A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu

***Abstract*—Emotion recognition in conversations (ERC), the task of recognizing the emotion of each utterance in a conversation, is crucial for building empathetic machines. Existing studies focus mainly on capturing context- and speaker-sensitive dependencies on the textual modality but ignore the significance of multimodal information. Different from emotion recognition in textual conversations, capturing intra- and inter-modal interactions between utterances, learning weights between different modalities, and enhancing modal representations play important roles in multimodal ERC. In this paper, we propose a transformer-based model with self-distillation (SDT) for the task. The transformer-based model captures intra- and inter-modal interactions by utilizing intra- and inter-modal transformers, and learns weights between modalities dynamically by designing a hierarchical gated fusion strategy. Furthermore, to learn more expressive modal representations, we treat soft labels of the proposed model as extra training supervision. Specifically, we introduce self-distillation to transfer knowledge of hard and soft labels from the proposed model to each modality. Experiments on IEMOCAP and MELD datasets demonstrate that SDT outperforms previous state-of-the-art baselines.**

***Index Terms*—Multimodal emotion recognition in conversations, intra- and inter-modal interactions, multimodal fusion, modal representation.**

I. Introduction

EMOTION recognition in conversations (ERC) aims to automatically recognize the emotion of each utterance in a conversation. The task has recently become an important research topic due to its wide applications in opinion mining [1], health care [2], and building empathic dialogue systems [3], etc. Unlike traditional emotion recognition (ER) on context-free sentences, modeling context- and speaker-sensitive dependencies lie at the heart of ERC.

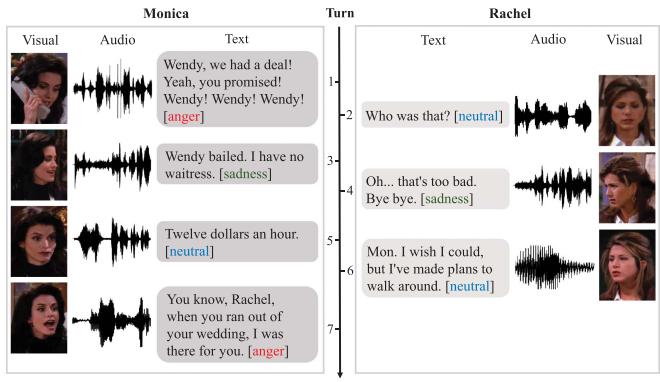


Fig. 1. Multimodal conversation example from the Friends TV series.

Existing mainstream works on ERC can generally be categorized into sequence- and graph-based methods. Sequence-based methods [4], [5], [6], [7], [8], [9], [10], [11] use recurrent neural networks or transformers to model long-distance contextual information in a conversation. In contrast, graph-based methods [12], [13], [14], [15] design graph structures for conversations and then use graph neural networks to capture multiple dependencies. Although these methods show promising performance, most of them focus primarily on textual conversations without leveraging other modalities (i.e., acoustic and visual modalities). According to Mehrabian [16], people express emotions in a variety of ways, including verbal, vocal, and facial expressions. Therefore, multimodal information is more useful for understanding emotions than unimodal information.

Unlike emotion recognition in textual conversations, we argure that three key characteristics are essential for multimodal ERC: intra- and inter-modal interactions between conversation utterances, different contributions of modalities, and efficient modal representations. An example is shown in Fig. 1. (1) To understand the importance of intra- and inter-modal interactions, let us focus on the single and multiple modalities, respectively. Recognizing "anger" emotion of the 7th utterance spoken by Monica is difficult using only "You know, Rachel, when you ran out of your wedding, I was there for you.", but it becomes easy when looking back to the textual expression of the 6th

Manuscript received 25 April 2022; revised 17 December 2022 and 3 February 2023; accepted 23 April 2023. Date of publication 27 April 2023; date of current version 24 July 2024. This work was supported in part by the Natural Science Foundation of China under Grant 62006034, in part by the Natural Science Foundation of Liaoning Province under Grant 2021-BS-067, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT21RC(3)015. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Mohammed Daoudi. *(Corresponding author: Jian Wang.)*

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, and Bo Xu are with the School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China (e-mail: huima@mail.dlut.edu.cn; wangjian@dlut.edu.cn; hflin@dlut.edu.cn; zhangbo1998@mail.dlut.edu.cn; xubo@dlut.edu.cn).

Yijia Zhang is with the School of Information Science and Technology, Dalian Maritime University, Dalian 116024, China (e-mail: zhangyijia@dlmu.edu.cn).

The code is available at https://github.com/butterfliesss/SDT.

Digital Object Identifier 10.1109/TMM.2023.3271019

utterance because Rachel has made plans to walk around. Additionally, we believe there are two types of inter-modal interactions: interactions between the same and different utterances. First, as stated above, it is hard to identify the emotion of the 7th utterance using its textual expression; however, it also would be easy when fused with its visual and acoustic expressions since they burst instantaneously. Second, we know that the textual expression of the 5th utterance shows "neutral" emotion, and hence it could be possible to identify "neutral" emotion of the 6th utterance by interacting this utterance's visual expression and the 5th utterance's textual expression. (2) To understand the importance of contributions of different modalities, let us focus on the 3rd utterance. The textual and acoustic expressions play more critical roles in recognizing "sadness" emotion than the visual expression because a smiling face usually means "joy" emotion. (3) To understand the importance of efficient modal representations, let us focus on the textual expression of the 1st utterance, which contains multiple exclamation points. If the learned representation does not include the meaning of "!", it is challenging to identify "anger" emotion.

Therefore, it is valuable to capture intra- and inter-modal interactions between utterances, dynamically learn weights between modalities, and enhance modal representations for multimodal ERC. However, existing studies of the task have some limitations in achieving these characteristics. On the one hand, most methods have drawbacks in modeling intra- and inter-modal interactions. For example, CMN [4], ICON [5], and DialogueRNN [6] concatenate unimodal features at the input level, and thus cannot capture intra-modal interactions explicitly. While DialogueTRM [10] designs hierarchical transformer and multi-grained interaction fusion modules to explore intra- and inter-modal emotional behaviors, it ignores inter-modal interactions between different utterances. MMGCN [13] and MMDFN [17] are graph-based fusion methods that require manually constructed graph structures to represent conversations. On the other hand, existing methods rely on the designed model to learn modal representations, but no work focuses on further improving modal representations using model-agnostic techniques for ERC.

In this work, a transformer-based model with self-distillation (SDT) is proposed to take into account the three aforementioned characteristics. First, we introduce intra- and inter-modal transformers in a modality encoder to capture intra- and inter-modal interactions, and take positional and speaker embeddings as additional inputs of these transformers to capture contextual and speaker information. Next, a hierarchical gated fusion strategy is proposed to dynamically fuse information from multiple modalities. Then, we predict emotion labels of conversation utterances based on fused multimodal representations in an emotion classifier. We call the above three components a transformer-based model. Finally, to learn more effective modal representations, we introduce self-distillation into the proposed transformer-based model, which transfers knowledge of hard and soft labels from the model to each modality. We treat the proposed model as the teacher and design three students according to three existing modalities. These students are trained by distilling knowledge from the teacher to learn better modal representations.

In summary, our contributions are as follows:

- We propose a transformer-based model for multimodal ERC that contains a modality encoder for capturing intra- and inter-modal interactions between conversation utterances and a hierarchical gated fusion strategy for adaptively learning weights between modalities.
- To learn more effective modal representations, we devise self-distillation that transfers knowledge of hard and soft labels from the proposed model to each modality.
- Experiments on two benchmark datasets show the superiority of our proposed model. In addition, several studies are conducted to investigate the impact of positional and speaker embeddings, intra- and inter-modal transformers, self-distillation loss functions, and hierarchical gated fusion strategy.

The rest of this paper is organized as follows: Section II discusses the related work; Section III formalizes the task definition and describes the proposed model; Section IV gives the experimental settings; Section V presents the experimental results and discussion; Finally, Section VI concludes the paper and provides directions for further work.

II. Related Work

A. Emotion Recognition in Conversations

ERC has attracted widespread interest among researchers with the increase in available conversation datasets, such as IMEOCAP [18], AVEC [19], and MELD [20], etc. Early studies primarily used lexicon-based methods [21], [22]. Recent works have generally resorted to deep neural networks and focused on modeling context- and speaker-sensitive dependencies. We divide the existing methods into two categories: speaker-ignorant and speaker-dependent methods, according to whether they utilize speaker information.

Speaker-ignorant methods do not distinguish speakers and focus only on capturing contextual information in a conversation. HiGRU [7] contains two gated recurrent units (GRUs) to model contextual relationships between words and utterances, respectively. AGHMN [23] uses a hierarchical memory network to enhance utterance representations and introduces an attention GRU to model contextual information. MVN [11] utilizes a multi-view network to model word- and utterance-level dependencies in a conversation. In contrast, speaker-dependent methods model both context- and speaker-sensitive dependencies. DialogueRNN [6] leverages three distinct GRUs to update speaker, context, and emotional states in a conversation, respectively. DialogueGCN [12] uses a graph convolutional network to model speaker and conversation sequential information. HiTrans [8] consists of two hierarchical transformers to capture global contextual information and exploits an auxiliary task to model speaker-sensitive dependencies.

However, most of them are proposed for the textual modality, ignoring the effectiveness of other modalities. Due to the

promising performance in the multimodal community, some approaches tend to address multimodal ERC. DialogueTRM [10] explores intra- and inter-modal emotional behaviors using hierarchical transformer and multi-grained interaction fusion modules, respectively. MMGCN [13] constructs a fully connected graph to model multimodal and long-distance contextual information, and speaker embeddings are added for encoding speaker information. MM-DFN [17] designs a graph-based dynamic fusion module to reduce redundancy and enhance complementarity between modalities. MMTr [24] preserves the integrity of main modal representations and enhances weak modal representations by using multi-head attention. UniMSE [25] performs modality fusion at syntactic and semantic levels and introduces inter-modality contrastive learning to differentiate fusion representations among samples. This paper focuses on exploring intra- and inter-modal interactions between utterances, learning weights between modalities, and enhancing modal representations for multimodal ERC.

B. Multimodal Language Analysis

Multimodal language analysis is a rapidly growing field and includes various tasks [26], such as multimodal emotion recognition, sentiment analysis, and personality traits recognition. The key in this area is to fuse multimodal information. Early studies on multimodal fusion mainly included early fusion and late fusion. Early fusion [27], [28] integrates features of different modalities at the input level. Late fusion [29], [30] constructs distinct models for each modality and then ensembles their outputs by majority voting or weighted averaging, etc. Unfortunately, as stated in [31], these two kinds of fusion methods cannot effectively capture intra- and inter-modal interactions.

Subsequently, model fusion has become popular and various models have been proposed. TFN [32] models unimodal, bimodal, and trimodal interactions explicitly by computing Cartesian product. LMF [31] utilizes low-rank weight tensors for multimodal fusion, which reduces the complexity of TFN. MFN [33] learns cross-modal interactions with an attention mechanism and stores information over time by a multi-view gated memory. MulT [34] utilizes cross-modal transformers to model long-range dependencies across modalities. Rahman et al. [35] fine-tuned large pre-trained transformer models for multimodal language by designing a multimodal adaptation gate (MAG). Self-MM [36] uses a unimodal label generation strategy to acquire independent unimodal supervision and then learns multimodal and unimodal tasks jointly. Yuan et al. [37] adopted transformer encoders to model intra- and inter-modal interactions between modality sequences. In order to capture intra- and inter-modal interactions between conversation utterances and meanwhile learn weights between modalities, we present a transformer-based model.

C. Knowledge Distillation

Knowledge distillation (KD) aims at transferring knowledge from a large teacher network to a small student network. The knowledge mainly includes soft labels of the last output layer (i.e., output-based knowledge) [38], features of intermediate layers (i.e., feature-based knowledge) [39], and relationships between different layers (i.e., relation-based knowledge) [40]. Depending on the learning schemes, existing methods on KD are categorized into three classes: offline distillation [41], [42], online distillation [43], [44], and self-distillation [45], [46]. In offline distillation, the teacher network is first trained and then the pre-trained teacher distills its knowledge to guide the student training. In online distillation, the teacher and student networks are updated simultaneously, and hence its training process is only one-phase. Self-distillation is a special case of online distillation that teaches a single network using its own knowledge.

Recently, KD has been used for multimodal emotion recognition. For example, Albanie et al. [47] transferred visual knowledge into a speech emotion recognition model using unlabelled video data. Wang et al. [48] proposed K-injection subnetworks to distill linguistic and acoustic knowledge representing group emotions and transfer implicit knowledge into the audiovisual model for group emotion recognition. Schoneveld et al. [49] applied KD to further improve performance for facial expression recognition. Most existing models belong to offline distillation, which requires training a teacher network. In contrast, self-distillation needs no extra network except for the network itself. While self-distillation has been successfully applied in computer vision and natural language processing [50], [51], [52], it focuses on unimodal tasks.

In this work, we adopt the idea of self-distillation to enhance modal representations for multimodal ERC. Moreover, output-based knowledge is used only due to the following reasons: (1) Soft labels can be used as training supervision which contain dark knowledge [38] and can provide effective regularization for the model [53]. (2) Intuitively, the features of different modalities vary widely, and hence matching fused multimodal features with unimodal features is inappropriate.[1] (3) Our teacher and student networks lying in the same model have different architectures that results in an inability to inject relationships between different layers of the teacher network into the student network [41]. Therefore, we adopt output-based knowledge rather than feature- and relation-based knowledge.

III. Methodology

A. Task Definition

A conversation is composed of N consecutive utterances $\{u_1, u_2, \ldots, u_N\}$ and M speakers $\{s_1, s_2, \ldots, s_M\}$. Each utterance u_i is spoken by a speaker $s_{\phi(u_i)}$, where ϕ is the mapping between an utterance and its corresponding speaker's index. Moreover, u_i involves textual (t), acoustic (a), and visual (v) modalities, and their feature representations are denoted as $\mathbf{u}_i^t \in \mathbb{R}^{d_t}$, $\mathbf{u}_i^a \in \mathbb{R}^{d_a}$, and $\mathbf{u}_i^v \in \mathbb{R}^{d_v}$, respectively. We represent textual, acoustic, and visual modality sequences of all utterances in the conversation as $\mathbf{U}_t = [\mathbf{u}_1^t; \mathbf{u}_2^t; \ldots; \mathbf{u}_N^t] \in \mathbb{R}^{N \times d_t}$, $\mathbf{U}_a = [\mathbf{u}_1^a; \mathbf{u}_2^a; \ldots; \mathbf{u}_N^a] \in \mathbb{R}^{N \times d_a}$, and $\mathbf{U}_v = [\mathbf{u}_1^v; \mathbf{u}_2^v; \ldots; \mathbf{u}_N^v] \in \mathbb{R}^{N \times d_v}$, respectively. The ERC task aims to predict the emotion label of each utterance u_i from pre-defined emotion categories.

[1] We tried to add feature-based knowledge, but the performance drops significantly.

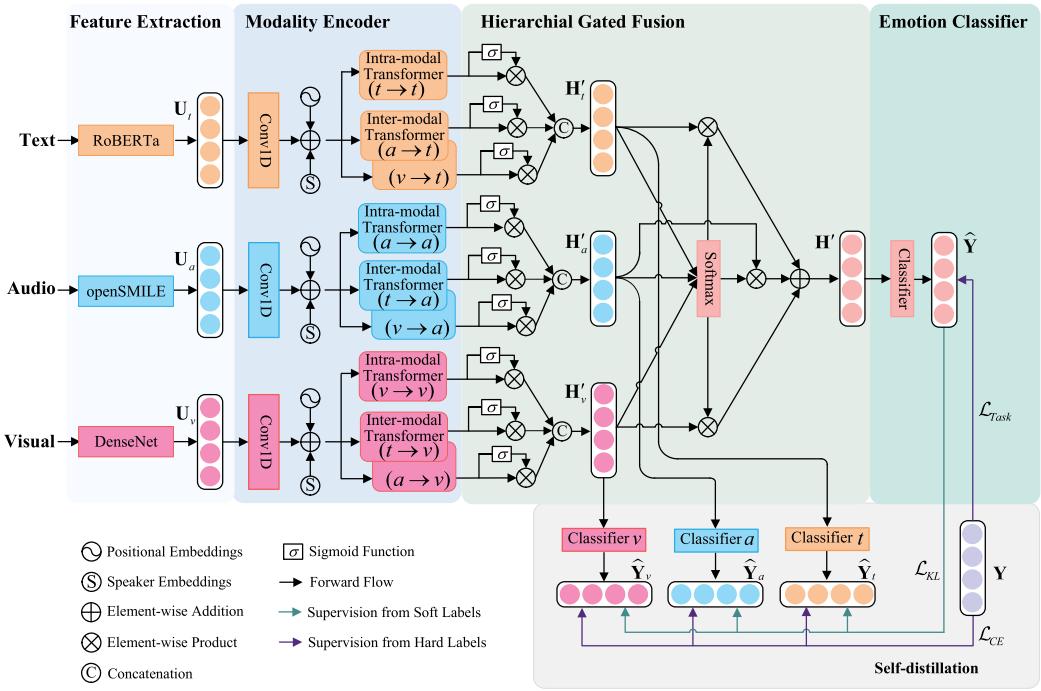


Fig. 2. Overall architecture of SDT. After extracting utterance-level unimodal features, it consists of four key components: Modality Encoder, Hierarchical Gated Fusion, Emotion Classifier, and Self-distillation.

B. Overview

Fig. 2 gives an overview of our proposed SDT. After extracting utterance-level unimodal features, the transformer-based model consists of three modules: a modality encoder module for capturing intra- and inter-modal interactions between different utterances, a hierarchical gated fusion module for adaptively learning weights between modalities, and an emotion classifier module for predicting emotion labels. Furthermore, we introduce self-distillation and devise two kinds of losses to transfer knowledge from our proposed model within each modality to learn better modal representations.

C. Modality Encoder

The modality encoder obtains modality-enhanced modality sequence representations that can learn intra- and inter-modal interactions between conversation utterances.

Temporal Convolution: To ensure that three unimodal sequence representations lie in the same space, we feed them into a 1D convolutional layer:

$$\mathbf{U}'_m = \text{Conv1D}\left(\mathbf{U}_m, k_m\right) \in \mathbb{R}^{N\times d},\ m \in \{t, a, v\}, \tag{1}$$

where k_m is the size of convolutional kernel for m modality, N is the number of utterances in the conversation, and d is the common dimension.

Positional Embeddings: To utilize positional and sequential information of the utterance sequence, we introduce positional embeddings [54] to augment the convolved sequence:

$$\mathbf{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right),$$

$$\mathbf{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right), \tag{2}$$

where pos is the utterance index and i is the dimension index.

Speaker Embeddings: To capture speaker information of the utterance sequence, we also design speaker embeddings to augment the convolved sequence. Speaker s_j in conversations is mapped into a vector:

$$\mathbf{s}_j = \mathbf{V}_s\mathbf{o}\left(s_j\right) \in \mathbb{R}^d,\ \ j = 1,\ 2,\ \ldots,\ M, \tag{3}$$

where M is the total number of speakers, $\mathbf{V}_s \in \mathbb{R}^{d\times M}$ is a trainable speaker embedding matrix, and $\mathbf{o}(s_j) \in \mathbb{R}^M$ is a one-hot vector of speaker s_j, i.e., 1 in the jth position and 0 otherwise.

Hence, speaker embeddings corresponding to the conversation can be represented as $\mathbf{SE} = [\mathbf{s}_{\phi(u_1)}; \mathbf{s}_{\phi(u_2)}; \ldots; \mathbf{s}_{\phi(u_N)}]$.

Overall, we augment positional and speaker embeddings to the convolved sequence:

$$\mathbf{H}_m = \mathbf{U}'_m + \mathbf{PE} + \mathbf{SE}. \tag{4}$$

Here, $\mathbf{H}_m$ is the low-level positional- and speaker-aware utterance sequence representation for m modality.

Intra- and Inter-modal Transformers: We introduce intra- and inter-modal transformers to model intra- and inter-modal interactions for the utterance sequence, respectively. These transformers adopt the transformer encoder [54], which contains three inputs, queries $\mathbf{Q} \in \mathbb{R}^{T_q\times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{T_k\times d_k}$, and values $\mathbf{V} \in \mathbb{R}^{T_k\times d_v}$. We denote the transformer encoder as Transformer$(\mathbf{Q}, \mathbf{K}, \mathbf{V})$.

For the intra-modal transformer, we take $\mathbf{H}_m$ as queries, keys, and values:

$$\mathbf{H}_{m\to m} = \text{Transformer}\left(\mathbf{H}_m, \mathbf{H}_m, \mathbf{H}_m\right) \in \mathbb{R}^{N\times d}, \quad (5)$$

where $m \in \{t, a, v\}$. The intra-modal transformer enhances m-modality sequence representation by itself and thus can capture intra-modal interactions between the utterance sequence.

For the inter-modal transformer, we take $\mathbf{H}_m$ as queries, and $\mathbf{H}_n$ as keys and values:

$$\mathbf{H}_{n\to m} = \text{Transformer}\left(\mathbf{H}_m, \mathbf{H}_n, \mathbf{H}_n\right) \in \mathbb{R}^{N\times d}, \quad (6)$$

where $m \in \{t, a, v\}$ and $n \in \{t, a, v\} - \{m\}$. The inter-modal transformer enables m modality to get information from n modality and hence can capture inter-modal interactions between the utterance sequence.

In summary, n-enhanced m-modality sequence representation, $\mathbf{H}_{n\to m}$, is obtained from the modality encoder module, where $n, m \in \{t, a, v\}$.

D. Hierarchical Gated Fusion

We design a hierarchical gated fusion module containing unimodal- and multimodal-level gated fusions to adaptively obtain enhanced single-modality sequence representation and dynamically learn weights between these enhanced modality representations, respectively.

Unimodal-level Gated Fusion: We first use a gated mechanism to filter out irrelevant information in $\mathbf{H}_{n\to m}$:

$$g_{n\to m} = \sigma\left(\mathbf{W}_{n\to m} \cdot \mathbf{H}_{n\to m}\right), \quad (7)$$

$$\mathbf{H}'_{n\to m} = \mathbf{H}_{n\to m} \otimes g_{n\to m}, \quad (8)$$

where $\mathbf{W}_{n\to m} \in \mathbb{R}^{d\times d}$ is a weight matrix, σ is the sigmoid function, $\otimes$ is the element-wise product, and $g_{n\to m}$ denotes the gate.

Then, we concatenate $\mathbf{H}'_{m\to m}$, $\mathbf{H}'_{n_1\to m}$, and $\mathbf{H}'_{n_2\to m}$, followed by a fully connected (FC) layer to obtain enhanced m-modality sequence representation:

$$\mathbf{H}'_m = \mathbf{W}_m \cdot \left[\mathbf{H}'_{m\to m}; \mathbf{H}'_{n_1\to m}; \mathbf{H}'_{n_2\to m}\right] + \mathbf{b}_m \in \mathbb{R}^{N\times d}, \quad (9)$$

where $m \in \{t, a, v\}$, n_1 and n_2 represent other two modalities, $\mathbf{W}_m \in \mathbb{R}^{3d\times d}$ and $\mathbf{b}_m \in \mathbb{R}^d$ are trainable parameters.

We set $\mathbf{H}'_m = [\mathbf{h}'_{m1}; \mathbf{h}'_{m2}; \ldots; \mathbf{h}'_{mN}]$, where $\mathbf{h}'_{mi}$ is enhanced m-modality representation for the utterance u_i.

Multimodal-level Gated Fusion: We also design a gated mechanism using the softmax function to dynamically learn weights between enhanced modalities for each utterance.

Specifically, the final multimodal representation of the utterance u_i is calculated by:

$$[\mathbf{g}_{ti}; \mathbf{g}_{ai}; \mathbf{g}_{vi}] = \text{softmax}\left([\mathbf{W}\cdot\mathbf{h}'_{ti}; \mathbf{W}\cdot\mathbf{h}'_{ai}; \mathbf{W}\cdot\mathbf{h}'_{vi}]\right), \quad (10)$$

$$\mathbf{h}'_i = \sum_{m\in\{t,a,v\}} \mathbf{h}'_{mi} \otimes \mathbf{g}_{mi}, \quad (11)$$

where $\mathbf{W} \in \mathbb{R}^{d\times d}$ is a weight matrix, $\mathbf{g}_{ti}$, $\mathbf{g}_{ai}$, and $\mathbf{g}_{vi}$ are learned weights of t, a, v modalities for the utterance u_i, respectively.

Thus, multimodal sequence representation of conversation utterances is obtained and denoted as $\mathbf{H}' = [\mathbf{h}'_1; \mathbf{h}'_2; \ldots; \mathbf{h}'_N]$.

E. Emotion Classifier

To calculate probabilities over C emotion categories, $\mathbf{H}'$ is fed into a classifier with an FC and softmax layer:

$$\mathbf{E} = \mathbf{W}_e \cdot \mathbf{H}' + \mathbf{b}_e \in \mathbb{R}^{N\times C}, \quad (12)$$

$$\hat{\mathbf{Y}} = \text{softmax}\left(\mathbf{E}\right), \quad (13)$$

where $\mathbf{W}_e \in \mathbb{R}^{d\times C}$ and $\mathbf{b}_e \in \mathbb{R}^C$ are trainable parameters. We set $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1; \hat{\mathbf{y}}_2; \ldots; \hat{\mathbf{y}}_N]$, where $\hat{\mathbf{y}}_i$ is the emotion probability vector for the utterance u_i. Finally, we choose $\text{argmax}(\hat{\mathbf{y}}_i)$ as the predicted emotion label for u_i.

Task Loss: We utilize the cross-entropy loss for estimating the quality of emotion predictions during training:

$$\mathcal{L}_{Task} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \mathbf{y}_{i,j} \log\left(\hat{\mathbf{y}}_{i,j}\right), \quad (14)$$

where N represents the number of utterances in the conversation, and C represents the number of emotion classes. $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ denote the ground-truth one-hot vector and probability vector for the emotion of u_i, respectively.

F. Self-Distillation

Soft labels containing informative dark knowledge can be used as training supervision; hence, we devise self-distillation to transfer knowledge of hard and soft labels to each modality, and guide the model in learning more expressive modal representations.

We treat our proposed transformer-based model as the teacher and design three students according to existing modalities. Specifically, a classifier consisting of an FC and softmax layer only used during training, is set after each unimodal-level gated fusion. During training, textual, acoustic, and visual modality encoders with their corresponding unimodal-level gated fusions and classifiers are trained as three students (i.e., student t, student a, student v) via distilling from the teacher.

The output of student m is its predicted emotion probabilities:

$$\mathbf{E}_m = \mathbf{W}'_m \cdot \text{ReLU}\left(\mathbf{H}'_m\right) + \mathbf{b}'_m \in \mathbb{R}^{N\times C}, \quad (15)$$

$$\hat{\mathbf{Y}}_m = \text{softmax}\left(\mathbf{E}_m\right),$$

$$\hat{\mathbf{Y}}^{\tau}_m = \text{softmax}\left(\mathbf{E}_{m/\tau}\right), \quad (16)$$

where $m \in \{t, a, v\}$, $\mathbf{W}'_m \in \mathbb{R}^{d\times C}$ and $\mathbf{b}'_m \in \mathbb{R}^C$ are trainable parameters. τ is the temperature to soften $\hat{\mathbf{Y}}_m$ (written as $\hat{\mathbf{Y}}^{\tau}_m$ after softened) and a higher τ produces a softer distribution over classes [38]. We set $\hat{\mathbf{Y}}_m = [\hat{\mathbf{y}}_{m1}; \hat{\mathbf{y}}_{m2}; \ldots; \hat{\mathbf{y}}_{mN}]$ and $\hat{\mathbf{Y}}^{\tau}_m = [\hat{\mathbf{y}}^{\tau}_{m1}; \hat{\mathbf{y}}^{\tau}_{m2}; \ldots; \hat{\mathbf{y}}^{\tau}_{mN}]$.

During training, we introduce two kinds of losses to train the student m to learn better enhanced m-modality sequence representation, where $m \in \{t, a, v\}$.

Cross Entropy Loss: We minimize the cross entropy loss between the predicted probability of the student m and the ground-truth:

$$\mathcal{L}^m_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \mathbf{y}_{i,j} \log\left(\hat{\mathbf{y}}_{mi,j}\right), \quad (17)$$

TABLE I
STATISTICS OF THE TWO DATASETS

Dataset	#Conversations		#Utterances		#Classes
	Train+Val	Test	Train+Val	Test	
IEMOCAP	120	31	5810	1623	6
MELD	1153	280	11098	2610	7

where $\hat{\mathbf{y}}_{mi}$ is the emotion probability vector of the student m for u_i. In this way, knowledge from hard labels is directly introduced to the student to learn better modal representations.

KL Divergence Loss: To make the output probability of the student m approximate the output of the teacher (i.e., soft labels), the Kullback-Leibler (KL) divergence loss between them is minimized:

$$\mathcal{L}_{KL}^{m} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\hat{\mathbf{y}}_{mi,j}^{\tau}\log\left(\frac{\hat{\mathbf{y}}_{mi,j}^{\tau}}{\hat{\mathbf{y}}_{i,j}^{\tau}}\right), \tag{18}$$

where $\hat{\mathbf{y}}_{mi}^{\tau}$ and $\hat{\mathbf{y}}_{i}^{\tau}$ are soften probability distributions of the student m and the teacher, respectively. In this way, knowledge from soft labels is transferred to the student to learn better modal representations.

With both hard and soft labels, the overall loss can be expressed as:

$$\mathcal{L} = \gamma_1\mathcal{L}_{Task} + \gamma_2\mathcal{L}_{CE} + \gamma_3\mathcal{L}_{KL}, \tag{19}$$

$$\mathcal{L}_{CE} = \sum_{m\in\{t,a,v\}}\mathcal{L}_{CE}^{m}, \tag{20}$$

$$\mathcal{L}_{KL} = \sum_{m\in\{t,a,v\}}\mathcal{L}_{KL}^{m}, \tag{21}$$

where γ_1, γ_2, and γ_3 are hyper-parameters that control the weights of the three kinds of losses. In experiments, we set $\gamma_1 = \gamma_2 = \gamma_3 = 1$.

IV. EXPERIMENTAL SETTINGS

A. Datasets and Evaluations

We use IEMOCAP [18] and MELD [20] datasets to evaluate the proposed model. The statistics of the two datasets are listed in Table I.

IEMOCAP: The dataset consists of two-way conversations of ten speakers, containing 153 conversations and 7,433 utterances. The dataset is divided into five sessions, where the first four sessions are used for training, while the last one is for testing. Each utterance is labeled with one of six emotions: happy, sad, neutral, angry, excited, and frustrated.

MELD: This is a multi-speaker conversation dataset collected from the Friends TV series, containing 1,433 conversations and 13,708 utterances. Each utterance is labeled with one of seven emotions: neutral, surprise, fear, sadness, joy, disgust, and anger.

Evaluation Metrics: Following previous works [6], [12], we report the overall accuracy and weighted average F1-score to measure overall performance, and also present the accuracy and F1-score on each emotion class.

B. Feature Extraction

We extract utterance-level unimodal features as follows.

Textual Modality: Following [55], we employ RoBERTa Large model [56] to extract textual features. RoBERTa, a pre-trained model using a multi-layer transformer encoder architecture, builds on BERT which can efficiently learn textual representations. We fine-tune RoBERTa for emotion recognition from conversation transcripts and then take [*CLS*] tokens' embeddings at the last layer as textual features. The dimensionality of textual feature representation is 1024.

Acoustic Modality: Following [13], we use openSMILE [57] for acoustic feature extraction. openSMILE, a flexible feature extraction toolkit for signal processing, provides a scriptable console application to configure modular feature extraction components. After using openSMILE toolkit, an FC layer reduces the dimensionality of acoustic feature representation to 1582 for IEMOCAP and 300 for MELD.

Visual Modality: Following [13], we use DenseNet [58] pre-trained on Facial Expression Recognition Plus dataset for visual feature extraction. DenseNet, an effective CNN architecture, consists of multiple dense blocks, each of which contains multiple layers. Finally, the dimensionality of visual feature representation is 342.

C. Baselines

We compare SDT with the following baseline models.

CMN [4]: It uses two GRUs and memory networks to model contextual information for both speakers, but it is only available for dyadic conversations.

ICON [5]: It is an extension of CMN that captures inter-speaker emotional influences using another GRU. Similar to CMN, the model is applied to dyadic conversations.

DialogueRNN [6]: It adopts three distinct GRUs to track the speaker, context, and emotional states in conversations, respectively.

The above models concatenate textual, acoustic, and visual features to obtain multimodal utterance representations.

MMGCN [13]: It constructs a conversation graph based on all three modalities and designs a multimodal fused graph convolutional network to model contextual dependencies across multiple modalities.

DialogueTRM [10]: It uses a hierarchical transformer to manage the differentiated context preference within each modality and designs a multi-grained interactive fusion for learning different contributions across modalities for an utterance.

MM-DFN [17]: It designs a graph-based dynamic fusion module to fuse multimodal context features, and this module could reduce redundancy and enhance complementarity between modalities.

MMTr [24]: It uses distinct bidirectional long short-term memory networks (Bi-LSTMs) to learn contextual representations at the speaker's self-context level and contextual context level, and designs a cross-modal fusion module to enhance weak modal representations.

TABLE II
RESULTS ON THE IEMOCAP DATASET; "*": BASELINES ARE RE-IMPLEMENTED USING OUR EXTRACTED FEATURES; BOLD FONT DENOTES THE BEST PERFORMANCE

Models	IEMOCAP													
	happy		sad		neutral		angry		excited		frustrated		ACC	w-F1
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1		
CMN	24.31	30.30	56.33	62.02	52.34	52.41	61.76	60.17	56.19	60.76	**72.44**	61.27	56.87	56.33
ICON	25.00	31.30	67.35	73.17	55.99	58.50	69.41	66.29	70.90	67.09	71.92	65.08	62.85	62.25
DialogueRNN	25.00	34.95	82.86	84.58	54.43	57.66	61.76	64.42	**90.97**	76.30	62.20	59.55	65.43	64.29
MMGCN	32.64	39.66	72.65	76.89	65.10	62.81	73.53	**71.43**	77.93	75.40	65.09	63.43	66.61	66.25
DialogueTRM	61.11	57.89	84.90	81.25	69.27	68.56	76.47	65.99	76.25	76.13	50.39	58.09	68.52	68.20
MM-DFN	44.44	44.44	77.55	80.00	71.35	66.99	75.88	70.88	74.25	76.42	58.27	61.67	67.84	67.85
MMTr	-	-	-	-	-	-	-	-	-	-	-	-	72.27	71.91
UniMSE	-	-	-	-	-	-	-	-	-	-	-	-	70.56	70.66
DialogueRNN*	57.64	57.64	77.96	80.25	75.52	70.56	68.24	64.99	73.91	75.95	59.06	62.41	69.38	69.37
MMGCN*	50.00	56.25	78.78	81.43	71.35	67.57	68.24	66.29	75.92	76.82	65.09	64.92	69.62	69.61
DialogueTRM*	72.22	62.84	**85.71**	83.33	69.27	68.12	**79.41**	66.67	67.22	75.00	57.22	63.28	69.87	69,93
MM-DFN*	57.64	52.87	84.49	**86.07**	76.04	71.66	70.59	65.04	73.24	75.26	55.91	62.19	69.87	69.91
SDT (Ours)	**72.71**	**66.19**	79.51	81.84	**76.33**	**74.62**	71.88	69.73	76.79	**80.17**	67.14	**68.68**	**73.95**	**74.08**
w/o self-distillation	71.53	58.52	79.59	79.43	69.27	70.65	69.41	67.05	69.23	77.09	67.98	68.07	70.73	71.10

TABLE III
RESULTS ON THE MELD DATASET

Models	MELD															
	neutral		surprise		fear		sadness		joy		disgust		anger		ACC	w-F1
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1		
DialogueRNN	82.17	76.56	46.62	47.64	0.00	0.00	21.15	24.65	49.50	51.49	0.00	0.00	48.41	46.01	60.27	57.95
MMGCN	84.32	76.96	47.33	49.63	2.00	3.64	14.90	20.39	56.97	53.76	1.47	2.82	42.61	45.23	61.34	58.41
DialogueTRM	83.20	79.41	56.94	55.27	12.00	17.39	27.88	36.48	60.45	60.30	16.18	20.18	51.01	49.79	65.10	63.80
MM-DFN	79.06	75.80	53.02	50.42	0.00	0.00	17.79	23.72	59.20	55.48	0.00	0.00	50.43	48.27	60.96	58.72
MMTr	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.64	64.41
UniMSE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.09	65.51
DialogueRNN*	**85.11**	79.60	54.09	56.72	10.00	12.66	29.81	38.63	62.94	63.81	22.06	27.27	53.62	53.24	66.70	65.31
MMGCN*	81.53	79.20	58.36	57.75	8.00	13.79	31.73	39.40	**69.90**	63.43	20.59	24.56	52.17	53.49	66.40	65.21
DialogueTRM*	83.44	79.54	54.45	57.09	24.00	**27.91**	33.17	40.95	60.45	62.79	22.06	28.04	**58.26**	53.96	66.70	65.76
MM-DFN*	83.52	79.65	**63.35**	58.17	**32.00**	26.67	26.44	35.71	63.68	**64.89**	19.12	24.76	49.28	52.15	66.55	65.48
SDT (Ours)	83.22	**80.19**	61.28	**59.07**	13.80	17.88	**34.90**	**43.69**	63.24	64.29	22.65	**28.78**	56.93	**54.33**	**67.55**	**66.60**
w/o self-distillation	82.01	80.00	57.65	57.96	20.00	23.81	32.21	41.61	65.17	64.22	**25.00**	27.42	57.97	54.05	66.97	66.26

UniMSE [25]: It uses T5 to fuse acoustic and visual modal features with multi-level textual features, and performs inter-modality contrastive learning to obtain discriminative multi-modal representations.

For a fair comparison, we re-run all baselines, except MMTr and UniMSE, whose source codes are not released.[2] In addition, we re-implement DialogueRNN, MMGCN, DiaglogueTRM, and MM-DFN with our extracted features, namely DialogueRNN*, MMGCN*, DiaglogueTRM*, and MM-DFN*. We use the same data splits to implement all models.

D. Implementation Details

We implement the proposed model using Pytorch[3] and use Adam [59] as optimizer with an initial learning rate of $1.0e-4$ for IEMOCAP and $5.0e-6$ for MELD. The batch size is 16 for IEMOCAP and 8 for MELD, and the temperature τ for the two datasets are set to 1 and 8, respectively. For the 1D convolutional layers, the number of input channels are set to 1024, 1582, and 342 for textual, acoustic, and visual modalities, respectively (i.e., their corresponding feature dimensions) on IEMOCAP. On MELD, these parameters are set to 1024, 300, and 342, respectively. In addition, the number of output channels and kernel size are set to 1024 and 1 respectively for all three modalities on the two datasets. For the transformer encoder, the hidden size, number of attention heads, feed-forward size, and number of layers are set to 1024, 8, 1024, and 1, respectively. To prevent overfitting, we set the L2 weight decay to $1.0e-5$ and employ dropout with a rate of 0.5. All results are averages of 10 runs.

V. RESULTS AND DISCUSSION

A. Overall Results

Table II and Table III present the performance of baselines and SDT on IEMOCAP and MELD datasets, respectively. On IEMOCAP dataset, SDT performs better than all baselines and outperforms MMTr by 1.68% and 2.17% in terms of overall accuracy and weighted F1-score, respectively. In addition, SDT achieves a significant improvement on most emotion classes in terms of F1-score. On MELD dataset, SDT achieves the best performance compared to all baselines in terms of overall accuracy and weighted F1-score, and outperforms UniMSE by 2.46%

[2]We carefully implemented DialogueTRM to explore its performance using our extracted features, since its source code is not available; MMTr uses basically same feature extractors as us, and therefore we did not implement it; UniMSE uses T5 to learn contextual information on textual sequences and embeds multimodal fusion layers into T5, and hence our extracted features cannot be used for UniMSE and we also did not implement it.

[3][Online]. Available: https://pytorch.org/

TABLE IV
RESULTS OF ABLATION STUDIES ON THE TWO DATASETS

	IEMOCAP		MELD	
	ACC	w-F1	ACC	w-F1
SDT	**73.95**	**74.08**	**67.55**	**66.60**
Transformer-based model				
w/o positional embeddings	72.27	72.39	66.86	66.20
w/o speaker embeddings	71.84	72.03	67.13	66.18
w/o intra-modal transformers	73.38	73.36	67.13	66.21
w/o inter-modal transformers	72.09	72.26	66.97	65.55
Self-distillation				
w/o $\mathcal{L}_{CE}$	73.07	73.32	67.39	66.37
w/o $\mathcal{L}_{KL}$	72.95	73.03	67.09	66.33
Modality				
Text	66.42	66.58	66.82	65.52
Audio	59.77	59.34	48.12	40.81
Visual	41.47	42.71	48.05	32.01
Text + Audio	72.52	72.75	67.05	66.24
Text + Visual	69.01	69.07	67.20	66.18
Audio + Visual	62.05	62.26	47.24	40.21

and 1.09%, respectively. Similar to IEMOCAP, SDT performs superior on most emotion classes in terms of F1-score.

Overall, the above results indicate the effectiveness of SDT. Furthermore, we have several similar findings on the two datasets: (1) DialogueTRM has a superior performance compared to DialogueRNN, MMGCN, and MM-DFN that use TextCNN [60] to extract textual features. This is because textual modality plays a more important role for ERC [13], and DialogueTRM extracts textual features using BERT [61], which is more powerful than TextCNN. (2) The baselines gain further improvement and achieve comparable results when using our extracted utterance features. The results show that our feature extractor is more effective and sequence- and graph-based baselines can achieve similar performance using our extracted features. (3) Even without self-distillation, our proposed model is still comparable to strong baselines, demonstrating the power of the proposed transformer-based model.

B. Ablation Study

We carry out ablation experiments on IEMOCAP and MELD. Table IV reports the results under different ablation settings.

Ablation on Transformer-based Model: Positional embeddings, speaker embeddings, intra-modal transformers, and inter-modal transformers are four crucial components of our proposed transformer-based model. We remove only one component at a time to evaluate the effectiveness of the component. From Table IV, we conclude that: (1) All components are useful because removing one of them leads to performance degradation. (2) Positional and speaker embeddings have considerable effects on the two datasets, which means capturing sequential and speaker information are valuable. (3) Inter-modal transformers are more important than intra-modal transformers on the two datasets. This indicates that inter-modal interactions between conversation utterances could provide more helpful information.

Ablation on Self-distillation Loss Functions: There are two kinds of losses (i.e., $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$) for self-distillation. To verify the importance of these losses, we remove one loss at a time.

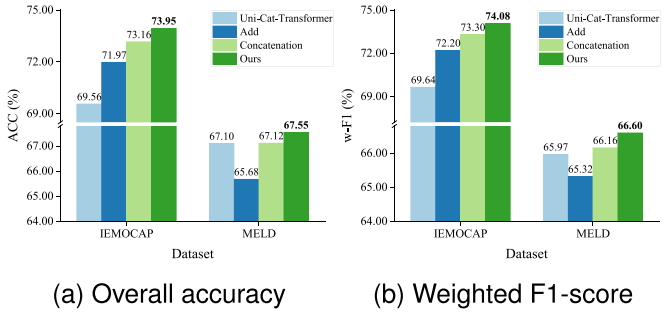


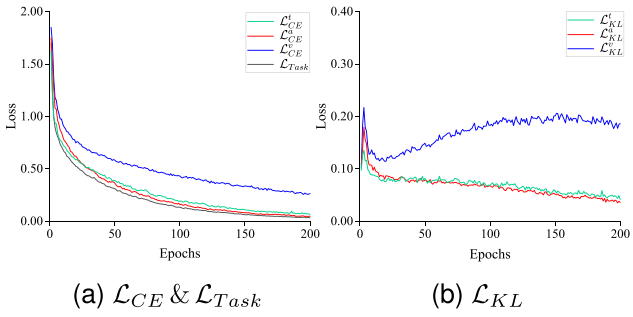
Fig. 3. Performance of different fusion methods on the two datasets. Bold font means that the improvement to all baselines is statistically significant (t-test with $p < 0.05$).

Table IV shows that $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$ are complementary and our model performs best when all losses are included. The result demonstrates that transferring knowledge of both hard and soft labels from the proposed transformer-based model to each modality can further boost the model performance.

Effect of Different Modalities: To show the effect of different modalities, we remove one or two modalities at a time. From Table IV, we observe that: (1) For unimodal results, the textual modality has far better performance than the other two modalities, indicating that the textual feature plays a leading role in ERC. This finding is consistent with previous works [10], [13], [17]. (2) Any bimodal results are better than its own unimodal results. Moreover, fusing the textual modality and acoustic or visual modality performs superior to the fusion of the acoustic and visual modalities due to the importance of textual features. (3) Using all three modalities gives the best performance. The result can validate that emotion is affected by verbal, vocal, and visual expressions, and integrating multimodal information is essential for ERC.

Effect of Different Fusion Strategies: To investigate the effect of our proposed hierarchical gated fusion module, we compare it with two typical information fusion strategies: (1) Add: representations are fused via element-wise addition. (2) Concatenation: representations are directly concatenated and followed by an FC layer. Add treats all representations equally, while Concatenation could implicitly choose the important information due to the FC layer. For a fair comparison, we replace the hierarchical gated fusion module of our model with hierarchical add and concatenation operations to implement the Add and Concatenation fusion strategies, respectively. In addition, we also compare SDT with a general transformer-based fusion method (i.e., unimodal features are concatenated and then fed into a transformer encoder) that we call Uni-Cat-Transformer.

As shown in Fig. 3, compared with other fusion strategies, our proposed hierarchical gated fusion strategy significantly outperforms them. The result indicates that directly fusing representations with Add and Concatenation is sub-optimal. Our proposed hierarchical gated fusion module first filters out irrelevant information at the unimodal level and then dynamically learns weights between different modalities at the multimodal level, which can more effectively fuse multimodal representations.



(a) $\mathcal{L}_{CE}$ & $\mathcal{L}_{Task}$ (b) $\mathcal{L}_{KL}$

Fig. 4. Trends of all losses during training on the IEMOCAP dataset.

In addition, our model achieves a significant performance improvement over Uni-Cat-Transformer that demonstrates the effectiveness of the proposed SDT in multimodal fusion. Interestingly, Uni-Cat-Transformer has poorer performance than Add and Concatenation on IEMOCAP; however, it shows an acceptable performance on MELD. This may be because interactions between modalities are not as complex on MELD as on IEMOCAP, and hence modeling modal interactions by multiple transformer encoders could generate some noise on MELD that makes Ui-Cat-Transformer gain comparable performance with Add and Concatenation. In contrast, SDT has superior performance than all baselines as it contains a hierarchical gated fusion module to filter out noise information, which further illustrates the usefulness of our proposed hierarchical gated fusion strategy.

C. Trends of Losses

During training, we illustrate the trends of all types of losses on IEMOCAP dataset to better understand how these losses work, and Fig. 4 displays the results.

From Fig. 4(a), we find that $\mathcal{L}_{Task}$, $\mathcal{L}^t_{CE}$, $\mathcal{L}^a_{CE}$, and $\mathcal{L}^v_{CE}$ keep descending in the whole training process. From Fig. 4(b), we can see $\mathcal{L}^t_{KL}$ and $\mathcal{L}^a_{KL}$ also have decreasing trends except for fluctuations at the beginning, and $\mathcal{L}^v_{KL}$ goes down during early training except for the fluctuation and then goes up and achieves stability. Therefore, all of the losses can converge. These show that all students can learn knowledge from hard and soft labels to improve the model performance. Besides, we find that losses of student v (i.e., $\mathcal{L}^v_{CE}$ and $\mathcal{L}^v_{KL}$) are larger than the other two students. This may due to a unsuitable learning rate for the student v. Hence, we would like to adaptively modify learning rates between different modalities to effectively optimize the proposed model in the future.

D. Multimodal Representation Visualization

We extract multimodal representations for each utterance on IEMOCAP from our proposed transformer-based model without and with self-distillation. Besides, pre-extracted unimodal representations are concatenated to produce original multimodal representations. Then, these multimodal representations are projected into two dimensions via the t-SNE algorithm [62].

Fig. 5 illustrates the visualization results with different emotion categories. Compared with original multimodal representations, representations learned by the proposed transformer-based model become more clustered even without self-distillation. However without self-distillation, multimodal representations of similar emotions (i.e., "happy" and "excited", "angry" and "frustrated") are difficult to separate; furthermore, representations of "neutral" emotion are intermingled with other emotions. By comparing Fig. 5(b) and (c), we observe that our model with self-distillation yields a better separation and representations of different emotions are less mixed together. Therefore, introducing self-distillation training could learn more effective multimodal representations.

On the other hand, we also show the visualization results with different genders of speakers in Fig. 6. Fig. 6(b) and (c) form two large clusters respectively corresponding to the gender of the speaker. This interesting finding indicates that with or without self-distillation, our model can distinguish the gender of the speaker, which may also be helpful for ERC.

E. Case Study

To demonstrate the efficacy of SDT, we present a case study. Fig. 7 shows a conversation that comes from MELD. SDT identifies the emotions of all utterances successfully, while DialogueRNN* and MMGCN* predict the 3rd utterance as "surprise" incorrectly, probably because a question mark "?" generally expresses "surprise". This could indicate the more powerful multimodal fusion capability of our proposed model. On the other hand, using only the textual modality, our model recognizes the 4th utterance as "neutral" wrongly. To explore the reason behind it, we visualize multi-head attention weights of SDT (only text) and SDT for the 4th utterance, respectively. For SDT, we find that the weights of textual features are obviously larger than acoustic and visual features for the 4th utterance by outputting their weights. Therefore, we visualize only attention weights of the transformers that form enhanced textual modality representation in Fig. 8, and other visualization results can be found in the appendix.

As can be seen from Fig. 8(a), the 4th utterance depends heavily on the 3rd and 5th utterances when using only the textual modality. The 3rd utterance, which expresses "neutral" emotion, may be more important due to a larger number of darkest attention heads; hence the 4th utterance is identified as the same emotion as the 3rd utterance, i.e., "neutral". The utterance can be correctly recognized as "disgust" by SDT for the following reasons: (1) According to Fig. 8(b), the text of the 4th utterance is influenced the most by the text of the 5th utterance whose emotion is "disgust". (2) From Fig. 8(c) and (d), we observe that the acoustic and visual expressions of the 2nd, 4th and 5th utterances are more valuable for the 4th utterance's textual expression, and the 4th and 5th utterances express "disgust" emotion. Overall, the results show that interactions between modalities are helpful in identifying emotion from different perspectives, and therefore it is necessary to use multimodal information. In addition, comparing Fig. 8(a) and (b), SDT learns better to correlate the 4th utterance with the 5th utterance. The finding illustrates that introducing self-distillation can learn more appropriate attention weights.

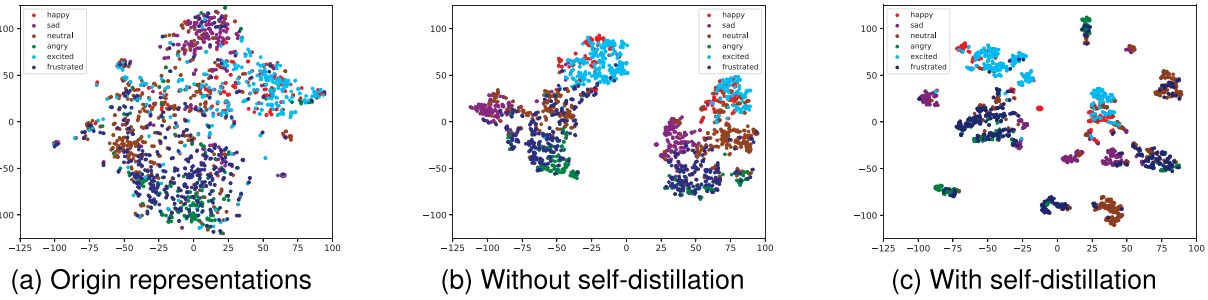


Fig. 5. t-SNE visualization of the multimodal representations with different emotion categories on the IEMOCAP dataset.

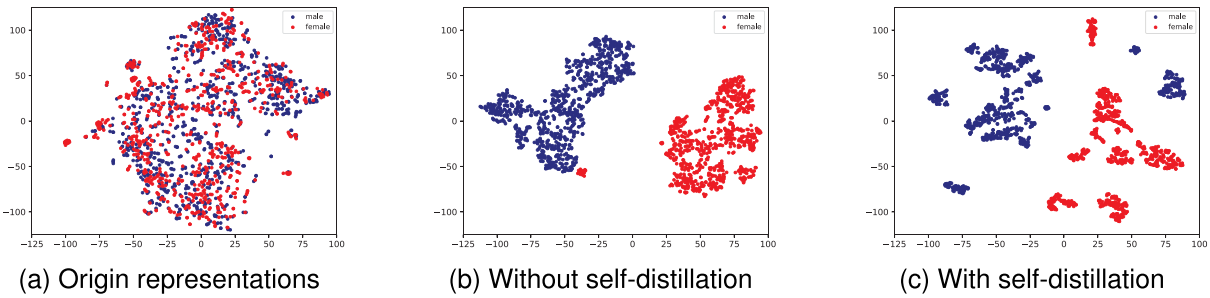


Fig. 6. t-SNE visualization of the multimodal representations with different genders of speakers on the IEMOCAP dataset.

Turn	Speaker	Visual	Audio	Text	Dialogue RNN*	MM GCN*	SDT (only text)	SDT	Ground Truth
1	Joey			Oh my god, you're back!	surprise	surprise	surprise	surprise	surprise
2	Phoebe			Ohh, let me see it! Let me see your hand!	surprise	surprise	surprise	surprise	surprise
3	Monica			Why do you want to see my hand?	*surprise*	*surprise*	neutral	neutral	neutral
4	Phoebe			I wanna see what's in your hand. I wanna see the trash.	disgust	disgust	*neutral*	disgust	disgust
5	Phoebe			Eww! Oh, it's all dirty. You should throw this out.	disgust	disgust	disgust	disgust	disgust

Fig. 7. Example of emotion recognition results in a conversation from the MELD dataset.

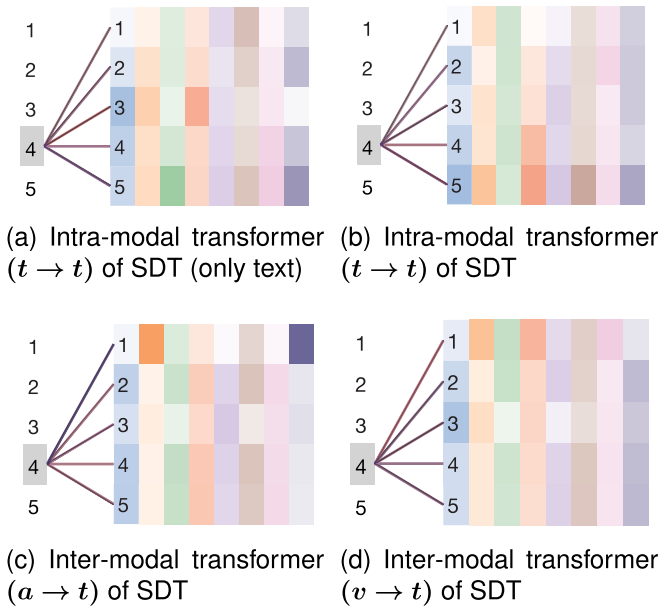


Fig. 8. Multi-head attention visualization for the 4th utterance in Fig. 7. There are 8 attention heads and different colors represent different heads. The darker the color, the more important for the 4th utterance.

F. Error Analysis

Although the proposed SDT achieves strong performance, it still fails to detect some emotions. We analyze confusion matrices of the test set on the two datasets. From Fig. 9, we see that: (1) SDT misclassifies similar emotions, like "happy" and "excited", "angry" and "frustrated" on IEMOCAP, and "surprise" and "anger" on MELD. (2) SDT also tends to misclassify other emotions as "neutral" on MELD due to that "neutral" is the majority class. (3) It is difficult to correctly detect "fear" and "disgust" emotions on MELD because the two emotions are minority classes. Thus, it is challenging to recognize similar emotions and emotions with unbalanced data.

Besides, we also investigate SDT performance on emotional shift (i.e., two consecutive utterances spoken by the same speaker have different emotions). As shown in Table V, we observe that SDT performs poorer on utterances with emotional shift than that without it,[4] which is consistent with previous

[4]In this paper, without emotional shift means two consecutive utterances spoken by the same speaker have same emotions.

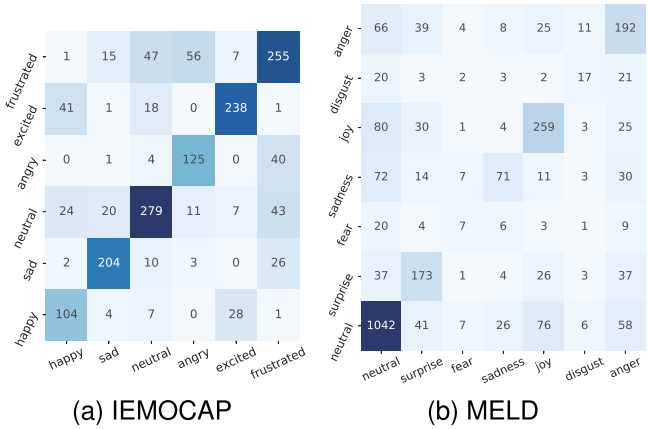


Fig. 9. Confusion matrices of the test set on the two datasets. The rows and columns represent true and predicted labels, respectively.

TABLE V
TEST ACCURACY OF SDT ON UTTERANCES WITH AND WITHOUT EMOTIONAL SHIFT

Dataset	Emotional Shift		w/o Emotional Shift	
	#Utterances	ACC	#Utterances	ACC
IEMOCAP	410	54.88	1151	80.71
MELD	1003	61.62	861	73.05

works. The emotional shift in conversations is a complex phenomenon caused by multiple latent variables, e.g., the speaker's personality and intent; however, SDT and most existing models do not consider these factors, which may result in poor performance. Further improvement on the case needs to be explored.

VI. CONCLUSION

In this paper, we propose SDT, a transformer-based model with self-distillation for multimodal ERC. We use intra- and inter-modal transformers to model intra- and inter-modal interactions between conversation utterances. To dynamically learn weights between different modalities, we design a hierarchical gated fusion strategy. Positional and speaker embeddings are also leveraged as additional inputs to capture contextual and speaker information. In addition, we devise self-distillation during training to transfer knowledge of hard and soft labels within the model to learn better modal representations, which could further improve performance. We conduct experiments on two benchmark datasets and the results demonstrate the effectiveness and superiority of SDT.

Through error analysis, we find that distinguishing similar emotions, detecting emotions with unbalanced data, and emotional shift are key challenges for ERC that are worth further exploration in future work. Furthermore, transformer-based fusion methods cause high computational costs as the self-attention mechanism of transformer has a complexity of $O(N^2)$ with respect to sequence length N. To alleviate the issue, Ding et al. [63] proposed sparse fusion for multimodal transformers. Similarly, we plan to design a novel multimodal fusion method for transformers to reduce computational costs in the future.

APPENDIX ATTENTION VISUALIZATION

Multi-head attention weights of the transformers in our SDT that form enhanced acoustic and visual modality representations are visualized in Fig. 10.

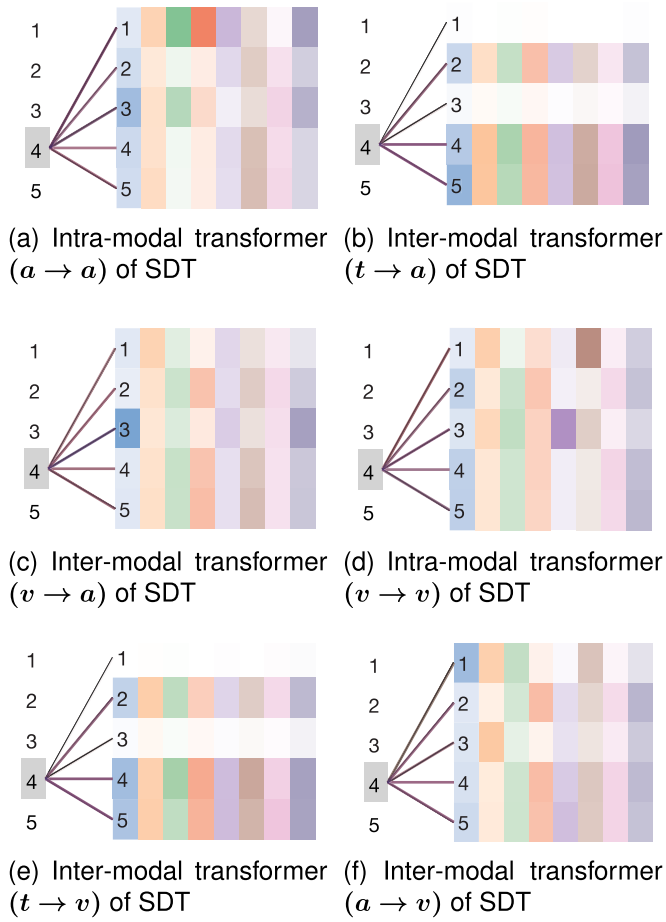


Fig. 10. Multi-head attention visualization for the 4th utterance in Fig. 7.

REFERENCES

[1] A. Kumar, P. Dogra, and V. Dabas, "Emotion analysis of twitter using opinion mining," in *Proc. 8th Int. Conf. Contemporary Comput.*, 2015, pp. 285–290.
[2] F. A. Pujol, H. Mora, and A. Martínez, "Emotion recognition to improve e-healthcare systems in smart cities," in *Proc. Res. Innov. Forum*, A. Visvizi and M. D. Lytras, Eds., 2019, pp. 245–254.
[3] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Comput. Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.
[4] D. Hazarika et al., "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, (Long Papers), 2018, vol. 1. pp. 2122–2132.
[5] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.
[6] N. Majumder et al., "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 6818–6825.
[7] W. Jiao, H. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, (Long and Short Papers), 2019, vol. 1, pp. 397–406.

[8] J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu, "HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4190–4200.

[9] H. Ma, J. Wang, L. Qian, and H. Lin, "HAN-ReGRU: Hierarchical attention network with residual gated recurrent unit for emotion recognition in conversation," *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2685–2703, 2021.

[10] Y. Mao, G. Liu, X. Wang, W. Gao, and X. Li, "DialogueTRM: Exploring multi-modal emotional dynamics in a conversation," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 2694–2704.

[11] H. Ma et al., "A multi-view network for real-time emotion recognition in conversations," *Knowl.-Based Syst.*, vol. 236, 2022, Art. no. 107751.

[12] 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 Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 154–164.

[13] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*(Long Papers), 2021, vol. 1, pp. 5666–5675.

[14] W. Nie, R. Chang, M. Ren, Y. Su, and A. Liu, "I-GCN: Incremental graph convolution network for conversation emotion detection," *IEEE Trans. Multimedia*, vol. 24, pp. 4471–4481, 2022.

[15] M. Ren, X. Huang, W. Li, D. Song, and W. Nie, "LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 4422–4432, 2022.

[16] A. Mehrabian et al., *Silent Messages*. Belmont, CA, USA: Wadsworth, 1971.

[17] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7037–7041.

[18] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[19] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2012: The continuous audio/visual emotion challenge - an introduction," in *Proc. 14th ACM Int. Conf. Multimodal Interact.*, 2012, pp. 361–362.

[20] S. Poria et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[21] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[22] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 801–804.

[23] W. Jiao, M. Lyu, and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 05, pp. 8002–8009.

[24] S. Zou, X. Huang, X. Shen, and H. Liu, "Improving multimodal fusion with main modal transformer for emotion recognition in conversation," *Knowl.-Based Syst.*, vol. 258, 2022, Art. no. 109978.

[25] G. Hu et al., "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 7837–7851.

[26] P. P. Liang, Z. Liu, A.B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 150–161.

[27] M. Wöllmer et al., "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May/Jun. 2013.

[28] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.

[29] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 284–288.

[30] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics* (Short Papers), 2018, vol. 2pp. 606–611.

[31] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics* (Long Papers), 2018, vol. 1, pp. 2247–2256.

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

[33] A. Zadeh et al., "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 5634–5641.

[34] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[35] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[36] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 12, pp. 10790–10797.

[37] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4400–4407.

[38] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[39] A. Romero et al., "Fitnets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Representations*, 2015.

[40] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4133–4141.

[41] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 268–284.

[42] T. Li, J. Li, Z. Liu, and C. Zhang, "Few sample knowledge distillation for efficient network compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14639–14647.

[43] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4320–4328.

[44] I. Chung, S. Park, J. Kim, and N. Kwak, "Feature-map-level online adversarial knowledge distillation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2006–2015.

[45] L. Zhang et al., "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3713–3722.

[46] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection cnns by self attention distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1013–1021.

[47] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 292–301.

[48] Y. Wang et al., "Implicit knowledge injectable cross attention audiovisual model for group emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 827–834.

[49] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, 2021.

[50] T. Moriya et al., "Self-distillation for improving CTC-transformer-based ASR systems," in *Proc. INTERSPEECH*, 2020, pp. 546–550.

[51] T. Zhou et al., "Automatic ICD coding via interactive shared representation networks with self-distillation mechanism," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*(Long Papers), 2021, vol. 1, pp. 5948–5957.

[52] X. Luo, Q. Liang, D. Liu, and Y. Qu, "Boosting lightweight single image super-resolution via joint-distillation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1535–1543.

[53] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3903–3911.

[54] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, vol. 30, 2017, pp. 5998–6008.

[55] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion identification in conversations," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2020, pp. 2470–2481.

[56] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[57] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 835–838.

[58] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[60] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.
[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, (Long and Short Papers), 2019, vol. 1, pp. 4171–4186.
[62] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[63] Y. Ding et al., "Sparse fusion for multimodal transformers," 2021, *arXiv:2111.11992*.

Hui Ma received the M.S. degree in 2019 from the Dalian University of Technology, Dalian, China, where she is currently working toward the Ph.D. degree with the School of Computer Science and Technology. Her research interests include natural language processing, dialogue system, and sentiment analysis.

Jian Wang received the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2014. She is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. Her research interests include natural language processing, text mining, and information retrieval.

Hongfei Lin received the Ph.D. degree from Northeastern University, Shenyang, China, in 2000. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. His research interests include natural language processing, text mining, and sentimental analysis.

Bo Zhang received the B.S. degree from Tiangong University, Tianjin, China, in 2019. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. His research interests include natural language processing, dialogue system, and text generation.

Yijia Zhang received the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2014. He is currently a Professor with the School of Information Science and Technology, Dalian Maritime University, Dalian. His research interests include natural language processing, bioinformatics, and text mining.

Bo Xu received the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2018. He is currently an Associate Professor with the School of Computer Science and Technology, Dalian University of Technology. His research interests include information retrieval, dialogue system, and natural language processing.