Contextual and Cross-Modal Interaction for Multi-Modal Speech Emotion Recognition

Dingkang Yang, Shuai Huang, Yang Liu, *Graduate Student Member, IEEE*, and Lihua Zhang, *Member, IEEE*

***Abstract*—Speech emotion recognition combining linguistic content and audio signals in the dialog is a challenging task. Nevertheless, previous approaches have failed to explore emotion cues in contextual interactions and ignored the long-range dependencies between elements from different modalities. To tackle the above issues, this letter proposes a multimodal speech emotion recognition method using audio and text data. We first present a contextual transformer module to introduce contextual information via embedding the previous utterances between interlocutors, which enhances the emotion representation of the current utterance. Then, the proposed cross-modal transformer module focuses on the interactions between text and audio modalities, adaptively promoting the fusion from one modality to another. Furthermore, we construct associative topological relation over mini-batch and learn the association between deep fused features with graph convolutional network. Experimental results on the IEMOCAP and MELD datasets show that our method outperforms current state-of-the-art methods.**

***Index Terms*—Contextual interaction, cross-modal interaction, graph convolutional network, speech emotion recognition.**

I. Introduction

EMOTION plays a role in human communication. In recent years, Speech Emotion Recognition (SER) has attracted widespread attention in speech and natural language communities. It has been an essential sub-task in building the intelligent systems in many fields, such as voice assistant [1], mental health monitoring [2], and human-computer interaction [3]. Benefiting from the excellent performance of deep learning methods in SER, various neural network models have been developed to extract emotion-related information from either handcrafted acoustic features or raw audio signals, such as Convolutional Neural Networks (CNN) [4], [5], Recurrent Neural Networks (RNN) [6], [7], and their variants [8], [9]. However, the prior methods are mainly devoted to modeling speech segments in isolation without considering contextual interaction. Psychological researches [10], [11] emphasized the importance of characterizing the transitions and co-occurrences of emotional states in the dialog. For example, the emotion of a happy speaker is affected by the interlocutor's context, not just the current utterance. Consequently, to better understand a target speaker's current emotional state, his own previous state and utterances from his interacting partners are two prime contributions in contextual interactions of emotion.

Furthermore, most existing studies [12], [13], [14] on SER only focused on acoustic information. However, the ambiguity of human expression results in difficulty to effectively learn the representation of emotion with isolated audio modality [15]. To this end, recent works [16], [17], [18], [19] aimed at fusing text and audio modalities to improve the performance of the SER system. The textual information is also crucial because in some cases, the emotion of an utterance can be determined by linguistic semantics. For instance, "It's really a terrible day" indicates that the speaker is in a negative mood. Specifically, Peng et al. [19] proposed multi-scale CNN with statistical pooling units to learn the text and audio modalities. Wu et al. [16] enforced the time synchronous and asynchronous branches to capture correlations between each word and its acoustic realisation. Yoon et al. [17] presented a multi-hop attention for adaptive computation of the relevance of textual data and audio signals. Nevertheless, these methods ignored long-term dependencies between elements from different modalities.

In addition, the high inter-class similarity of data samples in the interactive dataset such as IEMOCAP is due to the annotators' subjective consciousness and prejudice [20]. For example, the acoustic utterances in the dialog have different emotional states but similar prosodic and tonal information. Additionally, the high intra-class variability also leads to ambiguity in emotion expression [21]. To this end, the publisher of the existing dataset encourages exploring effective methods to mitigate the effect of the subjective assignment.

Motivated by the above observations, in this letter, we propose a multimodal SER method based on interaction awareness. Fig. 1 shows an illustration of the pipeline. The first interaction integrates the speaker's previous utterance and the interlocutor's utterance into the current utterance to learn contextual information. The second interaction performs the fusion and reinforcement of information between text and audio modalities via the cross-modal attention mechanism. When people perceive ambiguous expression, they tend to infer emotion by association [22]. Thus, we further present associative learning strategy to alleviate intra- and inter-class problems in training data samples and make the model more robust. The primary contributions are summarized below:

1) We first propose a contextual transformer module to enhance the emotion representation of current utterance in spoken dialog by introducing contextual information.

Manuscript received 18 March 2022; revised 17 July 2022; accepted 22 August 2022. Date of publication 29 September 2022; date of current version 20 October 2022. This work was supported in part by the National Key R&D Program of China under Grants 2021ZD0113502 and 2021ZD0113503, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0103, and in part by the National Natural Science Foundation of China under Grant 82090052. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sandro Cumani. *(Corresponding author: Lihua Zhang.)*

The authors are with the Academy for Engineering and Technology, Fudan University, Shanghai 200433, China (e-mail: dkyang20@fudan.edu.cn; shuaihuang20@fudan.edu.cn; yang_liu20@fudan.edu.cn; lihuazhang@fudan.edu.cn).

Digital Object Identifier 10.1109/LSP.2022.3210836

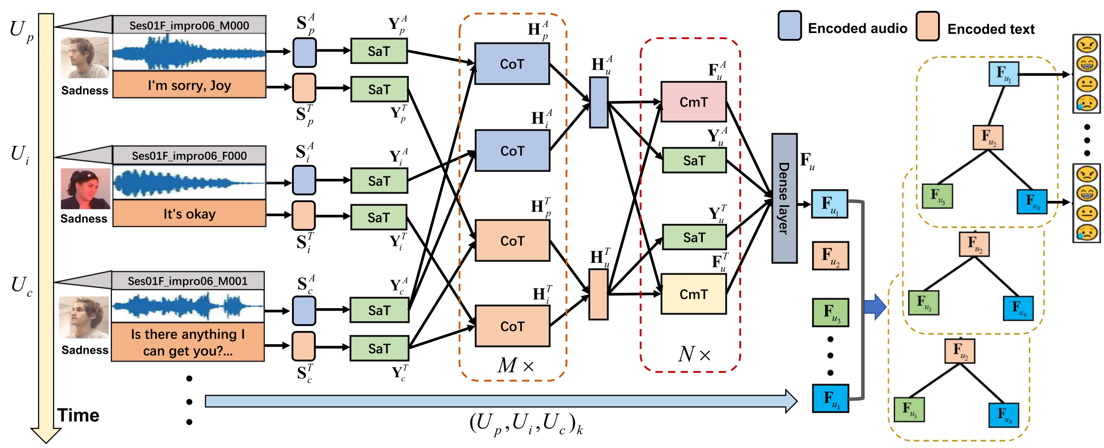

Fig. 1. The pipeline of our proposed method. "SaT" denotes self-attention transformer module. "CoT" denotes contextual transformer module. "CmT" denotes cross-modal transformer module. Interactive utterances from the IEMOCAP dataset.

Moreover, the proposed cross-modal transformer module facilitates feature fusion between different modalities.

2) We construct associative topological relation over mini-batch by similarity matrix with an adjacent regularization. Furthermore, the graph convolutional network is used for associative learning between deep features.
3) Experimental results on the IEMOCAP and MELD datasets demonstrate that our method outperforms previous state-of-the-art methods.

II. Proposed Method

A. Contextual Transformer Module

Our goal is to recognize a speaker's current utterance U_c by combining the speaker's previous utterance U_p and the interlocutor's utterance U_i. Therefore, (U_p, U_i, U_c) is a trainable data point with the label of U_c, containing both acoustic and textual information. We consider the sequences of two modalities (*i.e.*, audio as "A" and text as "T") corresponding to the above utterances as follows:

$$\left(\mathbf{S}_p^{\{A,T\}}, \mathbf{S}_i^{\{A,T\}}, \mathbf{S}_c^{\{A,T\}}\right) \in \mathbb{R}^{L_{\{A,T\}} \times d_{\{A,T\}}}, \tag{1}$$

where $L_{\{\cdot,\cdot\}}$ and $d_{\{\cdot,\cdot\}}$ denote sequence length and feature dimension, respectively. Immediately, we use a 1D temporal convolutional layer to process the input sequences and then augment them by the positional embedding [23]. The 1D convolutional layer projects the features of different modalities to the identical dimension. Then, the processed sequences are denoted as $\mathbf{S}_{\{p,i,c\}}^{\{A,T\}} \in \mathbb{R}^{L_{\{A,T\}} \times d}$.

In Fig. 1, the sequences of interactive utterances for both modalities begin with the parallel Self-attention Transformer module (SaT) [23] to capture the temporal dependencies, which are described as $\mathbf{Y}_{\{p,i,c\}}^{\{A,T\}} = \text{SaT}(\mathbf{S}_{\{p,i,c\}}^{\{A,T\}})$. After that, $\mathbf{Y}_c^{\{A,T\}}$ integrates the contextual information of $\mathbf{Y}_p^{\{A,T\}}$ and $\mathbf{Y}_i^{\{A,T\}}$ respectively by stacking M-layer Contextual Transformer modules (CoT). We present the details of the CoT as an example of integrating $\mathbf{Y}_i^A$ in the audio modality. In contextual attention, $\mathbf{Q}_c = \mathbf{Y}_c^A \mathbf{W}_{Q_c}^A$ is the embedded projection from the current utterance. $\mathbf{K}_i = \mathbf{Y}_i^A \mathbf{W}_{\mathbf{K}_i}^A$ and $\mathbf{V}_i = \mathbf{Y}_i^A \mathbf{W}_{\mathbf{V}_i}^A$ are the embedded projection from the interlocutor's utterance, where $\{\mathbf{W}_{Q_c}^A, \mathbf{W}_{K_i}^A, \mathbf{W}_{V_i}^A\} \in \mathbb{R}^{d \times d}$ are weights. The attention weights are obtained by applying the softmax function to scaled dot product of $\mathbf{Q}_c$ and $\mathbf{K}_i$. Then, the transmission of contextual information is expressed as:

$$\mathbf{F}_{i \to c}^A = \text{softmax}\left(\mathbf{Q}_c \mathbf{K}_i^T / \sqrt{d}\right) \mathbf{V}_i \in \mathbb{R}^{L_A \times d}. \tag{2}$$

Formally, the CoT computes feed-forwardly as follows:

$$\mathbf{H}_i^A = LN\left(\mathbf{Y}_c^A\right) + \mathbf{F}_{i \to c}^A, \tag{3}$$

$$\mathbf{H}_i^A = f_\theta\left(LN\left(\mathbf{H}_i^A\right)\right) + \mathbf{H}_i^A, \tag{4}$$

where LN means layer normalization, and $f_\theta(\cdot)$ is feed-forward network parametrized by θ. Following the same process, $\mathbf{H}_p^A$ is obtained by integrating contextual information $\mathbf{Y}_p^A$. Meanwhile, we can obtain both $\mathbf{H}_p^T$ and $\mathbf{H}_i^T$ from the text modality. Subsequently, $\mathbf{H}_u^A = [\mathbf{H}_p^A, \mathbf{H}_i^A] \in \mathbb{R}^{L_A \times 2d}$ and $\mathbf{H}_u^T = [\mathbf{H}_p^T, \mathbf{H}_i^T] \in \mathbb{R}^{L_T \times 2d}$ are the fused features obtained through concatenation operation, respectively.

B. Cross-Modal Transformer Module

Cross-modal Transformer module (CmT) focuses on guiding the transfer of one modality to another and learning the long-term dependencies between modalities. We adopt the potential adaptive process from audio to text to describe the details. Specifically, the three projection spaces for cross-modal attention are defined as $\mathbf{Q}_T = LN(\mathbf{H}_u^T)\mathbf{W}_{Q_u}^T$, $\mathbf{K}_A = LN(\mathbf{H}_u^A)\mathbf{W}_{K_u}^A$ and $\mathbf{V}_A = LN(\mathbf{H}_u^A)\mathbf{W}_{V_u}^A$, respectively, where $\{\mathbf{W}_{Q_u}^T, \mathbf{W}_{K_u}^A, \mathbf{W}_{V_u}^A\} \in \mathbb{R}^{2d \times 2d}$ are projection matrices. Further, the cross-modal interaction is defined as follows:

$$\mathbf{F}_u^{A \to T} = \text{softmax}\left(\mathbf{Q}_T \mathbf{K}_A^T / \sqrt{2d}\right) \mathbf{V}_A \in \mathbb{R}^{L_T \times 2d}. \tag{5}$$

The CmT is stacked with N layers to reinforce cross-modal fusion and information interaction progressively. Immediately, the forward computation is expressed as:

$$\mathbf{F}_u^T = LN\left(\mathbf{H}_u^T\right) + \mathbf{F}_u^{A \to T}, \tag{6}$$

$$\mathbf{F}_u^T = f_\varphi\left(LN\left(\mathbf{F}_u^T\right)\right) + \mathbf{F}_u^T, \tag{7}$$

where $f_{\varphi}(\cdot)$ is feed-forward network parametrized by φ. Another branch ($i.e.$, from text to audio) follows the same way to obtain $\mathbf{F}_u^A$. Moreover, $\mathbf{H}_u^A$ and $\mathbf{H}_u^T$ also enhance their own feature representations by parallel SaTs to obtain $\mathbf{Y}_u^A$ and $\mathbf{Y}_u^T$, respectively. After that, the above features are merged through dense layers parametrized by ϕ to obtain $\mathbf{F}_u$, which is expressed as $\mathbf{F}_u = f_{\phi}([\mathbf{F}_u^A, \mathbf{Y}_u^A, \mathbf{F}_u^T, \mathbf{Y}_u^T])$. In practice, all transformer modules enforce multi-head attention [23].

C. Associative Learning Based on GCN

Inspired by the research [24] that connected nodes are likely to share the same label, we assume that intra- and inter-class samples have similar and opposite high-level features respectively and use graph structures for object association. Formally, we construct associative relation in the form of adjacent matrix $\mathbf{A}$ over mini-batch by data-driven method. The data samples over mini-batch will be transformed to feature vectors $(\mathbf{F}_{u_1}, \mathbf{F}_{u_2}, \ldots, \mathbf{F}_{u_k})$ through the model described above. We calculate the cosine similarity matrix $\mathbf{C}_s$ as:

$$\mathbf{C}_s(i,j) = \frac{\mathbf{F}_{u_i} \cdot \mathbf{F}_{u_j}}{\|\mathbf{F}_{u_i}\| \left\|\mathbf{F}_{u_j}\right\|}, (i,j = 0, 1, \ldots, k). \tag{8}$$

Specifically, we use threshold γ to filter noisy edges, and the operation is described as follows:

$$\mathbf{A}(i,j) = \begin{cases} 1, & \text{if } \mathbf{C}_s(i,j) \geq \gamma, \\ 0, & \text{if } \mathbf{C}_s(i,j) < \gamma. \end{cases} \tag{9}$$

To further optimize $\mathbf{A}$, we consider features with same labels should be connected in graph, and the corresponding positions must be 1 in $\mathbf{A}$. Thus, the mask matrix $\mathbf{A}_m(i,j)$ can be constructed according to labels over mini-batch, $i.e.$, if $\text{label}_i = \text{label}_j$, then $\mathbf{A}_m(i,j) = 1$, otherwise is 0. Consider reducing the complexity of the model, we define graph convolution in the 2nd order Chebyshev expansion as:

$$g_\theta \star \mathbf{f} \approx \boldsymbol{\theta} \left(\mathbf{I}_N + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}\right) \mathbf{f}, \tag{10}$$

where $\mathbf{D} = \sum_j \mathbf{A}_{ij}$, θ is parameter of Graph Convolutional Network (GCN) and $\mathbf{f}$ is the feature. Hence, we can simplify $\mathbf{I}_N + \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ to $\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$, where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ and $\tilde{\mathbf{D}} = \sum_j \tilde{\mathbf{A}}_{ij}$. Finally, the graph convolution can be written as follows:

$$\mathbf{Z} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{F} \Theta. \tag{11}$$

We can model the complex inter-relationships of the nodes by stacking multiple GCN layers. The details can be found in [24]. Eventually, the loss as $\mathcal{L}_{\text{adj}}$ is posed to partly regularize the matrix $\mathbf{A}$, which is formulated as follows:

$$\mathcal{L}_{\text{adj}} = \lambda \left(\sum_{i,j} \mathbf{A}_m * \mathbf{A} - \sum_{i,j} \mathbf{A}_m\right)^2, \tag{12}$$

where λ is a weight coefficient. The standard cross-entropy loss $\mathcal{L}_{\text{class}}$ is used for classification, and total loss is expressed as $\mathcal{L} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{adj}}$.

III. Experiments

A. Datasets and Evaluation Metrics

1) IEMOCAP [20]: A standard dataset that is widely used in SER. It contains approximately 12 hours of audiovisual data from 10 speakers. For consistent comparison with the previous state-of-the-art methods, all utterances labeled *excitement* are merged with those labeled *happy*. This letter utilizes 5531 utterances containing four emotion categories: *sadness* (1084 utterances), *angry* (1103 utterances), *neutral* (1708 utterances) and *happy* (1636 utterances). Furthermore, a 10-fold cross-validation is performed with 8, 1, 1 in train, dev, test set, respectively. The Weighted Accuracy (WA) and Unweighted Accuracy (UA) are adopted as the evaluation metrics.

2) MELD [21]: A new multimodal dataset of 13,708 utterances with seven emotions of 1,433 dialogues from the classic TV-series Friends. The distribution of data samples in MELD is training 73%, validation 8%, and testing 19%. Following [25], [26], we report the weighted average F1 score.

B. Feature Extraction and Implementation Details

For the audio processing, we use COVAREP toolkit [29] for extracting 74-dimensional low level acoustic features. The features include 12 Mel-Frequency Cepstral Coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, etc. Meanwhile, we convert the ground-truth transcripts of both datasets into pre-trained Glove word embedding [30], and the embedding is a 300-dimensional vector. In the practical scenario where we may not have access to audio transcripts, the Google Speech-to-Text API [1] is used to generate the Automatic Speech Recognition (ASR) transcripts. The word error rate for the API on the IEMOCAP and MELD datasets are 5.80% and 7.56%, respectively. Our method is built on the Pytorch toolbox with four Nvidia Tesla V100 GPUs. The number of contextual and cross-modal transformer module are 4 and 5, respectively. All attention heads are 8. We minimize the loss using the Adam optimizer with a learning rate of $2e^{-3}$ on the IEMOCAP and $5e^{-4}$ on the MELD. We train all models 30 epochs on the IEMOCAP and 20 epochs on the MELD. Meanwhile, the batch size is 32 and 256, respectively.

C. Comparison With State-of-The-Art Methods

Table I shows the results of our method compared to previous state-of-the-art methods [16], [17], [18], [19], [27] on the IEMOCAP dataset. First, we train models with single modality ($i.e.$, utterances or ground-truth transcripts only). In this setting, we remove the CmT and keep only the SaT. The results show that the text modality achieves better performance than the audio modality. When performing multimodal learning using ground-truth transcripts, our method outperforms previous SOTA [19] by 3.44% and 2.87% on WA and UA, respectively. Furthermore, we observe a significant performance drop in all models when using ASR-processed transcripts. However, our method is the most competitive as the drop in both WA and UA is less than about 2%. Additionally, we evaluate the effectiveness of the proposed method on the MELD dataset. The results in Table II provide the following observations. First, the text modality still achieves better results in the unimodal setting. Following the

[1] Google Cloud Speech-to-Text: https://cloud.google.com/speech-to-text/

TABLE I
COMPARISON RESULTS ON THE IEMOCAP DATASET USING GROUND-TRUTH TRANSCRIPTS AND ASR-PROCESSED TRANSCRIPTS. 'A' AND 'T' DENOTE AUDIO AND TEXT MODALITY, RESPECTIVELY

Methods	Modality	WA	UA
Ground-Truth Transcripts			
Only Audio (ours)	A	66.26%	67.51%
Only Text (ours)	T	68.64%	69.28%
MDRE [18]	A+T	71.80%	-
Xu *et al.* [27]	A+T	72.50%	70.90%
MHA-2 [17]	A+T	76.50%	77.60%
TSB+TAB [16]	A+T	77.76%	78.30%
MSCNN-SPU-ATT [19]	A+T	80.30%	81.40%
Full (ours)	A+T	**83.74%**	**84.27%**
ASR-Processed Transcripts			
Only Text(ours)	T	63.77%	64.61%
MDRE [18]	A+T	69.10%	-
MHA-2 [17]	A+T	73.00%	73.90%
MSCNN-SPU-ATT [19]	A+T	78.00%	79.10%
Full (ours)	A+T	**82.13%**	**83.25%**

TABLE II
COMPARISON RESULTS ON THE MELD DATASET USING GROUND-TRUTH TRANSCRIPTS AND ASR-PROCESSED TRANSCRIPTS

Methods	Modality	F1 score
Ground-Truth Transcripts		
Only Audio (ours)	A	0.412
Only Text (ours)	T	0.465
cMKL [28]	A+T	0.555
Liang *et al.* [26]	A+T	0.561
MCSAN [25]	A+T	0.592
Full (ours)	A+T	**0.638**
ASR-Processed Transcripts		
Only Text (ours)	T	0.387
Full (ours)	A+T	0.614

TABLE III
ABLATION STUDIES ON THE IEMOCAP AND MELD DATASETS, RESPECTIVELY

Methods	IEMOCAP		MELD
	WA	UA	F1 score
Full (ours)	**83.74%**	**84.27%**	**0.638**
w/o CoT	80.26%	81.35%	0.587
w/o CmT	82.32%	82.47%	0.615
w/o GCN	81.68%	82.24%	0.620

prior methods [25], [26], [28] of using ground-truth transcripts, our method achieves better performance. When using ASR-processed transcripts, our model still outperforms the SOTA MCSAN [25] using ground-truth transcripts by 2.2% absolute value in terms of weighted average F1 score. The above results clearly demonstrate the superiority of the proposed method.

D. Ablation Studies

In Table III, we perform ablation studies to verify the necessity of the different components on the IEMOCAP and MELD datasets, respectively. Note that all experiments use ground-truth transcripts. First, we evaluate the effect of the proposed transformer modules. When the CoT modules are removed, we switch to training on each utterance sample. When the CmT modules are removed, we keep only the SaT modules for each

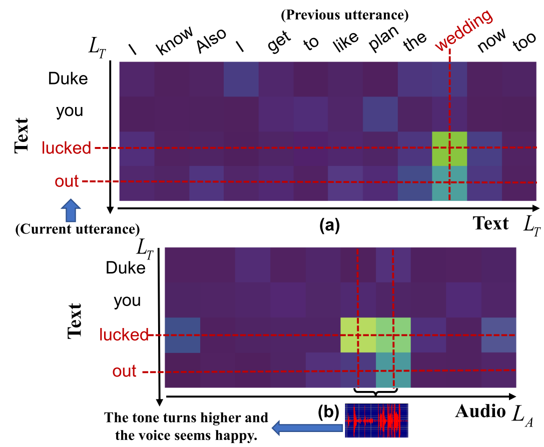

Fig. 2. Attention activation of the last contextual transformer module (a) and the last cross-modal transformer module (b). We observe that contextual and cross-modal attention can clearly capture emotion cues in the dialog context and across different modalities.

modality to update the features. In both cases, the model's performance decreased by 3.48%/1.42% in terms of WA and 2.92%/1.80% in terms of UA on the IEMOCAP dataset, and by 5.1%/2.3% in terms of F1 score on the MELD dataset. The above observations demonstrate that it is necessary to model both the contextual and cross-modal interactions. Further, we find that the model's performance when the CoT modules are removed is comparable to that of the previous SOTA [19], [25], suggesting that the introduction of contextual information in the dialogical SER system is essential. Finally, we evaluate the potential of the GCN-based associative learning strategy. When the GCN is removed, the results of about a 2% decrease in all metrics suggest that considering intra- and inter-class relationships among samples is indispensable.

E. Qualitative Results

In Fig. 2, we show the attention activation of the last contextual and cross-modal transformer modules, respectively. The example is from a dialog on the IEMOCAP dataset using ground-truth transcripts. We find that contextual attention focuses on the intersection of meaningful signals in the current and previous utterances in Fig. 2(a). Concretely, the spoken words "lucked," "out" and "wedding" suggest happy emotion. From Fig. 2(b), the emotion-related words successfully attend to the audio clips that contain the corresponding high tone. The above observations prove that our method can clearly capture emotion-related signals in different interactions.

IV. CONCLUSION

This letter presents a novel multimodal SER method for learning contextual and cross-modal interactions. The proposed contextual transformer module effectively combines contextual information to enhance the representation of the current utterance. Besides, the cross-modal transformer module facilitates the feature fusion and information exchange between audio and text modalities. We also introduce an associative learning strategy to further improve the performance of the SER system. Experimental results on the IEMOCAP and MELD datasets fully show the superiority of our method.

References

[1] J. C. Castillo, Á. Castro-González, F. Alonso-Martín, A. Fernández-Caballero, and M. Á. Salichs, "Emotion detection and regulation from personal assistant robot in smart environment," in *Personal Assistants: Emerging Computational Technologies*. Berlin, Germany: Springer, 2018, pp. 179–195.

[2] S. Ghosh, S. Sahu, N. Ganguly, B. Mitra, and P. De, "EmoKey: An emotion-aware smartphone keyboard for mental health monitoring," in *Proc. IEEE Int. Conf. Commun. Syst. Netw.*, 2019, pp. 496–499.

[3] G. Wadley et al., "The future of emotion in human-computer interaction," in *Proc. CHI Conf. Comput. Syst. Extended Abstr.*, 2022, pp. 1–6.

[4] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," *Sensors*, vol. 22, no. 4, 2022, Art. no. 1414.

[5] M. Ezz-Eldin, A. A. Khalaf, H. F. Hamed, and A. I. Hussein, "Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition," *IEEE Access*, vol. 9, pp. 19999–20011, 2021.

[6] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. W. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017. [Online]. Available: https://doi.org/10.1109/LSP.2017.2672753

[7] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Syst. Appl.*, vol. 173, 2021, Art. no. 114683.

[8] O. Kwon, I. Jang, C. Ahn, and H. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1383–1387, Sep. 2019. [Online]. Available: https://doi.org/10.1109/LSP.2019.2931673

[9] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018. [Online]. Available: https://doi.org/10.1109/LSP.2018.2860246

[10] L. F. Barrett and J. A. Russell, *The Psychological Construction of Emotion*. New York, NY, USA: Guilford Press, 2014.

[11] A. Manstead et al., *Gender and Emotion: Social Psychological Perspectives*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[12] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5734–5738.

[13] S. Kanwal and S. Asghar, "Speech emotion recognition using clustering based GA-optimized feature set," *IEEE Access*, vol. 9, pp. 125830–125842, 2021.

[14] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Appl. Acoust.*, vol. 179, 2021, Art. no. 108046.

[15] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription," in *Proc. Annu. Conf. Int. Speech. Commun. Assoc.*, 2019, pp. 3302–3306.

[16] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6269–6273.

[17] 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.

[18] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 112–118.

[19] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale CNN and attention," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 3020–3024.

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

[21] 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. Conf. Annu. Meet. Assoc. Comput. Linguist.*, 2019, pp. 527–536.

[22] Y. Du, D. Yang, P. Zhai, M. Li, and L. Zhang, "Learning associative representation for facial expression recognition," in *Proc. Int. Conf. Image Process.*, 2021, pp. 889–893.

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

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[25] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross-and self-attention network for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 4275–4279.

[26] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proc. 28th Int. Conf. Multimedia*, 2020, pp. 2852–2861.

[27] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Proc. Annu. Conf. Int. Speech. Commun. Assoc.*, 2019, pp. 3569–3573.

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

[29] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 960–964.

[30] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.