Interspeech 2024
1-5 September 2024, Kos, Greece

Cross-modal Features Interaction-and-Aggregation Network with Self-consistency Training for Speech Emotion Recognition

Ying Hu[1,3,*], *Huamin Yang*[1,3,*], *Hao Huang*[1], *Liang He*[1,2]

[1]College of Information Science and Engineering, Xinjiang University, Urumqi, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
[3]Key Laboratory of Signal Detection and Processing, Xinjiang, Urumqi, China

huying@xju.edu.cn,yanghuamin@stu.xju.edu.cn

Abstract

In recent years, much research has been into speech emotion recognition (SER) using multimodal data. Selective fusion of the features from different modalities is critical for multimodal SER. In this paper, we propose a cross-modal features interaction-and-aggregation network (CFIA-Net) with self-consistency training for SER. Specifically, we design a cross-modal features interaction-and-aggregation (CFIA) module to adaptively interact and integrate the features of audio and text modalities. Moreover, we introduce a self-consistency training strategy, which exploits the features from deeper layers to supervise those from shallower ones to obtain the SER task-related information. The experimental results show that compared with other bimodal SER methods, the CFIA-Net achieves the state-of-the-art performance on the weighted accuracy (WA) of 83.37% and unweighted accuracy (UA) of 83.67% on the IEMOCAP dataset.

Index Terms: speech emotion recognition, features interaction-and-aggregation, self-consistency training

1. Introduction

Speech emotion recognition (SER) is a critical research for human-computer interaction (HCI), which aims to assist machines in understanding the human emotions [1]. SER can be applied to a variety of intelligent devices, including call centers [2], smart voice assistants [3], customer service dialogues [4], and so on. In recent years, multimodal SER has received significant attention due to different modalities that can provide complementary information and achieve better performance than unimodal methods. This paper focuses on the audio and text modalities for SER task.

In the field of SER research, obtaining large annotated datasets is costly. Due to the time-consuming acquisition of artificially labeled emotional speech, the existing public speech emotional datasets are not enough to train robust supervised learning models [5]. A popular solution is to utilize pre-trained models for SER tasks. Transformer-based models from natural language processing (NLP) are extensively employed for capturing language representations from texts in SER tasks [6, 7], such as BERT [8]. For the audio modality, many researchers [7, 9, 10, 11, 12, 13] utilized speech-based pre-trained models as feature extractors, such as Wav2Vec [14], Wav2Vec2 [15], and WavLM [16]. However, these pre-trained models may be data-specific or model-constrained, which need to be fine-tuned before they are applied in the field of SER. Chen et al. proposed an improved emotion-specific pre-trained encoder named Vesper, focusing on the SER task [17]. Ma et al. proposed a universal emotion representation model named emotion2vec that can be used to extract the features of speech for various emotion tasks [18]. In this paper, BERT and emotion2vec are used to extract the features of text and audio modalities, respectively.

Multimodal feature extraction and fusion have been attracting much attention in the field of SER. Shen et al. proposed a multimodal fusion framework based on word-level interaction, which mainly uses two Long Short-Term Memory (LSTM) networks for fusing the features from the audio and text modalities [19]. Li et al. proposed several multimodal transformer-based methods to combine features from the audio and text modalities [20]. Tang et al. proposed an audio–text-interactional attention structure to facilitate the interaction and fusion of bimodal information [21]. Zhao et al. proposed a co-attention based fusion of the features from pre-trained for multimodal SER, which is integrated emotion-related knowledge into Bayesian co-attention modules [22]. Maji et al. proposed a cross-modal transformer block to capture interaction and temporal information between the audio and textual features [23]. Current researches tend to use the cross-attention to achieve the effective fusion of bimodal features. However, those methods may cause feature redundancy.

Moreover, knowledge distillation is one of the popular methods for SER tasks. The traditional knowledge distillation method involves training a larger teacher model to guide a smaller student model by minimizing the loss between the teacher and student models [24]. Feature distillation methods allow the features of the student model to imitate the intermediate features of the teacher model, thereby utilizing its knowledge more effectively, but require extra training budgets [25]. Therefore, Li et al. proposed a teacher-free feature distillation framework on the image classification and object detection tasks, which aims to reuse channel-wise and layer-wise meaningful features within the student network without an additional model [26].

Motivated by the above observations, we propose a cross-modal features interaction-and-aggregation network (CFIA-Net) with self-consistency training strategy for speech emotion recognition. The contributions of this paper are as follows:

i) We propose a cross-modal features interaction-and- aggregation network, which promotes adaptive integration of multi-level features from the text and audio modalities through multiple cross-modal features interaction-and-aggregation (CFIA) modules.

ii) We introduce a self-consistency training strategy that uses the features of the deep layer to supervise those of the shallow ones. This strategy directly calculates the consistency loss of cross-layer features to improve the capability of feature extraction without additional parameters.

*Equal contribution.

10.21437/Interspeech.2024-1733

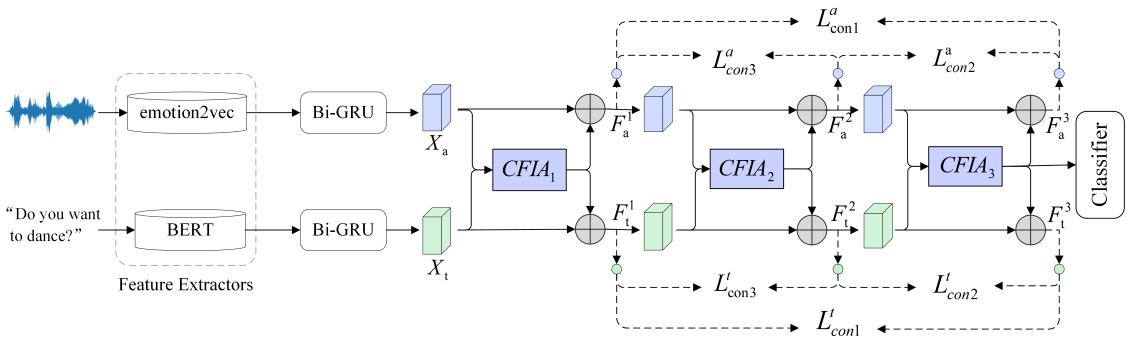

Figure 1: *The architecture of the proposed cross-modal feature interaction-and-aggregation network (CFIA-Net). The gray ⊕ indicates the averaging operation of the fused features from the CFIA module and the corresponding input. The dash lines represent that the procedures exist only during the training phase.*

iii) The experimental results show that our proposed CFIA-Net outperforms the compared state-of-the-art bimodal SER methods, achieving WA of 83.37% and UA of 83.67% on the IEMOCAP dataset.

2. Methodology

This section introduces our proposed CFIA-Net architecture, the feature extractors for audio and text data, and the CFIA module. Finally, we introduce the self-consistency training strategy.

2.1. Overall Architecture

The framework is shown in Figure 1. We first use two pre-trained models, BERT and emotion2vc, as the feature extractors to extract the text embeddings and frame-level audio embeddings, respectively. The embeddings of audio and text are fed into the bidirectional GRU (Bi-GRU) to capture the temporal dependencies of the features, respectively. The output features of two Bi-GRUs are undergone a linear layer to be mapped to the intermediate features with the same shapes of $C \times T \times F$. Then, the audio and text features are fed into three consecutive cross-modal features interaction-and-aggregation (CFIA) modules to realize the cross-modal fusion of the bimodal features, aggregating the information from two modalities.

In particular, the gray $\oplus$ indicates the averaging operation of the fused features from the CFIA module and the corresponding input, obtaining F_a^i and $F_t^i, i \in 1, 2, 3$.

$$F_a^1 = (F_{fusion} + X_a)/2, \tag{1}$$

$$F_t^1 = (F_{fusion} + X_t)/2, \tag{2}$$

where F_{fusion} represents the fused features from the CFIA module. Finally, the output features from the third CFIA module are fed into a classifier to predict the emotion classes.

2.2. Feature Extractors

We use a self-supervised speech emotion representation model, emotion2vec [18], to extract the emotion-related speech features from the waveforms for the audio modality. emotion2vec is a universal speech-based emotion representation model, which is obtained through self-supervised pre-training on 262 hours of open-source emotional data using an online distillation paradigm [18]. Additionally, utterance-level loss and frame-level loss are combined during pre-training phase. The popular BERT [8] model is adopted as a textual feature extractor, consisting of a tokenizer and 12 transformer encoders.

2.3. Cross-modal Features Interaction-and-Aggregation (CFIA) Module

Motivated by [27], we design a cross-modal features interaction-and-aggregation (CFIA) module, including two parts: features interaction and features aggregation blocks, as illustrated in Figure 2, CFIA module can effectively integrate complementary information from two modalities and unify the most informative cross-modal features into an effective representation.

Features Interaction. The feature interaction block is designed to promote the module to recalibrate and adaptively fuse the SER task related features of two modalities. The input features of audio and text are denoted as $X_a \in R^{C\times T\times F}$ and $X_t \in R^{C\times T\times F}$, respectively. The features of the two modalities are concatenated along the channel dimension and then fed into a global average pooling layer to obtain the features $X \in R^{2C\times 1\times 1}$. Subsequently, a multilayer perceptron (MLP) containing two linear layers and a $ReLu$, and a $sigmoid$ activation function are used to obtain the attention vectors W_a and W_t. Then, the recalibrated features X_a^1 can be obtained through a channel-wise multiplication operation $\odot$.

$$W_a = \sigma(f_{mlp}(X)), \tag{3}$$

$$X_a^1 = W_a \odot X_a, \tag{4}$$

where f_{mlp} represents the multilayer perceptron (MLP). σ is the sigmoid function scaling the weight value into $(0, 1)$. Through similar Equation 3 and Equation 4, we can get W_t and X_t^1.

Finally, we perform an interaction operation on the weighted features of the audio and text modalities to effectively aggregate across-modal channel-wise features.

$$I_a = X_a + X_t^1, \tag{5}$$

$$I_t = X_t + X_a^1, \tag{6}$$

where I_a and I_t denote the output features of interaction operation.

Features Aggregation. To fully integrate the spatial bimodal information, the features aggregation part uses the soft attention mechanism to control the information flow of each modality features after features interaction processing, thereby

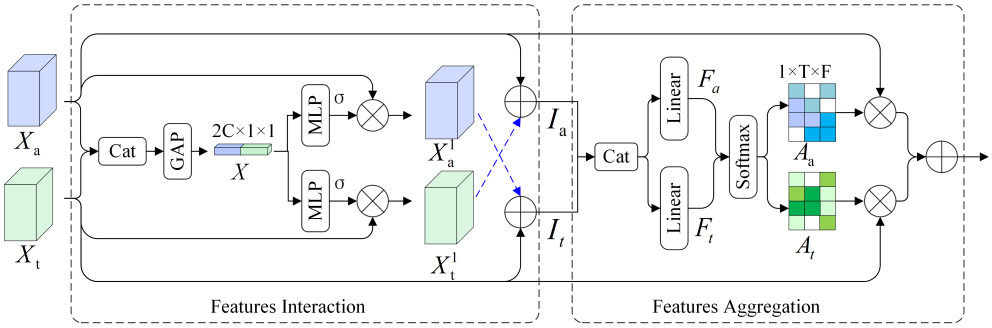

Figure 2: *Illustration of cross-modal features interaction-and-aggregation (CFIA) module. Cat, σ, and GAP denote the concatenation, global average pooling and sigmoid activation operations, respectively.*

realizing effective audio and text modalities features aggregation. We use audio and text features from the features interaction part as input, denoted as $I_a \in R^{C\times T\times F}$ and $I_t \in R^{C\times T\times F}$, respectively.

We first concatenate the features of the two modalities and define two mapping functions to map the high-dimensional features to two different spatial-wise features:

$$I_a \longrightarrow F_a \in R^{1\times T\times F}, \tag{7}$$

$$I_t \longrightarrow F_t \in R^{1\times T\times F}. \tag{8}$$

In practice, we use a 1D convolution to implement this mapping function. A softmax function is applied to these two feature spaces:

$$A_a^{(i,j)} = \frac{e^{F_a^{(i,j)}}}{e^{F_a^{(i,j)}} + e^{F_t^{(i,j)}}}, \tag{9}$$

$$A_t^{(i,j)} = \frac{e^{F_t^{(i,j)}}}{e^{F_a^{(i,j)}} + e^{F_t^{(i,j)}}}, \tag{10}$$

where A_a, $A_t \in R^{1\times T\times F}$, and $A_a^{(i,j)} + A_t^{(i,j)} = 1$. $A_a^{(i,j)}$ is the weight assigned to each position in the audio feature, and $A_t^{(i,j)}$ is the weight assigned to each position in the text feature. The final fusion feature F_{fusion} can be obtained by weighting the audio and text features:

$$F_{fusion} = X_a \cdot A_a^{(i,j)} + X_t \cdot A_t^{(i,j)}. \tag{11}$$

2.4. Self-consistency Training Strategy

The features of the deep layer contain more task-relevant semantic information [28]. Following the concept of features distillation presented in [26], we design a self-consistency training strategy, which utilizes the self-features from the deeper layers to supervise the shallow ones. These self-features are updated by the $l2$ loss ($\mathcal{L}_{\text{con}}$) during back propagation.

The loss function combines cross entropy loss ($\mathcal{L}_{\text{ce}}$) and self-consistency loss ($\mathcal{L}_{\text{con}}$). Note that $F_a^i, i \in 1,2,3$ and $F_a^i, i \in 1,2,3$ are frozen when updating losses. In particular, the self-consistency training strategy is only used during the training phase. The loss function can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \alpha\mathcal{L}_{\text{con1}} + \beta\mathcal{L}_{\text{con2}} + \gamma\mathcal{L}_{\text{con3}}, \tag{12}$$

where α, β, and γ are the weighting factors used to scale the the losses.

$$\mathcal{L}_{\text{con1}} = \mathcal{L}_{\text{con1}}^a + \mathcal{L}_{\text{con1}}^t, \tag{13}$$

$$\mathcal{L}_{\text{con2}} = \mathcal{L}_{\text{con2}}^a + \mathcal{L}_{\text{con2}}^t, \tag{14}$$

$$\mathcal{L}_{\text{con3}} = \mathcal{L}_{\text{con3}}^a + \mathcal{L}_{\text{con3}}^t, \tag{15}$$

where $\mathcal{L}_{\text{con(i)}}, i \in 1,2,3$ represents pairs of the consistency loss function for intermediate features of audio and text.

$$\mathcal{L}_{\text{con1}}^a = \mathcal{L}_{\text{con}}(T(F_a^3), T(F_a^1)), \tag{16}$$

$$\mathcal{L}_{\text{con2}}^a = \mathcal{L}_{\text{con}}(T(F_a^3), T(F_a^2)), \tag{17}$$

$$\mathcal{L}_{\text{con3}}^a = \mathcal{L}_{\text{con}}(T(F_a^2), T(F_a^1)), \tag{18}$$

where $T(.)$ denotes the channel alignment operation of the features. We utilize channel cropping to align features in the channel dimensions, avoiding the need for complex transformations. Similar to Equations 16, 17, and 18, we can get $\mathcal{L}_{\text{con1}}^t$, $\mathcal{L}_{\text{con2}}^t$, and $\mathcal{L}_{\text{con3}}^t$.

3. Experiments

3.1. Dataset

We performed experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [29], which is a mainstream dataset commonly used for speech emotion recognition. The corpus contains 12 hours of English conversations with 10 performers, divided into five sections, each including one male and one female speaker. To compare with previous studies [7, 30, 31], we performed a classification task on 5531 utterances containing four emotion classes: happy (1636 utterances, combined with excited), angry (1084 utterances), sad (1084 utterances), and neutral (1708 utterances). We only used audio and textual features in this paper.

3.2. Experimental Setup

The pre-trained models emotion2vec [18] and BERT [8] were used as feature extractors to extract 768-dimensional embedding features for audio and text, respectively. For the audio modality, the maximum length of each utterance is set to 512 frames. Longer utterances are cut in 512 frames, and shorter utterances are padding zero to 512 frames.

Adam was used as an optimizer, and the initial learning rate and batch size were set to 1e-4 and 32, respectively. α, β, and γ were set to 0.0005, 0.0003, and 0.0001, respectively. We conduct both leave-one-session-out cross-validation (5-fold CV) and leave-one-speaker-out cross-validation (10-fold CV) in the speaker-independent environment. Unweighted accuracy (UA), weighted accuracy (WA), and weighted average F1 (WF1) were adopted as the evaluation metrics.

3.3. Comparison with Other Methods

As shown in Table 1, we compared our proposed CFIA-Net with other bimodal SER methods in 5-fold CV and 10-fold CV set-

Table 1: *Comparison with seven bimodal SER methods on the IEMOCAP dataset keeping the same experimental settings with compared methods in 5-fold CV and 10-fold CV. "A" and "T" denote audio and text modalities, respectively.*

Methods	**Upstream(A)**	**Upstream(T)**	**WA(%)**	**UA(%)**
5-fold CV				
SMCN, 2022 [30]	-	BERT	75.60	77.60
MGM, (2023) [32]	-	Glove	74.50	75.00
BAM, 2023 [22]	Wav2Vec2	BERT	75.50	77.00
MNMF, 2023 [31]	Wav2Vec2	BERT	76.80	77.30
CFIA-Net	**emotion2vec**	**BERT**	**80.67**	**81.01**
10-fold CV				
ATIA, 2022 [21]	-	Glove	82.40	80.60
SWRR, 2023 [7]	WavLM	BERT	77.40	78.50
CMT-SA, 2023 [23]	Glove	-	80.63	81.49
CFIA-Net	**emotion2vec**	**BERT**	**83.37**	**83.67**

tings. "A" and "T" denote audio and text modalities, respectively. The results of compared methods are reported in their published papers. Compared with other methods, the performance of the proposed CFIA-Net is more competitive, both on the 5-fold CV and 10-fold CV. In particular, On the 5-fold CV, the CFIA-Net obtains 3.87% gains of WA than MNMF [31], and 3.41% gains of UA than SMCN [30]. CFIA-Net has been able to achieve the best performance with WA of 83.37% and UA of 83.67% on the 10-fold CV. Compared with ATIA [21], CFIA-Net achieves the improvements by 0.77% and 3.07% on WA and UA, respectively. Those show the advantages of CFIA-Net in interacting and aggregating bimodal features, which is helpful for the fusion of audio and text modalities to improve the performance of bimodal SER.

Table 2: *Comparison with ten unimodal (Audio) SER methods adopting pre-trained models in 5-fold CV.*

Proposed	**Upstream**	**Downstream**	**WA(%)**	**UA(%)**
Xia et al. [9]	Wav2Vec	classfier	63.80	65.80
Zou et al. [10]	Wav2Vec2	classfier	64.03	65.67
Ma et al. [18]	Hubert	classfier	64.92	-
Chen et al. [17]	Vesper-12	classfier	70.70	70.80
Chen et al. [16] *	WavLM	classfier	70.50	71.70
Ma et al. [18] *	emotion2vec	classfier	73.65	74.68
Li et al. [6]	Hubert	MSTR	70.30	71.60
Sun et al. [12]	Wav2Vec	EmotionNAS	72.10	69.10
Hu et al. [11]	Wav2Vec2	JointNet	72.48	73.32
Fang et al. [13]	WavLM	BAS	74.03	74.95
Ours	**emotion2vec**	**CFIA-Net**	**75.75**	**76.42**

* represents the results are obtained by our implementation.

We also evaluate the capabilities of different audio feature extractors, and the experimental results are shown in the upper part of Table 2. From the experimental results, emotion2vec is more suitable for extracting emotion representations. In addition, from the lower part of Table 2, it can be observed that when only using the audio modality, our proposed CFIA-Net is superior to four compared unimodal methods, which achieves the WA of 75.75% and UA of 76.42%. Note that the results in Table 2 are obtained by using only audio modality features, which indicates that the lower branch of CFIA-Net is not used.

Table 3: *Ablation study of self-consistency training strategy.*

L_{con1}	L_{con2}	L_{con3}	**WA(%)**	**UA(%)**	**WF1(%)**
			81.71	83.17	81.84
✓			81.93	81.69	82.31
	✓		82.74	83.85	82.73
		✓	82.01	81.10	82.14
✓	✓	✓	**83.37**	**83.67**	**83.43**

Table 4: *Ablation study of cross-modal features interaction-and-aggregation (CFIA) module.*

Model	**WA(%)**	**UA(%)**	**WF1(%)**
CFIA-Net w/o L_{con}	**81.71**	**83.17**	**81.84**
-$CFIA_3$ (i)	80.25	81.32	80.38
-$CFIA_2$ (ii)	79.00	81.19	78.91
-$CFIA_1$ (iii)	77.57	78.86	77.89

3.4. Ablation Study

All ablation experiments were performed in the 10-fold CV. We first explore the effectiveness of the self-consistency training strategy. Table 3 shows the individual efficacy of the different components of the self-consistency training strategy. $L_{con(i)}, i \in 1, 2, 3$ means pairs of the consistency loss function for audio and text intermediate features. Compared to the first line without any self-consistency loss function, the model with self-consistency loss function has led to varying improvement in performance, especially those with three achieving the best performance. According to the ablation experimental results in Table 3, using deeper features to supervise neighboring shallow features can make the network learn more emotion-related information.

Furthermore, we further conducted the impact of using different numbers of CFIA modules on SER performance. As shown in Table 4, the modules were discarded gradually. (i) denotes that the model didn't adopt the $CFIA_3$ module. (ii) denotes that the model didn't adopt the $CFIA_3$ and $CFIA_2$ modules. (iii) indicates that the model didn't employ the $CFIA$ module, the audio and text features output by Bi-GRU are added and fed into the classifier. In the Table 4, the ablation results indicate the effectiveness of the CFIA module, which can adaptively integrate features from different modalities.

4. Conclusions

In this paper, we propose a cross-modal features interaction-and-aggregation network (CFIA-Net) for bimodal speech emotion recognition, while designing a self-consistency training strategy. The experimental results show that CFIA-Net achieves superior performance to other bimodal methods on the IEMOCAP. The ablation study shows that adopting a self-consistency training strategy helps the network improve the SER performance, and our designed CFIA module can effectively fuse the features of two modalities. In the future, we can use the self-consistency training strategy for the multimodal SER.

5. Acknowledgements

This work is supported by the Xinjiang Uygur Autonomous Region Graduate Student Innovation Project (XJ2023G096).

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] M. Bojanić, V. Delić, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Applied Sciences*, vol. 10, no. 13, p. 4653, 2020.

[3] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[4] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[5] Y. Gao, C. Chu, and T. Kawahara, "Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining," in *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, 2023, pp. 3637–3641.

[6] Z. Li, X. Xing, Y. Fang, W. Zhang, H. Fan, and X. Xu, "Multi-scale temporal transformer for speech emotion recognition," in *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, vol. 2023, 2023, pp. 3652–3656.

[7] Z. Zhao, T. Gao, H. Wang, and B. Schuller, "Swrr: Feature map classifier based on sliding window attention and high-response feature reuse for multimodal emotion recognition," in *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, vol. 2023, 2023, pp. 2433–2437.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] Y. Xia, L.-W. Chen, A. Rudnicky, R. M. Stern *et al.*, "Temporal context in speech emotion recognition." in *Proc. INTERSPEECH 2021 – 22rd Annual Conference of the International Speech Communication Association*, vol. 2021, 2021, pp. 3370–3374.

[10] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. 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)*. IEEE, 2022, pp. 7367–7371.

[11] Y. Hu, S. Hou, H. Yang, H. Huang, and L. He, "A joint network based on interactive attention for speech emotion recognition," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1715–1720.

[12] H. Sun, Z. Lian, B. Liu, Y. Li, L. Sun, C. Cai, J. Tao, M. Wang, and Y. Cheng, "Emotionnas: Two-stream neural architecture search for speech emotion recognition," vol. 2023, pp. 3597–3601, 2023.

[13] Y. Fang, X. Xing, X. Xu, and W. Zhang, "Exploring downstream transfer of self-supervised features for speech emotion recognition," in *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, 2023, pp. 3627–3631.

[14] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. INTERSPEECH 2019 – 20rd Annual Conference of the International Speech Communication Association*, 2019, pp. 3465–3469.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–14, 2024.

[18] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[19] G. Shen, R. Lai, R. Chen, Y. Zhang, K. Zhang, Q. Han, and H. Song, "Wise: Word-level interaction-based multimodal fusion for speech emotion recognition." in *Proc. INTERSPEECH 2020 – 21rd Annual Conference of the International Speech Communication Association*, 2020, pp. 369–373.

[20] J. Li, S. Wang, Y. Chao, X. Liu, and H. Meng, "Context-aware multimodal fusion for emotion recognition," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, 2022, pp. 2013–2017.

[21] Y. Tang, Y. Hu, L. He, and H. Huang, "A bimodal network based on audio–text-interactional-attention with arcface loss for speech emotion recognition," *Speech Communication*, vol. 143, pp. 21–32, 2022.

[22] Z. Zhao, Y. Wang, and Y. Wang, "Knowledge-aware bayesian co-attention for multimodal emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] B. Maji, M. Swain, R. Guha, and A. Routray, "Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] Z. Ren, T. T. Nguyen, Y. Chang, and B. W. Schuller, "Fast yet effective speech emotion recognition with self-distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] L. Li, S.-N. Liang, Y. Yang, and Z. Jin, "Teacher-free distillation via regularizing intermediate representation," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 01–06.

[26] L. Li, "Self-regulated feature learning via teacher-free feature distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 347–363.

[27] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 561–577.

[28] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 7028–7036.

[29] C. Busso, M. Bulut, C.-C. Lee *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[30] M. Hou, Z. Zhang, and G. Lu, "Multi-modal emotion recognition with self-guided modality calibration," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4688–4692.

[31] D. Priyasad, T. Fernando, S. Sridharan, S. Denman, and C. Fookes, "Dual memory fusion for multimodal speech emotion recognition," in *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, vol. 2023, 2023, pp. 4543–4547.

[32] J. He, M. Wu, M. Li, X. Zhu, and F. Ye, "Multilevel transformer for multimodal emotion recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.