CTNet: Conversational Transformer Network for Emotion Recognition

Zheng Lian, Bin Liu, and Jianhua Tao, *Member, IEEE*

***Abstract*—Emotion recognition in conversation is a crucial topic for its widespread applications in the field of human-computer interactions. Unlike vanilla emotion recognition of individual utterances, conversational emotion recognition requires modeling both context-sensitive and speaker-sensitive dependencies. Despite the promising results of recent works, they generally do not leverage advanced fusion techniques to generate the multimodal representations of an utterance. In this way, they have limitations in modeling the intra-modal and cross-modal interactions. In order to address these problems, we propose a multimodal learning framework for conversational emotion recognition, called conversational transformer network (CTNet). Specifically, we propose to use the transformer-based structure to model intra-modal and cross-modal interactions among multimodal features. Meanwhile, we utilize word-level lexical features and segment-level acoustic features as the inputs, thus enabling us to capture temporal information in the utterance. Additionally, to model context-sensitive and speaker-sensitive dependencies, we propose to use the multi-head attention based bi-directional GRU component and speaker embeddings. Experimental results on the IEMOCAP and MELD datasets demonstrate the effectiveness of the proposed method. Our method shows an absolute 2.1$\sim$6.2% performance improvement on weighted average F1 over state-of-the-art strategies.**

***Index Terms*—Context-sensitive modeling, conversational transformer network (CTNet), conversational emotion recognition, multimodal fusion, speaker-sensitive modeling.**

I. Introduction

CONVERSATIONAL emotion recognition is an important research topic due to its wide applications in many tasks, such as dialogue generation [1], social media analysis [2] and intelligent systems [3]. The task of conversational emotion recognition requires understanding the way that humans express their emotions during conversations, and classifies each utterance into one of a fixed set of emotion categories [4].

Conventionally, conversational emotion recognition usually needs to model context-sensitive and speaker-sensitive dependencies [5], [6]. According to the emotion generation theory, humans perceive emotions not only through the current utterance but also from its surrounding utterances [7]. Meanwhile, a person's emotional state can be influenced by the interlocutor's behaviors [8]. To model these dependencies, previous works proposed various methods for conversational emotion recognition, including recurrent neural networks (RNNs) [9], memory networks [10] and graph-based models [11]. Although these works achieve promising results in emotion recognition, they generally do not leverage advanced fusion techniques to generate the multimodal representation of an utterance [12], [13]. Instead, they utilize a simple feature concatenation. In practice, high-dimensional concatenated features may easily suffer from the problem of data sparseness [14]. What's worse, these methods [9]–[11] have limitations in modeling intra-modal and cross-modal interactions [12], [13].

To deal with these drawbacks of feature concatenation, previous works mainly focused on the holistic utterance-level feature fusion and proposed some powerful fusion strategies [15]–[18]. For example, Chen *et al.* [15] fused utterance-level multimodal features via the attention mechanism. While Liu *et al.* [17] aggregated unimodal, bimodal and trimodal interactions of utterance-level multimodal features via 3-fold Cartesian product. However, recent works have found that humans can perceive emotions from small units (such as words) [19]. Integrating word-level features into utterance-level features will destroy the data structure, thus degrading the performance of emotion recognition. To address these problems, researchers explore towards word-level features fusion [20]–[22]. Typically, Gu *et al.* [21] obtained word-level alignment between modalities and explored time-restricted cross-modal interactions. Although these works [20]–[22] demonstrate their effectiveness, they mainly focus on the emotion estimation in individual utterances, ignoring speaker-sensitive and context-sensitive dependencies in conversations.

In this paper, we propose a multimodal learning framework for conversational emotion recognition, called "Conversational Transformer Network (CTNet)." As shown in Fig. 1, CTNet contains two key steps: *context-independent utterance-level feature extraction* and *context-dependent multimodal feature extraction*. We first extract word-level lexical features and segment-level acoustic features from each utterance. These multimodal

Manuscript received July 13, 2020; revised December 28, 2020; accepted January 5, 2021. Date of publication January 8, 2021; date of current version March 3, 2021. This work was supported in part by the National Key Research and Development Plan of China under Grant 2018YFB1005003, in part by the National Natural Science Foundation of China (NSFC) under Grants 61831022, 61771472, 61901473 and 61773379, and in part by Inria-CAS Joint Research Project under Grants 173211KYSB20170061 and 173211KYSB20190049. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Xie. *(Corresponding author: Jianhua Tao.)*

Zheng Lian and Bin Liu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: lianzheng2016@ia.ac.cn; liubin@nlpr.ia.ac.cn).

Jianhua Tao is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China (e-mail: jhtao@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TASLP.2021.3049898

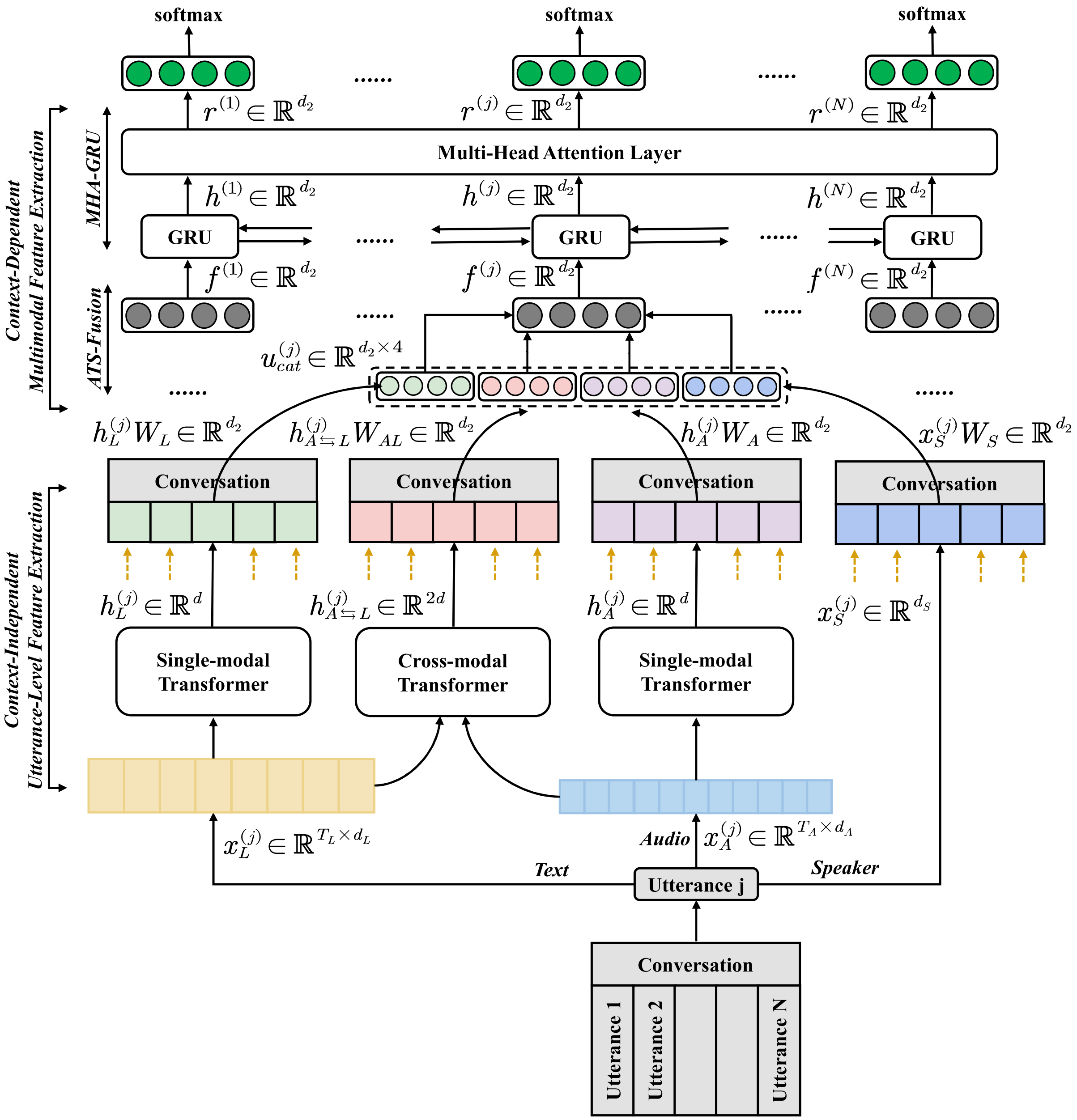

softmax
softmax
softmax
......
......
$r^{(1)} \in \mathbb{R}^{d_2}$
$r^{(j)} \in \mathbb{R}^{d_2}$
$r^{(N)} \in \mathbb{R}^{d_2}$
Multi-Head Attention Layer
$h^{(1)} \in \mathbb{R}^{d_2}$
$h^{(j)} \in \mathbb{R}^{d_2}$
$h^{(N)} \in \mathbb{R}^{d_2}$
GRU
GRU
GRU
$f^{(1)} \in \mathbb{R}^{d_2}$
$f^{(j)} \in \mathbb{R}^{d_2}$
$f^{(N)} \in \mathbb{R}^{d_2}$
$u_{cat}^{(j)} \in \mathbb{R}^{d_2 \times 4}$
MHA-GRU
ATS-Fusion
Context-Dependent Multimodal Feature Extraction
$h_L^{(j)} W_L \in \mathbb{R}^{d_2}$
$h_{A \leftrightarrows L}^{(j)} W_{AL} \in \mathbb{R}^{d_2}$
$h_A^{(j)} W_A \in \mathbb{R}^{d_2}$
$x_S^{(j)} W_S \in \mathbb{R}^{d_2}$
Conversation
Conversation
Conversation
Conversation
$h_L^{(j)} \in \mathbb{R}^{d}$
$h_{A \leftrightarrows L}^{(j)} \in \mathbb{R}^{2d}$
$h_A^{(j)} \in \mathbb{R}^{d}$
$x_S^{(j)} \in \mathbb{R}^{d_S}$
Single-modal Transformer
Cross-modal Transformer
Single-modal Transformer
Context-Independent Utterance-Level Feature Extraction
$x_L^{(j)} \in \mathbb{R}^{T_L \times d_L}$
Audio
$x_A^{(j)} \in \mathbb{R}^{T_A \times d_A}$
Text
Speaker
Utterance j
Conversation
Utterance 1
Utterance 2
Utterance N

Fig. 1. Overall structure of the CTNet. It consists of two key steps: Context-independent utterance-level feature extraction and Context-dependent multimodal feature extraction. (see Fig. 2 for the multihead attention layer, Fig. 3 for the single-modal transformer and Fig. 4 for the cross-modal transformer).

features have different sequence length and exhibit “unaligned” nature. To obtain optimal mapping between these features, we propose a transformer-based model-level fusion strategy in this paper. On the one hand, our method captures temporal dependencies among unimodal features by a transformer-based structure, called “single-modal transformer” (as shown in Fig. 3). On the other hand, our method learns cross-model interactions on the unaligned multimodal features by another transformer-based structure, called “cross-modal transformer” (as shown in Fig. 4). To model speaker-sensitive interactions, we take speaker embeddings as additional inputs. To dynamically focus on more trustful modalities, we propose to fuse multimodal features via the attention mechanism, called “Audio-Text-Speaker Fusion (ATS-Fusion).” Finally, we propose to use the bi-directional GRU layer to consider the contextual information in both directions, in combination with the multi-head attention mechanism to highlight the important contextual utterances. These layers are marked as “Multi-Head Attention based bi-directional GRU (MHA-GRU).”

Different from previous conversational emotion recognition strategies [9]–[11], we propose to use the transformer-based structure and ATS-Fusion to fuse multimodal features, thus enabling us to effectively model intra-modal and cross-modal interactions. Different from previous word-level multimodal fusion strategies [20]–[22], we propose to use MHA-GRU component to model context-sensitive dependencies and additional speaker

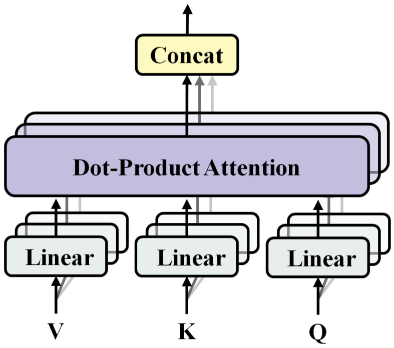


Fig. 2. Multihead attention mechanism: it consists of several dot-product attention layers running in parallel.

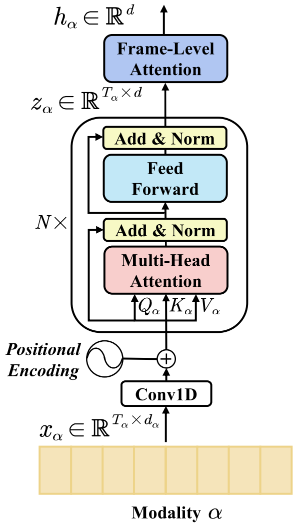

$h_\alpha \in \mathbb{R}^d$
Frame-Level Attention
$z_\alpha \in \mathbb{R}^{T_\alpha \times d}$
Add & Norm
Feed Forward
$N\times$
Add & Norm
Multi-Head Attention
Q_α K_α V_α
Positional Encoding
Conv1D
$x_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$
Modality α

Fig. 3. Architecture of the single-modal transformer for modality α.

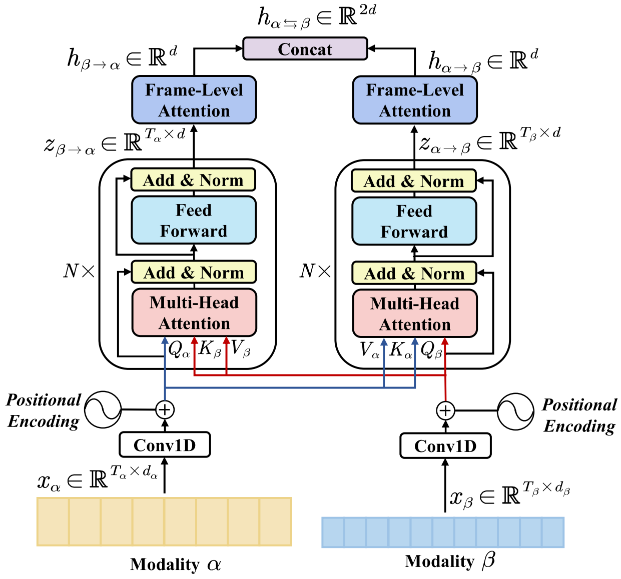

$h_{\alpha \leftrightarrows \beta} \in \mathbb{R}^{2d}$
Concat
$h_{\beta \to \alpha} \in \mathbb{R}^d$
$h_{\alpha \to \beta} \in \mathbb{R}^d$
Frame-Level Attention
Frame-Level Attention
$z_{\beta \to \alpha} \in \mathbb{R}^{T_\alpha \times d}$
$z_{\alpha \to \beta} \in \mathbb{R}^{T_\beta \times d}$
Add & Norm
Feed Forward
$N\times$
Add & Norm
Multi-Head Attention
Q_α K_β V_β
Add & Norm
Feed Forward
$N\times$
Add & Norm
Multi-Head Attention
V_α K_α Q_β
Positional Encoding
Positional Encoding
Conv1D
Conv1D
$x_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$
$x_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$
Modality α
Modality β

Fig. 4. Architecture of the cross-modal transformer between two time-series from modality α and β.

embeddings to model speaker-sensitive dependencies. The main contributions of this paper can be summarized as follows:

- We propose a novel framework for conversational emotion recognition, called CTNet. This model explores both intra-modal and cross-modal interactions of different modalities, filling the gap of current works on conversational emotion recognition.
- We systemically investigate and analyze the importance of each component in CTNet, including speaker embeddings, multimodal fusion and context-modeling strategies.
- Experimental results demonstrate that our CTNet outperforms other methods on two popular benchmark datasets, setting the new state-of-the-art record.

The remainder of this paper is organized as follows: In Section II, we provide a brief literature review on conversational emotion recognition. In Section III, we formalize the problem statement and describe our proposed method. In Section IV, we present the experimental datasets and setup in detail. The experimental results and analysis are illustrated in Section V. Finally, we conclude of the proposed work in Section VI.

II. Related Works

There are three key components in conversational emotion recognition: multimodal fusion, context-sensitive modeling and speaker-sensitive modeling. In this section, we will introduce the related works on these components, respectively.

Multimodal Fusion: Due to the complexity of emotion recognition, the single modality is difficult to meet the demand. To fuse multimodal features, previous works focused on two key steps: multimodal alignment and multimodal fusion [23]. Specifically, they first align multimodal features. Then fusion techniques are used to fuse these aligned features.

The target of multimodal alignment is to find correspondences between different modalities. There are mainly two strategies for multimodal alignment, namely explicit alignment and implicit alignment [23]: (1) *Explicit alignment:* This method assumes that multimodal features are already aligned in the utterance level [12], [17] or the word level [21], [22]. Previous works mainly focused on the utterance-level explicit alignment [12], [17]. Typically, Zadeh *et al.* [12] proposed a fusion method based on the utterance-level explicit alignment, called Tensor Fusion Network (TFN). TFN learned both intra-modality and inter-modality interactions via the Cartesian product. Later, Liu *et al.* [17] and Mai *et al.* [18] extended TFN. They attempted to improve efficiency and reduce trainable parameters. However, humans perceive emotions at the word level [19]. If we compress word-level features into utterance-level features, the temporal information and short-term information will be lost. To address this problem, a vast majority of works are explored toward word-level explicit alignment [20], [21]. For example, Gu *et al.* [21] leveraged word-level alignment between modalities and explored time-restricted cross-modal dynamic. (2) *Implicit alignment:* Explicit alignment only models multimodal interactions on the aligned parts, which ignores long-term cross-modal interactions. To overcome this problem, researchers propose implicit alignment that learns to latently align data during model training [22], [24]. Typically, Tsai *et al.* [22] proposed Multimodal Transformer to latently adapt features from one modality to another.

The target of multimodal fusion is to integrate information from different modalities. There are mainly three strategies for multimodal fusion, namely feature-level fusion, decision-level fusion and model-level fusion [14]. Feature-level fusion simply concatenates multimodal features into a joint feature vector at the input level. Although this method can improve the recognition performance, high-dimensional feature set may easily suffer from the problem of data sparseness [14]. Hence, the advantages of feature-level fusion will be limited. To address this problem, decision-level fusion uses unimodal decision values and fuses them by voting [25], averaging [26] or weighted sum [27]. However, this method ignores interactions and correlations between different modalities. To deal with these problems, a vast majority of works are explored toward model-level fusion. This method is a compromise between feature-level fusion and decision-level fusion, which is proposed to fuse intermediate representations of different modalities. In model-level fusion, prior works used kernel-based methods [28] and graphical models [29] to fuse multimodal features and showed performance improvement. Recently, more advanced approaches have utilized attention-based neural networks for model-level fusion [15], [16]. Typically, Poria *et al.* [16] fused multimodal features via the attention score for each modality, and achieved promising results for emotion recognition.

Context-sensitive Modeling: Humans perceive emotions not only through the current utterance but also from the contextual information in the surrounding utterances [7]. The contextual information can come from both local and distant contexts. Therefore, it is necessary to grasp the long-term contextual dependency for conversational emotion recognition. RNNs [9], [30], memory networks [5], [10] and self-attention mechanisms [31], [32] have been used in previous works. RNNs utilize the gating mechanisms to control the flow of information, thus modeling the long-term contextual dependency. Typically, Poria *et al.* [9] utilized a kind of RNNs, naming LSTMs [33] for context-sensitive modeling. However, RNNs suffer from unweighted influence from the context. That is to say, we cannot figure out the highest attended utterance of the current utterance. To deal with this problem, a vast majority of works are explored toward memory networks [5], [10]. Memory networks are often coupled with attention modules, allowing it to focus on only relevant memories in conversations. For example, Hazarika *et al.* [10] combined memory networks and attention modules for context-sensitive modeling in dyadic conversations. Memory networks repeatedly read and wrote to their memory cells, storing the contextual information in the conversational histories. Attention modules allowed the memory cells to only focus on the relevant context. Recently, the self-attention mechanism [34] has been verified to attend to longer sequences than many typical sequential models [34], [35]. This mechanism provides an opportunity for injecting the global context information into each input. Typical, Huang *et al.* [32] utilized the LSTM layer, in combination with the self-attention layer for emotion recognition.

Speaker-sensitive Modeling: A person's emotional state can be influenced by the interlocutor's behaviors in conversational dialogs [8]. Therefore, it is necessary to model speaker-sensitive interactions for conversational emotion recognition. The speaker-sensitive interactions are caused by two important factors: self-influence and inter-speaker influence [36]. Self-influence is a trait in conversations that the speaker's emotional state keeps consistent from one moment to another [37]. Inter-speaker influence is another trait in conversations that the interlocutor acts as an influencer in the speaker's emotional state [37]. To model the speaker-sensitive dependency, feature augmentation [20], [38], speaker-specific models [5], [30] and graph-based models [6] have been utilized in previous works. Feature augmentation is an approach that constructs speaker-dependent inputs to imitate the interaction patterns. For example, Chen *et al.* [20] explored the interlocutor influence when constructing the acoustic features. Later, Zhao *et al.* [38] extended the interaction strategies in [20]. Besides unimodal interaction strategies, they also explored and analyzed the influence of multimodal interaction strategies. In speaker-specific models, each speaker is modeled separately. Typically, Hazarika *et al.* [5] modeled separate contexts using two distinct GRUs for two speakers, and simultaneously used another GRU for explicit inter-speaker modeling. Majumder *et al.* [30] employed three GRUs to track individual speaker states, emotional states and global contexts during conversations. Recently, graph-based models have received growing attentions [39]. This model can well preserve the global structure information, and it is effective on tasks that have the rich relational structure. For example, Zhang *et al.* [6] utilized a kind of graph-based models, naming graph convolutional networks [40]. The nodes in the graph represented individual utterances or speakers. The edges between each utterance and its speaker bridged the speaker-sensitive dependency.

III. Proposed Method

In this section, we formalize the problem statement and describe our proposed CTNet in detail. Fig. 1 shows the overall structure of the CTNet, which consists of two key steps: (1) *Context-independent utterance-level feature extraction:* CTNet employs the single-modal transformer and cross-modal transformer to extract utterance-level unimodal representations and cross-modal representations, respectively. Meanwhile, CTNet uses speaker embeddings as additional inputs to model speaker-sensitive interactions; (2) *Context-dependent multimodal feature extraction:* We propose an Audio-Text-Speaker Fusion component (ATS-Fusion) for multimodal fusion and a Multi-Head Attention based bi-directional GRU component (MHA-GRU) for contextual feature extraction.

A. Problem Definition

Let us define a conversation $U = [u_1, u_2, \ldots, u_N]$, where u_j is the $j\mathrm{th}$ utterance in the conversation and N is the total number of utterances. And there are M participants $p_1, p_2, \ldots, p_M$ ($M \geqslant 2$ for the datasets we used) in the conversation. The task is to predict the discrete emotion labels (e.g., *happiness*, *sadness*, *neutral*, *anger*, *excitement*, *frustration*) of constituent utterances in the conversation.

B. Data Representation

Previous works [9]–[11] generally utilized utterance-level features for conversational emotion recognition. However, humans preserve emotions at the word level. If we integrate word-level features into utterance-level features, temporal information and short-term information will be lost. To deal with this problem, we utilize segment-level acoustic features and word-level lexical features as the inputs. To model speaker-sensitive interactions, utterance-level speaker embeddings are also utilized as the additional inputs. In this section, we will introduce these multimodal features in detail.

Furthermore, we conduct experiments with two popular feature normalization schemes, min-max normalization and z-normalization. However, we do not observe performance improvement with these schemes. The reason lies in that our CTNet has used layer normalization to normalize the activities of neurons [41], which exhibits a fast training convergence and a good result for emotion recognition. Therefore, no feature normalization scheme is utilized in this paper.

1) Segment-Level Acoustic Features: We extract segment-level acoustic features using the openSMILE toolkit [42]. Specifically, we use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) introduced by Eyben *et al.* [43]. We first extract low-level descriptors (such as energy, MFCC and pitch) with 10 ms frame shift and 25 ms frame size. Then we calculate statistics (such as mean, root quadratic mean) for 10 consecutive frames as our segment-level features. Totally, 88-dimensional segment-level acoustic features $x_A^{(j)} \in \mathbb{R}^{T_A \times d_A}$ are extracted for the utterance u_j. Here, T_A is used to represent the sequence length (corresponding to the number of segments in the waveform). $d_A = 88$ is used to represent the feature dimensions of acoustic features.

2) Word-Level Lexical Features: We extract word-level lexical features from the transcripts of the utterances. Specifically, we convert transcripts into the concatenation of vectors of the constituent words. These vectors are the public available word vectors trained on the Common Crawl and Wikipedia datasets [44]. Totally, 300-dimensional word-level lexical features $x_L^{(j)} \in \mathbb{R}^{T_L \times d_L}$ are extracted for the utterance u_j. Here, T_L is used to represent the sequence length (corresponding to the number of words in the transcript). $d_L = 300$ is used to represent the feature dimensions of lexical features.

3) Utterance-Level Speaker Embeddings: Each speaker has his own voice. With the help of the automatic speaker verification (ASV) system, we can verify the speaker's identity from his voice. In this paper, we extract speaker embeddings from the ASV system to distinguish speaker identities in conversations. Then these embeddings are utilized as additional inputs for speaker-sensitive modeling. Specifically, we employ the x-vector system [45], [46] to extract speaker embeddings from MFCCs via the Kaldi Speech Recognition Toolkit [47]. This system is trained on the NIST SREs dataset [48]. Totally, 512-dimensional utterance-level speaker embeddings $x_S^{(j)} \in \mathbb{R}^{d_S}$ are extracted for the utterance u_j. Here, $d_S = 512$ is used to represent the feature dimensions of speaker embeddings.

C. Context-Independent Utterance-Level Feature Extraction

Generally, the lexical features $x_L^{(j)} \in \mathbb{R}^{T_L \times d_L}$ and acoustic features $x_A^{(j)} \in \mathbb{R}^{T_A \times d_A}$ have different sequence length and exhibit "unaligned" nature. To capture temporal dependencies in unimodal features, we propose a "single-modal transformer" module. To learn optimal mapping between these unaligned features, we propose a "cross-modal transformer" module. The core component of the single-modal and cross-modal transformers is the multi-head attention mechanism [34]. In this section, we first introduce the structure of the multi-head attention mechanism. Then we present our method in detail.

1) Multi-Head Attention Mechanism: Compared with RNN-based models, multi-head attention splits one attention into several subspaces so that it can model the frame relationship in multiple aspects. Meanwhile, this mechanism directly builds the dependence between any frames, thus each input can consider the global context of the whole sequence. The overall structure of the multi-head attention is shown in Fig. 2.

We assume the query as $Q \in \mathbb{R}^{T_Q \times d_Q}$, the key as $K \in \mathbb{R}^{T_K \times d_K}$ and the value as $V \in \mathbb{R}^{T_V \times d_V}$. Here, T_Q, T_K and T_V are used to represent the sequence length of the query, key and value, respectively. d_Q, d_K and d_V are used to represent the feature dimensions of the query, key and value, respectively.

Firstly, we linearly project the query $Q \in \mathbb{R}^{T_Q \times d_Q}$, key $K \in \mathbb{R}^{T_K \times d_K}$ and value $V \in \mathbb{R}^{T_V \times d_V}$ m times with different linear projections [34]. Here, m is the number of heads:

$$\hat{Q} = Concat(QW_1^Q, \ldots, QW_i^Q, \ldots, QW_m^Q) \tag{1}$$

$$\hat{K} = Concat(KW_1^K, \ldots, KW_i^K, \ldots, KW_m^K) \tag{2}$$

$$\hat{V} = Concat(VW_1^V, \ldots, VW_i^V, \ldots, VW_m^V) \tag{3}$$

where $W_i^Q \in \mathbb{R}^{d_Q \times d_m}$, $W_i^K \in \mathbb{R}^{d_K \times d_m}$ and $W_i^V \in \mathbb{R}^{d_V \times d_m}$ are trainable parameters. Here, we assume the d_m as the output feature dimensions.

On each of these projected versions of queries $QW_i^Q \in \mathbb{R}^{T_Q \times d_m}$, keys $KW_i^K \in \mathbb{R}^{T_K \times d_m}$ and values $VW_i^V \in \mathbb{R}^{T_V \times d_m}$, we then perform the dot-product attention in parallel. Finally, the outputs of attention functions $head_i, i \in [1, m]$ are concatenated together as final values H_{head}, which are computed as follows (Note: T_K must be equal to T_V [34]):

$$head_i = \text{softmax}((QW_i^Q)(KW_i^K)^T)(VW_i^V) \tag{4}$$

$$H_{head} = Concat(head_1, \ldots, head_m) \tag{5}$$

where $head_i \in \mathbb{R}^{T_Q \times d_m}$ and $H_{head} \in \mathbb{R}^{T_Q \times md_m}$. Therefore, the sequence length of the outputs H_{head} is equal to the sequence length of the query Q.

2) Single-Modal Transformer: Based on the multi-head attention mechanism, we present the architecture of our single-modal transformer (as shown in Fig. 3). We assume the input sequence as $x_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$, where T_α and d_α are used to represent the sequence length and feature dimensions. To take the neighborhood information into account, we feed x_α into a

1-dimensional convolutional layer (Conv1D) [49], [50]:

$$\hat{x}_\alpha = Conv1D(x_\alpha) \tag{6}$$

where $\hat{x}_\alpha \in \mathbb{R}^{T_\alpha \times d}$. Here, we assume the d as the output feature dimensions.

To take the order of the sequence into consideration, information about the position of frames is also injected by triangle positional embeddings (PE) [34]:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10\,000^{2i/d}}\right) \tag{7}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10\,000^{2i/d}}\right) \tag{8}$$

where pos is the timestep index and i is the dimension index. The positional embeddings are added directly to the sequence, so that $Conv1D(x_\alpha) + PE$ encodes the elements' position information at every timestep. These features are treated as the query Q_α, the key K_α and the value V_α in Fig. 3.

Then we pass these features into a stack of N identical blocks. The outputs of the last block $z_\alpha \in \mathbb{R}^{T_\alpha \times d}$ are fed into the frame-level attention mechanism for utterance-level feature extraction. Specifically, each block contains two sub-modules: a multi-head attention layer and a feed-forward layer. We also employ a residual connection [51] around these two sub-modules, followed by layer normalization [41].

To extract utterance-level representations, the max-pooling layers and global-pooling layers are commonly utilized. However, these methods treat each frame equally and cannot select relevant frames for emotion recognition. To deal with this problem, we propose a attention mechanism to focus on informative words and attentive audio frames for each individual modality.

$$\alpha_{\text{fuse}} = \text{softmax}(z_\alpha W_z + b_z) \tag{9}$$

$$h_\alpha = \alpha_{\text{fuse}}^T z_\alpha \tag{10}$$

where $W_z \in \mathbb{R}^{d \times 1}$ and $b_z \in \mathbb{R}^1$ are trainable parameters. Here, $\alpha_{\text{fuse}} \in \mathbb{R}^{T_\alpha \times 1}$ and $h_\alpha \in \mathbb{R}^d$.

3) Cross-Modal Transformer: To model interactions on the unaligned multimodal sequences, we extend the single-modal transformer (Fig. 3) to the cross-modal transformer (Fig. 4). Specifically, we consider two modalities α and β, with two sequences from each of them denoted as $x_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $x_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$, respectively. Generally, T_α is different from T_β. We first equalize the dimensions of these features to size d using two Conv1D layers, and add PE (in 7∼8) to these sequences. To model every pair of cross-modal interactions, we pass the information from modality α to modality β (denoted as "$\alpha \rightarrow \beta$"). Meanwhile, we also pass the information from modality β to modality α (denoted as "$\beta \rightarrow \alpha$"). Specifically, in "$\alpha \rightarrow \beta$," x_α is transformed to the Key K_α and Value V_α to interact with the Query Q_β from the modality β via several attention blocks. The outputs of the last block are denoted as $z_{\alpha\rightarrow\beta}$. According to 1∼5, the sequence length of the outputs is equal to the sequence length of the Query. Therefore, $z_{\alpha\rightarrow\beta} \in \mathbb{R}^{T_\beta \times d}$. Then, we extract utterance-level representations $h_{\alpha\rightarrow\beta} \in \mathbb{R}^d$ using the frame-level attention mechanism (in 9∼10). Repeating such process, we can also gain utterance-level representations $h_{\beta\rightarrow\alpha} \in \mathbb{R}^d$ for "$\beta \rightarrow \alpha$". Finally, we concatenate these utterance-level representations together, denoted as $h_{\alpha\leftrightarrows\beta} = [h_{\alpha\rightarrow\beta}; h_{\beta\rightarrow\alpha}] \in \mathbb{R}^{2d}$.

D. Context-Dependent Multimodal Feature Extraction

In the conversational emotion analysis, the multimodal information provides us the benefit of understanding the way that humans express their emotions [14]. Meanwhile, the emotional state of the target utterance is temporally and contextually dependent to the surrounding utterances [9]. Therefore, we propose a framework to extract context-dependent multimodal features. This framework contains ATS-Fusion for multimodal fusion and MHA-GRU for contextual feature extraction. To verify the effectiveness of ATS-Fusion and MHA-GRU, we will compare our method with other approaches in Section V-E and Section V-F, respectively.

1) Multimodal Fusion (ATS-Fusion): Previous works [9]–[11] generally utilized a feature concatenation to generate multimodal representations. Although these works [9]–[11] can achieve promising results for conversational emotion recognition, high-dimensional concatenated features may easily suffer from the problem of data sparseness [14]. To deal with this problem, recent works are explored towards model-level fusion [14]. In this paper, we propose a model-level fusion strategy, called ATS-Fusion. Experimental results in Section V-E verify the effectiveness of our ATS-Fusion. Our method can achieve better recognition performance than feature concatenation.

As illustrated in Fig. 1, we extract utterance-level acoustic features $h_A^{(j)} \in \mathbb{R}^d$, lexical features $h_L^{(j)} \in \mathbb{R}^d$, cross-modal features $h_{A\leftrightarrows L}^{(j)} \in \mathbb{R}^{2d}$ and speaker embeddings $x_S^{(j)} \in \mathbb{R}^{d_S}$ for each utterance u_j (where u_j is the $j\text{th}$ utterance in the conversation U). Then we equalize feature dimensions of all four inputs to size d_2 and concatenate them together:

$$u_{cat}^{(j)} = [h_A^{(j)} W_A; h_L^{(j)} W_L; h_{A\leftrightarrows L}^{(j)} W_{AL}; x_S^{(j)} W_S] \tag{11}$$

where $W_A \in \mathbb{R}^{d \times d_2}$, $W_L \in \mathbb{R}^{d \times d_2}$, $W_{AL} \in \mathbb{R}^{2d \times d_2}$ and $W_S \in \mathbb{R}^{d_S \times d_2}$ are feature embedding matrices for acoustic features, lexical features, cross-modal features and speaker features, respectively. Here, $u_{cat}^{(j)} \in \mathbb{R}^{d_2 \times 4}$.

Different modalities have different contributions in emotion recognition. To focus on important modalities, we calculate the attention scores $\alpha_{att} \in \mathbb{R}^{1\times 4}$ over four inputs. The final multimodal features $f^{(j)} \in \mathbb{R}^{d_2}$ are generated as follows:

$$P_F = tanh(W_F u_{cat}^{(j)}) \tag{12}$$

$$\alpha_{att} = \text{softmax}(w_F^T P_F) \tag{13}$$

$$f^{(j)} = u_{cat}^{(j)} \alpha_{att}^T \tag{14}$$

where $W_F \in \mathbb{R}^{d_2 \times d_2}$ and $w_F \in \mathbb{R}^{d_2 \times 1}$ are trainable parameters. Here, $P_F \in \mathbb{R}^{d_2 \times 4}$. This multimodal representation is generated for all utterances in the conversation, marked as $F = [f^{(1)}, f^{(2)}, \ldots, f^{(N)}]$, where $F \in \mathbb{R}^{N \times d_2}$.

2) Contextual Feature Extraction (MHA-GRU): Inspired by the success of RNNs for context-sensitive modeling [9], [30], we extend RNN-based models and propose MHA-GRU to extract

TABLE I
EMOTIONAL DISTRIBUTION IN THE IEMOCAP AND MELD DATASETS

Datasets	Emotions								
	Anger	*Happiness/Joy*	*Sadness*	*Neutral*	*Excitement*	*Frustration*	*Fear*	*Surprise*	*Disgust*
IEMOCAP	1103	648	1084	1708	1041	1849	—	107	—
MELD	1607	2308	1002	6436	—	—	358	1636	361

contextual features. It consists of a bi-directional GRU layer and a multi-head attention layer. Experimental results on Section V-F verify the effectiveness of our MHA-GRU.

GRUs [52] are the gating mechanisms in RNNs. Since each GRU cell consists of an update gate and a reset gate to control the flow of information, it is capable of modeling long-term dynamic dependencies. As for the bi-directional GRU layer, the inputs $F = [f^{(1)}, f^{(2)}, \ldots, f^{(N)}]$ are given to two GRU components moving in opposite directions. For each timestep t, each GRU takes the hidden state from the previous timestep, along with the input from the current timestep $f^{(t)}$, and outputs a new hidden state. The final representations $h^{(t)}$ are obtained by concatenating hidden representations from both directions. Outputs are generated for all utterances in the conversation, marked as $H = [h^{(1)}, h^{(2)}, \ldots, h^{(N)}]$, where $H \in \mathbb{R}^{N \times d_2}$.

To highlight the important contextual utterances for emotion analysis, the outputs of the bi-directional GRU layer $H \in \mathbb{R}^{N \times d_2}$ are passed into a multi-head attention layer. Specifically, H is transformed to the key, query and value in Fig. 2, followed with 1$\sim$5 for high-level feature extraction. Outputs are generated for all utterances in the conversation, marked as $R = [r^{(1)}, r^{(2)}, \ldots, r^{(N)}]$, where $R \in \mathbb{R}^{N \times d_2}$.

E. Model Training

The final outputs R are fed into two fully-connected layers with a residual connection [51], followed by a softmax layer to calculate C emotion-class probabilities:

$$Z = R + ReLU(RW_R + b_R) \tag{15}$$

$$P = \mathrm{softmax}(ZW_Z + b_Z) \tag{16}$$

where $W_R \in \mathbb{R}^{d_2 \times d_2}$, $b_R \in \mathbb{R}^{d_2}$, $W_Z \in \mathbb{R}^{d_2 \times C}$ and $b_Z \in \mathbb{R}^C$ are trainable parameters. Here, $Z \in \mathbb{R}^{N \times d_2}$ and $P \in \mathbb{R}^{N \times C}$. We pick the most likely emotion class as the predicted label:

$$\hat{y}^{(j)} = argmax(P^{(j)}) \tag{17}$$

where $P^{(j)} \in \mathbb{R}^C$ and $\hat{y}^{(j)} \in \mathbb{R}^1$ are the emotion-class probabilities and the predicted label for the utterance u_j, respectively. We choose the categorical cross-entropy loss function during training. The calculation formula is shown as follows:

$$L = -\frac{1}{\sum_{i=1}^{M} L_i} \sum_{i=1}^{M} \sum_{j=1}^{L_i} \sum_{c=1}^{C} y_{i,c}^{(j)} log_2(\hat{y}_{i,c}^{(j)}) \tag{18}$$

where M is the number of conversations and L_i is the number of utterances in the $i\mathrm{th}$ conversation. $y_{i,c}^{(j)} \in \mathbb{R}^1$ and $\hat{y}_{i,c}^{(j)} \in \mathbb{R}^1$ are the original output and the predicted output of class c for the $j\mathrm{th}$ utterance in the $i\mathrm{th}$ conversation, respectively.

IV. EXPERIMENTAL DATABASES AND SETUP

A. Corpus Description

IEMOCAP [53] and MELD [54] are two popular benchmark datasets for conversational emotion recognition. A comparison of these datasets is listed in Table I and Table II. In Table I, we present the emotional distribution in the IEMOCAP and MELD datasets. In Table II, we summarize other statistical values (such as the number of speakers, the number of dialogues and average duration of an utterance) of the training, validation and testing sets in these datasets.

The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) consists of dyadic conversations, spanning 12.46 hours [53]. There are five sessions and a pair of skilled actors are grouped in a single session. All the conversations are split into small utterances, which are annotated by at least three annotators using the following discrete categories: *anger*, *happiness*, *sadness*, *neutral*, *excitement*, *frustration*, *fear*, *surprise* and *other*. To compare with state-of-the-art frameworks, we use their dataset split manner and conduct experiments on the 4-way [9], [10] and the 6-way conditions [5]. In the 4-way condition [9], [10], we consider the first four categories, where *happiness* and *excitement* categories are merged into the single *happiness* category. In the 6-way condition [5], we consider the first six categories. Followed with previous works [5], [9], [10], utterances from the first 8 speakers are used as the training and validation sets. While utterances from other speakers are used as the testing set. Since no predefined train/val split is provided in the IEMOCAP dataset, we use 10% of training dialogues as a held-out validation set for hyper-parameter tuning.

The Multi-modal EmotionLines Dataset (MELD) is a multimodal dataset for conversational emotion recognition, spanning 13.7 hours of various dialogue scenarios [54]. This dataset is created by extending the EmotionLines dataset [55] for the multimodal scenario. To bring conversations close to real-word dialogues, this dataset contains dialogues from the *Friends* TV shows. All the conversations are split into small utterances, which are annotated by three annotators using the following discrete categories: *anger*, *joy*, *sadness*, *neutral*, *disgust*, *fear* and *surprise*. Then a majority voting scheme is applied to select the final emotion label for each utterance. Different from IEMOCAP that contains dyadic conversations, MELD is a multi-party dataset where two or more speakers are involved in a conversation. To compare with state-of-the-art frameworks, we conduct experiments using the train/val/test splits in [6]. More details can be found in Table II.

B. Implementation Details

We use the validation set for hyper-parameter turning. To optimize the parameters, we use the Adam optimization scheme

TABLE II
Training, Validation and Testing Data Statistics in the 4-Way IEMOCAP, the 6-Way IEMOCAP and the MELD Datasets. No Predefined train/val Split Is Provided in the IEMOCAP Dataset, Hence We use 10% of the Training Dialogues as Validation Split

Statistics	IEMOCAP (4-way)			IEMOCAP (6-way)			MELD		
	train	val	test	train	val	test	train	val	test
Number of speakers	8		2	8		2	260	47	100
Number of dialogues	120		31	120		31	1039	114	280
Number of utterances	4290		1241	5810		1623	9989	1109	2610
Duration of all utterances	5.5h		1.6h	7.4h		2.1h	10.0h	1.1h	2.6h
Average duration of an utterance	4.6s		4.5s	4.6s		4.6s	3.6s	3.6s	3.6s
Average number of utterance per dialogue	35.8		40.0	48.4		52.4	9.6	9.7	9.3
Average/Max number of speakers per dialogue	2.0/2		2.0/2	2.0/2		2.0/2	2.7/9	3.0/8	2.6/8
Average/Max number of emotions per dialogue	2.4/4		2.4/4	3.5/6		3.4/6	3.3/7	3.3/7	3.2/6

with a learning rate of 0.001. We train the models for a maximum of 150 epochs, and stop training if the validation loss does not decrease for 10 consecutive epochs. L_2 regularization with the weight 0.00 001 and Dropout [56] with rate $p = 0.5$ are also used to alleviate over-fitting problems. To implement the proposed architecture in Tensorflow [57], all utterances of a mini-batch must have the same length. Meanwhile, different conversations of the same mini-batch need to have the same number of utterances. In view of variable-length utterances, we pad the utterances of the same mini-batch to the equal-length inputs, and the mini-batch size is set as 20. As the conversations do not have the same number of utterances, we pad the conversations of the same mini-batch to have the same number of utterances, and the mini-batch size is set as 20. To avoid the proliferation of noise within the network, bit masking is done on these padded utterances to eliminate their effect in the network.

Hyper-parameters are decided using a Random Search [58]. Based on the validation performance, we set hyper-parameters as follows: In the single-modal and cross-modal transformers, Conv1D layers map acoustic and lexical features into the fixed dimension of size $d = 30$, followed by 5 multi-head attention blocks (with 30 dimensional states and 5 attention heads). The outputs of the single-modal and cross-modal transformers are set to be $d = 30$ and $2d = 60$, respectively. ATS-Fusion equals the feature dimensions of acoustic features, lexical features, cross-modal features and speaker embeddings to size $d_2 = 100$. MHA-GRU contains a bi-directional GRU layer (50 units for each GRU component) and a multi-head attention layer (with 100 dimensional states and 4 attention heads).

C. Evaluation Metrics

In this section, we evaluate the performance of emotion recognition using the following four metrics: 1) weighted average accuracy (WAA); 2) unweighted average accuracy (UAA); 3) weighted average F1 (WAF1); 4) unweighted average F1 (UAF1). As shown in Table I, the inherent imbalance exists in the IEMOCAP and MELD datasets. In these four metrics, the last three metrics are commonly used when dealing with imbalanced classes [59], [60]. Therefore, we choose the last three metrics (UAA, WAF1 and UAF1) as our primary evaluation metrics and WAA as our secondary evaluation metric. These four metrics are calculated as follows:

Let there be C emotion classes in the dataset, and N_j denotes the number of samples of class $j, j \in [1, C]$. Accuracy_j and $F1_j$ represent the classification accuracy (or called Recall) and the F1 score of class j, respectively.

1) Weighted average accuracy (WAA) is a weighted mean accuracy over different emotion classes with weights proportional to the number of utterances in a class [60]:

$$WAA = \frac{\sum_{j=1}^{C} N_j * \text{Accuracy}_j}{\sum_{j=1}^{C} N_j} \tag{19}$$

2) Unweighted average accuracy (UAA) is defined as a mean of accuracies for different emotion categories [60]:

$$UAA = \frac{1}{C} \sum_{j=1}^{C} \text{Accuracy}_j \tag{20}$$

3) Weighted average F1 (WAF1) is a weighted mean F1 over different emotion categories with weights proportional to the number of utterances in a particular emotion class [60]:

$$WAF1 = \frac{\sum_{j=1}^{C} N_j * F1_j}{\sum_{j=1}^{C} N_j} \tag{21}$$

4) Unweighted average F1 (UAF1) is defined as a mean of F1-scores for different emotion categories [60]:

$$UAF1 = \frac{1}{C} \sum_{j=1}^{C} F1_j \tag{22}$$

D. Baseline Models

For comparison, we implement following state-of-the-art baseline approaches to comprehensively evaluate the performance of our proposed method.

SVM-ensemble [61]: This is a strong benchmark model, which is based on the automatically generated ensemble of trees with binary support vector machines (SVM) classifiers in the tree nodes. This model is context-independent as it does not use information from contextual utterances.

Memnet [62]: This is an end-to-end memory network. This model stores the historical context using the memories and updates in a multi-hop fashion. Finally, the outputs from the memory network are used for emotion recognition.

BC-LSTM [9]: To learn context-aware utterance-level representations, this model captures contextual content from the surrounding utterances using a bi-directional LSTM layer [33]. Finally, these context-aware representations are used for emotion classification. This model is speaker-independent as it does not model any speaker-level dependency.

TFN [12]: This model learns both the intra-modality and inter-modality dynamics in an end-to-end manner. Inter-modality dynamics are modeled by Tensor Fusion, which explicitly aggregates unimodal, bimodal and trimodal interactions. Intra-modality dynamics are modeled through Modality Embedding Subnetworks. Different from **BC-LSTM**, the contextual utterances are not considered.

MFN [13]: This model uses the multi-view sequential learning to model view-specific interactions and cross-view interactions. Similar to **TFN**, this approach considers neither context-sensitive dependency nor speaker-sensitive dependency. This is the state-of-the-art approach for multi-modal fusion.

CMN [10]: This model learns separate contexts for both speaker and listener using two distinct GRUs [52]. These contexts are stored as memories and combined with the test utterance via the attention mechanism. This model is a pioneer to explore inter-speaker interactions in conversations. In our implementation, we distinguish different speakers' utterances by distinct GRUs.

ICON [5]: This model is an extension of **CMN**. This model learns separate contexts for both speaker and listener using two distinct GRUs [52]. And simultaneously, this model uses another GRU to track the overall conversational flow. In our implementation, we distinguish different speakers' utterances by distinct GRUs.

DialogueRNN [30]: This is a strong benchmark model for conversational emotion detection. This model employs three GRUs [52] to model the emotional context in conversations. The spoken utterances are fed into the global GRU and the party GRU. The global GRU is utilized to update context, while the party GRU is used to update the speaker state. Finally, the updated speaker state is fed into the emotion GRU, which models the emotional information for classification.

ConGCN [6]: This model symbolizes the conversations as a large heterogeneous graph. This model represents each utterance and each speaker as a node. To model context-sensitive dependencies, it uses an undirected edge between two utterances nodes from the same conversation. To model speaker-sensitive dependencies, it uses an undirected edge between an utterance node and its speaker node.

DialogueGCN [11]: This model learns context-sensitive and speaker-sensitive dependencies via graph-based convolutional neural networks. The nodes in the graph represent individual utterances. The edges between a pair of nodes represent the dependence between the speakers of those utterances, along with their relative positions in the conversation. This is the state-of-the-art method for emotion detection in conversations.

V. Results and Discussion

In this section, we first present the comparisons between CTNet and existing works. Then, we study the importance of each modality. Following that, we compare CTNet with other approaches on speaker embeddings, multimodal fusion and context-modeling strategies. Finally, we visualize the attention scores of MHA-GRU and present the distribution of distances between the target utterance and its $(2nd)$ highest attended utterance. In the following experiments, results are an average of 10 runs with varied weight initializations. We assert significance when $p < 0.05$ under McNemar's test.

A. Comparison With Existing Works

To verify the effectiveness of our proposed method, we compare our CTNet with state-of-the-art baseline approaches on the IEMOCAP (4-way), IEMOCAP (6-way) and MELD datasets. In this section, we first compare CTNet with existing works on the overall performance using WAA, WAF1, UAA and UAF1. Then, we perform comparisons on classification accuracies and F1 scores for each emotion category. Finally, to release the impact of input features, we also compare our CTNet with other approaches using the same set of features.

1) Comparison on Overall Performance: In Table III$\sim$V, we show the performance of different approaches to emotion detection on the IEMOCAP (4-way), the IEMOCAP (6-way) and MELD datasets. Experimental results demonstrate the effectiveness of our CTNet. For the IEMOCAP (4-way) dataset, our method achieves the new state-of-the-art record, 83.6% on WAA and 83.3% in UAA, which shows an absolute improvement of 6.0% on WAA and 4.1% on UAA over currently advanced approaches (as shown in Table III). For the IEMOCAP (6-way) dataset, our method succeeds over currently advanced approaches by 4.0% on WAA, 4.0% on WAF1, 6.1% on UAA and 5.0% on UAF1 (as shown in Table IV). For the MELD dataset, our method succeeds over currently advanced approaches by 1.1% on WAF1 and 1.4% on UAF1 (as shown in Table V). These enhancements are caused by the following reasons:

1) Experimental results in Table III$\sim$V show that the proposed method outperforms **SVM-ensemble** and **BC-LSTM**. It is reasonable because these methods ignore speaker-sensitive dependencies, while our proposed method considers speaker-sensitive dependencies via additional speaker embeddings. These Speaker embeddings are necessary for capturing the self-influence and inter-speaker influence [36] among participants. Therefore, it is necessary to consider speaker-sensitive dependencies for emotion detection.

2) Experimental results in Table IV$\sim$V show that the proposed method outperforms **TFN** and **MFN** by 5.8%$\sim$7.6% on WAF1 and 8.7%$\sim$9.4% on UAF1. We argue that it is because **TFN** and **MFN** ignore context-sensitive dependency, while our proposed method considers this dependency via MHA-GRU. This module captures temporal dependencies via the bi-directional GRU layer and highlights the important context via the multi-head attention layer. These results prove that modeling contextual dependencies among utterances can improve performance for emotion recognition.

3) Experimental results in Table III$\sim$V show that our CTNet succeeds over **CMN**, **ICON** and **ConGCN**. The reason lies in that these methods generate the multimodal representation of an utterance through a simple feature concatenation, which

TABLE III
PERFORMANCE OF DIFFERENT APPROACHES ON THE IEMOCAP (4-WAY) TESTING SET. ALL RESULTS USE MULTIMODAL FEATURES. WE PRESENT SEPARATE EMOTION CATEGORY CLASSIFICATION ACCURACIES. MEANWHILE, WE ALSO REPORT SCORES USING WAA AND UAA. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

Methods	IEMOCAP: 4-way Emotion Categories					
	Happy	Sad	Neural	Angry	WAA	UAA
SVM-ensemble[a]	72.4	61.9	58.1	73.1	69.5	66.4
BC-LSTM[a]	74.2	76.5	66.3	75.7	74.3	73.2
Memnet[a]	75.0	76.4	66.5	81.6	75.1	74.9
CMN[a]	81.8	77.7	67.3	**89.9**	77.6	79.2
CTNet (ours)	**83.5**	**86.1**	**83.6**	80.0	**83.6**	**83.3**

[a]results come from [10].

TABLE IV
PERFORMANCE OF DIFFERENT APPROACHES ON THE IEMOCAP (6-WAY) TESTING SET. ALL RESULTS USE MULTIMODAL FEATURES. WE PRESENT SEPARATE EMOTION CATEGORY CLASSIFICATION ACCURACIES AND F1 SCORES. AND WE REPORT SCORES USING WAA, WAF1, UAA, AND UAF1. BOLD FRONT REPRESENTS THE BEST PERFORMANCE; ACC. = ACCURACY

Methods	IEMOCAP: 6-way Emotion Categories															
	Happy		Sad		Neural		Angry		Excited		Frustrated		WAA	WAF1	UAA	UAF1
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1				
Memnet[a]	24.4	33.0	60.4	69.3	56.8	55.0	67.1	66.1	65.2	62.3	68.4	63.0	59.9	59.5	57.1	58.1
BC-LSTM[a]	25.5	35.6	58.6	69.2	56.5	53.5	70.0	66.3	58.8	61.1	67.4	62.4	59.8	59.0	56.7	58.2
TFN[a]	23.2	33.7	58.0	68.6	56.6	55.1	69.1	64.2	63.1	62.4	65.5	61.2	58.8	58.5	56.3	57.7
MFN[a]	24.0	34.1	65.6	70.5	55.5	52.1	72.3	66.8	64.3	62.1	67.9	62.5	60.1	59.9	58.5	58.3
CMN[a]	25.7	32.6	66.5	72.9	53.9	56.2	67.6	64.6	69.9	67.9	71.7	63.1	61.9	61.4	59.6	59.8
ICON[a]	23.6	32.8	70.6	74.4	59.9	60.6	68.2	**68.2**	72.2	68.4	**71.9**	**66.2**	64.0	63.5	61.5	62.0
CTNet (ours)	**47.9**	**51.3**	**78.0**	**79.9**	**69.0**	**65.8**	**72.9**	67.2	**85.3**	**78.7**	52.2	58.8	**68.0**	**67.5**	**67.6**	**67.0**

[a]results come from [5].

TABLE V
PERFORMANCE OF DIFFERENT APPROACHES ON THE MELD (7-WAY) TESTING SET. ALL RESULTS USE MULTIMODAL FEATURES. WE PRESENT THE F1 SCORE OF EACH EMOTION CLASS. MEANWHILE, WE ALSO REPORT SCORES USING WAF1 AND UAF1. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

Methods	MELD: 7-way Emotion Categories								
	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	WAF1	UAF1
MFN[a]	40.8	0.0	0.0	46.7	76.2	13.7	40.7	54.7	31.2
BC-LSTM[a]	44.5	0.0	0.0	49.7	76.4	15.6	48.4	56.8	33.5
CMN[a]	44.7	0.0	0.0	44.7	74.3	23.4	47.2	55.5	33.5
ICON[a]	44.8	0.0	0.0	50.2	73.6	23.2	50.0	56.3	34.5
DialogueRNN[a]	45.6	0.0	0.0	53.2	73.2	24.8	51.9	57.0	35.5
ConGCN[a]	**46.8**	10.6	8.7	53.1	76.7	28.5	50.3	59.4	39.2
CTNet (ours)	44.6	**11.2**	**10.0**	**56.0**	**77.4**	**32.5**	**52.7**	**60.5**	**40.6**

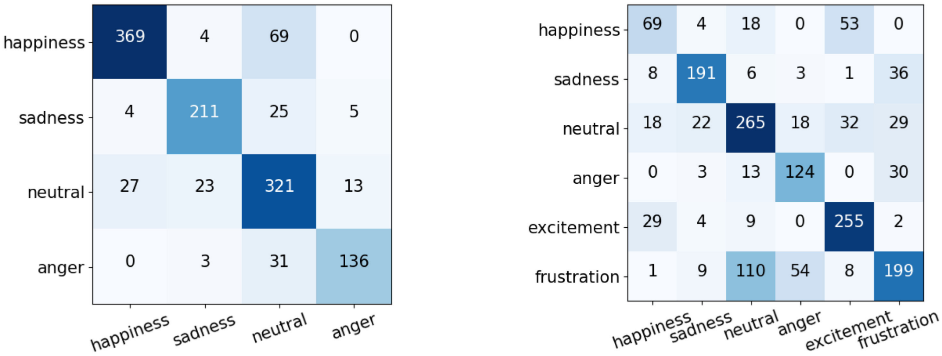
[a]results come from [6].

may suffer from the curse of dimensionality [14]. To deal with this problem, our method uses an attention-based model-level fusion strategy, ATS-Fusion. In the meantime, our CTNet uses single-modal and cross-modal transformers to capture long-term interactions on unaligned multimodal sequences. These results highlight the importance of multimodal fusion and multimodal interactions for emotion recognition.

2) Comparison on Performance for Each Class: In this section, we present classification accuracies and F1 scores for each emotion category in Table III∼V. We also visualize the confusion matrices of the testing set in Fig. 5.

For the IEMOCAP (4-way) dataset, experimental results in Table III demonstrate that our CTNet achieves improvements on classification accuracies in individual emotion recognition tasks in most cases, especially for the *happiness*, *sadness* and *neutral* emotions. In the meantime, we find *happiness*, *sadness* and *anger* emotions can be confused with the *neutral* emotion in some cases (as shown in Fig. 5(a)). The reason lies in that the *neutral* emotion is over-represented in the IEMOCAP (4-way) dataset [53]. As a result, CTNet tends to learn more about the *neutral* emotion.

For the IEMOCAP (6-way) dataset, the improvements on classification performance can be seen for most emotion categories over currently advanced approaches (as shown in Table IV). Specifically, CTNet outperforms other methods on the *happiness*, *sadness*, *neutral* and *excitement* emotions. This can be attributed to the improved ability of the model to identify features which are relevant for discriminating emotions. Meanwhile, we find *sadness*, *neutral* and *anger* emotions can be confused with the *frustration* emotion in some cases (as shown in Fig. 5(b)). Such phenomenon is related to imbalanced class distribution [53], as the majority of the utterances are labeled



(a)

(b)

Fig. 5. Visualization the confusion matrices of CTNet with multi-modality: (a) Confusion matrix of the testing set on the IEMOCAP (4-way) dataset; (b) Confusion matrix of the testing set on the IEMOCAP (6-way) dataset.

as the *frustration* emotion in the IEMOCAP (6-way) dataset. In the meantime, the *happiness* emotion can be confused with the *excitement* emotion in some cases (as shown in Fig. 5(b)). This result is the same with previous works [53]. As *excitement* and *happiness* emotions are close in the activation and valence domains [53], people have similar perception and interpretation of these emotions. Therefore, these emotions may be misclassified in some cases.

For the MLED dataset, experimental results in Table V demonstrate that our CTNet outperforms other methods on F1 scores for most emotion categories (including *joy*, *sadness*, *neutral*, *disgust*, *fear* and *surprise*). A slight decrease in the F1-score of the *anger* emotion can be attributed to better generalization for other emotions. Meanwhile, *disgust* and *fear* emotions are under-represented in MELD [54] (as shown in Table I). As a result, CTNet generally tends to learn less about them and has lower accuracies for these emotion categories.

3) Comparison Under the Same Set of Features: To release the impact of input features, we compare CTNet with other approaches (including **BC-LSTM** [9], **DialogueRNN** [30] and **DialogueGCN** [11]) using the same set of features. Details about our multimodal features can be found in Section III-B. We generate the multimodal representations for prior works [9], [11], [30] through feature concatenation. Since inherent imbalance exists in the IEMOCAP and MELD datasets [53], [54], we choose WAF1 as our primary evaluation metric and WAA as our secondary evaluation metric. Experimental results in Table VI demonstrate the effectiveness of our CTNet. For the IEMOCAP (4-way) dataset, our method succeeds over other methods by 2.1% on WAF1 and 1.7% on WAA. For the IEMOCAP (6-way) dataset, our method succeeds over other methods by 6.2% on WAF1 and 6.4% on WAA. For the MELD dataset, our method succeeds over other methods by 4.7% on WAF1 and 2.2% on WAA. The reason lies in that **BC-LSTM** [9], **DialogueRNN** [30] and **DialogueGCN** [11] utilize a simple feature concatenation for multimodal fusion, which has limitations in modeling intra-modal and cross-modal interactions. To deal with this problem, our CTNet proposes to use the transformer-bases structure and ATS-Fusion to fuse multimodal features. Experimental results demonstrate the effectiveness of our proposed method.

TABLE VI
PERFORMANCE OF DIFFERENT APPROACHES USING THE SAME SET OF FEATURES AS CTNET (IN SECTION III-B). WE REPORT SCORES USING WAA AND WAF1 ON THE TESTING SET. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

	IEMOCAP (4-way)		IEMOCAP (6-way)		MELD	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
BC-LSTM[a]	76.6	76.8	58.7	58.4	56.9	50.8
DialogueRNN[a]	75.9	76.1	55.0	53.1	58.0	54.1
DialogueGCN[a]	81.9	81.7	61.6	61.3	59.8	55.8
CTNet (Ours)	**83.6**	**83.8**	**68.0**	**67.5**	**62.0**	**60.5**

[a] Implementation available at https://github.com/SenticNet/conv-emotionhttps://github.com/SenticNet/conv-emotion.

TABLE VII
UNIMODAL AND BIMODAL RESULTS OF THE PROPOSED CTNET. HERE, L AND A DENOTES LEXICAL AND ACOUSTIC MODALITIES, RESPECTIVELY. WE REPORT SCORES USING WAA AND WAF1 ON THE TESTING SET. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

	IEMOCAP (4-way)		IEMOCAP (6-way)		MELD	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
L	80.8	80.6	63.5	63.7	60.6	58.3
A	67.5	67.6	56.0	55.2	46.9	38.2
L+A	**83.6**	**83.8**	**68.0**	**67.5**	**62.0**	**60.5**

B. Importance of the Modalities

To investigate the importance of each modality, we conduct experiments to compare the performance among unimodal and bimodal results. Experimental results are listed in Table VII. Since inherent imbalance exists in the experimental datasets, we choose WAF1 as our primary evaluation metric and WAA as our secondary evaluation metric. For unimodal results, experimental results in Table VII show that the lexical modality achieves better performance than that of the acoustic modality in all cases, which indicates the importance of spoken language in conversational emotion recognition. This result is the same with previous works [6], [9]. The text tends to have lesser noisy signals compared with the audio, thus learning more effective features in the joint representations.

TABLE VIII
PERFORMANCE OF TRANSFORMER-BASED MULTIMODAL FUSION STRATEGIES. ALL RESULTS USE MULTIMODAL FEATURES. WE REPORT SCORES USING WAA AND WAF1 ON THE TESTING SET. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

	IEMOCAP (4-way)		IEMOCAP (6-way)		MELD	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
Without-SCT	72.6	72.1	58.4	57.2	58.1	52.7
With-SCT (Ours)	**83.6**	**83.8**	**68.0**	**67.5**	**62.0**	**60.5**

Furthermore, experimental results in Table VII demonstrate that bimodal results outperform unimodal results in all kinds of experiments. Bimodal results succeeds over the lexical modality by 2.2%∼3.8% on WAF1 and 1.4%∼4.5% on WAA. And bimodal results succeeds over the acoustic modality by 12.3%∼22.3% on WAF1 and 12.0%∼16.1% on WAA. Emotion is not only related to spoken words but also to nonverbal behaviors from the audio. Multimodal fusion incorporates the information from different modalities, which improves the performance of emotion recognition.

C. Effectiveness of Single and Cross-modal Transformers

To verify the effectiveness of our transformer-based multimodal fusion strategies, one comparison system is implemented. Experimental results are listed in Table VIII.

- *Without Single-modal and Cross-modal Transformers (Without-SCT):* It comes from CTNet, but removing the single-modal and cross-modal transformers. Specifically, to extract utterance-level features, we calculate mean values of segment-level acoustic features and word-level lexical features in the utterance. Then, we combine these utterance-level features with additional speaker embeddings via the ATS-Fusion layer.
- *With Single-modal and Cross-modal Transformers (With-SCT):* It is our proposed method that utilizes the single-modal and cross-modal transformers for multimodal fusion.

Experimental results in Table VIII demonstrate that **With-SCT** shows an absolute improvement of 3.9%∼11.0% on WAA and 7.8%∼11.7% on WAF1 over **Without-SCT**. The reason lies in that **With-SCT** uses the single-modal transformer to capture temporal dependencies for unimodal sequence, and employs the cross-modal transformer to learn interactions for unaligned multimodal sequences. However, **Without-SCT** calculates mean values of segment-level acoustic features and word-level lexical features, which has limitations in modeling inter-modal and intra-modal interactions. These results verify the effectiveness of our transformer-based multimodal fusion strategies for conversational emotion recognition.

D. Effectiveness of X-Vector Speaker Embeddings

To verify the effectiveness of our x-vector speaker embeddings, three comparison systems are implemented. Experimental results are listed in Table IX.

TABLE IX
PERFORMANCE OF DIFFERENT SPEAKER EMBEDDINGS ON THE TESTING SET. ALL RESULTS USE MULTIMODAL FEATURES. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

	IEMOCAP (4-way)		IEMOCAP (6-way)		MELD	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
Without-SE	79.5	78.9	63.1	63.4	60.2	57.8
Onehot	81.1	81.4	64.2	64.4	61.8	59.7
I-vector	83.0	83.3	66.6	67.1	**62.5**	60.0
X-vector (Ours)	**83.6**	**83.8**	**68.0**	**67.5**	62.0	**60.5**

TABLE X
PERFORMANCE OF DIFFERENT MULTIMODAL FUSION STRATEGIES. ALL RESULTS USE MULTIMODAL FEATURES. WE REPORT SCORES USING WAA AND WAF1 ON THE TESTING SET. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

	IEMOCAP (4-way)		IEMOCAP (6-way)		MELD	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
Add	80.7	81.2	66.2	66.5	60.5	59.0
Concatenate	81.4	81.5	63.0	62.2	60.3	58.1
Tensor Fusion	82.0	81.8	65.9	64.3	60.1	58.2
ATS-Fusion (Ours)	**83.6**	**83.8**	**68.0**	**67.5**	**62.0**	**60.5**

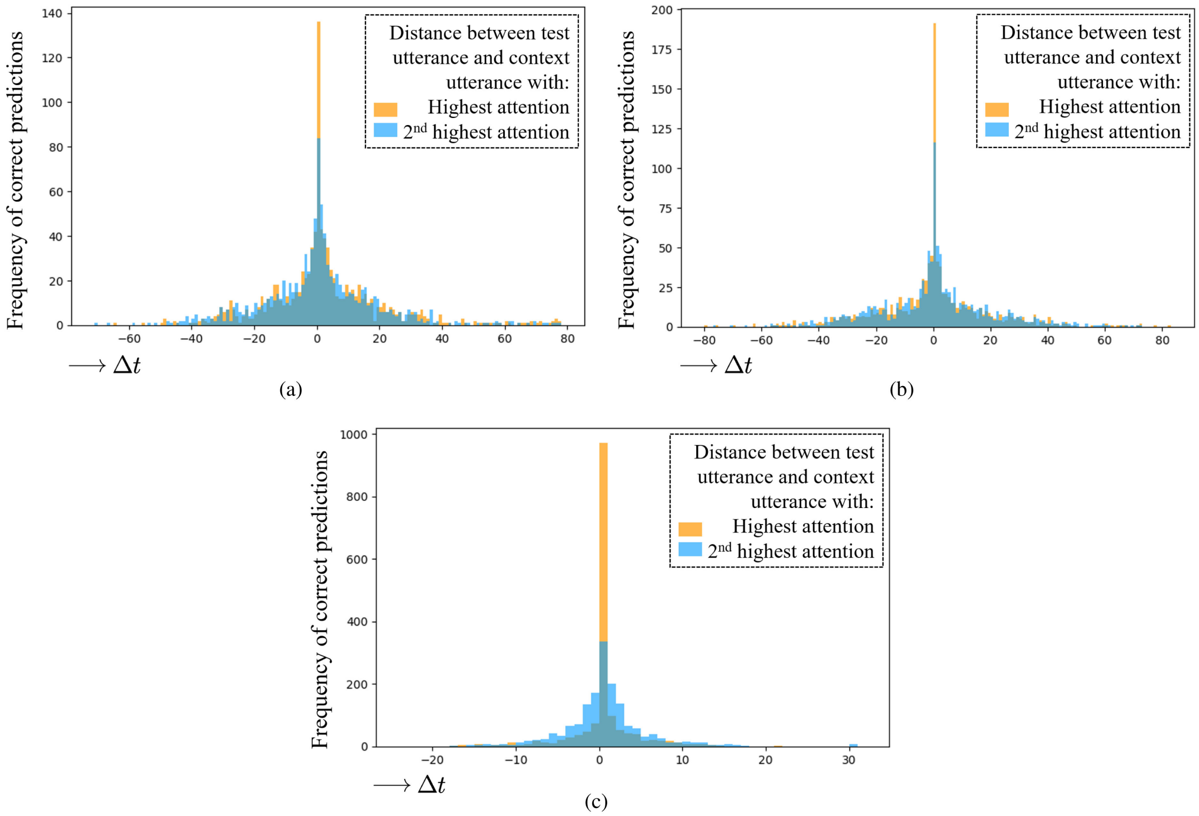
- *Without speaker embeddings (Without-SE):* It comes from the CTNet, but removing the speaker embeddings.
- *One-hot Encoding (Onehot):* It comes from the CTNet, but using the one-hot vector to identify speakers.
- *i-vector Encoding (I-vector):* Different from **Onehot**, we use i-vectors [63]–[65] to identify speakers.
- *x-vector Encoding (X-vector):* It is our proposed method that utilizes x-vectors for speaker identification.

Experimental results in Table IX demonstrate that **Without-SE** gains the worst performance in all cases. The speaker embeddings indicate the speaker's identity of each utterance, which are important for conversational emotion recognition [66]. Meanwhile, we find **X-vector** shows an absolute improvement of 0.8%∼3.2% on WAF1 over **Onehot**. Interestingly, there is no significant difference between **I-vector** and **X-vector**. Compared with **Onehot**, **I-vector** and **X-vector** transfer the knowledge from the speaker verification task to emotion recognition. Previous works [67] have verified that such process can improve emotion recognition performance.

E. Effectiveness of ATS-Fusion

In this section, we implement three comparison systems to verify the effectiveness of our ATS-Fusion. Experimental results are listed in Table X.

- *Add:* It comes from CTNet, but removing ATS-Fusion. Specifically, we first extract four representations after *Context-Independent Utterance-Level Feature Extraction* (as shown in Fig. 1). Then we equalize dimensions of these features via fully-connected layers. Finally, we simply add these multimodal representations together.
- *Concatenate:* Different from **Add**, we generate the multimodal representation through feature concatenation. Such strategy is often used in previous works [5], [10].



(a)

(b)

(c)

Fig. 6. Histogram of $\Delta t =$ distance between the target utterance and its $(2nd)$ highest attended utterance based on MHA-GRU's attention scores: (a) Histogram of IEMOCAP (4-way); (b) Histogram of IEMOCAP (6-way); (c) Histogram of MELD.

- *Tensor Fusion:* Different from **Add**, we generate the multimodal representation of an utterance through tensor fusion [12]. This method aggregates unimodal, bimodal and trimodal interactions via 3-fold Cartesian product.
- *ATS-Fusion:* It is our proposed method that uses ATS-Fusion for multi-modality fusion.

Experimental results in Table X demonstrate that our **ATS-Fusion** achieves the best performance compared with other fusion methods. For example, **ATS-Fusion** shows an absolute improvement of 1.0%∼2.6% on WAF1 over **Add**. The reason lies in that **Add** treats each modality equally and cannot select relevant modalities for emotion recognition. While **ATS-Fusion** can prioritize important modalities via the attention mechanism, thus improving the performance of emotion recognition. Meanwhile, experimental results in Table X demonstrate that **ATS-Fusion** shows an absolute improvement of 2.3%∼5.3% on WAF1 over **Concatenate**. The reason lies in that **Concatenate** generates multimodal representation via feature concatenation, which may suffer from the curse of dimensionality [14]. Compared with **Concatenate**, our method can deal with this problem by generating smaller-size multimodal features for emotion recognition. Additionally, **ATS-Fusion** shows an absolute improvement of 2.0%∼3.2% on WAF1 over **Tensor Fusion**. Compared with our method, **Tensor Fusion** is constrained by the exponential increase of cost in computation and memory introduced by using tensor representations, thus degrading the recognition performance. These results verify the effectiveness of **ATS-Fusion** for multimodal fusion.

TABLE XI
PERFORMANCE OF DIFFERENT CONTEXT-MODELING STRATEGIES. ALL RESULTS USE MULTIMODAL FEATURES. WE REPORT SCORES USING WAA AND WAF1 ON THE TESTING SET. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

	IEMOCAP (4-way)		IEMOCAP (6-way)		MELD	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
Without-CM	74.0	74.1	58.9	59.3	59.5	57.4
Uni-GRU	77.6	77.9	64.8	64.7	59.9	58.7
Bi-GRU	79.5	79.8	65.9	65.9	61.2	59.2
MHA-GRU (Ours)	**83.6**	**83.8**	**68.0**	**67.5**	**62.0**	**60.5**

F. Effectiveness of MHA-GRU

To verify the effectiveness of our MHA-GRU on context-sensitive modeling, three comparison systems are implemented. Table XI provides the results on this analysis.

- *Without context-modeling (Without-CM):* It comes from CTNet, but removing the MHA-GRU component. Specifically, we use a fully-connected layer for emotion recognition. Hence, no context-modeling strategy is utilized in this model.
- *Unidirectional GRU (Uni-GRU):* Different from **Without-CM**, we use the unidirectional GRU cells for context-modeling. Thus, an utterance can only get information from the utterances occurring before itself.
- *Bi-directional GRU (Bi-GRU):* Different from **Uni-GRU**, we use the bi-directional GRU cells for context-modeling.

Thus, an utterance can get information from the utterances occurring before and after itself.

- *MHA-GRU:* It is our proposed method that uses MHA-GRU for context-modeling, which contains a bi-directional GRU layer and a multi-head attention layer.

In Table XI, we find that **Without-CM** gains the worst performance in all cases. Such phenomenon reveals the importance of context-modeling strategies. Meanwhile, experimental results in Table XI demonstrate that our **MHA-GRU** shows an absolute improvement of 1.3%$\sim$4.0% on WAF1 and 0.8%$\sim$4.1% on WAA over **Bi-GRU**. Compared with **Bi-GRU**, **MHA-GRU** utilizes additional multi-head attention layer to highlight the important contextual utterances for emotion analysis of the target utterance, which is important for conversational emotion recognition. Additionally, we find that **Bi-GRU** achieves better performance than **Uni-GRU** in all kinds of experiments. The contextual information exists in both the historical and future utterances. Through considering the contextual information in both directions, **Bi-GRU** achieves better performance on emotion recognition. Therefore, our **MHA-GRU** that combines bi-directional GRU layer with the multi-head attention mechanism can achieve the best performance among other approaches.

G. Discussion on Contextual Distances

For all the test utterances correctly classified by CTNet, we summarize the distribution of distances between the target utterance and its ($2nd$) highest attended utterance in the conversation based on MHA-GRU's attention scores. Fig. 6 provides a summary of results for the IEMOCAP (4-way), IEMOCAP (6-way) and MELD datasets.

In Fig. 6, most of the correctly classified utterances focus on their local context. However, a significant portion is also present for the distant context. Compared with the highest attention, the dependence on distant context increases for the $2nd$ highest attention. These results are the same with previous works [30], which highlight the importance of the long-term emotional dependency. Meanwhile, we find that the contextual dependence exists both towards the historical utterances and the future utterances. Therefore, we need to consider the long-term contextual information in the whole conversation. These results support the necessity of MHA-GRU to model the long-term contextual information in conversations.

VI. Conclusion

In this paper, we propose the Conversational Transformer Network (CTNet), a multimodal multi-party framework for conversational emotion recognition. CTNet models inter-modality and intra-modality interactions for multimodal features. In the meantime, CTNet also considers context-sensitive and speaker-sensitive dependencies in the conversation. Experimental results on two popular benchmark datasets, IEMOCAP and MELD, demonstrate the effectiveness of CTNet. Our method sets the new state-of-the-art record for conversational emotion recognition. Additionally, experimental results on different modalities show the necessity of multimodal fusion. Meanwhile, the comparisons of speaker embeddings, multimodal fusion strategies and context-modeling strategies show the importance of each component in CTNet. Furthermore, through the discussion on contextual distances, we prove that capturing the long-term contextual dependency improves the performance of conversational emotion recognition.

In the future, besides acoustic and lexical modalities, we aim to further improve the classification accuracy using the visual information. Additionally, we will apply recently advanced approaches to transfer the knowledge from multimodal systems to unimodal systems, thus improving the performance for unimodal systems.

References

[1] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "Affect-lm: A neural language model for customizable affective text generation," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2017, pp. 634–642.

[2] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. Meng, "Towards discriminative representation learning for speech emotion recognition," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 5060–5066.

[3] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," in *Proc. ISCA Tut. Res. Workshop. Speech Emotion*, 2000, pp. 1–6.

[4] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[5] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proc. Confe. Empirical Methods Nat. Lang. Process.*, 2018, pp. 2594–2604.

[6] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2019, pp. 10–16.

[7] J. J. Gross and L. Feldman Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion Rev.*, vol. 3, no. 1, pp. 8–16, 2011.

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

[9] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 873–883.

[10] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 2122–2132.

[11] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2019, pp. 154–164.

[12] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[13] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[14] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Trans. Signal Inf. Process.*, vol. 3, 2014.

[15] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 571–575.

[16] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining.*, 2017, pp. 1033–1038.

[17] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[18] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 481–492.

[19] G. Aguilar, V. Rozgić, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 991–1002.

[20] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.

[21] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2225–2235.

[22] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[23] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[24] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2822–2826.

[25] E. Morvant, A. Habrard, and S. Ayache, "Majority vote of diverse classifiers for late fusion," in *Proc. Struct., Syntactic, Stat. Pattern Recognit. - Joint IAPR Int. Workshop*, 2014, pp. 153–162.

[26] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2016, pp. 160–170.

[27] M. Glodek *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2011, pp. 359–368.

[28] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, no. Jul, pp. 2211–2268, 2011.

[29] C. Sutton *et al.*, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2012.

[30] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6818–6825.

[31] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," in *Proc. Interspeech*, 2019, pp. 1936–1940.

[32] C. Huang, A. Trabelsi, and O. R. Zaïane, "Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert," in *Proc. 13th Int. Workshop Semantic Eval.*, SemEval, NAACL-HLT, 2019, pp. 49–53.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[35] P. J. Liu *et al.*, "Generating Wikipedia by summarizing long sequences," in *Proc. 6th Int. Conf. Learn. Representations*, 2018, pp. 1–18.

[36] M. W. Morris and D. Keltner, "How emotions work: The social functions of emotional expression in negotiations," *Res. Org. Behav.*, vol. 22, pp. 1–50, 2000.

[37] P. Kuppens, N. B. Allen, and L. B. Sheeber, "Emotional inertia and psychological maladjustment," *Psychol. Sci.*, vol. 21, no. 7, pp. 984–991, 2010.

[38] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proc. Audio/Vis. Emotion Challenge Workshop*, ACM. Oct. 2018, pp. 65–72.

[39] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[40] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017, pp. 1–13.

[41] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, 2016.

[42] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[43] F. Eyben *et al.*, "The geneva minimalistic acoustic parameter set (GEMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[44] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. Int. Conf. Lang. Res. Eval.*, 2018.

[45] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[46] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.

[47] D. Povey *et al.*, "The Kaldi speech recognition toolkit," *IEEE 2011 Workshop Automatic Speech Recognition Understanding*, IEEE Signal Processing Society, 2011.

[48] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Proc. Interspeech*, 2010, pp. 2726–2729.

[49] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 328–339, 1989.

[50] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[52] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2014, pp. 1724–1734.

[53] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.

[54] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[55] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. K. Huang, and L.-W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," in *Proc. 11th Int. Conf. Lang. Res. Eval.*, 2018.

[56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[57] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th {USENIX} Symp. Operating Syst. Des. Implementation*, 2016, pp. 265–283.

[58] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 281–305, 2012.

[59] Y. Yang, "An evaluation of statistical approaches to text categorization," *Inf. Retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[60] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[61] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf*, 2012, pp. 1–4.

[62] S. Sukhbaatar *et al.*, "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.

[63] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1-3, pp. 19–41, 2000.

[64] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2010.

[65] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Proc.*, vol. 1, 2006, pp. I–I.

[66] R. Zhang, A. Ando, S. Kobashikawa, and Y. Aono, "Interaction and transition model for speech emotion recognition in dialogue," in *Proc. Interspeech*, 2017, pp. 1094–1097.

[67] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7169–7173.

Zheng Lian received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016. He is currently working toward the Ph.D degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing, deep learning, and multimodal emotion recognition.

Bin Liu received the the B.S. and M.S. degrees from the Beijing Institute of Technology, Beijing, China, in 2007 and 2009, respectively, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include affective computing and audio signal processing.

Jianhua Tao (Member, IEEE) received the M.S. degree from Nanjing University, Nanjing, China, in 1996 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than eighty papers on major journals and proceedings including the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. His current research interests include speech recognition, speech synthesis and coding methods, humancomputer interaction, multimedia information processing, and pattern recognition. He is the Chair or Program Committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, NCMMSC, etc. He is also the Steering Committee Member for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, an Associate Editor for *Journal on Multimodal User Interface* and *International Journal of Synthetic Emotions*, and the Deputy Editor-in-Chief for *Chinese Journal of Phonetics*. He was the recipient of several awards from the important conferences, such as Eurospeech and NCMMSC.