INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

SWRR: Feature Map Classifier Based on Sliding Window Attention and High-Response Feature Reuse for Multimodal Emotion Recognition

Ziping Zhao [1,#], *Tian Gao* [1,#], *Haishuai Wang*[2], *Björn Schuller*[3,4]

[1]College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
[2]College of Computer Science, Zhejiang University, China
[3]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[4]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

ztianjin@126.com, 2111090038@stu.tjnu.edu.cn, haishuai.wang@zju.edu.cn, schuller@tum.de

Abstract

To achieve efficient feature fusion, existing research tends to employ cross-attention to control the contributions of different modalities in fusion. However, this inevitably causes high computational effort and introduces noise weights due to redundant computations. Therefore, this paper proposes sliding window attention (SliWa) to control the feature perception range and dynamically model the modality fusion at different granularities. In addition, we present a novel feature map classifier (FMC) based on high-response feature reuse (HRFR), which explicitly preserves the deep emotional feature structure, thus preventing the submersion of the crucial classification information after average flattening and the negative impacts of parameter flooding. We unify the mentioned modules in the SWRR framework, and the experimental results on the commonly used datasets IEMOCAP and CMU-MOSEI reveal the effectiveness of SWRR in improving the performance of emotion recognition.

Index Terms: multimodal emotion recognition, feature fusion, feature reuse, classifier

1. Introduction

As intelligent devices continue to advance, emotional intelligence technology is being integrated into applications such as escort robots and assisted driving [1]. Accurately identifying the user's emotional state is crucial to effectively meet user requirements. Human emotion expression often involves multimodality, making a multimodal emotion recognition (MER) system valuable in leveraging the complementarity of multimodal features to enhance robustness and performance [2].

Feature fusion is one of the important techniques in MER, different from previous studies that modeled each modality independently and then combined them at the classification level [3, 4], researchers are currently focusing on the interaction behavior between modalities and working on decoupling the complexity to design efficient fusion mechanisms. Cimtay et al. [5] propose a mixed feature-decision level fusion approach to jointly process physiologic modalities at varying time instances. Chen et al. [6] proposed a multi-stage multimodal dynamic fusion network with a unique design for unimodal, bimodal, and trimodal interactions, respectively. As more MER models are proposed, the self-attention-based [7] cross version further enhances the modalities interaction [8, 9, 10], Query, Key, and Value are assigned to different modal separately to obtain cross-correlation guiding weighted aggregations under fine granularity, in which, Peng et al. [8] proposed to use attention units to combine the outputs of different pooling methods; Sun et al. [10] utilize cross- and self-attention to accomplish inter- and intra-modal information propagation. However, obtaining the correlation matrix entails significant computational effort. Consequently, several labor-saving methods have been proposed, such as random, global [11], etc., but few studies have been conducted under multimodal cross-attention. And we also note that the cross-correlation matrix generates a large number of noise weights in the order of $[e^{-10}, e^{-20}]$ after $softmax$.

The classification algorithm will influence recognition performance to a great extent as well. Some traditional methods based on machine learning [12, 13] provide solutions for emotion recognition techniques. Due to the end-to-end processing in deep learning, it can be better adapted to unstructured data. Many deep learning-based classifiers have been developed to enhance the efficiency of emotion recognition [14]. Most studies [10, 15, 16] typically flatten the extracted deep features and employ fully-connected layers (FC) for classification, the structural features containing emotional-semantic information are often corrupted when the deep features are flattened to a one-dimensional sequence; additionally, the massive to-be-learned parameters introduced by FC make it difficult to effectively detect emotional features with category-critical information, and pose the risk of overfitting. To mitigate that, Huang et al. [17] suppress overfitting by adding label-smoothing regularization; Liu et al. [18] improve the distinguishability of emotion embeddings by jointly triplet loss and cross-entropy loss. However, in the MER tasks, there is less research on preventing structural feature collapse that impacts emotional-semantic features.

To solve the above issues, we propose a feature map classifier based on sliding window attention and high-response feature reuse (SWRR) for MER. Specifically, we summarize the contributions of this paper as follows:

1. In sliding window attention (SliWa) mechanism, the sliding window truncates redundant computation and weight noise so that the text and audio modalities are dynamically fused within the maximum-effective feature perception range.
2. We replace FC with the feature map classifier (FMC) to address the issue of submerging the emotion-critical information in feature maps due to the structure semantic corruption and propose high-response feature reuse (HRFR) to help the kernel obtain additional emotion-categories high-response information to enhance the category sensitivity and make the correct confidence mapping.
3. The experimental results indicate that SWRR achieves 77.4% WA and 78.5% UA on IEMOCAP; we further verify the generalization ability of SWRR on CMU-MOSEI and reach a superior performance of 52.4%, 84.8%, and 83.8% on 7-class accuracy, binary accuracy, and F1-score, respectively.

The present work is supported by the National Natural Science Foundation of China (No. 62071330).

[#] Both authors contributed equally to this work.

10.21437/Interspeech.2023-413

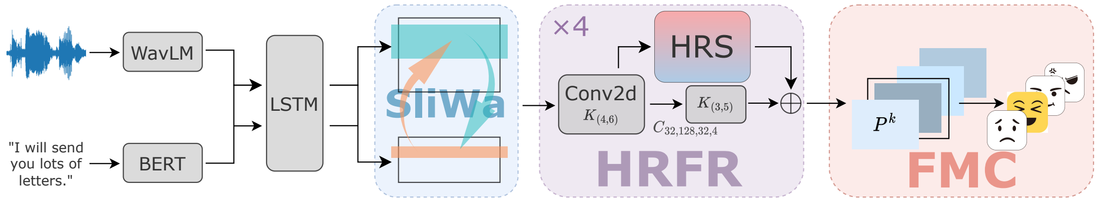


Figure 1: *Illustration of SWRR framework.* ***WavLM*** *[19] &* ***BERT*** *[20]: pre-trained audio & text models as feature extractors,* $\boldsymbol{K_{(h,w)}}$*: the convolution kernel size (h, w),* $\boldsymbol{C_n}$*: the number of feature maps,* $\oplus$*: element-wise addition.*

2. Methodology

The working mechanism of SWRR is illustrated in Fig.1. The pre-trained model extracts the embedding features from input modalities, sliding window attention (SliWa) is employed to model inter-modal interactions, and multi-level high-response feature reuse (HRFR) is utilized to continuously reinforce the emotion-categories high-response information in the feature map, finally, the feature map classifier (FMC) completes the sample label space mapping.

2.1. Redundant Filtering

Different modalities have intersectionality in representation, therefore we have to notice the issue of inter-modal feature redundancy [21]. To solve the issue, we consider that emotions are time-varying behaviors and there is a natural correlation between audio and text, so we design shared LSTM blocks to allow cross-modal weight sharing at the time step, and obtain audio and text filtering features $\mathrm{X} \in \mathbb{R}^{N\times D}$ and $\mathrm{Y} \in \mathbb{R}^{T\times D}$ under the joint influence of these two modalities.

2.2. Sliding Window Attention

To solve the problem of redundant computation and noise weight caused by the full-size cross-dot product, we propose the sliding window attention (SliWa) fusion technique, as shown in Fig.2, which aims to control the feature perception range at fine granularity by the sliding window mechanism, derive the correlation coefficients between modalities within the window, and dynamically model inter-modal information propagation.

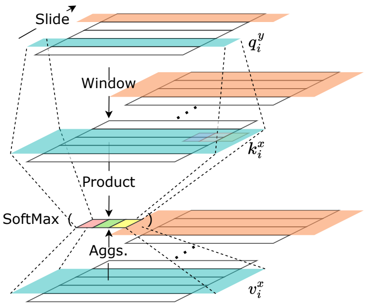


Figure 2: ***SliWa*** *block architecture,* ***Slide*** *is the sliding direction,* ***Window*** *is the window mapping.*

Specifically, the filtering features are first windowed in the sequence dimension:

$$\overline{m_i} = \{m_{(i-1)h_M+1}, \cdots, m_{(i-1)h_M+\omega_M}\} \tag{1}$$

$$\overline{N}_M = \left\lfloor \frac{N-\omega_M}{h_M} \right\rfloor + 1, \left(\overline{N}_X = \overline{N}_Y\right) \tag{2}$$

where $\mathrm{M} \in \{\mathrm{X}, \mathrm{Y}\}, m \in \{x, y\}, i \in \{1, 2, ..., \overline{N}_M\}$, $\overline{m_i} \in \mathbb{R}^{\omega_M \times D}$, ω_M is the window width, h_M is the window hop-length, $\lfloor\cdot\rfloor$ denotes rounding down, and $\overline{N}_M$ is the number of windows under ω_M and h_M divisions.

Since the sequence length of Y is much shorter compared to X, to adequately and efficiently model feature fusion, we set $\omega_Y = 1$ and $h_Y = 1$ when performing window attention calculations, and employ vector projections to obtain the query for text, as well as the key and value [7] for audio:

$$q_i^y = w_q \overline{y}_i \tag{3}$$

$$k_i^x, v_i^x = w_{k,v} \overline{x}_i \tag{4}$$

where $w_{q,k,v}$ is the learnable parameter, $q_i^y \in \mathbb{R}^{1\times d}, k_i^x, v_i^x \in \mathbb{R}^{\omega_X \times d}$

Following, the cross-dot product is performed at different granularities within the window and normalized to $[0, 1]$ with $Softmax$ as the correlation coefficient, which guides the weighted aggregations (Aggs.) of the value representations v_i^x within the current window ω_X, and the inter-modal information propagation through residual connections to obtain the complementary fusion feature $\overline{z}_i$:

$$z_i = Softmax\left(\frac{q_i^y\,(k_i^x)^{\mathbf{T}}}{\sqrt{d}}\right) v_i^x \tag{5}$$

$$\overline{z}_i = LayerNorm\left(\overline{y}_i + w_z z_i\right) \tag{6}$$

Notably, since SliWa is an algorithm based on matrix operations, we concatenate the divided windows $\overline{m_i}$ at the first dimension according to (7), which means that the operations (3) to (6) on $\overline{M}$ can be performed in parallel between the windows:

$$\overline{M} = \left\{\overline{m}_1, \overline{m}_2, \ldots, \overline{m}_{\overline{N}_M}\right\} \in \mathbb{R}^{\overline{N}_M \times \omega_M \times \mathrm{D}} \tag{7}$$

2.3. Feature Map Classifier

In the SWRR, convolutional neural networks (CNNs) act as the backbone network to further process the fusion features, which learn the emotional features from different perspectives and characterize them within the feature maps. However, using FC leads to the collapse of temporal- and structural-semantic information in feature maps, and crucial features with emotional discrimination are submerged in one-dimensional sequences after average flattening, making it difficult to be effectively explored by the numerous to-be-learned parameters.

To address these challenges, we utilize the feature map classifier (FMC) as the emotion discriminator, we define n kernel functions as filters for n-category emotions, where the k-category is mapped to the corresponding feature map P^k. Due to the continuous accumulation and overlap of the receptive field, the resolution of P^k is lower but semantic information is richer now, so the structural information can be globally aggregated and normalized as category confidence p^k to complete the emotion classification and optimized by cross-entropy loss.

2.4. High-Response Feature Reuse

The core requirement to improve the performance of FMC is that the confidence response of P^k should be higher than the non-k category. Therefore, how the emotion filter Φ^k makes P^k become a feature map with discriminative high-response is the bottleneck that limits the recognition efficiency of FMC.

We plan the mentioned issues uniformly and propose high-response feature reuse (HRFR). Since the FMC allows the model to optimize the emotion representation of the correct category toward high response, we design HRFR to consist of two parts: high-response feature map screening (HRS) and feature reuse. HRS is applied to capture the most effective category discriminative feature maps and subsequently reinforce the multi-level output's represent capability to emotional-critical information through feature reuse, thus facilitating Φ to produce high-response mappings for the correct categories.

Emotion expression is time-varying, so we believe that the content of category discriminative information should not be determined only by the global response of the feature map but also consider the number of frame features that are discriminative within it, which we name as emotional keyframes. Therefore, HRS is designed as a dual-branch structure integrating the global responses as well as local keyframes.

Avoid taking the value of the global average response directly, but employ (8) to (9) [22] to obtain the importance α as the global response branch score, which can prevent the misdecision due to the over-low response of certain local frame features:

$$z = \frac{\sum_{(i,j)\in F_m} Z_{(i,j)}}{\prod_{F_m}} \tag{8}$$

$$\alpha = F\left(z, W_{1,2}\right) = \sigma\left(W_2 \delta\left(W_1(z)\right)\right) \tag{9}$$

where $z, \alpha \in \mathbb{R}^C$, $F_m$is the pixel point range of the feature map, σ and δ are the $Sigmoid$ and $LeakyRelu$ activation functions. $W_{1,2} \in \mathbb{R}^{C\times\frac{C}{2}, \frac{C}{2}\times C}$ is the learnable parameter.

Same reasons as above, the number of keyframe blink points owned within a single frame is the criteria for judging emotional keyframes. First, we utilize the 1D-convolution shared among feature maps to conduct response filtering under the unified standard for all frame features within each feature map in Z to find $\widetilde{Z}$ ($\widetilde{Z}_{c,i,j}$ denotes the j-th response point in the i-th frame of the c-th feature map). Second, obtaining the mean value of the filtered response points, and we denote the points beyond that as keyframe blink points and their count as the importance score of a single frame. Finally, finding the mean of the importance scores of all frames as the keyframe threshold, attaining the total number of emotional keyframes within every feature map as the local branch score. We define the process as:

$$\beta = \sum_{i}^{H'} \varepsilon \left(\widetilde{Z}_i - \frac{\sum_i^{H'} \sum_j^{W'} \varepsilon\left(\widetilde{Z}_{i,j} - \frac{\Sigma_j^{W'} \widetilde{Z}_{i,j}}{W'} \right)}{H'} \right) \tag{10}$$

where ε denotes the Heaviside function, and $\beta \in \mathbb{R}^C$.

Lastly, screening the score-weighted key feature map $Z_{\gamma_{max}}$ determined jointly by the global response and local keyframes, and feature reuse is achieved at the next-level output through residual connections:

$$\gamma = H' \cdot \alpha + \beta \tag{11}$$

$$\gamma_{\max} = argmax\left(\frac{\exp\left(\gamma^i\right)}{\sum_{j=1}^{c} \exp\left(\gamma^j\right)}\right), i \in \{1, c\} \tag{12}$$

3. Experiment

3.1. Dataset

IEMOCAP [23] is a commonly used emotional dataset that contains approximately 12 hours of audio, video, transcription, and motion capture data, recorded by five male and five female actors. For this dataset, audio and transcription were selected as the input modalities, and a total of 5,531 utterances were evaluated using leave-one-speaker-out cross validation for four-category emotions: *happy* (merged with *excited*), *angry*, *sad*, and *neutral*. WA (weighted accuracy) and UA (unweighted accuracy) are used as evaluation metrics.

CMU-MOSEI [24] contains 3,228 videos of monologues performed by 1,000 people collected from YouTube for a total of 65 hours, which are further sliced into 23,453 sentences and transcriptions and labeled with sentiment scores of [-3,+3]. Similar to previous works, 16,326, 1,871, and 4,659 of these are used for training, validation, and testing. For the seven-category evaluation (ACC_7), we rounded the sentiment scores to seven discrete points, the binary accuracy (negative (<0), positive (>0); ACC_2) and the F1-score are also used as metrics.

3.2. Experiment settings

We employ pre-trained models WavLM [19] and BERT [20] to extract 768-dimensional embedding features for audio and text, respectively, with maximum sequence lengths of 255 and 50; the number of hidden units in the shared LSTM block is 384, h_X set to 5. HRS's response point and keyframe screening ratios are both 50% (H'= 0.5H, W'= 0.5W). SWRR is trained by PyTorch on 1 RTX 2080 Ti, using the Adam optimizer with $betas$ set to 0.9 and 0.99, the learning rate is $5e^{-4}$ on IEMOCAP and $1e^{-3}$ on CMU-MOSEI, and training 50 and 20 epochs respectively, every 50% past which the learning rate decays by a factor of 10.

3.3. Comparison

We conduct comparison experiments with other advanced methods on IEMOCAP. **FSER** [25]: utilizes dual spectrograms with custom spectrograms to solve the issue of disorder and proposes dynamic confidence to fuse each modal in the decision layer; **MPFU** [26]: proposes a data-driven multiplicative fusion method to combine the multi-modality for efficient predictions; **SWT** [27]: proposes a multimodal transformer with sharing weights in each layer to learn the mutual correlation; **TSIN** [28]: presents a Temporal and Semantic Interaction Network (TSIN) for keeping temporal and semantic consistency between audio and text; **MHA** [29]: proposal to fuse audio, text, and motion capture (MoCap) data via multi-head attention.

Table 1: *Evaluation results of IEMOCAP. (A: Audio; T: Text; F: Facial; M: MoCap; ^{3} is the trimodal version)*

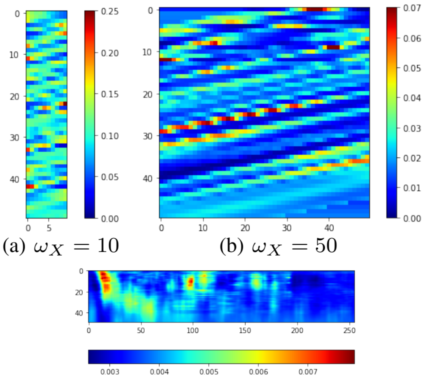
Methods	Features	WA$_{\%}$	UA$_{\%}$
FSER [25]	A + F	75.4	76.4
MPFU [26]	A + F	–	75.4
SWT [27]	A + T	76.8	77.1
TSIN [28]	A + T	76.2	78.1
MHA [29]	A + M + T	75.6	–
MPFU^3 [26]	A + F + T	–	78.2
SWRR	A + T	**77.4**	**78.5**

We also perform generalizability tests on CMU-MOSEI and compare with other methods in the literature. **MFRM** [30]: proposes a multi-fusion network with residual memory units for the long fusion sequence forgetting problem; **MISA** [31]: learning modality-invariant and -specific representations as a comprehensive and disentangled view to aid fusion; **self-MM** [32]: proposes modeling consistency and difference by unimodal supervision with multimodal joint training; **TETFN** [33]: proposes a text-enhanced transformer fusion network and retains the distinction by unimodal prediction; **LGCCT** [34]: presents an improved gating mechanism and transformer for cross-complementation of modalities.

Table 2: *Evaluation results of CMU-MOSEI*

Methods	**Features**	$ACC_{7\%}$	$ACC_{2\%}$	F1$_{\%}$
MFRM [30]	A + F + T	50.9	82.4	82.6
MISA [31]	A + F + T	52.2	83.6	83.8
self-MM [32]	A + F + T	- -	82.8	82.5
TETFN [33]	A + F + T	- -	84.3	**84.2**
LGCCT [34]	A + T	47.5	81.1	81.0
SWRR	A + T	**52.4**	**84.8**	83.8

We compare with advanced methods based on different models on both datasets and display them in Tables 1 and 2. Specifically, FSER [25] adopts decision-layer fusion and abandons explicit modeling of inter-modal interactions; SWRR improves by 1.8% compared to a tri-modal MHA [29] utilizing global attentional fusion and 3.1% compared to MPFU [26] employing multiplicative fusion. In addition, we further validate the effectiveness of SWRR on CMU-MOSEI, which also reaches a promising performance compared with the FC-based classifier [27, 28, 34] mentioned above, indicating that FMC can also perform MER tasks very well with the facilitation of HRFR. We should also consider that SWRR is a bimodal system employing audio and text, but still exhibits comparable or even better performance than the trimodal systems with the addition of MoCap [29] or Facial [30, 31, 32] features, and although it is lower than TETFN [33] in terms of F1-score, SWRR achieves optimal performance on the remaining evaluation metrics, demonstrating the superiority of our proposed model.



(a) $\omega_X = 10$ (b) $\omega_X = 50$

(c) ω_X set to full-size

Figure 3: *the thermogram of the mean correlation coefficient of ω_X at different widths*

3.4. Ablation and Variation Studies

We conduct the ablation study on IEMOCAP to validate several key components of SWRR and present the results in Table 3. First, we explore the performance of the unimodal version, which shows a significant decrease in both WA and UA, therefore, reasonably modeling multimodal information is one of the effective ways to improve recognition performance. To explore the effect of SliWa on the system performance, a full-size cross-dot product is used instead of the sliding window mechanism during the fusion phase, which leads to a 2.1% and 1.8% decreases in WA and UA, the thermogram of the mean correlation coefficient at ω_X of 10, 50, and full-size is shown in Fig.3, the introduced noise coefficient increases with the window width, which indicates that the sliding window can effectively prevent the inflowing truncated noise giving weak correlation coefficients to the audio single-frame features and participating in the aggregation, thus blurring the fusion representation. We also explore the role of HRFR and the advantages of its dual-branch structure design, the UA decreases by 2.3% when removing HRFR, revealing that the additional emotion-categories high-response information can effectively promote the kernel to make the correct category mapping and improve the system performance. The single-branch structure also causes varying degrees of performance degradation, thus considering both global response and local keyframes is a vital design to drive the success of HRS. Moreover, HRFR-based FMC can perform better with fewer parameters than linear classifiers, explicitly retaining temporal- and structural-semantic information, therefore, which can be applied as a new paradigm of classifiers for MER.

Table 3: *Module ablation studies on the IEMOCAP*

Models	**WA**$_{\%}$	**UA**$_{\%}$
w/o audio	71.7	72.3
w/o text	70.9	71.6
SliWa (full-size)	75.3	76.7
SliWa ($\omega_X = 50$)	76.2	76.9
w/o HRFR	75.6	76.2
HRFR (uni-global)	76.4	77.0
HRFR (uni-local)	75.9	76.8
Linear	76.5	77.4
SWRR	**77.4**	**78.5**

4. Conclusions

In this paper, we propose employing FMC to address the issue of one-dimensional average flattening causing the collapse of the feature structure. For further improving the performance of FMC, we present the SliWa fusion technique to reduce the redundant computation while truncating the noise coefficient and thus improving the fusion efficiency; the HRFR can strengthen the represent capability of feature maps to emotion-categories high-response information, thus facilitating the emotion filters to make correct confidence mapping ultimately. These proposals are integrated into SWRR and achieve superior performance on IEMOCAP and CMU-MOSEI, demonstrating the effectiveness of SWRR in MER tasks.

In future work, we will explore novel fusion mechanisms to incorporate facial modality and test the incentive effect of HRFR under the trimodal version.

5. References

[1] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59–73, 2021.

[2] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[3] T. Mittal, A. Bera, and D. Manocha, "Multimodal and context-aware emotion perception model with multiplicative fusion," *IEEE MultiMedia*, vol. 28, no. 2, pp. 67–75, 2021.

[4] Y. Lee, S. Yoon, and K. Jung, "Multimodal Speech Emotion Recognition Using Cross Attention with Aligned Audio and Text," in *Proc. 2020 Annual Conf. of the International Speech Communication Association (INTERSPEECH)*, Virtual Event / Shanghai, China, October 2020, pp. 2717–2721.

[5] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168 865–168 878, 2020.

[6] S. Chen, J. Tang, L. Zhu, and W. Kong, "A multi-stage dynamical fusion network for multimodal emotion recognition," *Cognitive Neurodynamics*, pp. 1–10, 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," in *Proc. 2021 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, Toronto, ON, Canada, June 2021, pp. 3020–3024.

[9] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition," *arXiv preprint arXiv:2207.04697*, 2022.

[10] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross-and self-attention network for speech emotion recognition," in *Proc. 2021 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*. Toronto, ON, Canada: IEEE, June 2021, pp. 4275–4279.

[11] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17 283–17 297, 2020.

[12] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proc. of the 25th ACM international conference on Multimedia (ACM MM)*, Mountain View, CA, USA, October 2017, pp. 478–484.

[13] P. Vasuki, "Design of hierarchical classifier to improve speech emotion recognition," *Computer Systems Science and Engineering*, vol. 44, no. 1, pp. 19–33, 2023.

[14] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[15] Y. Xia, L.-W. Chen, A. Rudnicky, R. M. Stern *et al.*, "Temporal context in speech emotion recognition." in *Proc. 2021 Annual Conf. of the International Speech Communication Association (INTERSPEECH)*, Brno, Czechia, August 2021, pp. 3370–3374.

[16] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, and T. Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure." in *Proc. 2021 Annual Conf. of the International Speech Communication Association (INTERSPEECH)*, Brno, Czechia, August 2021, pp. 1947–1951.

[17] J. Huang, J. Tao, B. Liu, and Z. Lian, "Learning utterance-level representations with label smoothing for speech emotion recognition." in *Proc. 2020 Annual Conf. of the International Speech Communication Association (INTERSPEECH)*, Virtual Event / Shanghai, China, October 2020, pp. 4079–4083.

[18] J. Liu and H. Wang, "A speech emotion recognition framework for better discrimination of confusions." in *Proc. 2021 Annual Conf. of the International Speech Communication Association (INTERSPEECH)*, Brno, Czechia, August 2021, pp. 4483–4487.

[19] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–14, 2022.

[20] 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.

[21] Z. Wang, X. Zhou, W. Wang, and C. Liang, "Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 4, pp. 923–934, 2020.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, Salt Lake City, UT, USA, June 2018, pp. 7132–7141.

[23] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[24] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018, pp. 2236–2246.

[25] N. Jia, C. Zheng, and W. Sun, "A multimodal emotion recognition model integrating speech, video and mocap," *Multimedia Tools and Applications*, vol. 81, no. 22, pp. 32 265–32 286, 2022.

[26] T. Mittal, A. Bera, and D. Manocha, "Multimodal and context-aware emotion perception model with multiplicative fusion," *IEEE MultiMedia*, vol. 28, no. 2, pp. 67–75, 2021.

[27] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition." in *Proc. 2021 Annual Conf. of the International Speech Communication Association (INTERSPEECH)*, Brno, Czechia, August 2021, pp. 4518–4522.

[28] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang, "Multimodal emotion recognition with temporal and semantic consistency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3592–3603, 2021.

[29] J. Zhang, L. Xing, Z. Tan, H. Wang, and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Computers Industrial Engineering*, vol. 168, p. 108078, 2022.

[30] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 320–334, 2020.

[31] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. of the 28th ACM international conference on Multimedia (ACM MM)*, Virtual Event / Seattle, WA, USA, October 2020, pp. 1122–1131.

[32] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. of the AAAI conference on artificial intelligence*, vol. 35, no. 12, Virtual Event, February 2021, pp. 10 790–10 797.

[33] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[34] F. Liu, S.-Y. Shen, Z.-W. Fu, H.-Y. Wang, A.-M. Zhou, and J.-Y. Qi, "Lgcct: A light gated and crossed complementation transformer for multimodal speech emotion recognition," *Entropy*, vol. 24, no. 7, p. 1010, 2022.