MULTI-MODAL EMOTION RECOGNITION WITH SELF-GUIDED MODALITY CALIBRATION

Mixiao Hou, Zheng Zhang[*]*, Guangming Lu*[*]

Harbin Institute of Technology, Shenzhen, China

ABSTRACT

Multi-modal emotion recognition aims to extract sentiment-related information from multiple sources and integrate different modal representations for sentiment analysis. Alignment is an effective strategy to achieve semantically consistent representations for multi-modal emotion recognition, while the current alignment models are jointly unable to maintain the dependence of word-to-sentence and independence of unimodal learning. In this paper, we propose a Self-guided Modality Calibration Network (SMCN) to realize multi-modal alignment which can capture the global connections without interfering with unimodal learning. While preserving unimodal learning without interference, our model leverages semantic sentiment-related features to guide modality-specific representation learning. On one hand, SMCN simulates human thinking by deriving a branch for acquiring knowledge of other modalities in unimodal learning. This branch aims to lean high-level semantic information of other modalities for realizing semantic alignment between modalities. On the other hand, we also provide an indirect interaction manner to integrate unimodal feature and calibrate features in different levels for avoiding unimodal features mixed with other clues. Experiments demonstrate that our approach outperforms the state-of-the-art methods on both IEMOCAP and MELD datasets.

Index Terms— Multi-modal Emotion Recognition, Feature Calibration, Indirect Interaction, Feature Fusion

1. INTRODUCTION

The perception of humans towards emotions is often accompanied by imagination and experience. When dealing with unimodal information sources, people often imagine more clues of other signals in their minds to enhance their intuitive feelings. Therefore, we can also simulate this characteristic of humans to deal with multi-source signals in the emotion recognition task. Similarly, unimodal feature learning can derive multiple semantic spaces to model the process of human emotion perception. In this work, we assume that each modality will generate a new semantic space for other modalities in the modality-specific feature learning to realize the purpose of emotional perception.

In recent years, multi-modal emotion recognition has received increasing attention [1, 2, 3, 4]. Since data are always collected from different sources and represented by the heterogeneous signals, alignment is one of the useful strategies to ensure the semantic consistency of multi-modal learning. There are two frequently-used modal alignment ways, that is, word-level alignment, and fine-grained alignment based on the attention mechanism. For example, Gu et al. [5] proposed a hierarchical multi-modal network for achieving the features fusion between each word and corresponding aligned acoustic frames. However, word-level alignment mainly focuses on the feature fusion of each word, and ignoring cross-word connections between sequences will also affect sentence emotion. Therefore, many works leverage attention mechanisms to capture the relationships between words and frames for obtaining the global influence in utterance level [6, 7, 8]. Xu et al. [9] introduced an attention to detect the alignment between acoustic frames and words for emotion recognition. Nevertheless, this alignment approach incorporates information of other modalities into the unimodal learning, resulting in the unimodal learning process being extremely disturbed.

In this paper, we propose a Self-guided Modality Calibration Network (SMCN) for multi-modal emotion recognition based on audio and text. SMCN employs information from other modalities to generate calibration features that are aligned with modality-specific features, and then integrates the calibration and unimodal features for multi-modal fusion. Specifically, our model consists of three modules: intra-modal feature learning, inter-modal calibration learning, and indirect interaction. The intra-modal feature learning focuses on specific information mining and maintaining the learning independence of a single modality. The inter-modal calibration learning is designed to learn semantic knowledge from other modalities in the unimodal encoding space. The indirect interaction can realize the aligned features fusion, so that the specific modality can supplement the information from the same semantic space influenced by other modalities.

The main contributions of this paper are summarised as follows: (1) A Self-Guided Modality Calibration Network (SMCN) is proposed to realize the semantic information alignment between multi-modalities. Our method ensures that frame-level and word-level features remain globally connected without interfering with unimodal learning. (2) We

[*] Corresponding authors: zhengzhang@hit.edu.cn, luguangm@hit.edu.cn.

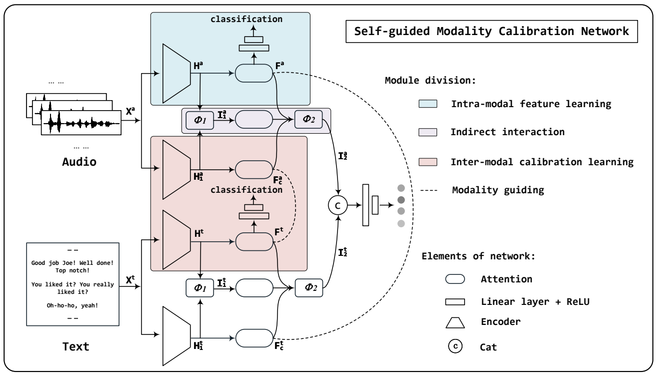

Fig. 1. The architecture of our SMCN. There are three branches in unimodal learning: intra-modal feature learning, inter-modal calibration learning, and indirect interaction.

introduce inter-modal learning module which can implement the learned features by intra-modal learning of other modals for guiding specific modality to calibrate semantic features with global properties. (3) We present an indirect interaction by integrating the modality-specific feature and calibration feature in utterance level. In this way, our model allows for the fusion of multi-modal representations while ensuring that specific modality only focuses on its own learning. (4) Experiments demonstrate that SMCN outperforms state-of the-art approaches on the IEMOCAP and MELD datasets.

2. METHOD

In this section, we will introduce SMCN from three modules: intra-modal feature learning, inter-modal calibration learning, and indirect interaction. The framework of the proposed method is shown in Fig.1. The intra-modal feature learning aims to capture specific characteristics of each modality, the inter-modal calibration learning can achieve an approximation of different modalities, and the indirect interaction will realize the fusion between modality-specific feature and aligned feature with other modal information.

2.1. Intra-modal Feature Learning

In this work, the inputs of the network for audio and text are frame-level and word-level features, respectively. The acoustic feature is described as $\mathbf{X}^a \in \mathbb{R}^{N \times f_a \times l}$, and the textual feature is presented as $\mathbf{X}^t \in \mathbb{R}^{N \times f_t \times d}$. Herein, N is the number of samples, and f_a and f_t indicate the number of frames and words, respectively. l is the number of the describers for acoustic frames, and d is the number of features about words in each sample. The intra-modal feature learning is the process of emotional analysis of the specific modality. Firstly, the sequential features from the preliminary sampling space are obtained by separate encoding:

$$\mathbf{H}^a = \overset{\leftrightarrows}{GRU}\left(\mathbf{X}^a\right), \tag{1}$$

$$\mathbf{H}^t = \overset{\leftrightarrows}{GRU}\left(\mathbf{X}^t\right), \tag{2}$$

where $\mathbf{H}^a \in \mathbb{R}^{N \times f_a \times 2hd}$ and $\mathbf{H}^t \in \mathbb{R}^{N \times f_t \times 2hd}$. hd represents the hidden cells of Bi-GRU. Then, the high-level semantic features $\mathbf{F}^i$ can be learned by attention mechanism in sentence units:

$$\mathbf{F}^i = Attention(\mathbf{H}^i; \mathbf{W}_h^i, \mathbf{b}_h^i), \tag{3}$$

where $i \in \{a, t\}$ and $\mathbf{F}^i$ is the weighted high-level semantic feature. Specifically,

$$c\left(h_m^i\right) = matmul\left(\tanh(\mathbf{W}_h^i h_m^i + \mathbf{b}_h^i), \mathbf{u}^i\right), \tag{4}$$

where $h_m^i \in \mathbf{H}^i$ and $\mathbf{H}^i \in \{\mathbf{H}^a,\ \mathbf{H}^t\}$. $\mathbf{W}_h^i$, $\mathbf{b}_h^i$ and $\mathbf{u}^i$ are learnable parameters. The attention weight α_m^i can be defined by:

$$\alpha_m^i = \frac{\exp\left(c(h_m^i)\right)}{\sum_{m=1}^{M} \exp\left(c(h_m^i)\right)}. \tag{5}$$

In this way, the output with attention learning can be obtained by multiplying features with weights:

$$f_k^i = \sum\nolimits_{m=1}^{M} \alpha_m^i h_m^i,\ k \in [1, 2hd] \tag{6}$$

where $M \in \{f_a, f_t\}$ is a scalar with a temporal dimension. Finally, the high-level semantic information of audio $\mathbf{F}^a \in \mathbb{R}^{N \times 2hd}$ and text $\mathbf{F}^t \in \mathbb{R}^{N \times 2hd}$ can be used for classification:

$$p^i = \mathit{softmax}\left(linear\left(\mathbf{F}^i; \mathbf{W}_y^i, \mathbf{b}_y^i\right)\right), \tag{7}$$

where $linear(.)$ indicates two linear layers. The corresponding Cross-Entropy loss function is written as follows:

$$Loss_i = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{nc} \log\left(p_{nc}^i\right), \tag{8}$$

where C is the total number of classes and y is annotated emotional label. The intra-modal feature learning can detect unimodal sentiment-related representations under the constraints of emotional labels.

2.2. Inter-modal Calibration Learning

The inter-modal calibration learning can learn other modal information in a modality-specific space for semantic alignment. The aligned time-level features can be encoded in the same feature space by:

$$\mathbf{H}_1^i = \overset{\leftrightarrows}{GRU}\left(\mathbf{X}^i\right). \tag{9}$$

This stage ensures that the basic semantics of the imaginary space is obtained from specific modality. The high-level semantic information is obtained by the attention mechanism:

$$\mathbf{F}_c^i = Attention\left(\mathbf{H}_1^i; \mathbf{W}_c^i, \mathbf{b}_c^i\right), \tag{10}$$

where $\mathbf{F}_c^i \in \mathbb{R}^{N \times 2hd}$, and $\mathbf{W}_c^i$ and $\mathbf{b}_c^i$ are parameters. In order to guarantee that specific modality can learn the semantic information of other modalities, we constrain the intra-modal representation and inter-modal representation with Mean Square Error (MSE) function:

$$Loss_{ij} = \frac{1}{N}\sum_{n}^{N}(\mathbf{F}^{j} - \mathbf{F}_{c}^{i})^{2}, \tag{11}$$

where $i, j \in \{a, t\}$ and $i \neq j$. In this way, we hope that the semantic information of other modalities can be learned in the encoding space of a particular modality. The inter-modal information has a mutual guidance effect to calibrate the derived representation of specific modality.

2.3. Indirect Interaction

Unlike interacting directly with representations obtained from different modalities, the indirect interaction combines the information of intra-modal feature learning and inter-modal calibration learning to derive a new representation with the same dimension. Without affecting the other two learning modules, the indirect interaction realizes the multi-modal interaction in two scales. After encoding by Bi-GRU, the first interaction fusion information $\mathbf{I}_1^i$ of the modality i is obtained by:

$$\mathbf{I}_1^i = \Phi_1(\mathbf{H}^i, \mathbf{H}_1^i), \tag{12}$$

where $\mathbf{I}_1^a \in \mathbb{R}^{N\times f_a \times 2hd}$ and $\mathbf{I}_1^t \in \mathbb{R}^{N\times f_t \times 2hd}$. $\Phi_1(.)$ can be any form of interactive fusion function that keeps the shape of representation unchanged. After attention learning, $\mathbf{I}_1^i$ is transformed into high-level semantic information:

$$\mathbf{I}_1^{i'} = Attention(\mathbf{I}_1^i; \mathbf{W}_a^i, \mathbf{b}_a^i), \tag{13}$$

where $\mathbf{I}_1^{i'} \in \mathbb{R}^{N\times 2hd}$, and $\mathbf{W}_a^i$ and $\mathbf{b}_a^i$ are also parameters. On this basis, the model performs the second interaction for accessing the final fusion features:

$$\mathbf{I}_2^i = \Phi_2(\mathbf{I}_1^{i'}, \mathbf{F}^i, \mathbf{F}_c^i), \tag{14}$$

where $\mathbf{I}_2^{'} \in \mathbb{R}^{N\times 2hd}$ and $\Phi_2(.)$ is interaction function of high-level semantic information. $\Phi_1(.)$ and $\Phi_2(.)$ are changeable in our model framework. For convenience, we apply simple addition to achieve fusion in this work. The fusion features of each modality are spliced together and sent to the linear layer to get the emotional prediction:

$$p = \mathit{softmax}\left(linear\left([\mathbf{I}_2^a, \mathbf{I}_2^t]; \mathbf{W}_p^i, \mathbf{b}_p^i\right)\right), \tag{15}$$

where $\mathbf{W}_p^i$ and $\mathbf{b}_p^i$ are parameters. $[\cdot,\cdot]$ is concentration operation. The loss function at this stage can be defined by:

$$Loss_p = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{nc}\log(p_{nc}). \tag{16}$$

Thus, the total loss function of the model is given as follows:

$$Loss = Loss_p + \sum_{i=a, j=t}^{t,a}(Loss_i + Loss_{ij}), i \neq j. \tag{17}$$

3. EXPERIMENTS

3.1. Dataset

In this session, we verify the effect of the proposed model on two datasets: the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [10] and the Multimodal EmotionLines Dataset (MELD) [11]. **IEMOCAP** is a dialogues dataset which performs improvised and scripts by 10 actors. The 10 actors are divided into 5 sessions, and every session consists of 1 male and 1 female. In the experiments, we select the 5531 commonly used samples that cover four emotions: happy (excited is also merged into happy), anger, sad and neutral for training and testing. Additionally, the 5-fold leave-one-session-out (LOSO) strategy is used for experiments. **MELD** is a multi-party conversational dataset and the number of speakers in each conversation is not fixed. This dataset has divided testing, training and verification, and it consists of a total of more than 13000 samples covering seven emotions: angry, sad, joy, neutral, surprise, fear, and disgust.

3.2. Implementation Details and Metric

For each acoustic sample, following the previous work [12], we extract 78-dimensional (f_a) log Mel-Bank features (covering first derivative and second derivative) and conduct mean pooling with 5 steps and 4 steps to get the 500 frames (l) on the IEMOCAP and MELD datasets. The pre-trained BERT model (uncased_L-12_H-768_A-12) is utilized to encode text and obtain 768-dimensional (d) features. We fix each utterance with 150 and 100 words (f_t) on the IEMOCAP and MELD datasets, respectively.

The learning rate and mini-batch sizes are set to 0.0001 and 32, and the hidden units for Bi-GRU are all set to 128. We run 2500 and 2000 steps for training on IEMOCAP and MELD datasets in the experiments. Since the commonly-used metrics on the two databases are different, we select the universal metrics on each dataset for measurement: unweighted accuracy (UA) and weighted accuracy (WA) for the IEMOCAP dataset, and accuracy (Acc) and weighted F1-score (W-F1) for the MELD dataset.

3.3. Results and Analysis

The comparison results of our method with other methods on the IEMOCAP are listed in Table 1. FAF is a word-level alignment fusion method, which is inferior to our algorithm due to the lack of word dependencies across sequences. Although FAF can realize more accurate semantic fusion, it ignores the effect between words. A-LSTM, GBAN, CMA+Raw waveform and CME are multi-modal alignment networks based on attention mechanism. These methods supplement modal information to specific modality by calculating the correlation between different modalities, which increases the computation and also generates useless relationships that interfere with learning. In contrast, SMCN completes feature alignment at the semantic level while maintaining that the dimension of features remains unchanged. It

Table 1. Comparison of methods on the IEMOCAP (%)

Method	UA	WA
FAF [5]	72.7	72.7
A-LSTM [9]	70.9	72.5
GBAN [13]	70.1	72.4
CMA + Raw waveform [14]	72.8	-
CME [15]	73.5	72.7
SMCN	**77.6**	**75.6**

Table 2. Comparison of methods on the MELD (%)

Method	Acc	W-F1
DialogueRNN [11, 17]	-	60.3
MMFA-RNN [18]	63.3	60.6
GNN+self-attention [19]	61.8	-
Semi-supervised MMER [20]	-	57.1
CTNet [21]	60.8	62.0
SMCN	**64.9**	**62.3**

is experimentally demonstrated that in the hypothetically aligned feature space, the fusion of features by our model enables effective classification.

Table 2 lists the comparison of algorithms on the MELD dataset. DialogueRNN, GNN+self-attention and Semi-supervised MMER lack interaction in the process of unimodal learning, which display unsatisfactory performances. MMFA-RNN and CTNet are based on the transformer[16], and they interact directly with other modalities on the basis of unimodal learning, which is better than the above methods without interaction. However, direct interaction often destroys the characteristics learned by each modality, resulting in the unimodal features being assimilated by other modalities. SMCN adopts indirect interaction by providing an intermediate learning between intra-modal feature learning and inter-modal calibration learning to integrate the characteristics between modalities. In general, SMCN can steadily learn the semantic information of other modalities with promising results while retaining the independence of unimodal learning.

3.4. Ablation Study

In order to explore the contribution of each module in SMCN, we conduct a series of ablation experiments, and the results are listed in Table 3. **Audio** and **Text** are unimodal emotion recognition by using Bi-GRU and attention mechanism, and **Cat** combines the high-level semantic features of the two modalities for the final emotion recognition. **Intra-modal learning** concentrates the semantic representations ($\mathbf{F}^a$ and $\mathbf{F}^t$) of intra-modal feature under the framework of SMCN for recognition. **Inter-modal calibration** leverages the representations ($\mathbf{F}_c^a$ and $\mathbf{F}_c^t$) of semantic calibration space for classification. **First interaction** only contains the first interaction process with temporal scale, while **Second interaction** only preserves the second interaction stage in utterance level.

Table 3. Comparison of different modules (%)

Method	IEMOCAP		MELD	
	UA	WA	Acc	W-F1
Audio	62.7	59.9	48.9	35.8
Text	68.6	67.6	62.5	60.9
Cat	75.4	73.2	63.6	60.4
Intra-modal learning	76.2	73.9	63.7	60.3
Inter-modal calibration	76.7	75.1	64.2	61.1
First interaction	76.6	74.6	64.4	61.3
Second interaction	76.9	75.2	64.0	60.5
SMCN	**77.6**	**75.6**	**64.9**	**62.3**

From table 3, we draw the following conclusions: (1) Compared with simple cat, the effect of interaction is greatly improved, which indicates the importance of considering interaction for multi-modal learning. (2) In our framework, the multi-modal interaction can enhance the recognition performance by utilizing simple addition operation, which shows that the idea of indirect interaction is completely feasible. (3) For the two learning spaces with different semantic tasks **Intra-modal learning** and **Inter-modal calibration**, their discrimination of emotion is also different. **Inter-modal calibration** is influenced by other modalities and is effective in distinguishing emotion, which verifies the feasibility of deriving this semantic alignment branch from unimodal learning. (4) The interaction of the two stages is indispensable, which implies that the modal information on different scales is sensitive to emotion recognition. Obviously, making full use of the information of different levels can clearly enhance the ability of emotion recognition.

4. CONCLUSION

In this paper, we propose a multi-modal sentiment recognition method based on mutual learning between modalities. This method hypothesizes the semantic information of another modality on the basis of the specific modality and achieves features calibration at different levels. In order to make the unimodal learning uninterrupted, we also present an indirect interaction strategy by fusing features within the modality and features learned from other modalities together for emotion recognition. Experiments prove that our model outperforms the state-of-the-art approaches.

5. ACKNOWLEDGMENTS

This work was supported in part by the NSFC fund 62176077, 62002085, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019Bl515120055, in part by the Shenzhen Key Technical Project under Grant 2020N046, in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20210324132210025, JCYJ20210324132212030, and JSGG20200103103401723.

6. REFERENCES

[1] Eesung Kim and Jong Won Shin, "Dnn-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. ICASSP*. IEEE, 2019, pp. 6720–6724.

[2] Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustín Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *Proc. ICASSP*. IEEE, 2020, pp. 6484–6488.

[3] Wen Wu, Chao Zhang, and Philip C Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *Proc. ICASSP*. IEEE, 2021, pp. 6269–6273.

[4] Bingzhi Chen, Cao Qi, Hou Mixiao, Zheng Zhang, Guangming Lu, and David Zhang, "Multimodal emotion recognition with temporal and semantic consistency," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2021.

[5] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. ACL*, 2018, vol. 2018, pp. 2225–2235.

[6] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. ICASSP*. IEEE, 2019, pp. 2822–2826.

[7] Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes, "Attention driven fusion for multi-modal emotion recognition," in *Proc. ICASSP*. IEEE, 2020, pp. 3227–3231.

[8] Zixuan Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," in *Proc. ICASSP*. IEEE, 2021, pp. 3020–3024.

[9] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li, "Learning alignment for multimodal emotion recognition from speech," in *Proc. INTERSPEECH*, 2019, pp. 3569–3573.

[10] Carlos Busso, Murtaza Bulut, ChiChun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–339, 2008.

[11] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL*, 2019, pp. 527–536.

[12] Mixiao Hou, Zheng Zhang, Qi Cao, David Zhang, and Guangming Lu, "Multi-view speech emotion recognition via collective relation construction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 218–229, 2022.

[13] Pengfei Liu, Kun Li, and Helen Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition.," in *Proc. INTERSPEECH*, 2020, pp. 379–383.

[14] Krishna D. N. and Ankita Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks," in *Proc. INTERSPEECH*, 2020, pp. 4243–4247.

[15] Hang Li, Wenbiao Ding, Zhongqin Wu, and Zitao Liu, "Learning fine-grained cross modality excitement for speech emotion recognition," in *Proc. INTERSPEECH*, 2021, pp. 3375–3379.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[17] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Proc. AAAI*, 2019, pp. 6618–6625.

[18] Ngoc-Huynh Ho, Hyung-Jeong Yang, Soo-Hyung Kim, and Gueesang Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[19] Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li, "Conversational emotion recognition using self-attention mechanisms and graph neural networks," in *Proc. INTERSPEECH*, 2020, pp. 2347–2351.

[20] Jingjun Liang, Ruichen Li, and Qin Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proc. ACM MM*, 2020, pp. 2852–2861.

[21] Zheng Lian, Bin Liu, and Jianhua Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.