Interspeech 2024
1-5 September 2024, Kos, Greece

MFDR: Multiple-stage Fusion and Dynamically Refined Network 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

Emotion recognition in conversation should not rely solely on discovering emotion keywords but also make comprehensive judgments after considering the context. To this end, we propose the MFDR to efficiently integrate acoustic and textual information. Specifically, acoustic-word combination and context perception are modeled sequentially in stages through the Sliding Adaptive Window Attention (SAWA) and Gated Context Perception Unit. More importantly, without additional memory overhead, SAWA allows the perception range to be adaptively adjusted according to the correlation strength to solve the misalignment and information loss caused by window truncation, modeling fusion under variable granularity. Furthermore, emotion refinement through Dynamic Frame Convolution strips out emotion-irrelevant frames, thereby generating a compact and emotionally discriminative fusion representation. The efficacy of MFDR is confirmed by IEMOCAP and CMU-MOSEI, where it demonstrates promising performance.

Index Terms: multi-stages fusion, multimodal emotion recognition, sliding window, emotion refinement

1. Introduction

Self-attention [1] based cross version (cross-attention) has been proven to be an effective strategy in fusing heterogeneous modalities to extract emotional representations[2, 3, 4]. Notably, existing studies using this technique to model audio-text interactions often allow the current word to establish associative connections with all acoustic frames [5, 6], obtaining a correlation matrix to guide weighted aggregation. However, humans can usually initially determine the speaker's emotional tendency by analyzing the emotional words in the utterance and combining them with the corresponding acoustic information such as loudness and pitch. In contrast, constructing a globally cross-modal correlation matrix inevitably leads to a significant computational load. And in this process, the redundant computation of acoustic-word irrelevance also introduces a large number of invalid noise weights, which in turn leads the words to incorrectly focus their attention on irrelevant acoustic information.

Recalling the process of human emotion recognition, we not only combine words and their corresponding acoustic information to mark obvious emotional expressions but also further consider contextual information in the global meaning, such as irony, transitivity, etc. It can be seen that acoustic-word interaction is a multi-stage perceptual process. Meanwhile, it has been evidenced that contextual information affects the brain's perception and judgment of emotion [7]. Although RNNs and their variant structures have excellent capabilities in modeling temporal information in sequences [8, 9], the structural property of only allowing a single stream of inputs at each time step limits their ability to synchronize the fusion and analysis of global contexts at the current time step.

We also need to note that fusion representations often have a relatively large dimension before feeding into the classifier. Linear layers [10, 11] can effectively compress on the feature dimension, but with the introduction of massive neurons will fall into the risk of overfitting. Considering that not every frame in a feature sequence is emotionally discriminative, therefore, some studies choose max- [5] or hybrid- [12] pooling without introducing additional parameters for compression in the sequence dimension, however, which also have some drawbacks in certain aspects. The former directly discards the original features, while the latter allows some important regions of information to be processed by averaging or other operations, which ultimately leads to weaker discriminative features. In addition, the de-parameterization also implies that they are model-agnostic and cannot be iteratively optimized to selectively retain emotionally critical information within each frame.

We propose a Multiple-stage Fusion and Dynamically Refined network (MFDR) to address the mentioned issues. The main contributions of this paper are summarized as follows:

1. We propose Sliding Adaptive Window Attention (SAWA) to model the acoustic-word combining stage, moreover, to prevent information loss caused by window truncation and the problem of temporal misalignment, we further endow the window with the ability to be dynamically adjusted according to the strength of feature correlation.
2. Gated Context Perception Unit (GCPU) enables parallel processing of the current frame and its global context, which can fully considers the impact of complex semantic information such as irony on emotion recognition.
3. Dynamic frame convolution (DFC) is used to effectively identify and weaken fine-grained information that is irrelevant to the emotion expression, and thus obtain a more compact and emotionally discriminative fusion representation.
4. The experimental results indicate that MFDR achieves 78.4% WA and 79.2% UA on IEMOCAP; we further verify the generalization ability of MFDR on CMU-MOSEI and reach a superior performance of 53.7%, 85.2%, and 84.4% on 7-class accuracy, binary accuracy, and F1-score, respectively.

2. Methodology

The architecture of MFDR is shown in Fig. 1, where the acoustic-word combination and context-perception stage are modeled by SAWA and GCPU respectively, and then the

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.2024-1735

model-knowable pooling dimensionality reduction is achieved by DFC. The design of each module is introduced subsequently.

2.1. Sliding Adaptive Window Attention

To overcome the redundant computation and noise weight associated with full-size dot products, we proposed the Sliding Window Attention (SliWa) [13] in our previous work, which can control the feature perception range between the acoustic frame and word, and then derive the cross-attention coefficients within the window, which are subsequently synchronized with different granularities to dynamically model the inter-modal information propagation. But, we have to realize that window truncation will inevitably lead to information loss, which is mainly manifested in two aspects: firstly, it will eliminate certain acoustic features that are more critical to the current word, especially when these features are at the window edges of ω_X; secondly, the temporal relationship between acoustic-word features will be more seriously misaligned as the window slides.

Therefore, we further optimize SliWa and propose the Adaptive Width Adjustment (AWA) strategy, which enables the perception range between frames and words to be dynamically adjusted according to the feature correlation strength, and the implementation details are described in Algorithm 1.

Algorithm 1 Adaptive Width Adjustment

Input: $att \in \mathbb{R}^{L\times\omega_X}$, $q^y \in \mathbb{R}^{L\times D}$, $k^x \in \mathbb{R}^{\overline{N}_X\times\omega_X\times D}$

Initialization: $diag\_att \in \mathbb{R}^{L\times T} \leftarrow$ Diagonalize att
$att\_key\_thr \leftarrow$ Mean value of each row in att as the threshold
$att\_key\_point \leftarrow$ Exceed the att_key_thr as the key point
$att\_keyWin \leftarrow$ Sum att_key_point by rows as correlation strength within each window
for i **in** $\overline{N}_X$ **do**
  if $att\_keyWin_i \geq 0.5*\omega_X$ **then**
    k_i^x = concat(k^x[(i-1), :0.5ω_X, :], k^x[(i+1), 0.5ω_X:, :])
    $add\_att = \mathrm{Softmax}(\frac{q_i^y (k_i^x)^T}{\sqrt{D}})$
    $diag\_att$[i, (i-1)*h_X:i*h_X] = add_att[:, :0.5ω_X]
    $diag\_att$[i, i*h_X+ω_X:i*h_X+1.5ω_X] = add_att[:, 0.5ω_X:]
    $new\_keyWin \leftarrow$ Perform 1∼3 on $diag\_att_i$
    if $new\_keyWin_i \geq \omega_X$) **then**
      k_i^x = concat(k^x[(i-2), :0.5ω_X, :], k^x[(i+2), 0.5ω_X:, :])
      $add\_att \leftarrow$ Perform 8
      $diag\_att$[i, (i-2)*h_X:(i-1)*h_X] = add_att[:, :0.5ω_X]
      $diag\_att$[i, i*h_X+1.5ω_X:i*h_X+2ω_X] = add_att[:, 0.5ω_X:]
    end if
  end if
end for ▷ Special processing for steps 0, 1, $\overline{N}_X$-2 and $\overline{N}_X$-1

Output: $diag\_att$

where L and T are the sequence lengths of text and audio features, ω_X and h_X are the window width and hop-length for the audio modality, and $\overline{N}_X$ is the number of divided windows, $diag\_att$ is used to aggregate the audio Value matrix v^x to generate the combining representation $z \in \mathbb{R}^{L\times D}$.

Notably, truncation information before and after the current window can be easily obtained by slicing the Key matrix k^x of the audio modality using the indexes instead of re-dividing it with a new window width, as shown in 7 in Algorithm 1, thus avoiding additional memory and computational overheads.

We organically integrate AWA with SliWa in Sliding Adaptive Window Attention (SAWA), which enables the window k_i^x with strong correlation to be expanded to three times of original range, thus providing richer acoustic contextual information for the current word, and effectively mitigating the information loss caused by window truncation and temporal misalignment.

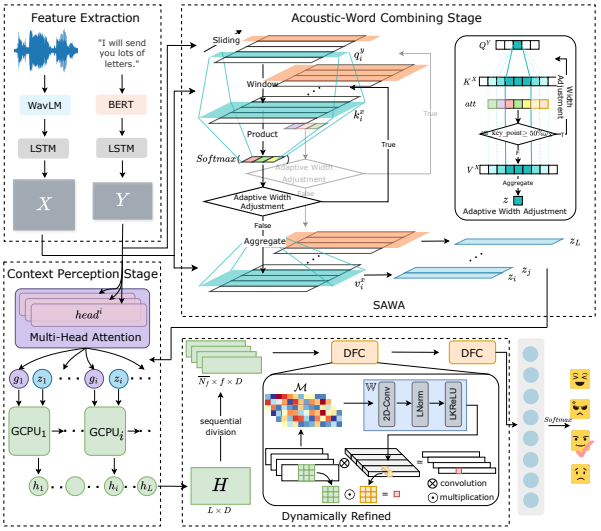


Figure 1: *The architecture of MFDR*

2.2. Gated Context Perception Unit

After completing the acoustic-word combination in our brain, it will also be placed in a wider context to synthesize complex linguistic phenomena such as irony. Inspired by this, we propose a recurrent structure called Gated Context Perception Unit (GCPU), which can simultaneously receive the combining representation z_t and the global contextual feature g_t at time step t, which means that it not only retains the capability of capturing temporal information but introduces the global understanding for z_t. Specifically, the GCPU contains two gates and two candidate states, as shown in Fig. 2, which are the reset gate r_t and update gate u_t as well as the combining representation candidate state $\tilde{z}_t$, the contextual information candidate state $\tilde{g}_t$.

Firstly, employing the multi-head attention [1] to obtain g_t, and information forgetting is controlled in u_t by combining the knowledge of h_{t-1}, z_t and g_t using Sigmoid (σ). Subsequently, h_{t-1}, which perpetuates useful information, is integrated with z_t and g_t to generate $\tilde{z}_t$ and $\tilde{g}_t$ respectively. $\tilde{z}_t$ emphasizes the affective meaning of the word when combined with acoustic information, while $\tilde{g}_t$ highlights the understanding of the global context. They complement each other in $\tilde{h}_t$ from different perspectives and together constitute a comprehensive understanding of the speaker's emotional state:

$$r_t = \sigma(W_{hr}h_{t-1} + W_{zr}z_t + W_{gr}g_t + b_r) \quad (1)$$

$$\tilde{z}_t = \mathrm{Tanh}(W_{z\tilde{z}}z_t + W_{h\tilde{z}}(r_t \odot h_{t-1}) + b_{\tilde{z}}) \quad (2)$$

$$\tilde{g}_t = \mathrm{Tanh}(W_{g\tilde{g}}g_t + W_{h\tilde{g}}(r_t \odot h_{t-1}) + b_{\tilde{g}}) \quad (3)$$

where $W \in \mathbb{R}^{D\times D}$ is the learnable weight matrix, b is the bias unit, and $\odot$ denotes the element-wise multiplication.

Unlike traditional GRU, in which the update gate is defined as a complementary relationship, which may not achieves effective update for a GCPU with two candidate states. This is because the $\tilde{z}_t$ and $\tilde{g}_t$ are independently generated by combining the z_t and g_t with the forgotten reset h_{t-1} respectively, summing them up to obtain the candidate hidden state $\tilde{h}_t$. And now, the knowledge redundancy between $\tilde{h}_t$ and h_{t-1} is unable to be represented by a weight vector generated by considering both z_t and g_t and achieving the complementary updating.

Therefore, we have redesigned u_t while controlling the computational complexity. Specifically, considering that $\tilde{z}_t$ and

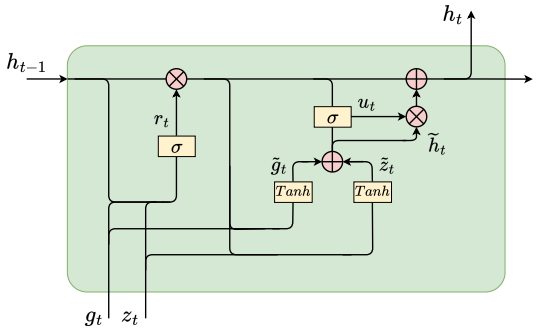


Figure 2: *The architecture of GCPU*

$\tilde{g}_t$ still introduce redundant or emotion-irrelevant information about the forgotten hidden state h_{t-1}, the update vector u_t integrally relies on $\tilde{z}_t$, $\tilde{g}_t$, and h_{t-1} to be generated. Finally, integrating the updated $\tilde{h}_t$ with the forgotten reset h_{t-1}, we get the $h_t \in H \in \mathbb{R}^{L\times D}$ at current step after considering both the local acoustic-word information and the global context knowledge:

$$\tilde{h}_t = \tilde{z}_t + \tilde{g}_t \tag{4}$$

$$u_t = \sigma(W_{\tilde{z}u}\tilde{z}_t + W_{\tilde{g}u}\tilde{g}_t + W_{hu}(r_t \odot h_{t-1}) + b_u) \tag{5}$$

$$h_t = r_t \odot h_{t-1} + u_t \odot \tilde{h}_t \tag{6}$$

The GCPU utilizes two gated mechanisms and two candidate states to dynamically adjust the information flow in the sequence, and by designing a dual-stream input structure, it achieves context perception at each time step simultaneously.

2.3. Dynamic Frame Convolution

Not every emotionally expressive moment in an utterance is equally discriminative, which is manifested in frame features is that certain detailed features are closely related to a particular emotion, while others are shown to be weakly or unrelated.

Therefore, inspired by the work on layout-aware processing [14], we propose Dynamic Frame Convolution (DFC) to achieve differentiated consideration for each frame.

Specifically, the DFC contains two crucial parts: the kernel generation map $\mathcal{M}$ and the kernel generation block $\mathbb{W}$. Firstly, obtaining the spliced frame features $H^{\mp} \in \mathbb{R}^{\overline{N}_f \times f \times D}$ by dividing H in sequence dimension:

$$\overline{N}_f = \left\lfloor \frac{L - \omega_f}{h_f} \right\rfloor + 1 \tag{7}$$

where ω_f is the frame window width, h_f is the frame hoplength, $\lfloor\cdot\rfloor$ denotes rounding down, and $\overline{N}_f$ is the number of frame features under ω_f and h_f divisions.

Subsequently, initializing $\mathcal{M}$ to the same size as H, $\mathbb{W}$ receives $\mathcal{M}$ as input and performs a series of refinement processes on it, including standard convolution, layer normalization and activation, which can generate private kernel function $\Phi \in \mathbb{R}^{\overline{N}_f \times f \times f \times D}$ for each spatial locations of the $H_i^{\mp} \in \mathbb{R}^{f\times D}$, so that they can be given different weights in the product accumulation process depending on the relevance of the emotion expression, ultimately achieving a model-knowable pooling effect. This process is repeated twice, aiming to further enhance the ability of the generative kernel Φ to capture different emotional details in the frame features:

$$\Phi = \mathbb{W}(\mathcal{M}) \tag{8}$$

3. Experiment

3.1. Dataset and evaluation metrics

IEMOCAP [15] 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 are selected as the input modalities, and a total of 5,531 utterances are evaluated using 5-fold and 10-fold 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 [16] 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 setting

We employ pre-trained WavLM [17] and BERT [18] to extract 768-dimensional audio and word embeddings, the T and L are set to 255 and 50. In SliWa, the window width ω_X and hoplength h_X are 10 and 5 for audio and are set to 1 for text. The hidden state mapping dimension of the GCPU is 256. In Ihe two-layer DFC, the ω_f is 3, h_f is set to 2 and 3 respectively, and the channel dimension of the middle layer is set to 64.

We implement MFDR in the PyTorch framework, training it with Adam optimizer on 1 RTX 2080 Ti, the learning rate and batch are $5e^{-4}$ and 32 on IEMOCAP, $1e^{-3}$ and 64 on CMU-MOSEI, training them 40 and 20 epochs respectively, every 50% past which the learning rate decays by a factor of 10.

3.3. Comparison

Considering that different division methods can significantly affect the evaluation results. Therefore, we adopted both 5-fold and 10-fold cross-validation to comprehensively evaluate the performance of MFDR and present them in Table 1, compared to the method adopting global cross-attention [6, 19], MFDR improved 3.7% and 2.8% on WA respectively, and 4.5% on UA; TSIN [20] effectively explores fine-grained inter-modal interactions and recalibrates the features, but directly concatenate the heterogeneous modalities before classification, which allows our method to achieve a better performance under both cross-validation strategies.

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

Methods	Features	validation	WA%	UA%
IA-MMTF [6]	A + T	5-fold	72.0	72.5
MER-HAN [2]	A + T	5-fold	73.3	74.2
TSIN [20]	A + T	5-fold	74.9	76.6
MFDR	A + T	5-fold	**75.7**	**77.0**
MHDNN [21]	A + T	10-fold	74.5	73.2
TSIN [20]	A + T	10-fold	76.2	78.1
MHA [19]	A + M + T	10-fold	75.6	–
MPFU [22]	A + F + T	10-fold	–	78.2
MFDR	A + T	10-fold	**78.4**	**79.2**

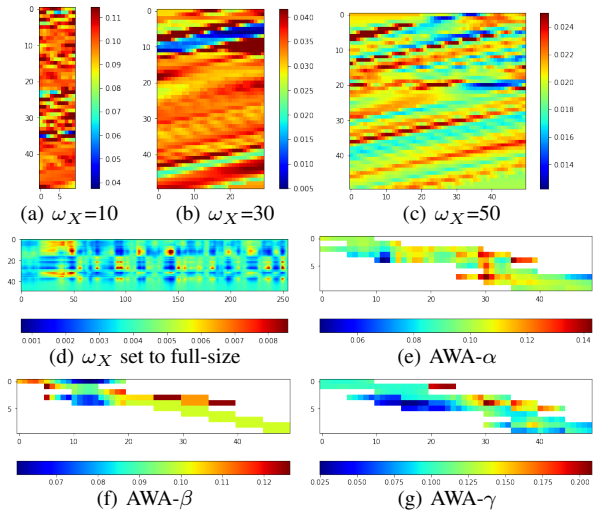
Table 2: *Evaluation results of CMU-MOSEI*

Methods	**Features**	$ACC_{7\%}$	$ACC_{2\%}$	F1$_{\%}$
MFRM [23]	A + F + T	50.9	82.4	82.6
BIMHA [24]	A + F + T	52.1	84.1	83.4
FmlMSN [25]	A + F + T	52.7	83.5	83.6
TETFN [26]	A + F + T	- -	84.3	**84.2**
LGCCT [27]	A + T	47.5	81.1	81.0
MFDR	A + T	**53.7**	**85.2**	84.1

In addition, to further evaluate the generalization ability of MFDR, we extended the range of experimental data and further tested it on the CMU-MOSEI, the experimental results are shown in Table 2. MFDR achieves comparable or even better performance in the situation where bi-modal are used as inputs (A + T) compared to the work with tri-modal (A + T + F/M) [19, 22, 23, 25]. In contrast to approaches that combine input features from more modalities to enhance the system performance, MFDR demonstrates the feasibility of also achieving this by optimizing the cross-modal fusion strategy.

3.4. Ablation and Variation Studies

To analyze the effect of each module, we conducted a series of ablation and variation studies as described in Table 3 on both datasets using WA and UA under 10-fold cross-validation as well as ACC_2 and F1-score as the main evaluation metrics.



(a) ω_X=10 (b) ω_X=30 (c) ω_X=50

(d) ω_X set to full-size (e) AWA-α

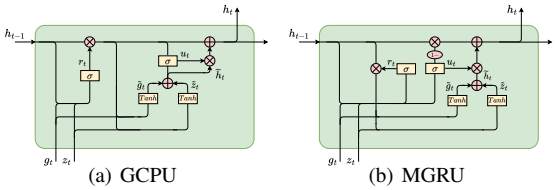
(f) AWA-β (g) AWA-γ

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

Firstly, we explored the effect of different ω_X on the acoustic-word combination, and the performance of MFDR displays certain fluctuations which indicate that at a fixed ω_X, too small will lead to information truncation, and too large will introduce massive noise weights as shown by (a), (b), (c), and (d) in Fig. 3. Therefore, we further propose the AWA strategy to incorporate the key frame features located at the window edges into the perception range and reduce the interference of irrelevant acoustic-word pairs, which can be clearly seen by (e), (f), and (g) in Fig. 3 (for perspective, we show the first ten windows), enabling the system performance to be further improved without introducing any new parameters. Subsequently, after replacing the GCPU with the complementary updated MGRU (as shown in Fig. 4 (a)), WA and UA decrease by 1.7% and 2.0%, and ACC_2 and F1-score reduce by 1.8% and 1.6% respectively. We can conclude that the design of u_t is effective by decoupling the complements and completing the information update after considering $\tilde{z}_t$, $\tilde{g}_t$, and the forgotten reset h_{t-1}. We further explore the importance for contextual modeling by using GRU to directly process z_t and acquiring g_t without multi-head attention. The performance degradation reveals that GCPU can effectively capture and integrate complex contexts. When utilizing model-agnostic max- and avg-pooling, the performance produces varying degrees of decline, which suggests that the Φ in DFC is able to more accurately identify and emphasize the key detail information in the frames through model-knowable fine-grained aggregation. Moreover, directly adopting the Linear layer (*w/o* DFC) leads to 1.1% and 1.2% declines in WA and UA, as well as 1.5% and 0.8%, decreases in ACC_2 and F1-score, which indicates that linear mapping cannot independently focus on different spatial locations. On the contrary, DFC is able to generate customized private kernel functions with the help of $\mathbb{W}$ and $\mathcal{M}$, which effectively improves the accuracy and generalization of MFDR.

Table 3: *Module ablation studies on both datasets*

Models	**WA**$_{\%}$	**UA**$_{\%}$	$ACC_{2\%}$	F1$_{\%}$
ω_X=10	76.2	77.3	83.9	83.2
ω_X=30	77.3	78.3	84.5	83.8
ω_X=50	76.8	77.9	83.7	83.2
full-size	77.5	78.4	84.1	83.6
w/o Context	76.6	77.6	84.3	83.7
GRU	75.4	76.1	82.6	81.1
MGRU	76.7	77.2	83.4	82.5
max-pooling	74.9	75.7	82.5	81.7
avg-pooling	76.2	76.6	82.9	82.0
w/o DFC	77.3	78.0	83.7	83.3
MFDR	**78.4**	**79.2**	**85.2**	**84.1**



(a) GCPU (b) MGRU

Figure 4: *Architecture comparison between GCPU and MGRU*

4. Conclusion

In this paper, we propose a Multi-stage Fusion and Dynamically Refined network (MFDR), in which, the SAWA can effectively control the inter-modal perception range and further adaptively adjust the window width according to the strength of feature correlation, thus effectively mitigating the information loss and temporal misalignment caused by window truncation. Subsequently, the parallelization of the current frame and global context by the GCPU enables MFDR to fully consider the effects of complex semantic information. Ultimately, emotion refinement of the fusion representation is achieved in the DFC by initializing and training a model-knowable private kernel function for the different spatial locations of the frame features. Extensive experiments on the IEMOCAP and CMU-MOSEI, as well as ablation studies for each module, demonstrated the effectiveness of MFDR in the multimodal emotion recognition task.

5. References

[1] 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.

[2] S. Zhang, Y. Yang, C. Chen, R. Liu, X. Tao, W. Guo, Y. Xu, and X. Zhao, "Multimodal emotion recognition based on audio and text by using hybrid attention networks," *Biomedical Signal Processing and Control*, vol. 85, p. 105052, 2023.

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

[4] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2021, pp. 4730–4738.

[5] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross-and self-attention network for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4275–4279.

[6] L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu, and S. Ding, "Emotion recognition with multimodal transformer fusion framework based on acoustic and lexical information," *IEEE MultiMedia*, vol. 29, no. 2, pp. 94–103, 2022.

[7] S. Paulmann and M. D. Pell, "Contextual influences of emotional speech prosody on face processing: How much is enough?" *Cognitive, Affective, & Behavioral Neuroscience*, vol. 10, no. 2, pp. 230–242, 2010.

[8] X. Wang, X. Chen, and C. Cao, "Human emotion recognition by optimally fusing facial expression and speech feature," *Signal Processing: Image Communication*, vol. 84, p. 115831, 2020.

[9] J. Wen, D. Jiang, G. Tu, C. Liu, and E. Cambria, "Dynamic interactive multiview memory network for emotion recognition in conversation," *Information Fusion*, vol. 91, pp. 123–133, 2023.

[10] Q. Lu, X. Sun, Z. Gao, Y. Long, J. Feng, and H. Zhang, "Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis," *Information Processing & Management*, vol. 61, no. 1, p. 103538, 2024.

[11] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "Dst: Deformable speech transformer for emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 3020–3024.

[13] 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*, vol. 2023, 2023, pp. 2433–2437.

[14] J. Chen, T. He, W. Zhuo, L. Ma, S. Ha, and S. G. Chan, "Tvconv: Efficient translation variant convolution for layout-aware visual processing," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 538–12 548.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] P. Singh, R. Srivastava, K. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, p. 107316, 2021.

[22] 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.

[23] 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.

[24] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowledge-Based Systems*, vol. 235, p. 107676, 2022.

[25] J. Peng, T. Wu, W. Zhang, F. Cheng, S. Tan, F. Yi, and Y. Huang, "A fine-grained modal label-based multi-stage network for multimodal sentiment analysis," *Expert Systems with Applications*, vol. 221, p. 119721, 2023.

[26] 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.

[27] 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.