Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation

Sidharth Anand[*][†]
f20191203@hyderabad.bits-pilani.ac.in
BITS Pilani
Hyderabad, Telengana, India

Naresh Kumar Devulapally[*][‡]
devulapa@buffalo.edu
State University of New York at Buffalo
Buffalo, New York, USA

Sreyasee Das Bhattacharjee
sreyasee@buffalo.edu
State University of New York at Buffalo
Buffalo, New York, USA

Junsong Yuan
jsyuan@buffalo.edu
State University of New York at Buffalo
Buffalo, New York, USA

ABSTRACT

Evaluating speaker emotion in conversations is crucial for various applications requiring human-computer interaction. However, co-occurrences of multiple emotional states (e.g. 'anger' and 'frustration' may occur together or one may influence the occurrence of the other) and their dynamic evolution may vary dramatically due to the speaker's internal (e.g., influence of their personalized socio-cultural-educational and demographic backgrounds) and external contexts. Thus far, the previous focus has been on evaluating only the dominant emotion observed in a speaker at a given time, which is susceptible to producing misleading classification decisions for difficult multi-labels during testing. In this work, we present *Se*lf-supervised *M*ulti-*La*bel *P*eer *C*ollaborative Distillation (SeMuL-PCD) Learning via an efficient *Multimodal Transformer Network*, in which complementary feedback from multiple mode-specific peer networks (e.g.transcript, audio, visual) are distilled into a single mode-ensembled fusion network for estimating multiple emotions simultaneously. The proposed *Multimodal Distillation Loss* calibrates the fusion network by minimizing the Kullback–Leibler divergence with the peer networks. Additionally, each peer network is conditioned using a self-supervised contrastive objective to improve the generalization across diverse socio-demographic speaker backgrounds. By enabling peer collaborative learning that allows each network to independently learn their mode-specific discriminative patterns, *SeMUL-PCD* is effective across different conversation environments. In particular, the model not only outperforms the current state-of-the-art models on several large-scale public datasets (e.g., MOSEI, EmoReact and ElderReact), but with around 17% improved weighted F1-score in the cross-dataset experimental settings. The model also demonstrates an impressive generalization ability across age and demography-diverse populations.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**; *Semi-supervised learning settings*; Cross-validation.

KEYWORDS

Emotion Analysis, Transformer, Knowledge Distillation, Multi-label Classification, Collaborative Learning

ACM Reference Format:
Sidharth Anand, Naresh Kumar Devulapally, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612517

1 INTRODUCTION

Emotion analysis in conversation has become a prominent area of research in recent years due to its widespread applications in various fields including healthcare, education, entertainment, and marketing industries. In fact, a reliable and accurate emotion recognition system can play a pertinent role in an intelligent interactive agent that may require to adaptively determine appropriate responses based on users' emotions. *Though there is evidence that humans can experience multiple emotions simultaneously [29, 40, 41, 46, 53], most existing methods emphasize only estimating a single dominant emotion [32, 35, 62], which limits its usability in a realistic conversation setting.* In fact, just inferring the dominant emotion in isolation may be severely misleading at times. For example, compare two situations in a classroom setting: a student exclusively experiencing 'boredom' vs. a student experiencing both 'boredom' and 'frustration'. While it is obvious that both situations require a personalized intervention to facilitate an uninterrupted and enjoyable learning experience for the student, they would need completely different types of intervention. *Nevertheless, evaluating a human expression to analyze the existence of multiple co-occurring emotion states is particularly difficult as the ground truth labels are often correlated, with some emotions being more visually prominent than*

[*] The authors contributed equally to this work.
[†] Work done as part of an undergraduate thesis at State University of New York at Buffalo
[‡] Work done as part of a graduate thesis at State University of New York at Buffalo
Code available at: https://github.com/devulapa/multimodal-emotion-recognition

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612517

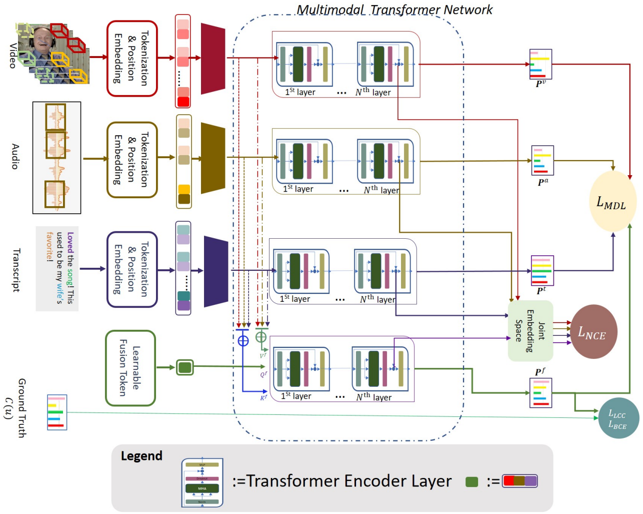

Multimodal Transformer Network
Video
Audio
Transcript
Loved the song! This used to be my wife's favorite!
Ground Truth
$C(u)$
Tokenization & Position Embedding
Tokenization & Position Embedding
Tokenization & Position Embedding
Learnable Fusion Token
1st layer
...
Nth layer
V^f
Q^f
K^f
P^v
P^a
P^t
P^f
L_{MDL}
Joint Embedding Space
L_{NCE}
L_{LCC}
L_{BCE}
Legend
MLP
Dropout
MHA
Norm
:=Transformer Encoder Layer
:=

Figure 1: Overview of the proposed *SeMUL-PCD* Model

other co-occurring ones. With limited training data describing such multi-label co-occurrence patterns, predicting a mixture of easy and difficult labels frequently may result in the model overfitting specifically due to over-training on those hard-to-learn labels.

Furthermore, it is important to note that emotion is performed by perceiving, recognizing, and interpreting human behavior and intent from various cues such as facial expressions, acoustic features, and text transcripts [34, 36]. As a human we use complementary information from multiple modalities to conclude an individual’s emotion [9]. Toward this, most of the existing methods [49, 58] leverage a static approach for multimodal fusion. However, *the importance of all modes is not always constant. In fact, an important challenge in this scenario is that the mode-specific contexts do vary across time, conversational context, and subject.* A lack of understanding of this mutual interplay may generate the risk of misclassification.

Lastly, a reliable emotion recognition system is expected to generalize well across users from diverse demographics (e.g. age, race, ethnicity) [16, 22, 43, 49, 61]. For example, analyzing emotion for individuals across ages is challenging due to the fact that aging causes facial shapes and appearances to change significantly [20, 25, 47]. A similar observation has also been made in analyzing the expression of emotions across other demographic variations as well [19, 37]. Few studies address the task within a broader spectrum of demographic variances [5, 38, 46, 72]. For an emotion recognition system to be reliable, its generalization capacity across different demographic backgrounds is critical to ensure an optimized decision quality and prevent systemic bias.

The proposed *S*elf-supervised *Mu*lti-*L*abel *P*eer *C*ollaborative *D*istillation (*SeMuL-PCD*) learning framework aims to address the aforementioned challenges. An overview of the framework is illustrated in Figure 1. The Primary Contributions of *SeMUL-PCD* include the followings:

(1) An effective *Multimodal Transformer Network* that enables its *multiple mode-specific networks to independently learn the complementary uni-modal perspectives and distill their individually perceived knowledge into its single attention-enhanced peer-ensembled fusion branch network.* In a self-supervised collaborative learning setting the proposed model is trained via an efficient cross-network attention fusion mechanism to facilitate robust decision making.
(2) The proposed *Label Consistency Calibration loss* enables calibration of the peer-ensembled fusion branches with respect to the difficulty of different emotion labels, thereby avoiding the network output being biased by the difficult labels.
(3) To address the lack of sufficient representative samples of difficult co-occurring emotions in the training collection, the proposed *Multmodal Distillation Loss* is specifically tailored to leverage the calibrated mode-specific peer network outputs for pseudo-supervising the peer-ensembled branch network. Thus, the label prediction uncertainties are learned via knowledge distillation, which may significantly relieve the manual labeling burden.
(4) The multi-label prediction is facilitated by the introduction of Binary Cross Entropy loss, which allows shortlisting sufficiently prominent emotional traits observed in an utterance. An intuitive emotion explanation combined with superior performance in several public datasets representing human behaviors across a variety of age/demographic backgrounds demonstrates an effective generalization capacity of *SeMUL-PCD* compared to the state-of-the-art models.

2 RELATED WORK

Multimodal emotion recognition in conversation replicates humans by combining information from multimodal heterogeneous behavioral, biological, and external cues to estimate an individual’s emotional state [55]. Several recent studies have studied multimodal interconnections of textual, visual, and acoustic data [35, 44, 48, 49, 52]. To combine cross-modal information, many of these methods either concatenate the processed unimodal feature vectors [35, 49] or by a predefined fixed combination (e.g., a weighted average of feature vectors) [57, 70]. Furthermore, To better estimate the emotion within an utterance, several existing works leverage the contextual information preserved within an utterance sequence [22, 23, 31, 35, 65]. While multimodal fusion methods seem to be promising [16], a static approach to fusion may not work equally well for all query and emotion patterns. *However, predicting emotion within an isolated utterance is still considered difficult.* For example, due to its reliance on the utterance sequence context, DialogueCRN [35] demonstrates a drastically deteriorated performance in predicting emotion in an utterance with significant polarity shifts[2].

An important challenge is that the fusion of mode-specific information may not be always uniform across instances or categories [18] and vary given an individual’s unique socio-cultural-demographic contexts. For example, many of the existing multimodal frameworks focus on estimating adult emotions and do not generalize well across ages [12, 27, 35, 46]. One major factor is the difference in emotional expressions (both verbal and non-verbal) between different age groups [20]. For instance, children and adults may express the same emotion differently due to differences in facial musculature, socialization, and cognitive development [46]. However, only a small fraction of public dataset [46, 68, 69] represent different emotions exhibited by the elder population, though the size constraint often prohibits effective training of large-scale computational models. In fact, only a limited few recent studies consider age in their decision-making task [46, 56]. In particular, models trained on a restricted age group may not be able to capture the variability of emotional expressions across different age groups[55] and thereby may not establish a reliable extent of generalization. *Toward this, SeMuL-PCD designs a Multimodal Transformer Network comprising of multiple mode-specific peer branch networks. In a self-supervised collaborative learning setting, these independent branch networks are trained to model a uni-modal perspective of the multimodal input, while simultaneously distilling their learned knowledge into a single peer-ensembled fusion branch network by means of an effective cross-network attention fusion.*

The other challenge that is frequently ignored in addressing the task of recognizing emotion in conversation video is that an utterance usually tends to generate more than one emotion. However, in most cases classifying multiple emotions in parallel is more difficult than just the dominant one. A set of recent works [71, 73, 77, 79] attempt to estimate multiple emotions simultaneously. While these models require a large-scale dataset for training, most of them do not generalize well across speakers' demographics. In fact, as we observe, among all emotions present in an individual's expression, some are more visible than others, which makes some emotions 'difficult' to recognize than others, and that makes the task of multi-label multimodal emotion recognition task furthermore challenging. *To address this the proposed Multimodal Transformer Network at the core of SeMuL-PCD, is trained using a multi-component loss function that simultaneously achieves the following objectives - enhances the model's ability to generalize across varying demographics (e.g., wider age range) with our cross-modal Self-supervised Noise Contrastive Objective, prevents label bias by improving the calibration via an effective Label Consistency Calibration Loss, and facilitates training in a limited data environment by leveraging learned uni-modal knowledge and distilling it into the fusion branch network decision via our Multmodal Distillation Loss.*

3 PROPOSED METHOD

In this section, we propose an effective *Se*lf-supervised *Mu*lti-*L*abel *P*eer *C*ollaborative *D*istillation (*SeMuL-PCD*) Emotion Learning approach by designing a *Multimodal Transformer Network*. The proposed multi-branch network shares complementary mode-specific distilled knowledge from a set of uni-modal peer branch networks with a single peer-ensembled fusion branch and in a collaborative learning environment, the entire multimodal network (with its multiple peer branches) is trained in parallel. Intuitively, since all these mode-specific branches usually contain semantically similar features regarding the input video, sharing them helps to reduce the overall training cost compared to traditional *Teacher-Student* network architectures and helps define an improved collaboration mechanism among mode-specific peers to enable defining a discriminative multimodal feature descriptor in a self-supervised manner. Figure 1 depicts the overview of the model architecture. As illustrated in the figure, a *Multimodal Transformer Network* is designed, in which the visual branch uses Tubelet embedding [6] for the spatial and temporal dimension factorization of the input video to develop the visual branch of the transformer, while a pre-defined tokenization scheme of text (and audio) branch is used to design the text-specific (and audio-specific) branch [1, 24]. Each mode-specific peer branch network makes its individual evaluation that indicates its respective probabilistic uncertainty score against each emotion state in consideration. These branch-specific estimates jointly attend the single peer-ensembled fusion branch evaluation to deliver an aggregated output estimate. The proposed *Multimodal Distillation Loss* calibrates the fusion branch network toward that of multiple mode-specific branch networks of the proposed *Multimodal Transformer Network* by minimizing their total pairwise Kullback–Leibler divergences. Additionally, the learning iterations of each branch network are conditioned using a self-supervised contrastive objective to improve the model's generalization capacity across diverse socio-demographic speaker backgrounds.

Problem definition: Given a conversation, represented in terms of video $\{u_j\}_j \in \mathcal{D}$, the objective of the proposed *SeMuL-PCD* model is to evaluate multiple emotion states ('happy', 'sad', 'neutral', 'angry', 'excited', and 'frustrated') present in the speaker expression along with their uncertainties. To introduce notation simplicity, from now on, we will omit the suffix j and an arbitrary element u will be denoted as u unless the suffix is specifically required otherwise. Each $u \in \mathcal{D}$ is represented as $= (\mathbf{x}_v, \mathbf{x}_a, \mathbf{x}_t)$, where $\mathbf{x}_v$ represents its visual content, $\mathbf{x}_a$ represents its acoustic content, and x_t represents its text (or semantic) content.

3.1 Multimodal Transformer Network Architecture

Tokenization and Positional Encoding: The backbone of the proposed *Multimodal Transformer Network* has three identical *mode-specific peer branch networks* and a *peer-ensembled fusion branch network*, each of which embeds their corresponding mode-specific (or a fused) component to obtain the input to a multimodal transformer layer. For example, we represent each video component $\mathbf{x}_v \in \mathbb{R}^{F\times H\times W\times C}$ into a sequence of tubelets $\mathbf{z}' \in \mathbb{R}^{n_F\times n_H\times n_W\times n_C}$, where each tubelet $\mathbf{z}'$ is of dimension $n\times h\times w$. Therefore, $n_F = \lfloor \frac{F}{n} \rfloor$, $n_H = \lfloor \frac{H}{h} \rfloor$ and $n_W = \lfloor \frac{W}{w} \rfloor$ tokens are extracted from the temporal, height, and width dimensions respectively to fuse spatiotemporal information during tokenization itself [6]. To capture the positional embedding for each tubelet, we follow the positional encoding scheme by Akbari et al.[3] that applies a linear projection to every tubelet to generate a d dimensional vector representation. The weight vector $\mathbf{W}_{v,pos} \in \mathbb{R}^{n_F\cdot n_H\cdot n_W\cdot n_C\times d}$ defining the linear projection is comprised of learnable parameters. For the subsequent self-attention operations within the proposed transformer network to be permutation invariant, we encode the position of these patches, and the dimension-specific sequence of positional embeddings is learned as follows:

$$\mathbf{pe}(i,j,k) = \mathbf{pe}_{F_i} + \mathbf{pe}_{H_j} + \mathbf{pe}_{W_k},$$

$$\mathbf{PE}_F \in \mathbb{R}^{n_F\times d}, \mathbf{PE}_H \in \mathbb{R}^{n_H\times d}, \mathbf{PE}_W \in \mathbb{R}^{n_W\times d}$$

and $\mathbf{pe}_i$ is the i^{th} row of $\mathbf{PE}$ so that $n_F \cdot n_H \cdot n_W$ video tubelets are encoded by using $n_F + n_H + n_W$ position embeddings. The audio component $\mathbf{x}_a$, defined as a 1D signal of length T is first segmented into $\lfloor \frac{T}{a'} \rfloor$ tokens each containing a' waveform amplitudes and the linear projection defined by a learnable vector $\mathbf{W}_{a,pos} \in \mathbb{R}^{a'\times d}$ to obtain each token in terms of a n_C dimensional representation. $\lfloor \frac{T}{a'} \rfloor$

positional embeddings are learned to encode the position of these tokens. For the text embedding, we leverage the widely adopted word representation technique [50] and for each component $\mathbf{x}_t$, a vocabulary of v words is created using the words present in the training collection. Each word present in $\mathbf{x}_t$ is mapped to a v-dimensional one-hot vector followed by a linear projection defined via a learnable weight $\mathbf{W}_{t,pos} \in \mathbb{R}^{v \times d}$.

The MultiModal Transformer Layer: For each mode-specific peer transformer branch in a multimodal transformer layer of the proposed *Multimodal Transformer Network*, we use the standard established transformer layer architecture [17] that has the standard self-attention [66] for the Multi-Head Attention (MHA) module, the Gaussian Error Linear Unit (GeLU) [33] activation for the feed-forward part of each layer, along with a pre-normalization [8] before the MHA and Multi-Layer Perceptron (MLP). For more details on the standard transformer architecture, readers are requested to refer [17]. Thus given $m \in \{v, a, t\}$, the sequence of a flattened stream of mode-specific tokens is defined as follows:

$$\mathbf{z}_{in}^{m} = [\mathbf{x}_0^m \mathbf{W}_{m,pos}; \mathbf{x}_1^m \mathbf{W}_{m,pos} .....\mathbf{x}_N^m \mathbf{W}_{m,pos}; \mathbf{x}_{agg}^m]$$

where $\mathbf{x}_n^m$ is the input token sequence. The term $\mathbf{x}_{agg}^m$ presents a learnable embedding for an aggregation token, whose corresponding output $\mathbf{z}_{out}^{m,0}$ defines a mode-specific aggregated output for each mode-specific peer transformer branch and is used for multi-label classification via joint space embeddings.

The peer-ensembled fusion branch of in a multimodal transformer layer adopts an architecture very similar to its mode-specific peer transformer branches, however with a simple yet effective cross-network attention enhanced information integration scheme defined with: the query $Q^f = z_{in}^f W_Q^f$, where $z_{in}^f = [x_{agg}^f]$ with $W_Q^f \in \mathbb{R}^{d \times d}$ is a learnable weight matrix; key $K^f = (\mathbf{z}_{in}^v W_v^k \oplus \mathbf{z}_{in}^a W_a^k \oplus \mathbf{z}_{in}^t W_t^k) W_0^k$, value $V^f = (\mathbf{z}_{in}^v W_v^{val} \oplus \mathbf{z}_{in}^a W_a^{val} \oplus \mathbf{z}_{in}^t W_t^{val}) W_0^{val}$ where $W_m^k, W_m^{val} \in \mathbb{R}^{(N+1)d \times d} \forall m \in \{v, a, t\}$ are learnable weights. The output $\mathbf{z}_{out}^{f,0}$ of the peer-ensembled fusion branch is used for multi-label classification via joint space embeddings A specific attention head of $\mathbf{z}_{out}^{f,0}$ is computed as:

$$\mathbf{z}_{out,[h]}^{f,0} = \text{linear}\Bigg(\text{softmax}(\frac{Q^f {K^f}^T}{\sqrt{d}}) V^f\Bigg)$$

The proposed cross-network attention-enhanced multi-modal fusion mechanism thus allows us to integrate multiple complementary mode-specific information into a single fusion token toward serving two-fold tasks: Preserving multiple co-occurring mode-specific cues that may help design a more robust perception model in a self-supervised setting; Evaluating the pairwise interactions between all spatiotemporal tokens within each mode-specific peer branch network as well as their aggregated views within the fusion branch that may facilitate modeling the long-range contextual relationships in videos, without having to define a computationally expensive hierarchical attention network [7].

3.2 Multi-Label Classification

Important to note that in addition to the external environmental or situational contexts, human expression significantly varies based on sociocultural and demographic specifications. For example, facial muscles undergo atrophy due to age, which restricts one's ability to generate emotional expression [21, 26]. This makes the task of the speaker's emotional evaluation during a conversation furthermore difficult. Therefore, toward enabling the design of a model with robust generalization ability, the proposed *SeMuL-PCD* with the *Multimodal Transformer Network* as its fundamental feature representation module, is trained end-to-end in a self-supervised learning setting. In a joint embedding space, the learning objective is designed as a weighted combination of four loss components: *Noise Contrastive Estimation (NCE) loss* ($\mathcal{L}_{\text{NCE}}$), *Label Consistency Calibration Loss* ($\mathcal{L}_{\text{LCC}}$); *Multimodal Distillation Loss* ($\mathcal{L}_{\text{MDL}}$); and the conventional *Binary Cross Entropy Loss* ($\mathcal{L}_{\text{BCE}}$)[14]. The complete weighted loss for *SeMul-PCD* is defined as:

$$\mathcal{L} = \mathcal{L}_{BCE} + \lambda_C \mathcal{L}_{LCC} + \lambda_K \mathcal{L}_{MDL} + \lambda_S \mathcal{L}_{NCE} \tag{1}$$

Next, we will discuss each of the loss functions separately.

Joint Embedding Space Projection: To eliminate the inherent heterogeneity observed in the semantic granularity of each mode, we adopt the approach by Alayrac et al.[4] and learn multiple mode-specific linear mappings to project $\mathbf{z}_{out}^{m,0}$ onto the single joint embedding space. In particular, we learn the linear projection heads $g_{v->c}(.), g_{a->c}(.), g_{t->c}(.)$, and $g_{f->c}(.)$ to respectively map $\mathbf{z}_{out}^{v,0}$, $\mathbf{z}_{out}^{a,0}$, $\mathbf{z}_{out}^{t,0}$, $\mathbf{z}_{out}^{f,0}$ onto a joint embedding space to directly compare them by using a similarity metric (e.g., cosine similarity).

Noise Contrastive Estimation (NCE) loss: Given a sample $u_j \in \mathcal{D}$, its mode-specific representatives ($\mathbf{z}_{out,j}^{v,0}, \mathbf{z}_{out,j}^{a,0}, \mathbf{z}_{out,j}^{t,0}$) as well as their fused counterpart $\mathbf{z}_{out}^{f,0}$ are aligned pairwise. We note that these pairs are composed by aggregating mode-specific descriptions of different temporal locations of the visual-audio-text stream $u \in \mathcal{D}$, which helps capture the dominant localized patterns and their co-occurrences across modalities. Given $\mathbf{f}_j^m := g_{m->c}(\mathbf{z}_{out,j}^{m,0})$ for $m \in \{v, a, t\}$ and $\mathbf{f}_j^f := g_{f->c}(\mathbf{z}_{out,j}^{f,0})$, in a self-supervised setting, we leverage an aggregated noise contrastive estimation ($\mathcal{L}_{ACE}$), which is as defined below:

$$\mathcal{L}_{\mathcal{ACE}} = \frac{1}{|\mathcal{D}|} \sum_{u_j \in \mathcal{D}} \frac{1}{4} \sum_{\substack{m \neq m_i \\ m, m_i \in \{v,a,t,f\}}} \mathcal{L}_{\text{NCE}}(\mathbf{f}_j^m, \mathbf{f}_j^{m_i}) \tag{2}$$

with

$$\mathcal{L}_{\text{NCE}}(\mathbf{f}_j^m, \mathbf{f}_j^{m_i}) = \Big[-log(\frac{P(\mathbf{f}_j^{m_i}|\mathbf{f}_j^{m_i})}{P(\mathbf{f}_j^{m_i}|\mathbf{f}_j^{m_i}) + \frac{|\mathcal{N}_j|}{|\mathcal{N}|}}) + \sum_{k \in \mathcal{N}_j} log(\frac{P(\mathbf{f}_k^{m_i}|\mathbf{f}_j^{m_i})}{P(\mathbf{f}_k^{m_i}|\mathbf{f}_j^{m_i}) + \frac{|\mathcal{N}_j|}{|\mathcal{N}|}}) - 1\Big]$$

that computes the probability of both features $\mathbf{f}_j^m$ and $\mathbf{f}_j^{m_i}$ representing the same instance u_j compared to other elements in a uniformly sampled negative set $\mathcal{N}_j$ and $\mathcal{N}$ represents the sample batch.

Label Consistency Calibration Loss: In the context of a multi-label classification task, the one-versus-all (OVA) loss [42] is popularly used to optimize the network weights of C independent binary classifiers, where C represents the total number of categories in consideration. However, in our problem context, generations of different emotions may have mutual influences on each other. Therefore, treating each emotion as an independent category by ignoring such influences may force a significant information loss resulting in a negative impact on the model's performance.

Therefore, to capture such difficult multi-label information patterns without having to experience the risk of model overfitting, we aim to propose model calibration for quantifying the model confidence against each prediction on a speaker's emotional state. More formally, if a model is well-calibrated, it should not just make some correct predictions, but also present an appropriate confidence for each prediction it makes. Therefore, the perfect calibration [51] is defined as:

$$\mathbb{P}(y_i = 1 | y_i^p = pr) = pr, \forall pr \in [[0, 1]$$

where y_i^p represents the predicted confidence for the i^{th} label. Intuitively the left-hand side of the above equation can be approximated by the model, the accuracy of which is dependent on the robustness/generalization ability of the model. However, the right-hand side of the equation can be visualized as the corresponding confidence. While due to the difficulty of the multi-label scenarios, the model may not always be expected to produce high confidence in each prediction, the deteriorated performance may be attributed to its miscalibration. Therefore, to enforce accurate model calibration, we propose a simple yet effective *Label Consistency Calibration Loss* defined as:

$$\mathcal{L}_{\text{LCC}} = -\frac{1}{|\mathcal{D}|} \sum_{u_j \in \mathcal{D}} \frac{1}{|C_j^p|} \sum_{c \in C_j^p} \log y_{j,c}^{p,f}$$

where C_j^p represents the set of all positive ground truth labels associated with $u_j \in \mathcal{D}$ and $y_{j,c}^{p,f}$ represents the predicted confidence on the c^{th} label for the sample u_j obtained from the peer-ensembled fusion branch network.

Multimodal Distillation Loss It is important to note that getting an accurate ground truth for difficult labels (e.g. precise confidence score against every emotion state co-occurring in a speaker's expression at a given time stamp) is itself another challenging task, which is not just time-consuming but also extremely expensive when considered for large scale annotation. To address this, we leverage the calibrated outputs via pseudo-supervision following a mode-specific peer knowledge transfer scheme. The proposed *Multimodal Distillation Loss* component leverages each prediction made by a uni-modal peer branch network with that of the peer-ensembled fusion branch network using Kullback–Leibler divergence. In particular, by introducing a target prediction distribution ($\mathbf{T}(.)$), we allow all the independent mode-specific peer branch of the proposed *Multimodal Transformer Network* to distill the uncertainty information with the peer-ensembled fusion branch for pairwise alignments. More formally, with each uni-modal peer branch being independent, we use the chain rule to derive the target prediction distribution as:

$$\mathbf{T}(u_j) = \prod_{m \in \{v,a,t\}} \mathbf{P}^m(u_j)$$

where $\mathbf{P}^m(u_j) := \left[y_{j,1}^{p,m}, ..., y_{j,|C|}^{p,m}\right]$ is the complete multi-label prediction inferred by the m-mode peer branch network. The total *Multimodal Distillation Loss* is then defined as:

$$\mathcal{L}_{\text{MDL}} = \frac{1}{|\mathcal{D}|} \sum_{u_j} \frac{1}{\tau_k |C|} \text{KL}(\mathbf{T}(u_j) || \mathbf{P}^f(u_j))$$

where τ_k is a learnable temperature parameter and $\mathbf{P}^f(u_j)$ is the complete multi-label prediction inferred by the peer-ensembled fusion branch network.

4 EXPERIMENTS

4.1 Datasets Used

As the main focus of this work is to design a model that may evaluate *multiple co-occurring emotion states observed in the utterance of a speaker from a wide age/demographic range*, the proposed *SeMuL-PCD* is evaluated using three publicly available datasets - Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) [76], a Multimodal Dataset for Recognizing Emotional Response in Aging Adults (ElderReact) [46], and a Multimodal Approach and Dataset for Recognizing Emotional Responses in Children (EmoReact) [56]. *Important to note that unlike several state-of-the-art models [35], [16], and [43], which take advantage of the past information on the speaker's evolving emotion patterns (as observed by tracking the sequence of their past utterances) to predict their present emotion state, in this work, we do not have the access to the speakers' past utterance detail. Therefore, we leverage information available only within a single utterance.*

Table 1: Comparison of *SeMuL-PCD* with other models using the weighted averaged F1 measure (wF1) and Accuracy scores on the MOSEI dataset. For category-specific performances please refer to the Appendix in the supplementary material.

Method	Accuracy	wF1
HHMPN (AAAI 2021) [80]	45.9	55.6
DRS2S (ACL 2019) [73]	-	87.90
MMS2S (EMNLP 2020) [78]	47.50	56.0
LMF (ACL 2018) [45]	82.0	82.1
MFM (ICLR 2019) [64]	84.4	84.3
SPC (EMNLP 2021) [15]	82.6	82.8
ICCN (AAAI 2020) [62]	84.2	84.2
MuIT (ACL 2019) [63]	82.5	82.3
MISA (ACM MM 2020) [32]	84.23	83.97
Self-MM (AAAI 2021) [74]	85.17	85.30
MAG-BERT (ACL 2020) [59]	84.70	84.50
MMIM (EMNLP 2021) [27]	85.97	85.94
DialogueCRN [35, 67]	70.1	-
UniMSE (EMNLP 2022) [36]	85.86	85.79
SeMUL-PCD (Text only)	84.49	84.75
SeMUL-PCD (Video only)	72.05	73.92
SeMUL-PCD (Audio only)	71.23	73.17
SeMUL-PCD (Video + Audio)	76.34	77.85
SeMUL-PCD (Text + Audio)	84.92	85.04
SeMUL-PCD (Text + Video)	85.26	86.41
***SeMUL-PCD* (Text+ Audio+ Video)**	**88.62**	**89.04**

The large-scale MOSEI dataset [76] expands its baseline MOSI dataset [75] by including a higher number of utterances, wider variety in samples, speakers, and topics over the MOSI dataset. MOSEI dataset contains $23,453$ annotated video segments (utterances), from $5,000$ videos of more than 65 hours of annotated video, $1,000$ distinct speakers, and 250 different topics. The labels in this dataset comprise six discrete emotions — anger, disgust, fear, happiness, sadness, and surprise. Even though samples in MOSEI have a slight bias towards positive sentiments, combinations of negative emotions like sadness-anger, and surprise-disgust are the most commonly occurring multi-label annotations, with the exception of happiness-surprise.

The EmoReact dataset [56] designs a challenging testbed for recognizing emotional responses in children. It is comprised of $1,102$ videos of 63 children in the age range of $4 - 14$ years. The videos are downloaded from Youtube and present a nearly balanced collection in terms of genders (i.e., 51% of the speakers is female). Each video represents children's reactions in the context of food, and technology. Each video is segmented into approximately 5 seconds clips, such that each clip covers only one child reacting. The children in the videos perform 5 tasks: being shown the context; being asked a question about it; answering a question about it; being told a fact about it; and explaining their opinion about it. The dataset was annotated by crowd-sourcing at Amazon Mechanical Turk (AMT), where the annotators were asked to label each clip independently in the following discrete emotion categories: neutral, disgust, fear, happiness, sadness, surprise, curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, frustration. Multiple emotions may be present in a clip of this dataset. While curiosity (defined as a need or design to learn something [13]) seems to co-occur with surprise, uncertainty, or happiness-like emotions, the correlation between curiosity-fear (or happiness-fear) is very low.

The ElderReact dataset is similar to EmoReact in terms of the way it was collected. There are 43 videos of 46 elderly subjects (46 female and 20 male) reacting to contexts like video games, social events, and online challenges downloaded from YouTube. The original videos were segmented into shorter clips of approximately 3 − 8 seconds in length. The crowd-sourced workers, from Amazon Mechanical Turk, were also used to annotate the discrete emotions: anger; disgust; fear; happiness; sadness; and surprise. Like Emoreact, an important feature of this dataset is that a clip in this collection may be annotated with multiple emotion labels. For example, as reported by the authors, in this dataset, happiness-surprise (or surprise-fear) often co-occur in a speaker's expression. On the other hand, as observed in this dataset, happiness-anger, disgust-fear, and happiness-sadness are very rare and do not seem to occur together.

Important to note that compared to both EmoReact and ElderReact datasets which have presented samples from 15 discrete emotion states, the CMU-MOSEI dataset is significantly larger and presents samples only from 6 emotion categories. Therefore, while results on the CMU-MOSEI dataset demonstrate the feasibility of the proposed *SeMUL-PCD* in a large-scale experimental setting, an equivalent performance on the EmoReact and ElderReact demonstrates the potential of *SeMUL-PCD* in handling difficult multi-label co-occurrence patterns without having to experience the risk of model-overfitting due to limited training set size.

4.2 Results & Comparative Study

Figure 2 shows some qualitative results, wherein Local Interpretable Model-Agnostic Explanations (LIME)[1] are used to explain each system decision. As illustrated in the figure, the proposal *SeMUL-PCD* demonstrates superior performance compared to its uni-modal counterparts. The performance of the proposed *SeMUL-PCD* is compared against competitive baselines in Table 1 and 2. These include: HHMPN [80] which models feature-to-label and modality-to-label dependencies, DRS2S [73] which takes a deep reinforcement learning approach, MMS2S [77] that uses the sequence-to-set approach of DRS2S to model modality-label depndencies, Low-rank Multimodal Fusion LMF [45]; Multimodal Factorization Model(MFM) [64]; sparse phased Transformer (SPC) [15]; interaction canonical correlation network (ICCN) [62]; multimodal Transformer (MulT) [63]; Modality-Invariant and -Specific representations (MISA) [32]; self-supervised multi-task learning for multimodal sentiment analysis (Self-MM) [74], MAG-BERT that integrates multimodal information in large pretrained transformers [59]; MMIM that hierarchically maximizes the mutual information [28]; UniMSE that Unifies Multimodal Sentiment Analysis and Emotion Recognition [36]; Facial Transformer Plugin (FACE-STN) [10]; DialogueCRN [35] that leverages historical information from an utterance sequence to gauge a speaker's expression, Support Vector Machine (SVM) [56]; Contrastive Adaptation Mechanism for Non-Universal Facial Expression Recognition (CIAO)[11] ; AlexNet EmoDB [54]. As observed in Table 1, the proposed *SeMUL-PCD* significantly outperforms (with around 3 − 4% improvement in Acc-2 and wF1) the best performing recent baseline UniMSE that fuses mode-specific information at the syntactic and semantic levels via contrastive learning to capture the differences between emotions. Having been able to capture the mode-specific complementary performances shared via a set of branch transformer networks in a collaborative learning environment, *SeMUL-PCD* ensures to maximize complementary and discriminative cues that can be leveraged to optimize the performance. By means of the proposed *Multimodal Distillation Loss* the *Multimodal Transformer Network* at the core of *SeMUL-PCD* facilitates the calibration of the fusion branch network toward that of multiple mode-specific branch networks by minimizing their total pairwise Kullback–Leibler divergences. This helps the model learn the nuances of cross-emotion correlations and thereby improves performance. As also reported in the table, while 'Text' appears to be the most reliable uni-modal feature, combining information from multiple modes is always helpful. In Table2, *SeMUL-PCD* demonstrates a similarly dominating performance against the existing literature in more recent datasets EmoReact and ElderReact, which not only justifies the model's effectiveness across a large range of age-demographic backgrounds but also exhibits its capacity in handling difficult multi-label scenarios, where more than one emotion state may be visible in a speaker's expression. In fact, an improvement of around 5% (6%) in EmoReact (and ElderReact) in the weighted averaged F1 score (wF1) clearly reveals the model potential for such challenging problem scenarios.

Table 2: Comparison of *SeMuL-PCD* with other models on EmoReact and ElderReact using wF1.

Method	**EmoReact**	**ElderReact**
SVM [56]	66.1	45.8
FaceSTN (IEEEAccess 2022) [10]	74.3	-
CIAO (ACII, 2022) [11]	76.0	-
AlexNet EmoDB (TENSYMP, 2019) [54]	86.0	-
Jannat et al. [39]	81.0	86.0
Hetterscheid et al [34]	-	86.0
SeMUL-PCD (Video only)	87.24	87.41
SeMUL-PCD (Audio only)	86.03	85.29
SeMUL-PCD **(Video + Audio)**	**91.24**	**92.56**

4.3 Cross Dataset Generalization Results

An important robustness aspect of an emotion evaluation system is defined by its cross-dataset generalization capacity, which is specifically to demonstrate the model's ability to estimate emotion across various demographic populations. For example, using ElderReact as the test collection in evaluating the performance of *SeMUL-PCD*, which has been trained on a set like EmoReact, would be important to understand the impact of age differences on the proposed computational model. *While only a limited existing works address this Cross Dataset Genaralization* task, following Kaixin et al. [46], we train the model in a given age population and test it on a completely different age population. Table 4 illustrates how SeMUL-PCD consistently outperforms other emotion recognition models by achieving improved results on 3 large public datasets while generalizing across multiple datasets that represent populations of different age and demographic backgrounds. For example, as observed the Table 4, Row 1 reports the performance, when *SeMUL-PCD* was trained using EmoReact that is specifically designed for recognizing the children's emotion and was tested using the ElderReact Dataset that is exclusively designed for recognizing elder emotion. We have used the publicly available codes of DialogueCRN [35] and MMIM [27] to perform the experiments, reported in the table. As reported in the table, while DialogueCRN [35] reports reasonable performance in evaluating emotions for a population demographically similar to what was seen during its training, in a cross-dataset experimental setting the performance deteriorates significantly. A similar performance pattern is also observed for MMIM [27]. In particular, across a variety of configurations adapted for the Training ($\mathcal{D}_{train}$) and Testing ($\mathcal{D}_{test}$) set pairs, the proposed *SeMUL-PCD* consistently report a significantly more robust and improved performance compared to the recent baselines and presents an average of 15 − 17% gain in the wF1 score. Using a weighted loss function that combines the complementary insights from multiple loss components during weight updates of the proposed *Multimodal Transformer Network*, *SeMUL-PCD* has evidently improved the precision performance in handling difficult multi-label scenarios as well as cross-demographic generalization. By enabling a self-supervised

[1] https://github.com/marcotcr/lime

contrastive learning scheme that allows each branch to independently learn their mode-specific discriminative patterns while parallelly sharing their respective insights to a single peer-ensembled fusion branch in a collaborative learning setting, *SeMUL-PCD* delivers a model that is simultaneously effective, efficient, and scalable across multiple data environments.

4.4 Ablation Study

Table 5 presents the results on all three datasets when the relative contributions of each loss are changed by means of different choices of values for the parameters λ_C, λ_K, and λ_S in Eqn 1. Especially of interest are rows 1 - 3 in the table, where we only include only one of the three additional losses ($\mathcal{L}_{\text{NCE}}$, $\mathcal{L}_{\text{LCC}}$, $\mathcal{L}_{\text{MDL}}$) along with $\mathcal{L}_{BCE}$. Comparing these three rows, we conclude that the proposed *Label Consistency Calibration Loss* ($\mathcal{L}_{LCC}$) is essential in improving multi-label classification performance. As the row 3, which shows the result of combining $\mathcal{L}_{\text{LCC}}$ with the conventional cross-entropy loss $\mathcal{L}_{BCE}$, reports an improvement of around $7-8\%$ in wF1 score compared to the performances reported in row-1 and row-2. Note that in row-1 and row-2 we select only $\mathcal{L}_{\text{MDL}}$ and $\mathcal{L}_{\text{NCE}}$ respectively to combine with $\mathcal{L}_{BCE}$. Looking at Row 4, it is also apparent that our *Multimodal Distillation Loss* ($\mathcal{L}_{MDL}$) improves performance, especially on EmoReact and ElderReact, which has only limited samples to represent each emotion category. As we note that aligning the fusion probability predictions to mode-specific predictions allows the model to be much more resilient to outliers and converge faster. Various combinations of values of λ_C, λ_K, and λ_S in Eqn 1 are used to analyze the performance. For example, comparing row 6 and row 13, we find that introduction of $\mathcal{L}_{MKL}$ helps the model report an improvement of around 4% in MOSEI and around 9% in smaller datasets like EmoReact and ElderReact. Overall, the performance remains nearly stable across similar value choices, the best performance was attained using λ_C=1.0, λ_K=0.5, and λ_S=1.0.

As described in Section 3.1, we adopt a tubelet-based tokenization (introduced by Anurag et al. [6]) for a video, so that each video component can be represented by a sequence of tubelets. The table 6 compares the performance of the proposed *SeMUL-PCD* as the underlying tokenization scheme is altered to another tokenization scheme *Uniform frame sampling* [6], in which each individual frame from an input video is embedded independently and later concatenated to preserve the sequence information. We note that the tubelet tokenization scheme that extracts non-overlapping, spatio-temporal "tubes" from the input video, is more effective than the *Uniform frame sampling*.

In Table 3 we explore the contribution of individual losses to this robustness of *SeMUL-PCD* across multiple datasets. As seen in Table 3, most of the robustness of *SeMUL-PCD* compared to other models derives from its novel use of *Noise Contrastive Estimation* as it performs consistently better when $\lambda_S = 1, \lambda_K = 0, \lambda_C = 0$ (the last column of Table 3). Table 3 also allows us to explore the contrast in the contribution of losses from Table 5 which reports performance when the model is trained and tested on the same dataset. Unlike in Table 5, where $\mathcal{L}_{LCC}$ contributes the most to performance, here the robustness is acquired from $\mathcal{L}_{NCE}$

Table 3: Cross dataset generalization performance contribution of each loss, evaluated by weighted average F1 (wF1), and setting the weight for the loss indicated by that column to 1 and other weights to 0 (λ_C - weight of $\mathcal{L}_{LCC}$, λ_K - weight of $\mathcal{L}_{MDL}$, and λ_S - weight of $\mathcal{L}_{NCE}$)

$\mathcal{D}_{train}$	$\mathcal{D}_{test}$	λ_C	λ_K	λ_S
EmoReact	ElderReact	0.75	0.73	0.85
ElderReact	EmoReact	0.73	0.72	0.8
ElderReact	MOSEI	0.67	0.61	0.71
MOSEI	ElderReact	0.76	0.74	0.83
MOSEI	EmoReact	0.71	0.69	0.81
EmoReact	MOSEI	0.63	0.61	0.76

Table 4: The Cross dataset Generalization Performance of *SeMUL-PCD* different Training ($\mathcal{D}_{train}$) and Testing ($\mathcal{D}_{test}$) set pairs. The performance is reported using Weighted Average F1 (wF1) as the evaluation metric and it is compared against the following State of the Art models, for which either the results or the codes were available.

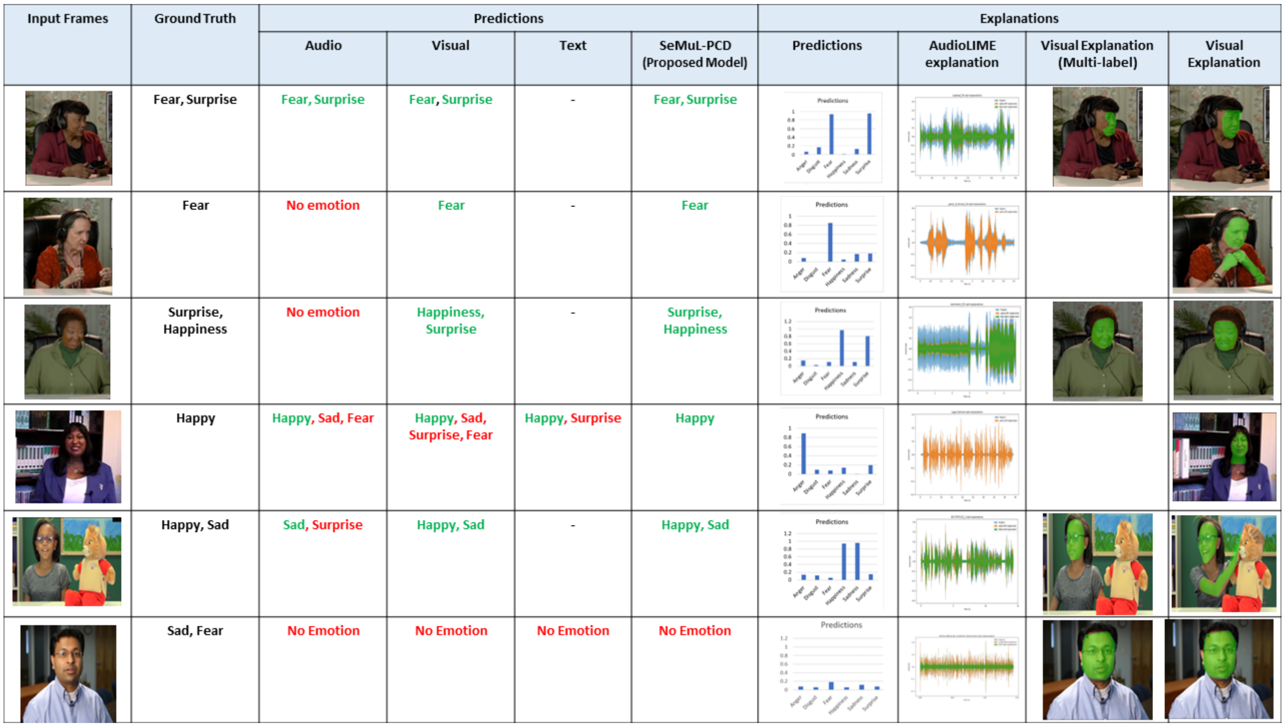
Method	Train-Test Configurations	wF1
DialogueCRN [35]	$\mathcal{D}_{train}$ = *EmoReact*, $\mathcal{D}_{test}$ = *ElderReact*	0.64
	$\mathcal{D}_{train}$ = *ElderReact*, $\mathcal{D}_{test}$ = *EmoReact*	0.56
	$\mathcal{D}_{train}$ = *ElderReact*, $\mathcal{D}_{test}$ = *MOSEI*	0.59
	$\mathcal{D}_{train}$ = *MOSEI*, $\mathcal{D}_{test}$ = *ElderReact*	0.60
	$\mathcal{D}_{train}$ = *MOSEI*, $\mathcal{D}_{test}$ = *EmoReact*	0.74
	$\mathcal{D}_{train}$ = *EmoReact*, $\mathcal{D}_{test}$ = *MOSEI*	0.61
MMIM [27]	$\mathcal{D}_{train}$ = *EmoReact*, $\mathcal{D}_{test}$ = *ElderReact*	0.74
	$\mathcal{D}_{train}$ = *ElderReact*, $\mathcal{D}_{test}$ = *EmoReact*	0.73
	$\mathcal{D}_{train}$ = *ElderReact*, $\mathcal{D}_{test}$ = *MOSEI*	0.65
	$\mathcal{D}_{train}$ = *MOSEI*, $\mathcal{D}_{test}$ = *ElderReact*	0.75
	$\mathcal{D}_{train}$ = *MOSEI*, $\mathcal{D}_{test}$ = *EmoReact*	0.71
	$\mathcal{D}_{train}$ = *EmoReact*, $\mathcal{D}_{test}$ = *MOSEI*	0.64
RBF-SVM [46]	$\mathcal{D}_{train}$ = *EmoReact*, $\mathcal{D}_{test}$ = *ElderReact*	0.27
	$\mathcal{D}_{train}$ = *ElderReact*, $\mathcal{D}_{test}$ = *EmoReact*	0.33
SeMUL-PCD	$\mathcal{D}_{train}$ = *EmoReact*, $\mathcal{D}_{test}$ = *ElderReact*	**0.88**
	$\mathcal{D}_{train}$ = *ElderReact*, $\mathcal{D}_{test}$ = *EmoReact*	**0.83**
	$\mathcal{D}_{train}$ = *ElderReact*, $\mathcal{D}_{test}$ = *MOSEI*	**0.76**
	$\mathcal{D}_{train}$ = *MOSEI*, $\mathcal{D}_{test}$ = *ElderReact*	**0.89**
	$\mathcal{D}_{train}$ = *MOSEI*, $\mathcal{D}_{test}$ = *EmoReact*	**0.87**
	$\mathcal{D}_{train}$ = *EmoReact*, $\mathcal{D}_{test}$ = *MOSEI*	**0.79**

4.5 Implementation Details

For all three datasets, for video, we sample 32 input frames at 10 fps resized to a size of 640x480. We do a random resize crop of these frames to 224x224, followed a by random horizontal flip with a probability of 0.5 and color augmentation. We sample single-channel mono audio in sync with the video frames at 48kHz. Both audio and video are normalized between [-1, 1]. To generate the patches we use a patch size of 4x16x16 for video and 128 for audio. For text, we cap the sentence size to 48 words and use a maximum dictionary size of 2^{10} words. For the model architecture, we use d = 2048, with 25 Transformer layers in each mode-specific transformer branch and in the peer-ensembled fusion branch. Each of these layers uses 16 attention heads with an internal representation size f = 4096. All transformer encoder layers and double linear layers use the GeLU activation function. We choose e = 1024 for the common space projection before NCE. Both temperature parameters τ_k and τ_s are initialized to 0.08 and are trainable. The weights used to balance the losses are λ_C = 1.0, λ_K = 0.5 and λ_S = 1.0. The model was trained using the Adam optimizer with half-period cosine annealing for learning rate scheduling ranging from 1e-4 to 1e-5 for 750 epochs with 5000 warmup steps. In experiments, we use 0.5 as a confidence threshold to predict the occurrence of an emotion in an utterance. We choose all 6 labels from MOSEI and ElderReact and exclude the 9 'complex' emotions from the EmoReact dataset since the 6 basic emotions are common to all three datasets. For each cross-testing experiment, we use the same training and validation splits from $\mathcal{D}_{train}$ and the test split from $\mathcal{D}_{test}$ used in experiments in Tables 1 and 2 to provide meaningful comparison. These tests prove the efficacy of the self-supervised features learned by our framework and its generalization capacity across age variations is investigated via experiments reported in Table 4.

5 CONCLUSION

We present *SeMUL-PCD* with a *Multimodal Transformer Network*, which enables effective knowledge sharing from multiple mode-specific peer networks into a single mode-ensembled fusion branch network to facilitate



Input Frames	Ground Truth	Predictions				Explanations			
		Audio	Visual	Text	SeMuL-PCD (Proposed Model)	Predictions	AudioLIME explanation	Visual Explanation (Multi-label)	Visual Explanation
	Fear, Surprise	Fear, Surprise	Fear, Surprise	-	Fear, Surprise				
	Fear	No emotion	Fear	-	Fear				
	Surprise, Happiness	No emotion	Happiness, Surprise	-	Surprise, Happiness				
	Happy	Happy, Sad, Fear	Happy, Sad, Surprise, Fear	Happy, Surprise	Happy				
	Happy, Sad	Sad, Surprise	Happy, Sad	-	Happy, Sad				
	Sad, Fear	No Emotion	No Emotion	No Emotion	No Emotion				

Figure 2: Example results to explain the performance of *SeMUL-PCD* in comparison with Uni-modal counterparts with users from diverse demographic backgrounds. We provide mode-specific explanations for audio, video, and text. Labels highlighted in Green indicate correct predictions, while labels highlighted in Red indicate incorrect ones. For audio explanations, we use audioLIME [30] that is based on Local Interpretable Model-agnostic Explanations (LIME) [60] extended by a musical definition of locality. The original audio waveplot is indicated in Blue. We indicate features that contribute towards the prediction for uni-label instances with Orange color, while multi-label instances are explained using a different color (Orange and Green in the above examples) for each label. For visual explanations, we use LIME[60] to explain regions (indicated in Green) to indicate features contributing towards the model predictions.

Table 5: Ablation study on different loss weights (wF1) scores for different values of λ_C (weight of $\mathcal{L}_{LCC}$), λ_K ($\mathcal{L}_{MDL}$), and weight of λ_S (weight of $\mathcal{L}_{NCE}$) in Eqn 1

λ_C	λ_K	λ_S	MOSEI	EmoReact	ElderReact
0.0	0.0	1.0	78.48	80.32	81.26
0.0	1.0	0.0	76.91	80.45	80.21
1.0	0.0	0.0	85.28	87.31	86.06
1.0	1.0	0.0	86.59	89.15	88.21
0.0	1.0	1.0	80.25	80.93	81.39
1.0	0.0	1.0	85.97	82.45	83.18
0.5	0.5	0.5	88.31	89.88	90.36
1.0	1.0	1.0	88.85	90.08	90.59
1.0	0.5	0.5	88.06	90.98	91.26
0.5	1.0	0.5	88.90	90.29	91.84
1.0	1.0	0.5	88.97	91.02	91.78
0.5	1.0	1.0	88.07	90.68	91.83
1.0	0.5	1.0	**89.04**	**91.24**	**92.56**

robust decision-making. In a self-supervised collaborative learning setting the proposed model is trained via an efficient cross-network attention fusion mechanism to ensure an impressive performance that can generalize across diverse user groups. We propose *Multimodal Distillation Loss* for calibrating the fusion branch network to better handle the difficult emotions in the multi-label emotion categorization task. As evident from the experiments, *SeMUL-PCD* trained with a weighted loss function that includes *Multimodal Distillation Loss* component, generalizes better than its baseline versions in handling different difficulty levels. Furthermore, the *Label Consistency Calibration Loss* can also be used to guide the network to produce an appropriate confidence score for each prediction.

Table 6: Performance Comparison for the proposed *SeMUL-PCD* under two different video tokenization schemes: *Uniform frame sampling* and *Spatio-Temporal Tubelet*

Method	MOSEI	EmoReact	ElderReact
Uniform frame sampling	87.92	90.21	89.98
Spatio-Temporal Tubelet	89.04	91.24	92.56

6 ACKNOWLEDGEMENT

We sincerely think Dr. Yu-Ping Chang[2], Associate Dean for Research and Professor, The School of Nursing, The State University of New York at Buffalo for sharing her insightful thoughts on the datasets, identified emotion categories, and the significance of their co-occurrences in a human's expression.

[2] https://nursing.buffalo.edu/about-us/ub-son-pride.host.html/content/shared/nursing/modules/our-students--faculty-and-alumni/alumni-students-faculty/yu-ping-chang.detail.html

REFERENCES

[1] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review* (2021), 1–41.

[2] Harsh Agarwal, Keshav Bansal, Abhinav Joshi, and Ashutosh Modi. 2021. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. *arXiv preprint arXiv:2112.01938* (2021).

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178* (2021).

[4] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems* 33 (2020), 25–37.

[5] Gerard F Anderson and Peter Sotir Hussey. 2000. Population Aging: A Comparison Among Industrialized Countries: Populations around the world are growing older, but the trends are not cause for despair. *Health affairs* 19, 3 (2000), 191–203.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.

[7] Balaji Arumugam, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2022. Multimodal Attentive Learning for Real-time Explainable Emotion Recognition in Conversations. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1210–1214.

[8] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[10] Pablo Barros and Alessandra Sciutti. 2022. Across the Universe: Biasing Facial Representations Toward Non-Universal Emotions With the Face-STN. *IEEE Access* 10 (2022), 103932–103947.

[11] Pablo Barros and Alessandra Sciutti. 2022. Ciao! a contrastive adaptation mechanism for non-universal facial expression recognition. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[12] Charles E. Hughes Louis-Philippe Morency Behnaz Nojavanasghari, Tadas Baltrusaitis. 2016. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. *International Conference on Multimodal Interfaces(ICMI)* (2016).

[13] Daniel E Berlyne. 1960. Conflict, arousal, and curiosity. (1960).

[14] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.

[15] Junyan Cheng, Iordanis Fostiropoulos, Barry Boehm, and Mohammad Soleymani. 2021. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2447–2458.

[16] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4652–4661.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. , 4171–4186 pages. https://doi.org/10.18653/v1/n19-1423

[18] Sarah E Donohue, Lawrence G Appelbaum, Christina J Park, Kenneth C Roberts, and Marty G Woldorff. 2013. Cross-modal stimulus conflict: the behavioral effects of stimulus input timing in a visual-auditory Stroop task. *PloS one* 8, 4 (2013), e62802.

[19] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2021. Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness. *Scientific reports* 11, 1 (2021), 1–11.

[20] Mara Fölster, Ursula Hess, and Katja Werheid. 2014. Facial age affects emotional expression decoding. *Frontiers in psychology* 5 (2014), 30.

[21] Maxi Freudenberg, Reginald B Adams Jr, Robert E Kleck, and Ursula Hess. 2015. Through a glass darkly: facial wrinkles affect our processing of emotion in the elderly. *Frontiers in psychology* 6 (2015), 1476.

[22] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. arXiv:2010.02795 [cs.CL]

[23] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540* (2019).

[24] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[25] Sarah A Grainger, Julie D Henry, Louise H Phillips, Eric J Vanman, and Roy Allen. 2017. Age deficits in facial affect recognition: The influence of dynamic cues. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 72, 4 (2017), 622–632.

[26] Sabrina N Grondhuis, Angela Jimmy, Carolina Teague, and Nicolas M Brunet. 2021. Having difficulties reading the facial expression of older individuals? Blame it on the facial muscles, not the wrinkles. *Frontiers in Psychology* 12 (2021), 620768.

[27] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9180–9192. https://doi.org/10.18653/v1/2021.emnlp-main.723

[28] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412* (2021).

[29] Susan Harter and Nancy Rumbaugh Whitesell. 1989. Developmental changes in children's understanding of single, multiple, and blended emotion concepts. (1989).

[30] Verena Haunschmid, Ethan Manilow, and Gerhard Widmer. 2020. audioLIME: Listenable Explanations Using Source Separation. *CoRR* abs/2008.00582 (2020). arXiv:2008.00582 https://arxiv.org/abs/2008.00582

[31] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2594–2604.

[32] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1122–1131.

[33] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[34] K.J.T. Hetterscheid. 2020. Detecting agitated speech : A neural network approach. http://essay.utwente.nl/82014/

[35] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978* (2021).

[36] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. *arXiv preprint arXiv:2211.11256* (2022).

[37] Isabelle Hupont, Songül Tolan, Pedro Frau, Lorenzo Porcaro, and Emilia Gómez. 2023. Measuring and fostering diversity in Affective Computing research. *IEEE Transactions on Affective Computing* (2023).

[38] Sk Rahatul Jannat and Shaun Canavan. 2021. Expression Recognition Across Age. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–5.

[39] Sk Rahatul Jannat and Shaun Canavan. 2021. Expression Recognition Across Age. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. 1–5. https://doi.org/10.1109/FG52635.2021.9667062

[40] Sébastien Lallé, Rohit Murali, Cristina Conati, and Roger Azevedo. 2021. Predicting co-occurring emotions from eye-tracking and interaction data in MetaTutor. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I*. Springer, 241–254.

[41] Jeff T Larsen and A Peter McGraw. 2014. The case for mixed emotions. *Social and Personality Psychology Compass* 8, 6 (2014), 263–274.

[42] Michael Leben. 2012. *Email Classification with Contextual Information*. Ph. D. Dissertation. Hasso-Plattner-Institute.

[43] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. *arXiv preprint arXiv:2203.13504* (2022).

[44] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 985–1000.

[45] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064* (2018).

[46] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M Girard, and Louis-Philippe Morency. 2019. ElderReact: a multimodal dataset for recognizing emotional response in aging adults. In *2019 international conference on multimodal interaction*. 349–357.

[47] Carol Magai, Nathan S Consedine, Yulia S Krivoshekova, Elizabeth Kudadjie-Gyamfi, and Renee McPherson. 2006. Emotion experience and expression across the adult life span: insights from a multimodal assessment study. *Psychology and aging* 21, 2 (2006), 303.

[48] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems* 161 (2018), 124–133.

[49] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. arXiv:1811.00405 [cs.CL]

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[51] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 15682–15694.

[52] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 1359–1367.

[53] Rohit Murali, Cristina Conati, and Roger Azevedo. 2023. Predicting Co-occurring Emotions in MetaTutor when Combining Eye-Tracking and Interaction Data from Separate User Studies. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 388–398.

[54] Bhalaji Nagarajan and V Ramana Murthy Oruganti. 2019. Cross-domain transfer learning for complex emotion recognition. In *2019 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 649–653.

[55] Shini Girija Naveed Ahmed, Zaher Al Aghbari. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications* (2023).

[56] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*. 137–144.

[57] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep Multimodal Fusion for Persuasiveness Prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 284–288. https://doi.org/10.1145/2993148.2993176

[58] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 439–448.

[59] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2020. NIH Public Access, 2359.

[60] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). arXiv:1602.04938 http://arxiv.org/abs/1602.04938

[61] Piao Shi, Min Hu, Fuji Ren, Xuefeng Shi, and Liangfeng Xu. 2022. Learning modality-fused representation based on transformer for emotion analysis. *Journal of Electronic Imaging* 31, 6 (2022), 063032.

[62] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8992–8999.

[63] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[64] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).

[65] Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022. Sentiment-Emotion-and Context-guided Knowledge Selection Framework for Emotion Recognition in Conversations. *IEEE Transactions on Affective Computing* (2022).

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[67] Jingyao Wang, Luntian Mou, Lei Ma, Tiejun Huang, and Wen Gao. 2023. AMSA: Adaptive Multimodal Learning for Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3s (2023), 1–21.

[68] KX Wang, QL Zhang, and SY Liao. 2014. A database of elderly emotional speech. In *Proc. Int. Symp. Signal Process. Biomed. Eng Informat.* 549–553.

[69] Kunxia Wang, ZongBao Zhu, Shidong Wang, Xiao Sun, and Lian Li. 2016. A database for emotional interactions of the elderly. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. IEEE, 1–6.

[70] Chung-Hsien Wu and Wei-Bin Liang. 2011. Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels. *IEEE Transactions on Affective Computing* 2, 1 (2011), 10–21. https://doi.org/10.1109/T-AFFC.2010.16

[71] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 466–475.

[72] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 506–523.

[73] Pengcheng Yang, Fuli Luo, Shuming Ma, Junyang Lin, and Xu Sun. 2019. A Deep Reinforced Sequence-to-Set Model for Multi-Label Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5252–5258. https://doi.org/10.18653/v1/P19-1518

[74] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10790–10797.

[75] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[76] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.

[77] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. 3584–3593.

[78] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal Multi-label Emotion Detection with Modality and Label Dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3584–3593. https://doi.org/10.18653/v1/2020.emnlp-main.291

[79] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14338–14346.

[80] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 16 (May 2021), 14338–14346. https://doi.org/10.1609/aaai.v35i16.17686

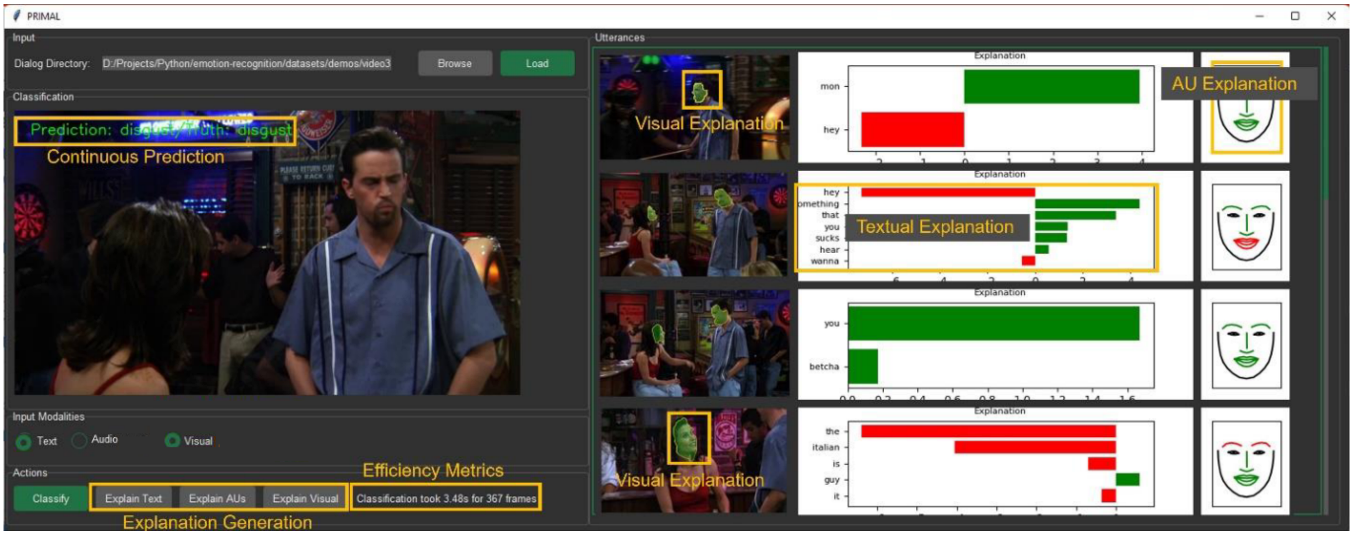
A USER INTERFACE TO EXPLAIN MULTIMODAL FEATURES

Figure 3 shows a screenshot of the User Interface (UI) for *SeMUL-PCD* that leverages text and visual modes as user-selected modes to predict speaker emotion in a video. While some qualitative results are shown in Figure 2 of the main submission, wherein the test samples from the *MOSEI*, *EmoReact* and *ElderReact* datasets were used to illustrate the results, to evaluate the generalized performance of the model, we also use random videos downloaded from YouTube. The figure shows one such result and it is noteworthy to mention that *SeMUL-PCD* produces a semantically meaningful prediction. Moreover, as seen in the screenshot, the user is allowed to choose different modes to obtain the mode-specific prediction and their corresponding mode-specific explainability. The UI also allows users to choose different combinations of modes (i.e. text-image, text-audio, text-audio, visual) to obtain system prediction with respect to their chosen modes.

B AN ABLATION STUDY ON CLASS-WISE, MODALITY-SPECIFIC PERFORMANCE

The existing literature mostly just reports the overall performances. The overall performances using *MOSEI*, *EmoReact*, and *ElderReact* datasets were reported in Table 1 and Table 2. However, in the Tables 7, 8 and 9 below, we also present the class-specific classification accuracy of various configurations of *SeMUL-PCD* on the *MOSEI*, *EmoReact* and *ElderReact* datasets respectively for each emotion present in the datasets. As observed, *SeMUL-PCD* produces remarkably good results, especially on the minority classes of all three datasets - *Fear* and *Disgust*, two emotion categories which have often been treated as difficult due to their limited representative samples.

Figure 3: Screenshot of the User Interface (UI) for *SeMUL-PCD* with the description of its different segments in annotations, where two modes (text, visual) are selected for enabling a multimodal analysis.



PRIMAL

Input

Dialog Directory: D:/Projects/Python/emotion-recognition/datasets/demos/video3

Classification

Prediction: disgust/Truth: disgust

Input Modalities

● Text ○ Audio ● Visual

Actions

Classification took 3.48s for 367 frames

Table 7: Classwise results on *MOSEI* evaluated by classification accuracy (*Acc*) for each combination of available modalities

	MOSEI					
Modality	**Happiness**	**Sadness**	**Anger**	**Surprise**	**Disgust**	**Fear**
(Text only)	92.35	84.16	85.29	79.01	67.32	64.76
(Video only)	81.26	71.28	70.91	68.08	59.61	57.41
(Audio only)	80.30	69.28	68.43	66.31	57.61	53.41
(Video + Audio)	88.26	74.26	73.79	72.39	63.19	59.83
(Audio + Text)	93.47	84.16	85.29	80.47	67.32	64.76
(Text + Video)	94.01	85.28	85.29	80.47	69.48	68.74
All	95.97	88.85	88.01	84.23	72.45	70.25

Table 8: Classwise results on *EmoReact* evaluated by classification accuracy (*Acc*) for each combination of available modalities

	EmoReact					
Modality	**Happiness**	**Sadness**	**Surprise**	**Fear**	**Disgust**	**Anger**
Video	93.57	87.39	86.12	69.16	70.37	84.29
Audio	92.58	85.14	86.12	67.41	65.48	83.90
All	98.73	90.57	91.82	70.17	70.98	87.28

Table 9: Classwise results on *ElderReact* evaluated by classification accuracy (*Acc*) for each combination of available modalities

	ElderReact					
Modality	**Anger**	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Video	85.21	73.19	70.51	94.16	84.19	80.57
Audio	84.16	70.88	67.82	92.45	84.19	78.85
All	90.75	73.97	71.69	98.47	91.65	86.20