynamic temporal alignment of speech to lips," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 3980–3984.

[41] J. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian Conference on Computer Vision (ACCV 2016 Workshop)*, ser. Lecture Notes in Computer Science, C. Chen, J. Lu, and K. Ma, Eds. Taipei, Taiwan: Springer Berlin Heidelberg, November 2016, vol. 10117, pp. 251–263.

[42] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Audio visual emotion recognition with temporal alignment and perception attention," *ArXiv e-prints (arXiv:1603.08321)*, pp. 1–8, March 2016.

[43] J. Wang, Z. Fang, and H. Zhao, "AlignNet: A unifying approach to audio-visual alignment," in *IEEE Winter Conference on Applications of Computer Vision (WACV 2020)*, Snowmass, CO, USA, March 2020, pp. 3298–3306.

[44] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *AAAI Conference on Artificial Intelligence (AAAI 2020)*, vol. 34, New York, NY, USA, February 2020, pp. 1359–1367.

[45] J. Chen and A. Zhang, "HGMF: Heterogeneous graph-based fusion for multimodal data with incompleteness," in *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*, Online, July 2020, pp. 1295–1305.

[46] F. Ma, S. L. Huang, and L. Zhang, "An efficient approach for audio-visual emotion recognition with missing labels and missing modalities," in *IEEE International Conference on Multimedia and Expo (ICME 2021)*, Shenzhen, China, July 2021, pp. 1–6.

[47] F. Ma, X. Xu, S.-L. Huang, and L. Zhang, "Maximum likelihood estimation for multimodal learning with missing modality," *ArXiv e-prints (arXiv:arXiv:2108.10513)*, August 2021.

[48] C. Du, C. Du, H. Wang, J. Li, W.-L. Zheng, B.-L. Lu, and H. He, "Semi-supervised deep generative modelling of incomplete multimodality emotional data," in *ACM international conference on Multimedia (MM 2018)*, Seoul, Republic of Korea, October 2018, pp. 108–116.

[49] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7357–7361.

[50] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, Seattle, WA, USA, June 2020, pp. 130–139.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, MA, USA, June 2015, pp. 1–9.

[52] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC 2015)*, Swansea,UK, September 2015, pp. 1–12.

[53] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. Wong, and L. Chao, "Learning deep transformer models for machine translation," in *Association for Computational Linguistics (ACL 2019)*, Florence, Italy, July 2019, pp. 1810–1822.

[54] A. Baevski and M. Auli, "Adaptive input representations for neural language modeling," in *International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA, May 2019, pp. 1–11.

[55] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[56] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[57] E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect data set," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 395–409, October-December 2015.

[58] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[59] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *IEEE Conference on Automatic Face and Gesture Recognition (FG 2018)*, Xi'an, China, May 2018, pp. 59–66.

[60] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, October 2016.

[61] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[62] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[63] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[64] G. Paraskevopoulos, S. Parthasarathy, A. Khare, and S. Sundaram, "Multimodal and multiresolution speech recognition with transformers," in *Association for Computational Linguistics (ACL 2020)*, Online, July 2020, pp. 2381–2387.

[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR 2021)*, Vienna, Austria, May 2021, pp. 1–12.

[66] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC, Canada, December 2019, pp. 1–13.

[67] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[68] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Interspeech 2022*, Incheon, South Korea, September 2022, pp. 1168–1172.

[69] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 2871–2875.

[70] ——, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 6447–6451.

[71] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multi-task learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1–11, January 2021.

Lucas Goncalves (S'22) received his BS in Electrical Engineering from University of Wisconsin - Platteville, in 2018. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. At UTD, he is a Research Assistant at the Multimodal Signal Processing (MSP) laboratory. In 2022, he was awarded the Excellence in Education Doctoral Fellowship from the Erik Jonsson School of Engineering and Computer Science. His research interests include areas related to affective computing, deep learning, and multimodal signal processing. He is also a student member of the IEEE Signal Processing Society.

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and in 2022 (with Yannakakis and Cowie). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017 and ICMI 2021. He is a member of ISCA, AAAC, and a senior member of ACM and IEEE.