AMuSE: *A*daptive *Mu*ltimodal Analysis for *S*peaker *E*motion Recognition in Group Conversations

Sidharth Anand, Naresh Kumar Devulapally, Sreyasee Das Bhattacharjee, Junsong Yuan, Yu-Ping Chang
The State University of New York at Buffalo.

***Abstract*—Analyzing individual emotions during group conversation is crucial in developing intelligent agents capable of natural human-machine interaction. While reliable emotion recognition techniques depend on different modalities (text, audio, video), the inherent heterogeneity between these modalities and the dynamic cross-modal interactions influenced by an individual's unique behvioral patterns make the task of emotion recognition very challenging. This difficulty is compounded in group settings, where the emotion and its temporal evolution are not only influenced by the individual but also by the external contexts like audience reaction and context of the ongoing conversation. Towards this, we propose a *Multimodal Attention Network* (*MAN*) that captures cross-modal interactions at various levels of spatial abstraction by jointly learning its interactive bunch of mode-specific *Peripheral* and *Central* networks. The proposed *MAN* "injects" cross-modal attention via its *Peripheral* key-value pairs within each layer of a mode-specific *Central* query network. The resulting cross-attended mode-specific descriptors are then combined using an *Adaptive Fusion* (*AF*) technique that enables the model to integrate the discriminative and complementary mode-specific data patterns within an instance-specific multimodal descriptor. Given a dialogue represented using a sequence of utterances, the proposed *AMuSE* (*A*daptive *Mu*ltimodal Analysis for *S*peaker *E*motion) model condenses both spatial (within-mode and within-utterance) and temporal (across-mode and across-utterances in the sequence) features into two dense descriptors: speaker-level and utterance-level. This helps not only in delivering better classification performance ($3-5\%$ improvement in Weighted-F1 and $5-7\%$ improvement in Accuracy) in large-scale public datasets (MELD and IEMOCAP) but also helps the users in understanding the reasoning behind each emotion prediction made by the model via its *Multimodal Explainability Visualization* module.**

***Index Terms*—Artifical Intelligence, Supervised Learning, Emotion Recognition**

I. Introduction

Understanding the emotional nuances within conversations has become a pivotal task, with applications ranging from sentiment analysis in social media to affect-aware human-robot interactions. The complexities inherent in multi-party conversations, involving interactions between multiple speakers via various modalities like text, video, and audio, pose significant challenges for accurate emotion recognition. Extracting the subtle interplay of emotions from these diverse modalities necessitates a comprehensive approach that can effectively capture the spatio-temporal evolution of heterogegeneous co-occurring mode-specific patterns and their mutual interactions across various modes, while also taking into consideration of the unique and dynamic nature of the speaker's expressions.

In fact, important to note that the affective state of each individual evolves continuously during conversation. Such transitions depend on various intra (e.g. personal background, unique behavioral traits, and habits) and inter (e.g. audience behavior, their interpreting conducts) personal contexts as well as other environmental circumstances. While each speaker's visual and auditory cues offer valuable insights into their emotional states, the relationships between these modalities can be subtle and context-dependent. Additionally, the importance of all modes is not always constant and may vary in an instance specific manner. For example, for a speaker, facial expression may be less explicitly representative of their emotion and its temporal evolution compared to their acoustic signal. In another scenario, a speaker may have restrictions in expressing their emotion in a given conversational setting like public gathering. In such cases, surrounding contexts may provide more important cues to facilitate accurate inference of the speaker's emotion at that setting. Thus the challenges related to *the presence of strong heterogeneity among such cross-modal representations* and *the influence of an individual's intra-personal and other situational contexts toward learning a variety of the cross-modal interaction patterns* are less studied for this task.

To address these challenges, we propose a two-level information integration technique. First, a Multimodal Attention Network (*MAN*) is trained to bridge the heterogeneity gaps among the mode-specific representations by capturing the cross-modal interactions at various levels of spatial abstraction. *MAN* incorporates an interactive bunch of mode-specific *Peripheral* and *Central* networks, where the utterance-level mode-specific emotion patterns at every layer of a *Central* network is attended by injected feedback from the *Peripheral* networks. This enables each mode-specific *Central* network to prioritize on the mode-invariant spatial (within utterance) details of the emotion patterns, while also retaining its mode-exclusive aspects within the learned model. Intuitively, this is necessary because while representing a specific emotion state, various modalities like text, audio, and video exhibit correlations that may be apparent at different levels of abstraction. While these correlating patterns are important, certain mode-exclusive cues (e.g., non-verbal response of audience) may also convey some useful insights about the speaker's evolving emotion state at the next time-stamp. Second, an Adaptive Fusion (*AF*) technique is employed, recognizing that not all modalities contribute equally to the process of emotion recognition for every query instance, the resulting

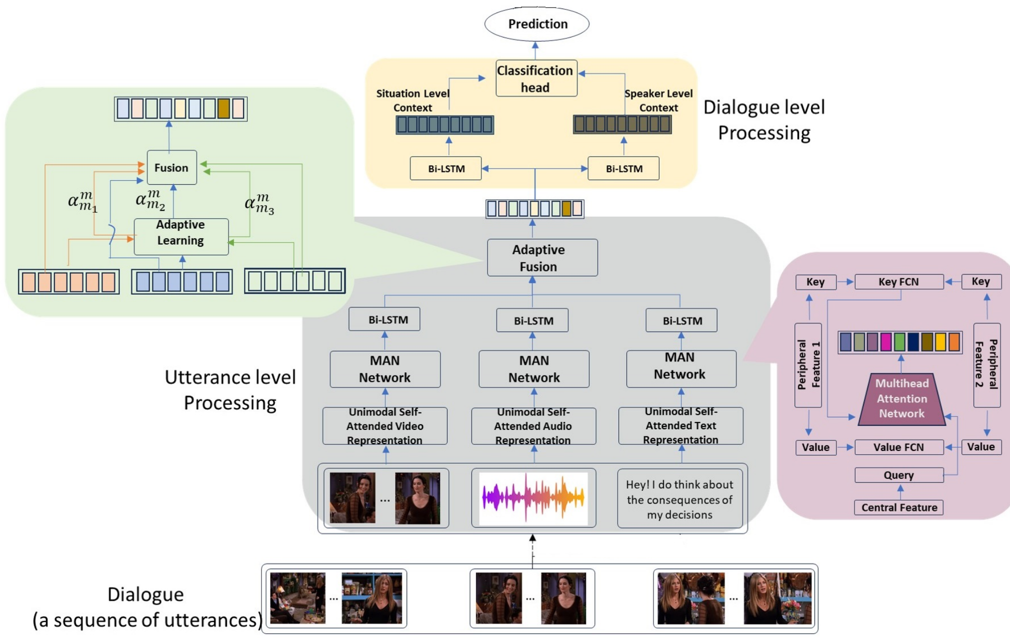


Fig. 1. Proposed *AMuSE* Architecture that captures cross-modal interactions and their spatio-temporal evolution to predict the speaker's emotion in a conversation.

cross-attended uni-modal feature descriptors derived from the mode-specific *Central* networks are then interpolated via *AF* in an instance specific manner. These interpolated descriptors from the *AF* are analyzed on both the dialogue level and on a speaker level, allowing us to track emotional changes for the entire group and for each individual speaker.Thus, the key contributions of proposed model for the *A*daptive *Mu*ltimodal Analysis for *S*peaker *E*motion (aka *AMuSE*), are:

1) **Cross-Attended Feature Representation via Multimodal Attention Network** that models cross-modal interaction by "injecting" features from multiple *Peripheral* networks into the layers of a *Central* network. This helps the model to prioritize on the mode-invariant spatial (within utterance) details of the emotion patterns, while also retaining its mode-exclusive aspects at various levels of abstraction.
2) **Adaptive Fusion (AF)** that interpolates the cross-attended mode-specific descriptors to combine the novel instance-specific and category-specific utterance-level spatial patterns within the learned multimodal descriptor.
3) **Extensive Evaluations with Explainable Visualization** using publicly available (MELD [29], IEMOCAP [3]) datasets not only demonstrate an impressive classification performance ($3-5\%$ improvement in Weighted-F1 and $5-7\%$ improvement in Accuracy) of *AMuSE*, its user-friendly *interface* also facilitate the multimodal reasonings behind a specific prediction made by the model to deliver improved reliability on its decision

spatio-temporalat various levels

II. Related Work

Traditionally, works on Emotion Recognition in Conversations (ERC) have focused heavily on unimodal techniques, primarily due to the strength of natural language transcriptions or descriptors as a strong emotional indicator [7], [10], [23]. However, while text serves the purpose in simple scenarios, it often struggles to evaluate more complex human responses involving sarcasm or confusion, where more information can be gleaned about the speaker's emotional state by studying the tone, face, posture, and gestures [39]. Recent works [5], [19] demonstrate the superiority of multimodal techniques. Most current ERC research has focused on modeling cross-modal interactions by either concatenating the processed unimodal feature vectors [13], [23], [28] or by a predefined fixed combination (e.g., a weighted average of feature vectors) [25], [34]. While better than unimodal approaches, these techniques ignore various levels of information that may be critical for the comprehensive modeling of their spatial and temporal multimodal relationships.

Though promising, many existing methods [16], [31] suffer from the weakness in an insufficient fusion of cross-modal interactions, unlike popular beliefs, which may not be uniform across instances or categories [8] and vary given an individual's unique socio-behavioral responses. Furthermore, the precision vs. explainability trade-off continues to pose challenges for the systems. Toward this, we leverage multimodal information from each sample to track the evolution of emotion over the conversation both for individual speakers and the group as a whole, while also simultaneously providing meaningful insights that attempt to explain the model's decision.

III. Proposed Method

Problem Definition: Given a multi-party conversation represented as a sequence of utterances $\{u_j\}_j \in \mathcal{D}$, the objective is to evaluate the dominant emotional state of the speaker for each utterance u_j. For brevity, we will henceforth omit the suffix j, and an arbitrary utterance u_j will be represented as u unless the suffix is specifically required. Each $u \in \mathcal{D}$ contains (T, V, A), where T is the *text transcription*, V is the *video* and A is the *audio*.

A. Unimodal Self-Attended Feature Representation

To capture the spatial evolution of information within each utterance, we propose a mode-specific unimodal feature representation scheme as described below

1) Text Representation: To derive a compact descriptor for the text component T represented as a sequence of p words, i.e. $T = \{\omega_1, \omega_2, ..., \omega_p\}$, we employ the pretrained model [32] to obtain the fixed language embedding $\mathbf{f}^T \in \mathbb{R}^{p \times d_t}$ for the text component T. The masked and permuted language modelling (MPNet) inherits the advantages of BERT [6] and XLNet [37] by leveraging the dependency within the predicted tokens through permuted language modelling and utilizes the auxiliary position information to mitigate the position discrepancy. In fact, to explicitly capture the contextual meaning of each word in an utterance, the initial MPNet-based word embeddings are used as input to a Bi-directional-Long Short-Term Memory (Bi-LSTM) followed by the embedding layer to produce a derived word representation vector $\mathbf{h}_i$ for each ω_i, which in turn develops a derived text representation vector $\mathbf{w}_0 = [\mathbf{h}_1, \mathbf{h}_2, ...\mathbf{h}_p]$ for the text component T. Toward attaining an attention-aware text descriptor, $\mathbf{w}_0$ is further processed through a M-layered attention F_t^{SA} network computed as $\mathbf{w}_M = F_t^{SA}(\mathbf{w}_0, M) \in \mathbb{R}^{p \times d_t}$. An intermediate m^{th} layer output in F_t^{SA} is computed as $\mathbf{w}_{m+1} = \text{linear}\Bigg(\text{softmax}\big(\frac{\mathbf{w}_m \mathbf{w}_m^T}{\sqrt{d_t}}\big)\mathbf{w}_m\Bigg)$ and the resulting attention-enhanced average pooled text descriptor is defined as $\mathbf{f}^T = \mathbf{w}_M$.

2) Video Representation: For the visual component V of each utterance $u \in \mathcal{D}$, FFmpeg is used to identify n keyframes and MTCNN [41] is applied to extract the aligned faces from each keyframe. To represent the facial expression information within the context of the individuals' environment, each keyframe is then decomposed into two components: "face frames", which is a derived frame containing only the face regions of the keyframes; "back frames", which captures the background environment by removing all identified faces. JAA-Net [30], which jointly performs Action Units (AU) detection and facial landmark detection, is employed to extract AUs from each of these "face frames". Thus, the visual content V of u is represented in terms of two equal-sized derived frame sequences: $\mathbf{v}^{face} = \{\mathbf{au}_1, \mathbf{au}_2, ..., \mathbf{au}_n\}$ and $\mathbf{v}^{back} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n\}$, where each $\mathbf{au}_j$ and $\mathbf{b}_j$ represent a learned descriptor describing the j^{th} element in $\mathbf{v}^{face}$ and $\mathbf{v}^{back}$ respectively. Two identical Bi-LSTM-based sequence representation modules, which take $\mathbf{v}^{face}$ or $\mathbf{v}^{back}$ as inputs, are employed to obtain the initial regional descriptors $\mathbf{v}^{face} \in \mathbb{R}^{n \times d_v^{hid}}$ or $\mathbf{v}^{back} \in \mathbb{R}^{n \times d_v^{hid}}$, where d_v^{hid} is the number of the final embedding layer units in the Bi-LSTM model. Similar to the approach followed in the text feature representation process, a stacked self-attention layer network (F_v^{SA}), which retrieves the multi-view attention between $\mathbf{v}^{face}$ and $\mathbf{v}^{back}$ to derive a self-attended visual descriptor $\mathbf{f}^v \in \mathbb{R}^{2n \times d_v}$ for V.

3) Audio Representation: Patchout fast (2-D) spectrogram transformer (PASST) [17] model, which is initialized from an ImageNet vision transformer model, and further pretrained on 10s audio from AudioSet [9], is used to represent the audio component A of the utterance u. Each segment is then represented in terms of their PASST descriptor, so $A = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_e\}$, where $\mathbf{a}_i \in \mathbb{R}^{d_{passt}}$ is the PASST feature of the i^{th} audio frame. Similar to visual and text representations discussed above, a Bi-LSTM network followed by an M-layered self-attention module F_a^{SA} is leveraged to obtain attention enhanced descriptor $\mathbf{f}^a \in \mathbb{R}^{e \times d_a}$.

B. Multimodal Attention Network (MAN) for Cross Attended Feature Representation

While the intra-modal feature discriminability can be addressed by the proposed technique above, toward integrating the contents across modalities, heterogeneity of the data patterns across multiple modes is often a bottleneck. For example, an utterance by a speaker may reflect an impression on the speaker's face as well as on that of other participants in the conversation. Similarly, the transcript of the utterance should also be relevant to the visual background context or may have been a response to another utterance by a previous speaker. With this intuition, we propose a Multimodal cross-attended feature representation learning using a Multimodal Attention Network (*MAN*) that takes $\mathbf{f}^m$ $(m \in \{v, t, a\})$ as inputs, and models cross-modal interactions at various levels of detail. As observed in Figure 1, each layer of our *MAN* models cross-modal interaction by "injecting" features from multiple *Peripheral* networks into a *Central* network.

More specifically, for each mode m in consideration, its mode-specific *Central* query network is designed using h dense layers followed by a Softmax layer [18], which takes $\mathbf{f}^m$ as input and its intermediate l^{th} dense layer output $\mathbf{g}_m^l$, is cross-attended by one or more pairs of *Peripheral* key $(\mathbf{K}_{m_i}^l)$ and value $(\mathbf{V}_{m_i}^l)$, such that $m_i \neq m$. Each key-value pair is generated via linear mappings. Thus, we have $\mathbf{K}_{m_i}^l = (W_{m_i}^{l,K})^T \mathbf{f}^{m_i}$ and $\mathbf{V}_{m_i}^l = (W_{m_i}^{l,V})^T \mathbf{f}^{m_i}$ for $\mathbf{W}_{m_i}^{l,V}, \mathbf{W}_{m_i}^{l,K} \in \mathbb{R}^{d_{m_i} \times d_l}$. In a multi-head attention learning framework, a particular head for the cross-attended *Central* network output from its l^{th} layer is computed as:

$$\mathbf{g}_m^l = \mathbf{g}_m^{l-1} + \frac{1}{|\mathcal{M}|} \sum_{m_i \in \mathcal{M} \setminus \{m\}} \text{softmax}(linear(\frac{\mathbf{g}_m^{l-1}(\mathbf{K}_{m_i}^{l-1})^T}{\sqrt{d_{m_i}}}))\mathbf{V}_{m_i}^{l-1} \tag{1}$$

Where $\mathcal{M}$ represents the set of all modes in consideration. Such responses from multiple heads are average pooled to derive the final output of the *Central* network.

While *MAN* architecture is generic and can be extended for any number of modalities, as described in Section III-A, in our experiments, we use information from three different modes (i.e. $\mathcal{M} \subseteq \{t, v, a\}$). As shown in Figure 1, each mode-specific *Central* query network for each $m \in \{a, v, t\}$) of

*MAN*s thus produces an average pooled cross-attended mode-specific descriptor $\mathbf{f}^m_{CA} \in \mathbb{R}^d$ for the uni-mode components T, V, and A for u. We will discuss the learning algorithm later in Section III-E.

C. Adaptive Fusion (AF)

Most of the existing methods [2], [14], [24] leverage a static approach for multimodal feature fusion. Here we make an important observation: The relative importance of each modality is not uniform and varies across samples exhibiting different emotions. For example, speaker emotion may reflect an influence of several contexts like audience reaction or surrounding environment. However, the transition of an individual's expression of emotion is continuous. Therefore, slight changes in the utterance representations should not have caused drastic changes in the utterance's emotion labels. Keeping this in mind, we propose an *A*daptive *F*usion (*AF*) function $\mathcal{A}$ that learns a linear combination of the mode-specific representations ($\mathbf{f}^m_{CA}$), to derive a comprehensive *spatial multimodal descriptor* $\mathcal{A}(u)$ for an utterance u as follows:

$$\mathcal{A}(u) = \bigoplus_{m\in\mathcal{M}} \left(\frac{1}{|\mathcal{M}|} \sum_{m_i \in \mathcal{M}\setminus m} \alpha^m_{m_i} \mathbf{f}^{m_i}_{CA} + (1-\alpha^m_{m_i})\mathbf{f}^{m_i}_{CA} \right) \tag{2}$$

where $0 \leq \alpha^m_{m_i} \forall m, m_i \in \mathcal{M} \leq 1$ are learnable parameters and $\bigoplus$ represents the concatenation operator. Thus, the proposed fusion function $\mathcal{A}$ provides a flexible multimodal representation mechanism, by which the resulting multimodal descriptor $\mathcal{A}(u)$ for an utterance u is able to retain category-specific discriminative data patterns, however not completely disregarding the unique instance-specific data patterns observed in the utterance.

D. Emotion Classification

Given a conversational dialogue represented using a sequence of n utterances $\{u_j\}^n_j \in \mathcal{D}$, our task is to evaluate the emotion of a user-identified speaker s by utilizing the spatio-temporal contexts observed in the dialogue. To attain this objective, we design a speaker specific representation of the dialogue by using two parallel utterance sequences: *Dialogue Context*, which describes the entire sequence $\{\mathcal{A}(u_j)\}_j$; *Speaker Context*, a derived utterance sub-sequence $\{\mathcal{A}(u_{s_j})\}_j$, where the sub-sequence $\{u_{s_j}\}_{s_j\in[1,n]}$ is generated from the dialogue and includes only those utterances, in which s vocally contributes to the conversation. Given the voice of a speaker identified by the user in the first keyframe, we use Librose library function to match it across utterances to identify this derived subsequence. Two parallel Bi-LSTMs are trained to capture the spatio-temporal contexts independently from these contexts' perspectives. In our experiments, Bi-LSTMs were found to be more useful than LSTm due to ability to capture the bidirectional temporal contexts. For example using the *Speaker Context*, the pertinent spatio-temporal representation $\mathbf{s}_l \in \mathbb{R}^s$ of an utterance u^s_l is reasoned from the representation $\mathbf{s}_{(l-1)}$ of $u^s_{(l-1)}$, while also considering the current state of utterance u^s_l denoted as $\mathbf{z}^s_l$ (which are initially null) - as $\mathbf{s}_l = \overleftrightarrow{\mathrm{LSTM}}^s(\mathbf{s}_{(l-1)}, \mathbf{z}^s_l)$. Similarly, using the *Dialogue Context*, the pertinent spatio-temporal representation $\mathbf{d}_l \in \mathbb{R}^s$ of an utterance u_l is reasoned from the representation $\mathbf{d}_{(l-1)}$ of $u_{(l-1)}$, while also considering the current state of utterance u_l denoted as $\mathbf{w}^d_l$ (which are initially null) - as $d_l = \overleftrightarrow{\mathrm{LSTM}}^d(\mathbf{d}_{(l-1)}, \mathbf{w}^s_l)$. The final representation of each utterance in the *Speaker Context* is then derived as, $e_l = \mathbf{s}_l \oplus \mathbf{d}_l$, which is used as an input to a simple neural network comprising of a linear layer followed a softmax activation to estimate the occurance of an emotion in the utterance.

E. Learning

The learning algorithm of *AMuSE* includes two independent learning objectives. A loss objective ($\mathcal{L}$) for *MAN* learning and an *AF* objective to optimize the values of $\alpha^m_{m_i}$.

1) MAN Learning: Given an utterance $u_j \in \mathcal{D}$, the proposed Multimodal Attention Network (*MAN*) jointly learns the cross-attended representations $\mathbf{f}^t_{CA,j}$, $\mathbf{f}^v_{CA,j}$, and $\mathbf{f}^a_{CA,j}$ with twofold objectives: 1) preserving Instance-level discriminability [11], [26] within them. This is equivalent to obtaining the learned representations such that they must be more discriminative of u compared to the other samples in $\mathcal{D}$. To incorporate this intuition, we use Noise Contrastive Estimation (NCE) [11] loss ($\mathcal{L}_{NCE}$); 2) preserving the 'category-level' information within their learned representations, such that the predictions obtained from each of these learned mode-specific representations may also align with the ground-truth labels. We leverage Focal Loss [22] for this purpose.

We leverage an aggregated noise contrastive estimation ($\mathcal{L}_{ACE}$) and an averaged focal loss ($\mathcal{L}_{fl}$), as defined below:

$$\mathcal{L}_{\mathcal{ACE}} = \frac{1}{|\mathcal{D}|} \sum_{u_j\in\mathcal{D}} \frac{1}{|\mathcal{M}|} \sum_{\substack{m\neq m_i \\ m,m_i\in\mathcal{M}}} \mathcal{L}_{\mathrm{NCE}}(\mathbf{f}^m_{CA,j}, \mathbf{f}^{m_i}_{CA,j}) \tag{3}$$

with

$$\mathcal{L}_{\mathrm{NCE}}(\mathbf{f}^m_{CA,j}, \mathbf{f}^{m_i}_{CA,j}) = \big[-log(\frac{P(\mathbf{f}^{m_i}_{CA,j}|\mathbf{f}^{m_i}_{CA,j})}{P(\mathbf{f}^{m_i}_{CA,j}|\mathbf{f}^{m_i}_{CA,j}) + \frac{|\mathcal{N}_j|}{|\mathcal{N}|}}) + \sum_{k\in\mathcal{N}_j} log(\frac{P(\mathbf{f}^{m_i}_{CA,k}|\mathbf{f}^{m_i}_{CA,j})}{P(\mathbf{f}^{m_i}_{CA,k}|\mathbf{f}^{m_i}_{CA,j}) + \frac{|\mathcal{N}_j|}{|\mathcal{N}|}}) - 1\big] \tag{4}$$

that computes the probability of both features $\mathbf{f}^m_{CA,j}$ and $\mathbf{f}^{m_i}_{CA,j}$ representing the same instance u_j compared to other elements in a uniformly sampled negative set $\mathcal{N}_j$.

The averaged Focal Loss [22], specifically effective for an imbalanced dataset like ours, is defined below:

$$\mathcal{L}_{fl} = \frac{1}{|\mathcal{M}|} \frac{1}{|\mathcal{D}|} \sum_{m\in\mathcal{M}} \sum_{u_j\in\mathcal{D}} (1 - p^m_c(u_j))^\gamma p^m_c(u_j) \tag{5}$$

where $p^m_c(u_j)$ is the predicted class-membership probability score for the sample u_j by the m^{th} mode-specific *Central* query network and γ is a tunable parameter. We use a combined loss function $\mathcal{L} = \mathcal{L}_{ACE} + \mathcal{L}_{fl}$ to jointly learn its mode-specific *Central* query networks.

TABLE I

PERFORMANCE COMPARISON OF DIFFERENT METHODS USING THE WEIGHTED AVERAGE F1 MEASURE (W-AVG F1) ON THE MELD DATASET WITH UNI (T:=TEXT, A:=AUDIO, AND V:= VIDEO) AND MULTI-MODAL DATA REPRESENTATIONS. DUE TO THE IMBALANCED CLASS DISTRIBUTION OF THE DATASET, THE 'FEAR' AND 'DISGUST' CLASSES ARE REPRESENTED AS THE MINORITY CLASSES, THE PROPOSED METHOD WAS ALSO COMPARED AGAINST OTHER 5 MAJORITY CLASSES ('NEUTRAL', 'SURPRISE', 'SADNESS', 'JOY', AND 'ANGER') IN THE DATASET AND THE RESULTS ARE REPORTED IN COLUMN 'W-AVG F1 5 CLS'. 'FEATURE CONCAT' IN ROW-12 AND ROW-13 DESCRIBES THE CONCATENATION OF MULTIPLE UNI-MODE DESCRIPTORS TO DEFINE A MULTIMODAL DESCRIPTOR.

Method	Mode	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	w-Avg F1	w-avg F1 5-CLS
MFN [38]	T + A	0.762	0.407	0.0	0.137	0.467	0.0	0.408	0.547	0.5732
ICON [12]	T	0.762	0.462	0.0	0.189	0.485	0.0	0.301	0.546	0.5718
	A	0.669	0.0	0.0	0.0	0.086	0.0	0.315	0.377	0.3947
	T + A	0.736	0.500	0.0	0.232	0.502	0.0	0.448	0.563	0.5897
DialogueRNN [24]	T	0.737	0.449	0.054	0.234	0.476	0.0	0.415	0.551	0.5759
	A	0.53	0.156	0.0	0.083	0.112	0.051	0.321	0.34	0.3542
	T + A	0.732	0.519	0.0	0.248	0.532	0.0	0.456	0.57	0.5971
ConGCN [40]	T	0.749	0.498	0.065	0.226	0.524	0.088	0.432	0.574	0.5969
	A	0.641	0.254	0.047	0.193	0.155	0.030	0.341	0.422	0.44
	T + A	0.767	0.503	0.087	0.285	0.531	0.106	0.468	0.594	0.6175
DialogueCRN [14]	T + A	-	-	-	-	-	-	-	0.6073	-
EmoCaps [20]	T + A + V	0.7712	0.6319	0.0303	0.4254	0.5750	0.0769	0.5754	0.6400	-
M2FNet [5]	T + A + V	-	-	-	-	-	-	-	0.6785	-
Cross-Modal Distribution Matching [21]	T + A	-	-	-	-	-	-	-	0.571	-
Transformer Based Cross-modality Fusion [35]	T + A +V	-	-	-	-	-	-	-	0.64	-
Hierarchical Uncertainty for Multimodal Emotion Recognition [4]	T + A +V	-	-	-	-	-	-	-	0.59	-
Shape of Emotion [1]	T + A +V	-	-	-	-	-	-	-	0.63	-
UniMSE [15]	T + A +V	-	-	-	-	-	-	-	0.66	-
Proposed Uni-mode Feature Rep. (Section III-A) +Classifier (Section III-D)	T	0.7439	0.6191	0.0209	0.3914	0.5178	0.0613	0.5036	0.6041	0.6306
	A	0.3838	0.3581	0.0209	0.3286	0.3617	0.0613	0.3529	0.3537	0.3684
	V	0.5562	0.4905	0.0209	0.3374	0.4098	0.0613	0.3713	0.4615	0.4813
Proposed Uni-mode Feature Rep.(Section III-A) + Feature Concat. + Classifier (Section III-D)	T + A	0.7627	0.6318	0.0241	0.4214	0.5316	0.0613	0.5597	0.6265	0.6540
	T + V	0.7427	0.6218	0.0241	0.4214	0.5316	0.0613	0.5597	0.6158	0.6428
	A + V	0.5562	0.5796	0.0209	0.3610	0.4098	0.0613	0.4318	0.4810	0.5017
	T + A + V	0.7671	0.6518	0.0319	0.4629	0.5291	0.0691	0.5713	0.6356	0.6632
Proposed *MAN*-based Feature Rep.(Section III-B) + Feature Concat +Classifier (Section III-D)	T + A + V	0.8359	0.7094	0.0674	0.4468	0.6297	0.0891	0.6389	0.6992	0.7286
AMuSE	T + A + V	**0.8469**	**0.7283**	0.0674	**0.4632**	**0.6481**	0.0891	**0.6574**	**0.7132**	**0.7431**

2) AF Learning: As observed in Eqn.2, the multimodal descriptor $\mathcal{A}(u)$ interpolates all cross-attended mode-specific descriptors ($\mathbf{F}_{CA}^{m_i}$) to reveal all the discriminative feature information by leveraging the changes in the model behavior in response to varying inputs. As such, it is intuitive to note that a slight change in the feature representation should not cause any observable change in the model's decision. Nevertheless, manual selection for any of the interpolation coefficients (however small it is) $\boldsymbol{\alpha}_{m_i}^{u}$ may not be equally effective across all samples. Thus, the threefold approximation task specific to our scenario is solved in a pairwise manner. In other words, we perform the learning of these interpolation parameters by first approximating $\boldsymbol{\alpha}_1^{'}$ for $\mathbf{u}_1^{'} = \boldsymbol{\alpha}_1^{'}\mathbf{f}_{CA}^{m_1} + (1-\boldsymbol{\alpha}_1^{'})\mathbf{f}_{CA}^{m_2}$ followed by approximating $\boldsymbol{\alpha}_2^{'}$ for $\mathbf{u}_2^{'} = \boldsymbol{\alpha}_2^{'}\mathbf{u}_1^{'} + (1-\boldsymbol{\alpha}_2^{'})\mathbf{f}_{CA}^{m_3}$. Then by setting $F^{AMuS}(u) = \frac{1}{3}(\mathbf{u}_1^{'} + \mathbf{u}_2^{'})$ and equating the coefficients of like terms, we obtain $\boldsymbol{\alpha}_{m_1}^{u} = \boldsymbol{\alpha}_1^{'}.\boldsymbol{\alpha}_2^{'}$, $\boldsymbol{\alpha}_{m_2}^{u} = \boldsymbol{\alpha}_2^{'}.(1-\boldsymbol{\alpha}_1^{'})$, and $\boldsymbol{\alpha}_{m_3}^{u} = 1-\boldsymbol{\alpha}_2^{'}$. For optimizing the interpolation parameter $\boldsymbol{\alpha}_1^{'}$ (and similarly $\boldsymbol{\alpha}_2^{'}$), we adopt the optimization approach of [27], which is as follows:

$$\boldsymbol{\alpha}_1^{'} \approx \epsilon \frac{\|(\mathbf{f}_{CA}^{m_1} - \mathbf{f}_{CA}^{m_2})\|_2 \nabla_{\mathbf{f}_{CA}^{m_2}} \mathcal{L}(Q_{m_2}(\mathbf{f}^{m_2}), y^*)}{\|\nabla \mathbf{f}_{CA}^{m_2} \mathcal{L}(Q_{m_2}(\mathbf{f}^{m_2}), y^*)\|_2} \oslash (\mathbf{f}_{CA}^{m_1} - \mathbf{f}_{CA}^{m_2}) \tag{6}$$

where y^* is the ground truth label for the sample u, $\oslash$ is the element-wise division, ϵ is the hyper-parameter that controls the amount of interpolation in the result, and Q_{m_i} represents the i^{th} mode-specific *Central* query network. To facilitate the learning process, we randomly identify a set of informative samples from the validation pool for which the loss due to a small interpolation indeed is indeed affected (i.e. system prediction indeed changes by a slight change in the interpolation parameters) to use in the following training epochs.

IV. EXPERIMENTS

A. Datasets

Derived from the TV series F.R.I.E.N.D.S, *MELD* [29] is a multi-party multimodal conversation dataset comprising 7 emotions - 'Anger', 'Disgust', 'Sadness', 'Joy', 'Surprise', 'Fear', and 'Neutral'. *IEMOCAP* [3] is a dyadic conversational dataset, with recordings of professional actors performing scripted and improvised scenarios comprising 6 emotions - 'Happy', 'Sad', 'Neutral', 'Angry', 'Excited', 'Frustrated'.

B. Results & Comparative Study

Figure 2 provides some qualitative results. Table I and Table II present the results on MELD and IEMOCAP test sets, respectively by using F1-score [24] as the evaluation metric. The results are compared against several state-of-the-art algorithms [2], [5], [12], [14], [20], [21], [24], [36], [38], [40], [42]. For each emotion category, we also evaluate the classification performance using their weighted averages across all emotion classes. As observed in Table I, while 'Text' appears to be the most reliable uni-modal feature,

TABLE II

PERFORMANCE COMPARISON OF DIFFERENCE METHODS USING THE WEIGHTED AVERAGE F1 MEASURE (W-AVG F1) ON THE IEMOCAP DATASET WITH WITH UNI (T:=TEXT, A:=AUDIO, AND V:= VIDEO) AND MULTI-MODAL DATA REPRESENTATIONS. 'FEATURE CONCAT' IN ROW 13 AND ROW 14 DESCRIBE THE CONCATENATION OF MULTIPLE UNI-MODE DESCRIPTORS TO DEFINE A MULTIMODAL DESCRIPTOR.

Method	Mode	Happy	Sad	Neutral	Angry	Excited	Frustrated	w-Avg F1
MFN [38]	T + A	-	-	-	-	-	-	0.3490
ICON [12]	T + A + V	0.3280	0.7440	0.6060	0.6820	0.6840	0.6620	0.6350
DialogueRNN [24]	T + A + V	0.3318	0.7880	0.5921	0.5128	0.7186	0.5891	0.6275
MMGCN [33]	T + A + V	0.4235	0.7867	0.6173	0.6900	0.7433	0.6232	0.6622
DialogueCRN [14]	T + A	0.6261	0.8186	0.6005	0.5849	0.7517	0.6008	0.6620
ERLDK [42]	T + A	0.4730	0.7919	0.5642	0.6054	0.7444	0.6385	0.6390
Hierarchical Uncertainty for Multimodal Emotion Recognition [4]	T + A + V	-	-	-	-	-	-	0.6598
DAG-ERC+HCL [36]	T	-	-	-	-	-	-	0.6803
M2FNet [5]	T + A + V	-	-	-	-	-	-	0.6986
Multimodal Attentive Learning [2]	T + A + V	-	-	-	-	-	-	0.6540
Proposed Uni-mode Feature Rep. (Section III-A)	T	0.2991	0.6141	0.5251	0.5728	0.5918	0.5969	0.5526
	A	0.2991	0.3894	0.3951	0.2749	0.326	0.3316	0.3417
	V	0.3038	0.5329	0.5619	0.2749	0.326	0.431	0.4260
Proposed Uni-mode Feature Rep. (Section III-A) + Feature Concat.	T + A	0.3038	0.6368	0.5619	0.598	0.6027	0.6069	0.5727
	T + V	0.3359	0.6368	0.5885	0.598	0.6027	0.6069	0.5815
	A + V	0.3038	0.5592	0.6328	0.321	0.326	0.5293	0.4782
	T + A + V	0.3917	0.6368	0.6354	0.6374	0.6027	0.6399	0.6117
Proposed *MAN*-based Feature Rep.(Section III-B)+ Feature Concat	T + A + V	0.6591	0.8106	0.7248	0.6599	0.7769	0.6734	0.7147
AMuSE	T + A + V	**0.7025**	**0.8418**	**0.7548**	**0.6748**	**0.7935**	**0.6923**	**0.7391**

combining information from multiple modes is always helpful. To this end, as we compare the last sub-row of row-12 and row-13, the proposed *MAN* based cross-attention appears to be extremely beneficial in improving the weighted F1-score **(w-avg F1) by around** 6%. Finally, using a flexible and efficient fusion approach, the proposed *AMuSE* facilitates further improvement in the performance by reporting $\sim 74\%$ **w-avg F1- yet another** $\sim 2\%$ **improvement** compared to the results reported in the baseline row-13 scenario. Row-13 reports the experiment results, wherein cross-attended mode-specific feature descriptors (Section III-A) are simply fused using equal values of the interpolation parameters in Eqn 2 (i.e. $\boldsymbol{\alpha}^{s_1}_{m_i} = \boldsymbol{\alpha}^{s_2}_{m_j} \forall m_i, m_j \in \mathcal{M}, \forall u^{s_1}, u^{s_1} \in \mathcal{D}$). As we compare this performance (in row-14) with row-8 and row-9 of Table **??**, we observe that ***AMuSE* reports around** $4-7\%$ **improved performance** compared to the best performing existing methods [5], [20]. A similar performance pattern is also observed in Table II, wherein *AMuSE* is compared against several baseline methods using the IEMOCAP dataset. By comparing the last sub-row of row-13 and row-14, we find that the proposed ***MAN*-based cross-attention enables the mode to attain an impressive** 10% **improvement** over its baseline test scenario, in which only the mode-specific feature descriptors (Section III-A) are simply concatenated to define a multimodal descriptor. Finally, by employing *AF* for feature fusion, **the model attains** $\sim 74\%$ **w-avg F1**, which overshoots some of the best-performing baselines [5], [20] **by around** $2-4\%$. While most of the works use F1-score as the evaluation metric, compared to a handful few recent works [14], [42], which have also reported classification accuracy of their method, proposed *AMuSE* reports an impressive performance. Compared to one of the best-performing baselines M2FNet [5] that reports 66.71 accuracy in the MELD dataset and 69.69% accuracy in the IEMOCAP dataset, *AMuSE* reports ***around*** 7% ***(i.e.*** 73.28% ***accuracy score in MELD dataset) and*** 5% ***(i.e.*** 74.49% ***accuracy score in IEMOCAP dataset) improvement respectively***.

C. Ablation Study

As observed in Table III, compared to the other testing scenarios (where either the $\{\boldsymbol{\alpha}^u_{m_i} \forall m_i \in \mathcal{M}, \forall u \in \mathcal{D}\}$ parameters were chosen at random or were fixed to the same value for all samples in $\mathcal{D}$ and modes $\mathcal{M}$) the proposed *AMuSE* shows an improved performance in Test-3 experiment setting, wherein it leverages the learning algorithm for the *AF* interpolation parameters, introduced in Section **??**, to optimize the choices of these parameters in mode-specific and input-specific manner. This makes the model more adaptable to the newer data patterns, observed in analyzing a speaker's emotions from a socio-racial background, compared to the speakers/participants population available in the training collection. In the other set of ablation Study experiments, we choose different values for the tunable parameter γ in the focal loss function defined in Eqn. 3. Again as observed in Table IV, in both datasets, the chosen value of $\gamma = 1$ produces a slightly better W-Avg F1 score, compared to the other values of γ. In fact, the performance of *AMuSE* remains mostly stable over a range of values in $[0.75, 1.25]$, which highlights the system stability in the performance over the choice of γ values. Finally, in Table V we also report ablation study results for the number of *MAN* layers in the model. The performance remains fairly consistent when using 3, 4, or 5 layers and peaks at 4, which is the number of layers we have chosen in the model.

D. Explainability

The proposed explainability analysis approach uses Local Interpretable Model-Agnostic Explanations (LIME)[1] to ex-

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

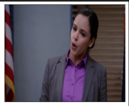
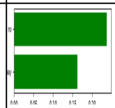
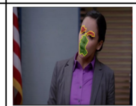
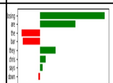
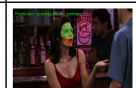
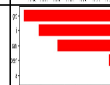
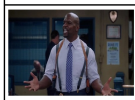
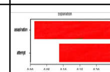
Query Utterance	Query Utterance Transcript	Ground Truth	Prediction				Explanation		
			DRNN	DCRNN	Uni-mode (Section 3.1)	AMuS-HAL	T	AU	V
	No! Because he passed 4 million dollars in fraudulent bills Jake!	Angry	Disgust	Disgust	Angry (T), Disgust (A), Angry (V)	Angry			
	Chris says they're closing down the bar	Sadness	Neutral	Sadness	Neutral (T), Sadness (A), Sadness (V)	Sadness			
	Can I get a beer?	Sadness	Neutral	Disgust	Neutral (T), Disgust (A), Disgust (V)	Disgust			
	there's been an assassination attempt	Anger	Sadness	Sadness	Sadness(T), Anger (A), Anger (V)	Anger			

Fig. 2. Some example results with 3 mode-specific explainability analysis, wherein explanation columns regions/texts contributing to the model decision are highlighted in Green. The detracting regions are highlighted in Red

TABLE III
ABLATION STUDY ON THE *AF* FUNCTION PARAMETERS ($\boldsymbol{\alpha}^u_{m_i}$ FOR $m_i \in \mathcal{M}$) WAS PERFORMED IN SEVERAL TESTING SCENARIOS: *Test-1* IN WHICH WE CHOOSE THESE PARAMETERS AT RANDOM SUCH THAT $\boldsymbol{\alpha}^u_{m_i} \neq \boldsymbol{\alpha}^u_{m_j} \forall m_i, m_j \in \mathcal{M}$ AND $\boldsymbol{\alpha}^{u^{s1}}_{m_i} = \boldsymbol{\alpha}^{u^{s2}}_{m_i} \forall u^{s1}, u^{s2} \in \mathcal{D}$; *Test-2* IN WHICH WE CHOOSE THESE PARAMETERS SUCH THAT $\boldsymbol{\alpha}^{s1}_{m_i} = \boldsymbol{\alpha}^{s2}_{m_j} \forall m_i, m_j \in \mathcal{M}, \forall u^{s1}, u^{s1} \in \mathcal{D}$; *Test-3* IN WHICH WE LEARN THE PARAMETERS FOLLOWING THE APPROACH (*AF Learning*) DISCUSSED IN SECTION III-E2. THE TABLE REPORTS THE WEIGHTED AVERAGE F1 MEASURE (W-AVG F1) OVER ALL CLASSES IN THE DATASETS.

Dataset	*Test-1*	*Test-2*	*Test-3*
MELD [29]	70.86	71.10	71.32
IEMOCAP [3]	72.07	72.89	73.91

TABLE IV
ABLATION STUDY ON THE TUNABLE PARAMETER γ IN THE FOCAL LOSS FUNCTION DEFINED IN EQN. 3.THE TABLE REPORTS THE WEIGHTED AVERAGE F1 MEASURE (W-AVG F1) OVER ALL CLASSES IN THE DATASETS.

Dataset	$\gamma = 0.5$	$\gamma = 0.75$	$\gamma = 1.0$	$\gamma = 1.25$
MELD [29]	70.71	71.08	71.32	70.97
IEMOCAP [3]	71.46	73.47	73.91	73.12

plain system decisions. LIME provides Interpretable, Model-Agnostic Visual explanations for any classifier by treating the classification model as a black box. LIME approximates the classifier model locally in the neighborhood of the prediction. In Figure 2, we present the explainability analysis in various modes: Textual explanation with the words that contribute the most (or against) the prediction; face landmarks and Action Unit (AU) based explanation that illustrates the regions in the speaker's face that contribute (using Green) and distract (using Red) to the prediction. Finally, we present visual regions of interest in the image responsible for the model's decision.

V. CONCLUSION

We present *AMuSE* with a *Multimodal Attention Network*, which enables effective knowledge sharing from multiple interactive mode-specific branches to facilitate robust decision-making. Following a multi-loss learning framework, the proposed *Adaptive Fusion* allows *AMuSE* model to learn the

TABLE V
THE ABLATION STUDY SHOWING THE EFFECT OF CHANGING THE NUMBER OF *MAN* LAYERS

Model	MELD	IEMOCAP
1-layer	65.37	68.14
3-layer	69.26	70.98
5-layer	71.03	72.57
AMUSE (4 layers)	**71.32**	**73.91**

relative contributions of each mode in an effort to learn both category-specific discriminative details and instance-specific contrast-enhanced discriminative cross-modal correspondence within the learned multimodal descriptor. As evident from the experiments, *AMuSE* delivers a significantly improved performance compared to the baselines. Furthermore, the *Interactive Explainaibility Visualization* also guides the user and produces appropriate mode-wise reasoning for its classification.

ACKNOWLEDGMENT

Computational support is provided by the Center for Computational Research at the University at Buffalo.

REFERENCES

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

[2] Balaji Arumugam, Sreyasee Das Bhattacharjee, and Junsong Yuan. Multimodal attentive learning for real-time explainable emotion recognition in conversations. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1210–1214. IEEE, 2022.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335, Nov 2008.

[4] Feiyu Chen, Jie Shao, Anjie Zhu, Deqiang Ouyang, Xueliang Liu, and Heng Tao Shen. Modeling hierarchical uncertainty for multimodal emotion recognition in conversation. *IEEE Transactions on Cybernetics*, 2022.

[5] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661, 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

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

[9] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[10] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations, 2020.

[11] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[12] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[13] Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *CoRR*, abs/2106.01978, 2021.

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

[15] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*, 2022.

[16] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[17] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *CoRR*, abs/2110.05069, 2021.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[19] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[20] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504*, 2022.

[21] Jingjun Liang, Ruichen Li, and Qin Jin. Semi-supervised multimodal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2852–2861, 2020.

[22] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[23] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations, 2019.

[24] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations, 2019.

[25] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 284–288, New York, NY, USA, 2016. Association for Computing Machinery.

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[27] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing, 2022.

[28] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448, 2016.

[29] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019.

[30] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaa-net: Joint facial action unit detection and face alignment via adaptive attention. *CoRR*, abs/2003.08834, 2020.

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

[32] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297, 2020.

[33] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1437–1445, New York, NY, USA, 2019. Association for Computing Machinery.

[34] Chung-Hsien Wu and Wei-Bin Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21, 2011.

[35] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors*, 21(14):4913, 2021.

[36] Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. Hybrid curriculum learning for emotion recognition in conversation. In *AAAI Conference on Artificial Intelligence*, 2021.

[37] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[38] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[39] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5415–5421. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[40] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5415–5421. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[42] Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021.