DWFORMER: DYNAMIC WINDOW TRANSFORMER FOR SPEECH EMOTION RECOGNITION

Shuaiqi Chen[1], *Xiaofen Xing*[1*], *Weibin Zhang*[2], *Weidong Chen*[1], *Xiangmin Xu*[3 4]

[1]School of Electronic and Information Engineering, South China University of Technology, China
[2] VoiceAI Technologies, Shenzhen, China [3] Pazhou Laboratory, China
[4] School of Future Technology, South China University of Technology, China

ABSTRACT

Speech emotion recognition is crucial to human-computer interaction. The temporal regions that represent different emotions scatter in different parts of the speech locally. Moreover, the temporal scales of important information may vary over a large range within and across speech segments. Although transformer-based models have made progress in this field, the existing models could not precisely locate important regions at different temporal scales. To address the issue, we propose Dynamic Window transFormer (DWFormer), a new architecture that leverages temporal importance by dynamically splitting samples into windows. Self-attention mechanism is applied within windows for capturing temporal important information locally in a fine-grained way. Cross-window information interaction is also taken into account for global communication. DWFormer is evaluated on both the IEMOCAP and the MELD datasets. Experimental results show that the proposed model achieves better performance than the previous state-of-the-art methods.

Index Terms— speech emotion recognition, transformer, speech signal processing

1. INTRODUCTION

Speech Emotion Recognition (SER) is the key to human-computer interaction. To make human-computer interaction more natural, it is essential for machines to precisely capture emotions and respond in an appropriate manner.

SER has been studied for decades. In recent years, transformer-based models have fostered huge improvement in SER field [1, 2, 3, 4]. The vanilla transformer [5] is outstanding in modeling long-range dependencies in speech sequences. However, its core mechanism, global self-attention mechanism, is vulnerable to noise and may not be able to focus on the same areas as the location of the emotion [1]. This limits the effectiveness of the transformer model. Sound events that prominently represent emotions, such as changes in intonation and speed, laughs and sighs, are located in local regions. Furthermore, the scales of important information are varied over a large range within and across speech segments (see Fig. 1). [2] applies local window attention mechanism to enable models to focus more on local changes. However, immutable window lengths limit these models to capture sentiment information that varies with different temporal scales.

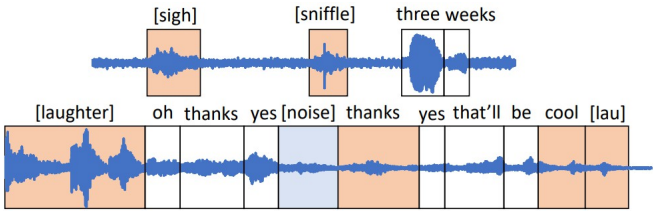

Fig. 1. Two examples are selected from IEMOCAP[6]. [lau] represents laughter. Important sound events that indicates different emotions, such as laughter, sigh, sniffle and positive semantics etc., exist in local regions of speech and their duration varies.

In computer vision field, dynamic designs [7, 8, 9] allow models to have a flexible perceptual field so that different shapes of targets can be captured. In SER field, [10, 11, 12] propose local-global architectures to capture important temporal information. Inspired by them, a new architecture named Dynamic Window transFormer (DWFormer) is proposed to solve the aforementioned problem. The core of the proposed architecture, the DWFormer block, is composed of a Dynamic Local Window Transformer (DLWT) module and a Dynamic Global Window Transformer (DGWT) module. DLWT dynamically divides the input feature into several scales of windows and captures local important information in each window. DGWT remeasures the importance between windows after DLWT. The combination of DLWT and DGWT helps the model discover task-relevant regions. The main contributions of this paper are as follows:

(i) A new architecture, named Dynamic Window transFormer (DWFormer), is proposed to provide insights into the problem of capturing important temporal information of variable lengths for SER.

(ii) We evaluate DWFormer on both the IEMOCAP and MELD [13] datasets and demonstrate that DWFormer substantially outperforms the vanilla transformer and the fixed window transformer. Besides, DWFormer achieves comparable results to the conventional studies and the state-of-the-art

*Corresponding Author. xfxing@scut.edu.cn

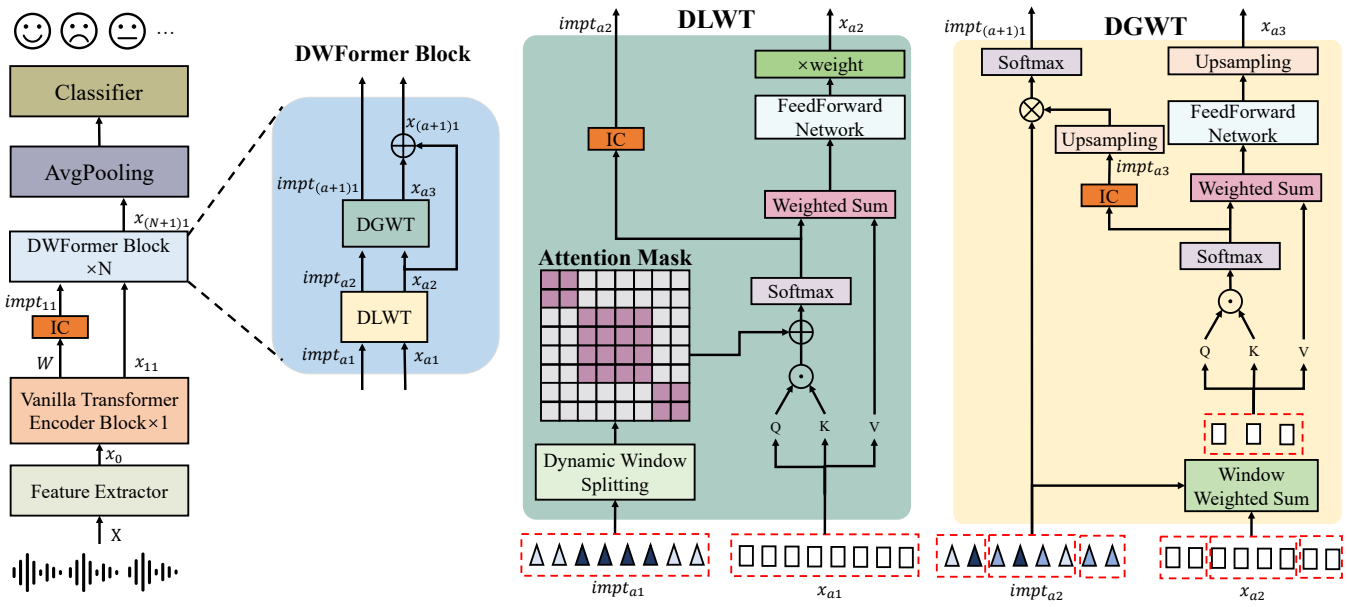

Fig. 2. Model architecture of DWFormer. For simplicity, the residual connection and layer normalization are not plotted in the figure. IC represents Importance Calculation module (detailed in Fig. 3). The triangular sequence $impt$ means the important weights. The deeper the color, the more important the token is. The rectangle sequence x represents feature map.

approaches. The code will be published.[1]

2. METHODOLOGY

The architecture of DWFormer is shown in Fig. 2. The core of the model architecture is the DWFormer block, which is made up of a Dynamic Local Window Transformer module and another Dynamic Global Window Transformer module. The components of the model are introduced below.

The input audio signal is first fed into the feature extractor to extract the features $x_0 \in \mathbb{R}^{T\times D}$, where T represents the number of feature tokens, D represents the feature dimension. Then x_0 is passed through a vanilla transformer encoder layer. The outputs of the encoder layer consists of the hidden feature x_{11} and attention weights $W \in \mathbb{R}^{H\times T\times T}$ where H represents the number of heads. W is sent into the Importance Calculation module to obtain an temporal importance estimation which is necessary for the 1st DWFormer block.

2.1. Importance Calculation Module

The Importance Calculation (IC) module is proposed to measure the importance of token. Inspired by [1], IC module utilize the attention weights obtained from transformer for calculation. The process is shown in Fig. 3, which is described as:

$$impt = Softmax[\sum_{1}^{T_1}(\frac{1}{H}\sum_{s=1}^{H} aw_s)] \quad (1)$$

where aw_s represents attention weight from s-head, T_1 is the row length of the averaged matrix aw_{avg}. Softmax function is used for normalization.

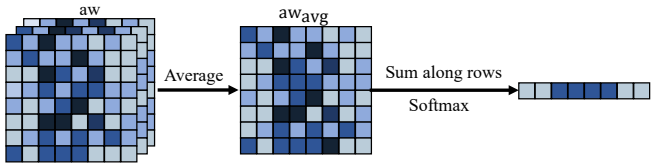

Fig. 3. The IC module calculates the importance from the attention weights.

The importance score of each token $impt_{11}$ obtained by IC module, together with the hidden feature x_{11} are then transferred to N stacked DWFormer blocks for further evaluation.

2.2. DWFormer Block

2.2.1. *Dynamic Local Window Transformer Module*

The Dynamic Local Window Transformer (DLWT) module dynamically partitions regions for input feature and captures important information through local relationship modeling. The procedure is elaborated as bellows.

Firstly, utilizing **dynamic window splitting**(DWS) operation, feature tokens are dynamically split into unequal-length windows according to their importance values obtained from the IC component. As shown in Fig. 4, based on the importance scores calculated from the former block, tokens with importance scores above/below the threshold are grouped chronologically into several strong/weak emotional correlation windows. The threshold is set to the median of all the importance values. A strong emotional correlation windows

[1] https://github.com/scutcsq/DWFormer

and B strong emotional correlation windows are obtained from x_{a1}.

To process data in batches, the window division results are implemented by attention mask mechanism:

$$M_{ij} = \begin{cases} 0, & (b_{w_k} \leq i \leq e_{w_k}, b_{w_k} \leq j \leq e_{w_k} \\ & k = 1, ..., A+B) \\ -\infty, & otherwise \end{cases} \quad (2)$$

where M_{ij} is the value of ith row and jth column of the attention mask $M \in \mathbb{R}^{T \times T}$. b_{w_k} and e_{w_k} are the begin and the end indexes of the row and column of the kth window.

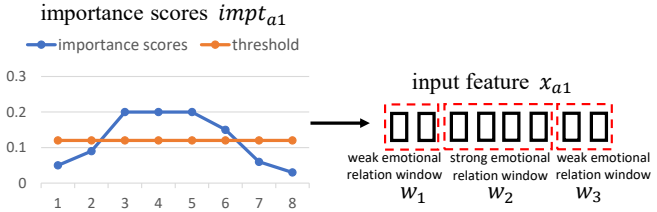

Fig. 4. The operation of dynamic window splitting.

Then, each window passes through a transformer encoder for intra-window information communication, which is defined as:

$$DLWT(x_{a1}) = FFN(Softmax(\frac{Q_{a1}K_{a1}^{\mathrm{T}}}{\sqrt{d_h}} + M)V_{a1}) \quad (3)$$

where FFN represents Feed Forward Network, Q_{a1}, K_{a1}, V_{a1} are the projection mapping of the feature x_{a1}, T means transposition operation, d_h is a scaled factor.

For the weak emotional correlation windows, the prior knowledge learned from the former block indicates that they have a high probability to be redundant for emotion recognition, so the features of the tokens located in them are multiplied by a weight $\lambda (\leq 1)$, while those in strong emotional correlation windows are multiplied by 1. The output is defined as $x_{a2} \in \mathbb{R}^{T \times D}$.

Temporal importance of each token within window is calculated by IC module. Calculation results of all windows are then concatenated together into a sequence along the chronological order, which is noted as $impt_{a2} \in \mathbb{R}^{T}$. x_{a2} and $impt_{a2}$ are passed to Dynamic Global Window Transformer module for further operation.

2.2.2. *Dynamic Global Window Transformer Module*

Dynamic Global Window Transformer (DGWT) module takes a holistic approach to remeasure the importance relationship between windows after DLWT. In detail, each window firstly generates a token through **Window Weighted Sum** operation, which is defined as:

$$wt_k = \sum_{p=b_{w_k}}^{e_{w_k}} impt_{a2_p} \times x_{a2_p} \quad (4)$$

where p is the index of the token. $wt \in \mathbb{R}^{(A+B) \times D}$ is the sequence of window tokens.

Then the sequence wt is passed through a transformer encoder for global interaction, which is defined as:

$$DGWT(wt) = FFN(Softmax(\frac{Q_{wt}K_{wt}^{\mathrm{T}}}{\sqrt{d_h}})V_{wt}) \quad (5)$$

where Q_{wt}, K_{wt}, V_{wt} are the projection mapping of wt.

Next, each window token is upsampled to the same length of the corresponding window by copying the vectors of the window. Then these tokens are concatenated together into a sequence, which is noted as $x_{a3} \in \mathbb{R}^{T \times D}$. The output of a DWFormer block $x_{(a+1)1}$ is the summation of x_{a2} and x_{a3} so that each token obtains both local and global information.

The importance scores between windows $impt_{a3} \in \mathbb{R}^{A+B}$ are calculated by IC. Through a DWFormer block, the importance of each token $impt_{(a+1)1}$ is remeasured by:

$$impt_{(a+1)1} = Softmax(impt_{a2} \times Upsampling(impt_{a3})) \quad (6)$$

the upsampling operation is the same as mentioned above.

In the next DWFormer block, $x_{(a+1)1}$ is split into windows based on $impt_{(a+1)1}$. Finally, the emotion classification is performed by applying the temporal average pooling layer on the output feature $x_{(N+1)1}$ of the Nth DWFormer block, followed by a multi layer perception classifier.

3. EXPERIMENT

3.1. Experiment Setup

We evaluate DWFormer on IEMOCAP and MELD datasets. On IEMOCAP dataset, DWFormer is evaluated using 5-fold leave-one-section-out cross validation. 4 emotions (happy &excited, angry, sad and neutral) are selected for classification. Weighted Accuracy (WA) and Unweighted Accuracy (UA) are the measuring metrics. On MELD corpus which contains 7 emotions (anger, disgust, fear, joy, neutral, sadness, surprise), the Weighted F1 (WF1) score is reported on test set.

The output feature of the 12th transformer encoder layer of Pre-trained WavLM-Large[14] model is used as the audio feature. The number of DWFormer blocks for IEMOCAP is 3 and for MELD is 2. The number of heads is 8. The activation function is ReLU. The number of batchsize is 32. The learning rate is initialized to be 3e-4 for IEMOCAP, while 5e-4 for MELD. The value of λ is 0.85. We employ an SGD optimizer for 120 epochs using a cosine decay learning rate scheduler with cosine warm-up scheduler. The optimization function is Cross Entropy Loss.

3.2. Comparison to Baseline Networks

The vanilla transformer, together with fixed window transformer which splits input feature into equal-length windows

and applies self-attention within each window, are selected as the baseline networks. The parameters of baseline networks are the same as DWFormer. The window length of fixed window transformer is the same as the average length of the windows in the DWFormer. To verify the validity of the modules from DWFormer, ablation experiments are also conducted.

Results in Table 1 demonstrate that DWFormer outperforms the Vanilla transformer and the fixed window transformer on both IEMOCAP and MELD datasets. Meanwhile, removing either the DLWT or the DGWT modules from DWFormer causes a significant decrease.

Model	IEMOCAP		MELD
	WA(%)	UA(%)	WF1(%)
Vanilla Transformer	70.7	71.9	47.1
Fixed Window Transformer	71.2	72.3	47.6
DWFormer (Ours) w/o DLWT	71.5	72.4	47.8
DWFormer (Ours) w/o DGWT	71.5	72.7	47.7
DWFormer (Ours)	**72.3**	**73.9**	**48.5**

Table 1. Comparison results to baseline networks.

3.3. Comparison to conventional research& Visualization

Since our model employs the local-global architecture, we have conducted the comparison experiment with the conventional studies. [10] is open-source, so we first reproduce the results of [10] to ensure the correctness of the codes, and then we test the model under our experimental settings (**Exp 1**). Since the codes of [11, 12] are not publicly available, we test our model under the experimental settings described in their papers(**Exp 2**: randomly split to 80% training set and 20% testing set). Experimental results are shown in Table 2, which demonstrate the superiority of our model compared with the other conventional studies.

Experimental Setting	Model	IEMOCAP		MELD
		WA(%)	UA(%)	WF1(%)
Exp 1	[10]	59.6	60.5	39.8
	DWFormer (Ours)	**72.3**	**73.9**	**48.5**
Exp 2	[11]	69.4	70.2	-
	[12]	70.3	70.8	-
	DWFormer (Ours)	**76.3**	**77.2**	-

Table 2. Comparison results to the other conventional studies.

In addition, Visualization results are shown in Fig. 5. As shown in Fig. 5, Vanilla Transformer, Fixed Window Transformer and ATDA are not as good as ours in locating important temporal information.

3.4. Comparison to Previous State-of-the-art Methods

Table 3 shows the comparison results between previous state-of-the-art methods and DWFormer. Experimental results prove that our method outperforms previous state-of-the-art methods.

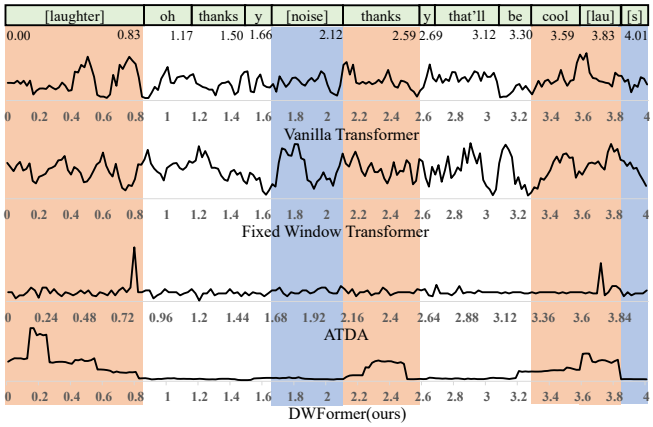

Fig. 5. Visualization results of vanilla transformer, fixed window transformer, ATDA and DWFormer are shown. Horizontal axis indicates the chronological order and vertical axis is of importance score. [s] means silence, [lau] means laughter, y means 'yes'. Areas that are required to focus on are indicated by yellow borders, such as laughter, accent ('thanks'), positive semantic('cool'). Blue borders represent the area that should not be attended to, such as noise and silence.

Dataset	Model	WA(%)	UA(%)	WF1(%)
IEMOCAP	[Chen et al., 2022][2]	62.9	64.5	-
	[Li et al. 2022][15]	68.0	68.2	-
	[Zou et al., 2022][16]	69.8	71.1	-
	DWFormer(Ours)	**72.3**	**73.9**	-
MELD	[Chudasama et al., 2022][17]	-	-	35.8
	[Chen et al., 2022][2]	-	-	41.9
	[Lian et al., 2022][18]	-	-	45.2
	DWFormer (Ours)	-	-	**48.5**

Table 3. Comparison results to previous state-of-the-art methods.

4. CONCLUSIONS

We propose a new transformer-based framework, DWFormer, which aims at capturing important temporal regions at variable scales within and across samples in SER field. We empirically demonstrate that DWFormer outperforms the previous state-of-the-art methods. Ablation study proves the effectiveness of DLWT and DGWT modules. With the ability to locate important information, we plan to apply DWFormer in the pathological speech recognition field to assist researchers in understanding the impact of disease on pronunciation.

5. ACKNOWLEDGEMENT

The work is supported in part by the Natural Science Foundation of Guangdong Province 2022A1515011588; the National Key R&D Program of China (2022YFB4500600); the Science and Technology Project of Guangzhou 202103010002; the Science and Technology Project of Guangdong Guangdong 2022B0101010003; the National Natural Science Foundation of China under Grant U1801262; Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

6. REFERENCES

[1] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, Jichen Yang, and Jianxin Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[2] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du, "SpeechFormer: A Hierarchical Efficient Framework Incorporating the Characteristics of Speech," in *Proc. Interspeech 2022*, 2022, pp. 346–350.

[3] Xianfeng Wang, Min Wang, Wenbo Qi, Wanqi Su, Xiangqian Wang, and Huan Zhou, "A novel end-to-end speech emotion recognition network with stacked transformer layers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6289–6293.

[4] Yuhua Wang, Guang Shen, Yuezhu Xu, Jiahang Li, and Zhengdao Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition.," in *Interspeech*, 2021, pp. 4518–4522.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[7] Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, and Xiaojun Chang, "Beyond fixation: Dynamic window visual transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11987–11997.

[8] Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, and Gao Huang, "Glance and focus: a dynamic approach to reducing spatial redundancy in image classification," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 2432–2444, Curran Associates, Inc.

[9] Sheng Wan, Chen Gong, Ping Zhong, Bo Du, Lefei Zhang, and Jian Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3162–3177, 2020.

[10] Lu-Yao Liu, Wen-Zhe Liu, Jian Zhou, Hui-Yuan Deng, and Lin Feng, "Atda: Attentional temporal dynamic activation for speech emotion recognition," *Knowledge-Based Systems*, vol. 243, pp. 108472, 2022.

[11] Jiaxing Liu, Zhilei Liu, Longbiao Wang, Lili Guo, and Jianwu Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7174–7178.

[12] Jiaxing Liu, Zhilei Liu, Longbiao Wang, Yuan Gao, Lili Guo, and Jianwu Dang, "Temporal attention convolutional network for speech emotion recognition with latent representation.," in *INTERSPEECH*, 2020, pp. 2337–2341.

[13] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[14] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[15] Jinchao Li, Shuai Wang, Yang Chao, Xunying Liu, and Helen Meng, "Context-aware Multimodal Fusion for Emotion Recognition," in *Proc. Interspeech 2022*, 2022, pp. 2013–2017.

[16] Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7367–7371.

[17] Mixiao Hou, Zheng Zhang, and Guangming Lu, "Multimodal emotion recognition with self-guided modality calibration," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4688–4692.

[18] Zheng Lian, Bin Liu, and Jianhua Tao, "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.