VERSATILE AUDIO-VISUAL LEARNING FOR HANDLING SINGLE AND MULTI MODALITIES IN EMOTION REGRESSION AND CLASSIFICATION TASKS

Lucas Goncalves, Seong-Gyun Leem, Wei-Cheng Lin, Berrak Sisman, Carlos Busso
Multimodal Signal Processing (MSP) Lab
Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
{goncalves, seong-gyun.leem, wei-cheng.lin, berrak.sisman, busso}@utdallas.edu

ABSTRACT

Most current audio-visual emotion recognition models lack the flexibility needed for deployment in practical applications. We envision a multimodal system that works even when only one modality is available and can be implemented interchangeably for either predicting emotional attributes or recognizing categorical emotions. Achieving such flexibility in a multimodal emotion recognition system is difficult due to the inherent challenges in accurately interpreting and integrating varied data sources. It is also a challenge to robustly handle missing or partial information while allowing direct switch between regression and classification tasks. This study proposes a *versatile audio-visual learning* (VAVL) framework for handling unimodal and multimodal systems for emotion regression and emotion classification tasks. We implement an audio-visual framework that can be trained even when audio and visual paired data is not available for part of the training set (i.e., audio only or only video is present). We achieve this effective representation learning with audio-visual shared layers, residual connections over shared layers, and a unimodal reconstruction task. Our experimental results reveal that our architecture significantly outperforms strong baselines on both the CREMA-D and MSP-IMPROV corpora. Notably, VAVL attains a new state-of-the-art performance in the emotional attribute prediction task on the MSP-IMPROV corpus. Code available at: https://github.com/ilucasgoncalves/VAVL

Keywords multimodal emotion recognition, audio-visual modeling, transformers, versatile learning, handling missing modalities

1 Introduction

Effective human interactions often include the expression and perceptions of emotional cues conveyed across multiple modalities to accurately comprehend and convey a message. During human interactions, there are two modalities that stand out as particularly significant: speech and facial expressions. These modalities play a crucial role in facilitating communication, making it essential for emotion recognition systems to incorporate speech-based cues [1, 2] and facial expression-based cues [3–5]. These modalities are intrinsically connected [6], proving complementary information [7]. Therefore, emotion recognition systems can be more accurate if they incorporate audio-visual solutions [8], mirroring the way humans interact in natural, real-world settings.

It is important to note that while humans primarily rely on these two modalities to recognize emotional states, there are scenarios in which only one modality may be utilized. In some cases, visual cues might be unavailable or occluded, leaving individuals to rely solely on acoustic information. Conversely, there might be situations where acoustic cues are insufficient or absent, requiring the exclusive use of visual cues for effective communication. As a result, it is vital for *human computer interaction* (HCI) systems to not only be capable of simultaneously handling both modalities, but also to adapt to unimodal situations. The straightforward approach is to have separate unimodal and multimodal solutions. However, it is more computationally effective to have versatile systems that can be adapted according to the available

information. This flexibility ensures that HCI systems can accommodate a wide range of communication contexts and maintain their effectiveness in various real-world conditions.

Artificial intelligent (AI) methods have been explored for audio-visual learning to achieve high performance in either multimodal or unimodal scenarios [9–11]. Most conventional models rely on building separate modality-specific branches [12–14] or employing ensemble-like techniques [15], which can lead to convoluted and complex architectures. With the introduction of transformer frameworks [16], many solutions have been developed to avoid training with modality-specific strategies. These models even implement formulations that unify the cross-modality relationships [17, 18] into a single, comprehensive model.

Despite significant recent advancements in the construction of simpler multimodal models, there are still open research questions to be considered when constructing a unified multimodal model. One of these challenges is that current methods tend to solely focus on one type of *machine learning* (ML) task [19]. For example, they consider either classification or regression problems, without considering the potential need for versatility across different problem types. In emotion recognition, this problem is especially important, since emotions can be alternatively described with emotional attributes (e.g., arousal, valence, dominance) and categorical emotions (e.g., anger, happiness, sadness) [20]. Depending on the setting, therefore, it is important to have a model that can be utilized in both regression (emotional attributes) and classification (categorical emotions) settings. Moreover, current methods often disregard the importance of maintaining distinct representations for each modality to capture modality-specific [21] features and characteristics within shared layers, which could affect performance in unimodal settings [12]. By failing to account for the unique characteristics of each modality, the models may not optimally leverage the strengths of each input type, ultimately limiting their effectiveness. Moreover, the complexity of many existing models can be a drawback, as it may lead to increased computational demands and reduced interpretability [22], making it more challenging for researchers and practitioners to analyze and adapt the models for various applications.

This study addresses these aforementioned challenges to further our understanding of effective versatile multimodal learning approaches. Our main contribution is the proposal of a *versatile audio-visual learning* (VAVL) model, which unifies multimodal and unimodal learning in a single framework that can be used for emotion regression and classification tasks. Our approach utilizes branches that separately process each modality. In addition, we introduce shared layers, residual connections over shared layers, and a unimodal reconstruction task. The shared layers encourage learning representations that reflect the connections between the two modalities. The addition of the auxiliary reconstruction task and unimodal residual connections over the shared layers helps the model learns representations that reflect the heterogeneity of the modalities. Collectively, these components are added to our framework to help with the multimodal feature representation, which is a core challenge in multimodal learning [23]. We implement this audio-visual framework so it can be trained even when audio and visual paired data is not available for part of the data. The proposed approach is attractive as it enables the training of multimodal systems using incomplete information from multimodal databases (e.g., audio-only or visual-only information is available for some data-points), as well as unimodal databases.

We quantitatively evaluate and compare the performance of our proposed model against strong baselines. The results demonstrate that our architecture achieves significantly better results on both the CREMA-D [24] and MSP-IMPROV [25] corpora compared to the baselines. For example, VAVL achieves a new state-of-the-art result for the emotional attribute prediction task on the MSP-IMPROV corpus. Also, our architecture is able to sustain strong performance in audio-only and video-only settings, compared to strong unimodal baselines. Additionally, we conduct ablation studies into our model to understand the effects of each component on our model's performance. These results show the benefits of our proposed framework for audio-visual emotion recognition.

2   Background

Emotion recognition has been widely explored using speech [2, 26–29], facial expressions [30, 31], and multimodal solutions [10, 32–34]. However, most of these models have focused on solving either emotion classification tasks [28, 30, 34] or emotional attribute prediction tasks [26, 27]. Furthermore, these studies are designed for either unimodal or multimodal solutions. Although some recent studies [14] have proposed approaches for audio-visual emotion recognition that can handle both tasks, they still rely on a single modality. There is a need for solutions that can handle multiple modality settings. Recent advancements in deep learning have resulted in the development of unified frameworks that aim to move away from unimodal implementations. This section focuses on relevant studies that address similar problems investigated in this paper.

2.1 Unified Multimodal Framework

Since the introduction of the transformer framework [16], many studies have explored multimodal frameworks that can still work even when only one modality is available [35]. Some proposed architectures employ training schemes that involve solely utilizing separate branches for different modalities, as seen in Baevski *et al.* [36]. These approaches construct branches containing unimodal information specific to each modality. In contrast, other studies explore the use of shared layers that encapsulate cross-modal information from multiple modalities [37, 38]. The aforementioned studies incorporate aspects that laid the foundation for modeling a more unified framework, but they still have some limitations. For example, common solutions require independent modality training, assume that all the modalities are available during inference, or fail to fully address the heterogeneity present in a multimodal setting. Recent studies have proposed audio-visual frameworks that can handle both unimodal and multimodal settings [39], but they still rely on having additional unimodal network branches. Gong *et al.* [18] proposed a unified audio-visual framework for classification, which incorporates some effective capabilities, such as independently processing audio and video, and including shared audio-visual layers into the model. However, the model focused only on a classification task.

2.2 Formulations for Emotion Recognition

Although an individual's cultural and ethnic background may impact their manner of expression, studies have indicated that certain basic emotions are very similar regardless of cultural differences [40]. These basic emotions are described as: happiness, sadness, surprise, fear, disgust, and anger. These basic emotional states are often the most widely explored task in emotion recognition, formulating the problem as a six-class classification problem [41, 42]. In fact, many of the affective corpora include all or a subset of these classes, including the CREMA-D [24], AffectNet [43], and AFEW [44] databases.

An alternative way to describe emotions is to leverage the continuous space of emotional attributes. The approach identifies dimensions to describe expressive behaviors. The most common attributes are arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) [45]. Several databases have explored the emotional attribute annotation of audio-visual stimuli, such as the IEMOCAP [46], and MSP-IMPROV [25] corpora. Studies have proposed algorithms to predict emotional attributes using audio-visual stimuli [47, 48], facial expression [49], and speech [2, 26, 50].

2.3 Versatile Task Modeling

For audio-visual unified models [18] and audio-visual models that have multiple modality branches and can handle unimodal scenarios [11], the final multimodal prediction is often obtained by averaging the outputs from different branches, receiving the contributions from all the modalities. In emotion classification tasks, the model output is a probability distribution over a set of discrete categories. Averaging predictions from multiple unimodal branches can help capture different aspects of the data, leading to improved classification accuracy. However, this approach may not be effective for emotion attribute regression tasks, as regression involves predicting a continuous numerical value rather than a discrete categorical emotion label. Therefore, the average of the predictions from different unimodal branches may not be appropriate. This is particularly true when considering the performance gap that often exists between speech and visual models for predicting arousal, valence, and dominance, as facial-only features are generally less effective than acoustic features [10]. Simply averaging these two results could lead to suboptimal outcomes. A potentially appealing alternative is to use a weighted combination of the predictions. However, this approach also has its limitations, as the prediction discrepancy may not be consistent across all data points. To address this issue, our study proposes training an audio-visual prediction layer, in addition to visual and acoustic prediction layers, used when only data from one modality is available. The audio-visual prediction layer is optimized only when both modalities are available. This layer can automatically learn appropriate weights to combine the modality representations and fuse them to generate a single multimodal prediction. This approach works well for both emotion classification and regression settings.

2.4 Relation to Prior Work

Our proposed VAVL model represents a significant contribution in comparison to previous works. This study investigates the use of the transformer architecture to build a versatile framework. Specifically, we employ conformer layers [51] as our encoders, which are transformer-based encoders augmented with convolutions. The closest study to our framework is the work of Gong *et al.* [18], which proposed a unified audio-visual model for classification. Their framework involved the independent processing of audio and video features, with shared transformer and classification layers for both modalities. Prior to the shared transformer layers, they have modality-specific feature extractors and optional modality-specific transformers for each modality. In contrast, our framework employs conformer layers instead of
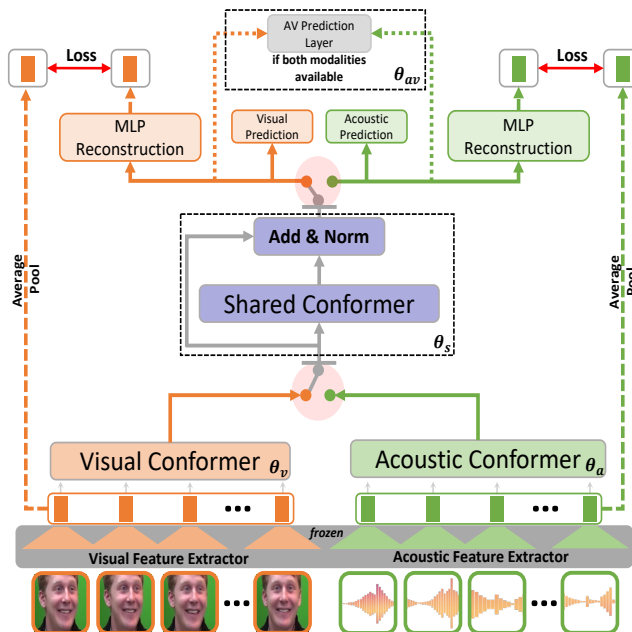
Figure 1: Overview of our proposed *versatile audio-visual learning* (VAVL) framework. The orange branches represent the visual information and are parameterized by the weights θ_v. The green branches represent the acoustic information and are parameterized by the weights θ_a. The purple modules are the shared layers where both modalities flow through, which are parameterized by the weights θ_s. Finally, the gray module is the audio-visual prediction layer, which is parameterized by the weights θ_{av}.

vanilla transformers. Moreover, we introduce three important components to our framework. First, an audio-visual prediction layer is utilized solely when audio-visual modalities are available during inference, enhancing prediction accuracy and model versatility for both regression and classification tasks. Second, a unimodal reconstruction task is implemented during training to ensure that the shared layers capture the heterogeneity of the modalities. Third, residual connections are incorporated over the shared layers to ensure that the unimodal representations are not forgotten in the shared layers. This strategy maintains high performance even in unimodal settings. These novel additions represent significant contributions to the field of multimodal emotion recognition.

3 Proposed Approach

Figure 1 illustrates the architecture of the proposed approach, which consists of four major components. Each of these components is represented in the figure by the set of weights θ_v, θ_a, θ_s, and θ_{av}. The orange branches represent parts of the model where only visual information flows. This region is parameterized by the weights θ_v, and includes the visual conformer encoder layers that encode the visual input representations, the visual prediction layer for visual-only predictions, and the MLP reconstruction layers for the averaged visual input features. The green branches represent parts of the model where only acoustic information flows. This region is parameterized by the weights θ_a, and consists of the acoustic conformer encoder layers that encode the acoustic input representations, the acoustic prediction layer for acoustic-only predictions, and the MLP reconstruction layers for the averaged acoustic input features. The shared layers, depicted in purple in Figure 1, are parameterized by the weights θ_s. This block processes both acoustic and visual inputs to learn intermediate multimodal representations from both modalities. It also contains residual connections from the unimodal branches to the shared layers to preserve information from the unimodal representations. Lastly, the audio-visual prediction layer is parameterized by the weights θ_{av}, which focuses on processing audio-visual representations from the shared layers when the model receives paired audio-visual data for training or inference. The following sections describe these components in detail.

3.1 Acoustic and Visual Layers

As shown in Figure 1, the acoustic and visual layers mirror each other. Both layers share the same basic structure and training mechanism. The main components of these layers are conformer encoders [51], which process all the

sequential video or acoustic frames in parallel, depending on the modality available during either training or inference. To obtain inputs for our model, we utilize pre-trained feature extractors that produce a 1,408D feature vector for each visual frame and a 1,024D feature vector for each acoustic frame. Section 4.2 describes these feature extractor modules in detail. We apply a 1D temporal convolutional layer at the frame level before entering the feature vectors to their corresponding conformer encoder layers. This step ensures that each element in the input sequences is aware of its neighboring elements. This approach also allows us to project both feature vectors into a 50D feature representation, matching their dimensions.

As seen in the top region of Figure 1, there are two additional components separately implemented for each modality (acoustic and visual) that are implemented to (1) predict the emotional attributes, and (2) reconstruct the unimodal feature representations. For the prediction of the emotional attributes, the acoustic and visual layers contain a set of prediction layers, referred to as visual prediction and acoustic prediction in Figure 1. They are responsible for generating unimodal predictions, which are utilized when our model operates in a unimodal setting. For the reconstruction of the unimodal feature representation, the model has a *multilayer perceptron* (MLP) head that serves as an auxiliary reconstruction task. The reconstruction task is used to have the model reconstruct the average-pooled input representations for the modality being used during training at that moment. The input of the reconstruction is the average-pooled representations obtained at the output of the shared layers (Sec. 3.2). Equation 1 shows the total loss function of the model,

$$\mathcal{L} = \mathcal{L}_{pred}(y, pred) + \alpha\mathcal{L}_{MSE}(x_{pool}, Rec_M) \qquad (1)$$

where $\mathcal{L}_{pred}$ is the emotion task-specific loss (i.e., cross-entropy – Eq. 2, or *concordance correlation coefficient* (CCC) – Eq. 3), y is the input label, $pred$ is the emotional prediction of the model, α is the scaling weight for the reconstruction loss, $\mathcal{L}_{MSE}$ is *mean squared error* (MSE) loss between the average-pooled input value x_{pool} and the reconstructed input Rec_M for modality M.

The reconstruction task is included in our model to promote the learning of more general features that can be applied to both modalities while preserving separate information for each modality, as the reconstruction from the shared layers requires the model to retain information from both modalities.

3.2 Shared Layers

The shared layers, depicted in purple in Figure 1, mainly comprise a conformer encoder, following a similar structure to the acoustic and visual layers. These shared layers are modality-agnostic, meaning that during training and inference, the features from both modalities will pass through these layers whenever each modality is available. The purpose of this block is to ensure that the shared layers maintain information from both modalities incorporated in our model.

We also introduce a residual connection over the shared layers (gray arrow in model θ_s shown in Fig. 1). This residual connection from the unimodal branches over the shared layers ensures that the model retains unimodal separable information. This mechanism complements the reconstruction tasks mentioned in Section 3.1 to increase the robustness of the system when only one modality is available. The approach allows the gradients to flow directly from the shared layers to the unimodal branches, preserving modality-specific information in our model.

3.3 Audio-Visual Prediction Layer

When only one modality is available, the system will use the visual or acoustic prediction blocks. When both modalities are available for training or inference, we use the audio-visual prediction layer, highlighted in gray in Figure 1. This layer plays a crucial role in making our approach versatile. Unlike other methods that rely on averaging predictions from different branches in their models, the audio-visual layers effectively utilize the contributions of acoustic and visual inputs in the multimodal space. The audio-visual prediction layer consists of two fully connected layers that feed into a final audio-visual prediction head. We added this simple structure to ensure that representations from both modalities, obtained from the average-pooled output of each modality from the shared layers, are properly combined to obtain a final audio-visual prediction. During training, when audio-visual data is available, all other layers of the model are frozen after updating with acoustic and visual data. Only this layer is separately optimized to learn the required weights for audio-visual predictions. This layer ensures proper combination of audio-visual representations for robust predictions in both classification and regression settings.

3.4 Versatile Model Training

We train our proposed VAVL model using Algorithm 1. The training process can involve either a single modality or both modalities at any given iteration. If both modalities are available, the model first backpropagates the errors using

---

Algorithm 1 - VAVL (Training and Inference)

Require: $\mathcal{D} = \{A, V, Label\}$, $\mathcal{M} = \{\theta_a, \theta_v, \theta_s, \theta_{av}\}$

Training$(\mathcal{D}, \mathcal{M}, \alpha)$

 1: **while** $i <$ *max train iterations* **do**
 2:     sample a batch of $\{A_i, V_i, Label_i\}$
 3:     **if** $A_i \neq$ None **then**
 4:         $Pred_a, A_{i-pool}, Rec_a = \mathcal{M}_{\{\theta_a, \theta_s\}}(A_i)$
 5:         $\mathcal{L} = \mathcal{L}_{pred}(Label_i, Pred_a) + \alpha\mathcal{L}_{MSE}(A_{i-pool}, Rec_a)$
 6:         backpropagate and update $\{\theta_a, \theta_s\}$
 7:     **if** $V_i \neq$ None **then**
 8:         $Pred_v, V_{i-pool}, Rec_v = \mathcal{M}_{\{\theta_v, \theta_s\}}(V_i)$
 9:         $\mathcal{L} = \mathcal{L}_{pred}(Label_i, Pred_v) + \alpha\mathcal{L}_{MSE}(V_{i-pool}, Rec_v)$
10:         backpropagate and update $\{\theta_v, \theta_s\}$
11:     **if** $A_i \neq$ None **and** $V_i \neq$ None **then**
12:         $freeze\ \{\theta_a, \theta_v, \theta_s\}$
13:         $Pred_{av} = \mathcal{M}_{\{\theta_a, \theta_v, \theta_s, \theta_{av}\}}(A_i, V_i)$
14:         $\mathcal{L} = \mathcal{L}_{pred}(Label_i, Pred_{av})$
15:         backpropagate and update $\{\theta_{av}\}$
16: **return** $\mathcal{M}$

Inference $(\{A, V\}, \mathcal{M})$

17: **if** $A \neq$ None **and** $V \neq$ None **then**
18:     $Pred = \mathcal{M}_{\{\theta_a, \theta_v, \theta_s, \theta_{av}\}}(A, V)$
19: **else if** one modality is not available **then**
20:     $Pred = \mathcal{M}_{\{\theta_a, \theta_s\}}(A)$ when $V == None$
21:     $Pred = \mathcal{M}_{\{\theta_v, \theta_s\}}(V)$ when $A == None$
22: **return** $Pred$

---

the acoustic predictions to optimize the acoustic and shared weights (θ_a, θ_s). Then, it backpropagates the errors using the visual predictions to optimize the visual and shared weights (θ_v, θ_s). Next, it freezes all the parameters of the models other than the audio-visual prediction block (i.e., $\theta_v, \theta_a,$ and θ_s are frozen). Then, it backpropagates the errors using the audio-visual predictions to optimize the audio-visual prediction weights (θ_{av}). This method facilitates the use of unpaired and paired audio-visual data for training. If only one modality is available, we only update either the visual and shared weights (θ_v, θ_s) or acoustic and shared weights (θ_a, θ_s).

At inference time, the model utilizes the available information to make predictions, and in cases where both modalities are present, the model outputs the predictions from the audio-visual prediction layer. When only one modality is available during inference, the model outputs the prediction from the corresponding visual or acoustic prediction blocks.

4 Experimental Settings

4.1 Emotional Corpora

In this study, we use the CREMA-D [24] and MSP-IMPROV [25] corpora. The CREMA-D corpus is an audio-visual dataset with high-quality recordings from 91 ethically and racially diverse actors (48 male, 43 female). Actors were asked to convey specific emotions while reciting sentences. Videos were recorded against a green screen background, with two directors overseeing the data collection. One director worked with 51 actors, while the other worked with 40 actors. Emotional labels were assigned by at least seven annotators. In total, 7,442 clips were collected and rated by 2,443 raters. We use the perceived emotions from the audio-visual modality in the CREMA-D corpus for our classification task. We consider six emotional classes: anger, disgust, fear, happiness, sadness, and neutral state.

The MSP-IMPROV corpus [25] is the second audio-visual database used in this study. The corpus was collected to study emotion perception [52]. The corpus required sentences with identical lexical content but conveying different emotions. Instead of actors reading sentences, a sophisticated protocol elicited spontaneous target sentence renditions. The corpus includes 20 target sentences with four emotional states, resulting in 80 scenarios and 652 speaking turns. It also contains the interactions that prompted the target sentence (4,381 spontaneous speaking turns), natural interactions

during breaks between the recording of dyadic scenarios (2,785 natural speaking turns), and read recordings expressing the target emotional classes (620 read speaking turns). In total, the MSP-IMPROV corpus contains 7,818 non-read speaking turns and 620 read sentences. The corpus was annotated using a crowdsourcing protocol, monitoring the quality of the workers in real-time [53]. Each sentence was annotated for arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) by five or more raters using a five-point Likert scale. We employ these three emotional attributes for our regression task.

4.2 Acoustic and Visual Features

Our acoustic feature extractor is based on the "wav2vec2-large-robust" architecture [54], which has shown superior recognition performance compared to other variants of the Wav2vec2.0 model [55], as demonstrated in the study by Wagner *et al.* [50]. The downstream head of our model consists of two fully connected layers with 1,024 nodes, layer normalization, and *rectified linear unit* (ReLU) activation function, followed by a linear output layer with three nodes for predicting emotional attribute scores (arousal, dominance, and valence). We import the pre-trained "wav2vec2-large-robust" model from the Hugging Face library [56]. We aggregate the output of the transformer encoder using average pooling and feed them to the downstream head. To regularize the model and prevent overfitting, we utilize dropout with a rate of $p = 0.5$ applied to all hidden layers. To fine-tune the model, we use the training set of the MSP-Podcast corpus (release of v1.10) [57]. The ADAM optimizer [58] is employed with a learning rate set to 0.0001. We update the model with mini-batches of 32 utterances for 10 epochs. With the fine-tuned "wav2vec2-large-robust" model, we extract frame-level representations from the given audio, which are then used as the acoustic features.

To obtain visual features, we used the *multi-task cascaded convolutional neural network* (MTCNN) face detection algorithm [59] to extract faces from each image frame in the corpora using bounding boxes. Following the extraction of bounding boxes, we resize the images to a predetermined dimension of $224 \times 224 \times 3$. After face extraction, we utilize the pre-trained EfficientNet-B2 model [31] to extract emotional feature representations, which has established state-of-the-art results on the AffectNet corpus [43]. The representation obtained from the EfficientNet-B2 model is retrieved from the last fully connected layer before classification, with an array dimension of 1,408. This representation is then concatenated row-wise with all other frames within each clip from the datasets, serving as the input for the visual branch of our framework.

Table 1: Comparison between the VAVL model and the audio-visual and unimodal baselines. The table reports the average performance metrics across five trials. The symbol $*$ indicates that the VAVL model is significantly better than the other baselines on the MSP-IMPROV and CREMA-D datasets.

| | MSP-IMPROV | | | | | | | | | CREMA-D | | | | | |
| | Audio-Visual | | | Acoustic | | | Visual | | | Audio-Visual | | Acoustic | | Visual | |
Model	Aro.	Val.	Dom.	Aro.	Val.	Dom.	Aro.	Val.	Dom.	F1-Ma	F1-Mi	F1-Ma	F1-Mi	F1-Ma	F1-Mi
VAVL	**0.856**$*$	**0.814**$*$	**0.876**$*$	**0.853**$*$	0.783	0.858	**0.422**$*$	**0.375**$*$	**0.631**$*$	**0.779**$*$	**0.826**$*$	**0.628**	**0.701**$*$	**0.738**$*$	**0.787**
UAVM	0.471	0.544	0.687	0.578	0.637	0.705	0.274	0.296	0.522	0.749	0.769	0.614	0.554	0.617	0.672
Auxformer	0.672	0.652	0.820	0.722	0.730	0.789	0.363	0.293	0.581	0.698	0.763	0.593	0.648	0.560	0.626
MulT	0.775	0.761	0.778	—	—	—	—	—	—	0.644	0.692	—	—	—	—
Uni. (A)	—	—	—	0.841	**0.820**	**0.878**	—	—	—	—	—	0.625	0.690	—	—
Uni. (V)	—	—	—	—	—	—	0.383	0.321	0.598	—	—	—	—	0.725	0.783

4.3 Implementation Details

We implemented the conformer blocks with an encoder hidden layer set to 512D, with 8 attention heads. We set a dropout rate to $p = 0.1$. The number of layers in the acoustic, visual, and shared conformer layers were set to three, three, and two, respectively. The acoustic/visual prediction layers and the reconstruction layers are implemented using a fully-connected structure with three layers. The first two layers are implemented with 512 nodes and 256 nodes, respectively. For the acoustic and visual prediction modules, the third layer is the output layer, where the size depends on the emotion recognition task (i.e., classification or regression). For the reconstruction task, the third layer is the target representation to be reconstructed, so it has 1,024 nodes for the acoustic features and 1,408 nodes for the visual features. The fully connected layers are implemented with dropout, with the rate set to $p = 0.2$. Similarly, the audio-visual prediction layers have a mostly identical fully-connected hidden layer structure to the unimodal prediction layers. The only difference is that the input layers are now 1,024D, as they need to take both unimodal representations from the shared layers as inputs. We trained the models for 20 epochs using the ADAM optimizer, with the ReLU as the

activation function. The learning rate is set to 5e-5. We used a batch size of 32 for the CREMA-D dataset experiments, and a batch size of 16 for the MSP-IMPROV dataset experiments, since the MSP-IMPROV has longer sentences. The model was implemented in PyTorch and trained using an Nvidia Tesla V100.

The recordings in each database are divided into train, development, and test sets, with approximately 70%, 15%, and 15% of the data in each set, respectively. The splits were carried out in a speaker-independent manner, ensuring that no speaker appeared in more than one set. We trained each model five times, with different splits each time. All models were trained using the train set, and the best-performing model on the development set was selected and used to make predictions on the test set.

4.4 Cost Functions and Evaluation Metrics

Two of the most prominent tasks in affective computing are classification tasks for categorical emotions and regression models for emotional attributes. Our model is designed to be utilized and evaluated in both formulations.

For emotion classification, our training objective is based on the multiclass cross-entropy loss, as seen in equation 2,

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{2}$$

where M is number of classes, log is the natural log, c is the correct label for observation o, and p is the predicted probability that observation o belong to class c. We predict the emotional state/value for each input sequence in our test set and report the 'micro' and 'macro' F1-scores. The 'micro' F1-score is computed by considering the total number of true positives, false negatives, and false positives, making it sensitive to class imbalance. The 'macro' F1-score calculates the F1-score for each class separately and aggregated the scores with equal weight, so the performance on the minority class is as important as the performance in the majority class.

For regression models, we use the *concordance correlation coefficient* (CCC), which measures the agreement between the true and predicted emotional attribute scores. Equation 3 illustrates the CCC measurements,

$$\mathcal{L}_{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3}$$

where μ_x and μ_y are the means of the true and predicted scores, σ_x and σ_y are the standard deviation of the true and predicted scores, and ρ is their Pearson's correlation coefficient. We train our model to maximize CCC so that the predicted scores have a high correlation with the true scores, while reducing their errors in the prediction. We use the CCC metric for model evaluation of the predictions of arousal, valence, and dominance.

We conducted each experiment in this work five times with different partitions and reported the average metrics. Additionally, we performed statistical analyses to evaluate our model's performance against the baseline models. We used a two-tailed t-test, asserting a significance level at p-value$<$0.05.

4.5 Baselines

To evaluate the performance of our proposed model, we conducted experiments with three strong audio-visual frameworks and one unimodal framework. Baseline models 1, 2, and 3 were implemented using the code provided in their respective repositories.

4.5.1 Baseline 1: UAVM

Gong *et al.* [18] proposed a unified audio-visual framework for classification, which independently processes audio and video features. The framework includes a shared transformer component and a classification layer, whose weights are shared between the modalities.

4.5.2 Baseline 2: AuxFormer

Goncalves and Busso [39] proposed an audio-visual transformer-based framework that creates cross-modal representations through transformer layers, sharing representations from query inputs from one modality to the keys and values of another modality. Additionally, their architecture includes unimodal auxiliary networks and modality dropout during training to enhance the robustness to missing modalities, allowing the model to be used in both multimodal and unimodal settings.

Table 2: Ablation analysis of the proposed VAVL model using the MSP-IMPROV and CREMA-D corpora. Ablation 1 removes the residual connections over the shared layers. Ablation 2 removes reconstruction step from the framework. Ablation 3 removes audio-visual prediction layer, estimating the output by averaging the unimodal predictions.

| | MSP-IMPROV | | | | | | | | | CREMA-D | | | | | |
| | Audio-Visual | | | Acoustic | | | Visual | | | Audio-Visual | | Acoustic | | Visual | |
Model	Aro.	Val.	Dom.	Aro.	Val.	Dom.	Aro.	Val.	Dom.	F1-Ma	F1-Mi	F1-Ma	F1-Mi	F1-Ma	F1-Mi
VAVL	**0.856**	**0.814**	**0.876**	**0.853**	**0.783**	0.858	**0.422**	**0.375**	**0.631**	**0.779**	**0.826**	**0.628**	**0.701**	**0.738**	**0.787**
Ablt. 1	0.716	0.705	0.779	0.682	0.712	0.753	0.176	0.133	0.341	0.751	0.807	0.397	0.528	0.727	0.783
Ablt. 2	0.810	0.788	0.873	0.791	0.780	0.864	0.172	0.227	0.577	0.761	0.816	0.604	0.684	0.718	0.775
Ablt. 3	0.711	0.659	0.843	0.784	0.776	**0.870**	0.374	0.265	0.617	0.762	0.814	0.629	0.691	0.713	0.764

4.5.3 Baseline 3: MulT

Tsai *et al.* [60] proposed a multimodal transformer architecture for human language time-series data. Their model uses a cross-modal transformer framework that generates pairs of bimodal representations, where the keys and values of one modality interact with the queries of a target modality. Vectors with similar target modalities are concatenated and passed to another transformer layer that generates representations used for prediction. The original model considered textual, visual, and acoustic features. We adapt the model from the original study by removing the dependencies on the textual branch, focusing only on visual and acoustic features.

4.5.4 Baseline 4: Unimodal Acoustic and Visual Model

The unimodal baseline model has a similar structure to the model proposed in this study. Like the acoustic and visual layers of our model, it uses conformer encoder layers [51] to process all sequential video or acoustic frames in parallel. The output of the conformer layers is then average-pooled and fed into a network that contains two fully connected layers to generate the prediction. We build the unimodal network with five conformer layers to match the structure of the full VAVL network.

5 Experimental Results

5.1 Comparison with Baselines

This section compares our proposed model with the multimodal and unimodal baselines explored in this study. All models were trained using the details discussed in Section 4.3. Table 1 presents the average results for all the models across the five trials on each corpus. The models' performances are evaluated based on three modalities: audio-visual, acoustic, and visual. For the MSP-IMPROV dataset, the performance is assessed using the CCC predictions for arousal (Aro.), valence (Val.), and dominance (Dom.). For the CREMA-D dataset, we report the F1-Macro (F1-Ma) and F1-Micro (F1-Mi) scores.

In the MSP-IMPROV dataset, the VAVL model demonstrates strong performance in both the audio-visual and visual modalities, achieving the highest scores in arousal, valence, and dominance. The VAVL method consistently outperforms all the multimodal baselines across all the emotional attributes. In the acoustic setting, the VALV model achieves the best CCC values for the prediction of arousal. The average CCC values achieved by the VAVL model for valence (0.783) and dominance (0.858) show close performance to the unimodal model (i.e., Uni. (A)), but it does not outperform it. The results for the audio-visual setting on the MSP-IMPROV corpus reinforce our hypothesis that the use of averaging unimodal predictions, as employed by other baselines, does not work well for regression tasks. Our model is the only one to hold stronger predictions in audio-visual settings over unimodal settings across all baselines.

On the CREMA-D dataset, the VAVL model consistently outperforms the other baseline models across all three modalities. For the audio-visual modality, VAVL achieves an F1-Macro score of 0.779 and an F1-Micro score of 0.826, significantly surpassing the performance of other baselines (i.e., UAVM, Auxformer, and MulT). In the acoustic and visual modalities, VAVL continues to outperform other models, although the difference in performance becomes less pronounced.
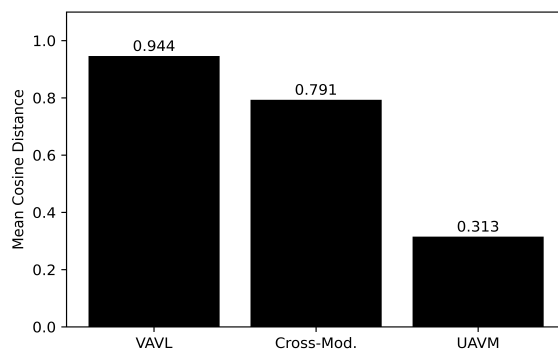
9

Figure 2: Comparison of average cosine distances between acoustic and visual representations generated by shared layers of the VAVL, AuxFormer, and UAVM models.

5.2 Ablation Analysis of the VAVL Framework

Table 2 presents the results of an ablation analysis of the VAVL model on the MSP-IMPROV and CREMA-D databases, comparing the performance of the full model with its ablated versions. Three ablated models are considered: Ablation 1 (Ablt. 1) removes the residual connections over the shared layers, Ablation 2 (Ablt. 2) removes the reconstruction step from the framework, and Ablation 3 (Ablt. 3) removes the audio-visual prediction layer, using the average of the unimodal predictions as the audio-visual prediction.

In the MSP-IMPROV dataset, the full VAVL model outperforms all the ablated versions in terms of arousal, valence, and dominance for the audio-visual and visual modalities. In the acoustic modality, VAVL shows the best performance for arousal and valence. However, Ablation 2 and 3 slightly surpass the VAVL results for dominance. These results indicate that the residual connections, reconstruction step, and audio-visual prediction layer all contribute to the strong performance of the VAVL model.

On the CREMA-D dataset, VAVL consistently achieves the highest F1-Macro and F1-Micro scores across multimodal and unimodal settings, indicating that the full model is superior to its ablated versions. Ablation 1 has the lowest performance in the acoustic modality, while Ablations 2 and 3 show competitive results in some cases, albeit not surpassing the results of the full VAVL model. These results are consistent with the MSP-IMPROV ablation results and suggest that the combination of all components in the VAVL model leads to the best performance, with each component playing a significant role in the overall success of the model.

5.3 Shared Embedding Analysis

In this section, we perform analysis on the output embedding representations for each modality at the output of the shared layers from our VAVL model and the baselines. This experiment is conducted to investigate whether the shared layers of a multimodal model are capable of producing different representations for each modality. Using shared layers in a multimodal model allows the architecture to learn a common representation across different modalities, which can help in tasks that require integrating information from multiple sources. However, maintaining distinct representations for each modality in a multimodal model is crucial for several reasons. First, it allows the model to capture modality-specific information, which is essential for optimal performance. Second, it enhances interpretability by making it easier to analyze each modality's contribution to the final output or decision. Lastly, distinct representations ensure the model's adaptability, enabling it to perform well even when some modalities are missing or incomplete. To verify our proposed model's capability of generating distinct separable representations for each modality from the shared layers, we obtain separate output embeddings from the shared layers of the trained model using either acoustic or visual information. Then, we calculate the cosine distance between their respective embeddings. On the one hand, a high cosine distance indicates that the embeddings are more dissimilar, and thus capturing modality-specific information. On the other hand, a low cosine distance implies that the embeddings are more similar, potentially indicating that the model is not effectively capturing the unique features of each modality, and thus not effectively utilizing the multimodal input.

Figure 2 presents a comparison of the average cosine distances between acoustic and visual representations generated by the shared layers of our proposed framework (VAVL). For comparison, we use the shared models from the AuxFormer

architecture [39], which are implemented with cross-modal layers (the same type of layers used in the MulT model [60]), and from the UAVM model [18]. Based on the average cosine distance values obtained, the VAVL shared layers is the most effective model in generating distinct representations for the acoustic and visual modalities. The cross-modal layers show moderate effectiveness, while the UAVM shared layers seem to struggle in separating the modalities. These results correlate with our model's superior overall performance in unimodal settings compared to the baselines.

6   Conclusions

This study introduced the VAVL model, a novel, unified, and versatile framework for audio-visual, audio-only, and video-only emotion recognition. The proposed approach is built with shared layers, residual connections over shared layers, and a unimodal reconstruction task. It attains high emotion recognition performance and can be trained even when paired audio and visual data is unavailable for part of the recordings. Furthermore, the model allows for direct application across both emotion regression and classification tasks. In the MSP-IMPROV dataset, the VAVL model consistently outperforms the CCC predictions of all multimodal baselines across all emotional attributes, achieving 0.856, 0.814, and 0.876 for arousal, valence, and dominance in audio-visual settings. Our results demonstrate that our method is able to better leverage visual representations to improve both unimodal and audio-visual performance, particularly in the visual-only setting. In the CREMA-D dataset, VAVL consistently surpasses baselines across all settings, achieving the highest F1-Macro (0.779) and F1-Micro (0.826) scores in the audio-visual modality. In addition to high performance on emotion recognition tasks, we show through the experimental evaluation that the shared layers of the proposed architecture are able to better capture the distinct features contained in acoustic and visual inputs, when compared to the shared layers of the baseline models.

Future work for this study is to explore the addition of more modalities to the framework (e.g., text) and utilize this framework on other audio-visual tasks such as sound recognition [61], scene classification [62], event localization [63], and audio-visual speech recognition [64, 65]. Furthermore, it would be interesting to explore the application of this model to more naturalistic data.

References

[1] C. Busso, M. Bulut, and S.S. Narayanan. Toward effective automatic recognition systems of emotion in speech. In J. Gratch and S. Marsella, editors, *Social emotions in nature and artifact: emotions in human and human-computer interaction*, pages 110–127. Oxford University Press, New York, NY, USA, November 2013.

[2] B. Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, April 2018.

[3] A. N. Salman and C. Busso. Style extractor for facial expression recognition in the presence of speech. In *IEEE International Conference on Image Processing (ICIP 2020)*, pages 1806–1810, Abu Dhabi, United Arab Emirates (UAE), October 2020.

[4] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[5] S. Mariooryad and C. Busso. Facial expression recognition in the presence of speech using blind lexical compensation. *IEEE Transactions on Affective Computing*, 7(4):346–359, October-December 2016.

[6] C. Busso and S.S. Narayanan. Interplay between linguistic and affective goals in facial expression during emotional utterances. In *7th International Seminar on Speech Production (ISSP 2006)*, pages 549–556, Ubatuba-SP, Brazil, December 2006.

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Sixth International Conference on Multimodal Interfaces ICMI 2004*, pages 205–211, State College, PA, October 2004. ACM Press.

[8] S. K. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, 47(3):1–36, April 2015.

[9] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18:351–376, June 2021.

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

[11] L. Goncalves and C. Busso. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170, October-December 2022.

[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *International conference on machine learning (ICML2011)*, pages 689–696, Bellevue, WA, USA, June-July 2011.

[13] R. Arandjelović and A. Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision (ICCV 2017)*, pages 609–617, Venice, Italy, October 2017.

[14] P. Antoniadis, I. Pikoulis, P.P. Filntisis, and P. Maragos. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2021)*, pages 3638–3644, Montreal, BC, Canada, October 2021.

[15] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao. Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. *Neurocomputing*, 391:42–51, May 2020.

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

[17] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Conference on Neural Information Processing Systems (NeurIPS 2021)*, pages 24206–24221, Virtual Conference, December 2021.

[18] Y. Gong, A. H. Liu, A. Rouditchenko, and J. Glass. UAVM: Towards unifying audio and visual models. *IEEE Signal Processing Letters*, 29:2437–2441, November 2022.

[19] T. Baltrušaitis, C. Ahuja, and L. P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, February 2019.

[20] M. El Ayadi, M.S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, March 2011.

[21] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency. Multi-attention recurrent network for human communication comprehension. In *AAAI Conference on Artificial Intelligence (AAAI 2018)*, volume 32, pages 5642–5649, New Orleans, LA, USA, February 2018.

[22] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 1955–1960, Austin, TX, USA, November 2016.

[23] P.P. Liang, A. Zadeh, and L.-P. Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ArXiv e-prints (arXiv:2209.03430)*, pages 1–36, September 2022.

[24] 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 Transactions on Affective Computing*, 5(4):377–390, October-December 2014.

[25] 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*, 8(1):67–80, January-March 2017.

[26] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 6447–6451, Singapore, May 2022.

[27] S. Parthasarathy, R. Lotfian, and C. Busso. Ranking emotional attributes with deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, pages 4995–4999, New Orleans, LA, USA, March 2017.

[28] R. Lotfian and C. Busso. Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning. In *Interspeech 2018*, pages 951–955, Hyderabad, India, September 2018.

[29] W.-C. Lin and C. Busso. An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks. In *Interspeech 2020*, pages 2322–2326, Shanghai, China, October 2020.

[30] A. N. Salman and C. Busso. Privacy preserving personalization for video facial expression recognition using federated learning. In *ACM International Conference on Multimodal Interaction (ICMI 2022)*, pages 495–503, Bangalore, India, November 2022.

[31] A. V. Savchenko, L. V. Savchenko, and I. Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, October-December 2022.

[32] J. Li, Y. Chen, X. Zhang, J. Nie, Z. Li, Y. Yu, Y. Zhang, R. Hong, and M. Wang. Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. *ArXiv e-prints (arXiv:2303.09164)*, pages 1–7, MArch 2023.

[33] J. Yu, J. Zhu, W. Zhu, Z. Cai, G. Xie, R. Li, and G. Zhao. A dual branch network for emotional reaction intensity estimation. *ArXiv e-prints (arXiv:2303.09210)*, pages 1–6, March 2023.

[34] L. Goncalves and C. Busso. Learning cross-modal audiovisual representations with ladder networks for emotion recognition. In *IEEE International Conference onicassp Acoustics, Speech and Signal Processing (ICASSP 2023),*, Rhodes island, Greece, 2023.

[35] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In M. Meila, , and T. Zhang, editors, *International Conference on Machine Learning (ICML 2021)*, volume 139, pages 4651–4664. Proceedings of Machine Learning Research (PMLR), Virtual, July 2021.

[36] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning (ICML 2022)*, volume 162, pages 1298–1312. Proceedings of Machine Learning Research (PMLR), Honolulu, HI, USA, July 2022.

[37] Y. Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *ArXiv e-prints (arXiv:2205.06126)*, pages 1–14, May 2022.

[38] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. Feris, D. Harwath, J. Glass, and H. Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 19988–19997, New Orleans, LA, USA, June 2022.

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

[40] P. Ekman and W.V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, March 1971.

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

[42] E. Ghaleb, M. Popa, and S. Asteriadis. Metric learning-based multimodal audio-visual emotion recognition. *IEEE MultiMedia*, 27(1):37–48, January-March 2020.

[43] 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*, 10(1):18–31, January-March 2019.

[44] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 19(3):34–41, July-September 2012.

[45] M.M. Bradley and P.J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, March 1994.

[46] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, December 2008.

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

[48] B. T. Atmaja and M. Akagi. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pages 4482–4486, Barcelona, Spain, May 2020.

[49] A.T. Wasi, K. Šerbetar, R. Islam, T.H. Rafi, and D.-K. Chae. ARBEx: Attentive feature extraction with reliability balancing for robust facial expression learning. *ArXiv e-prints (arXiv:2305.01486)*, pages 1–12, May 2023.

[50] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B.W. Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *ArXiv e-prints (arXiv:2203.07378)*, pages 1–25, March 2022.

[51] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040, Shanghai, China, October 2020.

[52] E. Mower Provost, Y. Shangguan, and C. Busso. UMEME: University of Michigan emotional McGurk effect data set. *IEEE Transactions on Affective Computing*, 6(4):395–409, October-December 2015.

[53] A. Burmania, S. Parthasarathy, and C. Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, October-December 2016.

[54] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *ArXiv e-prints (arXiv:2104.01027)*, pages 1–9, April 2021.

[55] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 12449–12460, Virtual, December 2020.

[56] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, and Q. Lhoest amd A.M. Rush. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv e-prints (arXiv:1910.03771v5)*, pages 1–8, October 2019.

[57] R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October-December 2019.

[58] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–13, San Diego, CA, USA, May 2015.

[59] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016.

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

[61] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pages 721–725, Barcelona, Spain, May 2020.

[62] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen. A curated dataset of urban scenes for audio-visual scene analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, pages 626–630, Toronto, ON, Canada, June 2021.

[63] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *European Conference on Computer Vision (ECCV 2018)*, volume 11206 of *Lecture Notes in Computer Science*, pages 252–268. Springer Berlin Heidelberg, Munich, Germany, September 2018.

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

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