



GIA-MIC: Multimodal Emotion Recognition with Gated Interactive Attention and Modality-Invariant Learning Constraints

Jiajun He¹, Jinyi Mi¹, Tomoki Toda²

¹Graduate School of Informatics, Nagoya University, Japan

²Information Technology Center, Nagoya University, Japan

jiajun.he@g.sp.m.is.nagoya-u.ac.jp, mi.jinyi@g.sp.m.is.nagoya-u.ac.jp,
tomoki@icts.nagoya-u.ac.jp

Abstract

Multimodal emotion recognition (MER) extracts emotions from multimodal data, including visual, speech, and text inputs, playing a key role in human-computer interaction. Attention-based fusion methods dominate MER research, achieving strong classification performance. However, two key challenges remain: effectively extracting modality-specific features and capturing cross-modal similarities despite distribution differences caused by modality heterogeneity. To address these, we propose a gated interactive attention mechanism to adaptively extract modality-specific features while enhancing emotional information through pairwise interactions. Additionally, we introduce a modality-invariant generator to learn modality-invariant representations and constrain domain shifts by aligning cross-modal similarities. Experiments on IEMO-CAP demonstrate that our method outperforms state-of-the-art MER approaches, achieving WA 80.7% and UA 81.3%.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Multimodal emotion recognition (MER) aims to leverage information from multiple perceptual modalities, such as visual, acoustic, and textual expressions, to accurately identify human emotions. MER has broad applications in fields such as human-computer interaction and intelligent customer service [1].

Although MER has demonstrated significant promise, it still faces two key challenges: effective feature extraction from each modality and robust multimodal fusion. The first challenge lies in extracting meaningful modality-specific features. Early MER approaches primarily relied on training models from scratch, utilizing architectures such as recurrent neural networks (RNNs) [2, 3], convolutional neural networks (CNNs) [4, 5], and Transformers [6, 7]. These methods improved emotion recognition by capturing latent patterns from video frames, audio spectrograms, and text sequences. More recently, inspired by the success of pretrained models in other speech-related tasks, researchers have discovered that deep pretrained features offer more robust and generalizable representations than hand-crafted features [8]. Consequently, various pretrained models have been adopted for MER, including visual-based models (e.g., CLIP [9, 10], DINO [11]), acoustic-based models (e.g., HuBERT [12], WavLM [13]), and text-based models (e.g., DeBERTa [14], RoBERTa [15]). These models have significantly enhanced MER performance by leveraging rich feature representations learned from large-scale datasets.

The second challenge is effectively integrating information across modalities. Attention mechanisms have played a pivotal role in MER research owing to their ability to selectively focus on emotion-relevant features. Traditional approaches of-

ten employ self-attention, where both the query and key originate from the same modality [16]. While effective, this method lacks direct cross-modal interaction, limiting its ability to fully exploit complementary information across modalities. To address this limitation, cross-attention mechanisms have been introduced, enabling one modality to attend to another by deriving queries and keys from different modalities, and utilizing values to retrieve the corresponding weighted information from the attended modality [17]. Although cross-attention improves information exchange, it still struggles with modality heterogeneity, which can lead to inconsistencies and imbalances in feature alignment.

To address these challenges, we propose a novel method, GIA-MIC, which integrates gated interactive attention (GIA) for modality-specific representation (MSR) learning and modality-invariant learning constraints (MIC) to enhance modality-invariant representation (MIR) learning. Specifically, GIA facilitates adaptive extraction of modality-specific representations by dynamically regulating intermodal interactions, thereby capturing fine-grained emotional cues. Meanwhile, MIC encourages cross-modal alignment by constraining domain shifts between different modalities and learning modality-invariant representations through similarity-based constraints. We conducted extensive experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, and the results demonstrate the effectiveness of GIA-MIC, achieving new state-of-the-art (SOTA) performance in multimodal emotion recognition. Our main contributions are as follows:

- We propose a GIA mechanism that dynamically regulates the importance of intermodal interactions, enabling the adaptive extraction of modality-specific representations.
- We introduce a modality-invariant generator (MIG) module along with MIC to learn robust modality-invariant representations. By capturing cross-modal similarities and enforcing constraints that reduce modality discrepancies, MIC enhances the generalization ability of the model, making it more resilient to domain shifts between different modalities and further improving MER performance.
- Our proposed GIA-MIC method outperforms state-of-the-art MER methods on the IEMOCAP dataset.

2. Proposed Method

2.1. Problem Formulation

The MER task is defined as $f(V, S, T) = L$, where V , S , and T represent the video, speech, and text modalities, respectively. The goal is to fuse these modalities for emotion classification, yielding $L \in \{l_1, l_2, \dots, l_e\}$, where e is the number of emotion categories.

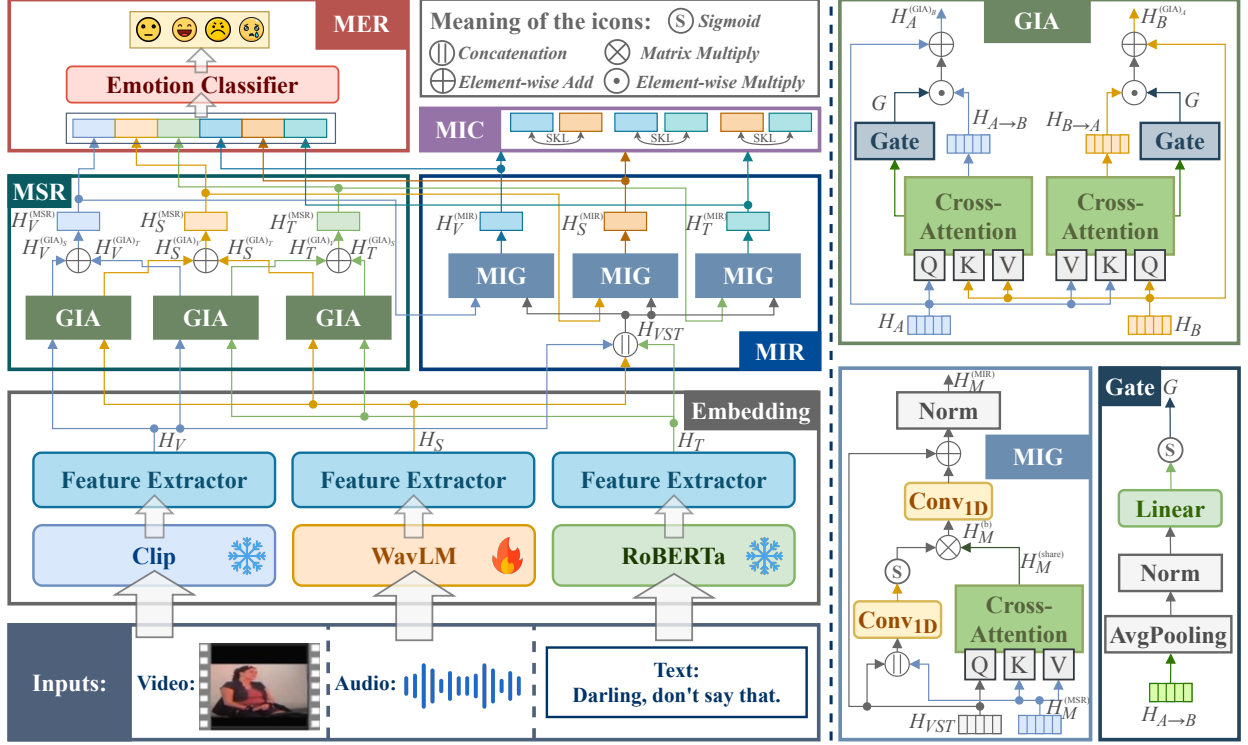


Figure 1: The overall architecture of our proposed method.

2.2. Embedding Module

We utilize pretrained encoders CLIP, WavLM, and RoBERTa to extract features from the visual, speech, and text modalities, respectively. To ensure consistency in feature dimensions across modalities, we apply a linear transformation layer to the speech features. These extracted features are then passed through a Feature Extractor, which consists of a single-layer Transformer encoder designed to refine contextual representations and enhance the quality of the learned embeddings. The resulting preliminary modality representations are formulated as:

$$H_V \in \mathbb{R}^{k \times d}, \quad H_S \in \mathbb{R}^{m \times d}, \quad H_T \in \mathbb{R}^{n \times d}, \quad (1)$$

where k, m, n denote the sequence lengths of the video, speech, and text modalities, respectively, and d represents the unified feature dimension.

2.3. Multimodal Interaction Module

Our multimodal interaction module consists of MSR and MIR modules. We describe these two modules in detail below.

2.3.1. Modality-specific Representations (MSR) Module

The MSR module is designed to capture modality-specific emotional features while dynamically regulating intermodal interactions. To achieve this, we introduce the gated interactive attention (GIA) mechanism, which adaptively controls the information flow between different modalities.

GIA Block. The GIA mechanism enables adaptive information fusion between modalities by learning modality-specific gating factors. Given two modalities A and B , the cross-attention mechanism computes the attention-weighted representation of A based on B :

$$H_{A \rightarrow B} = \text{Softmax} \left(\frac{Q_A K_B^T}{\sqrt{d}} \right) V_B \in \mathbb{R}^{t_A \times d}, \quad (2)$$

where $Q_A = W_Q H_A \in \mathbb{R}^{t_A \times d}$, $K_B = W_K H_B \in \mathbb{R}^{t_B \times d}$, and $V_B = W_V H_B \in \mathbb{R}^{t_B \times d}$ are the query, key, and value matrices for modalities A and B , with t_A and t_B as their sequence lengths, respectively. Here, $W_Q \in \mathbb{R}^{d \times d}$, $W_K \in \mathbb{R}^{d \times d}$, and $W_V \in \mathbb{R}^{d \times d}$ are learnable weight matrices for the query, key, and value transformations.

To control the information flow, we introduce a modality-specific gating function, which learns to regulate the influence of modality B on modality A :

$$G = \sigma(W_g(\text{Norm}(\text{AvgPooling}(H_{A \rightarrow B})) + b_g)) \in \mathbb{R}^{t_A \times d}, \quad (3)$$

where $W_g \in \mathbb{R}^{d \times d}$ and $b_g \in \mathbb{R}^d$ are learnable parameters, “ $\sigma(\cdot)$ ” is the sigmoid activation function. The final MSR of modality A is computed as:

$$H_A^{(\text{GIA})B} = G \odot H_{A \rightarrow B} + (1 - G) \odot H_A \in \mathbb{R}^{t_A \times d}, \quad (4)$$

where “ \odot ” represents element-wise multiplication. Similarly, we can obtain the gated representation $H_B^{(\text{GIA})A} \in \mathbb{R}^{t_B \times d}$ for modality B . The MSR module is composed of three GIA blocks, each designed to capture pairwise interactions between modalities. Specifically, the interactions are formulated as:

$$\begin{aligned} \{H_V^{(\text{GIA})S}, H_S^{(\text{GIA})V}\} &= \text{GIA}(H_V, H_S), \\ \{H_S^{(\text{GIA})T}, H_T^{(\text{GIA})S}\} &= \text{GIA}(H_S, H_T), \\ \{H_T^{(\text{GIA})V}, H_V^{(\text{GIA})T}\} &= \text{GIA}(H_T, H_V), \end{aligned} \quad (5)$$

where $H_V^{(\text{GIA})S}, H_V^{(\text{GIA})T} \in \mathbb{R}^{k \times d}$, $H_S^{(\text{GIA})V}, H_S^{(\text{GIA})T} \in \mathbb{R}^{m \times d}$, $H_T^{(\text{GIA})S}, H_T^{(\text{GIA})V} \in \mathbb{R}^{n \times d}$. For each modality, we obtain its final MSR by summing the outputs from the corre-

sponding GIA blocks:

$$\begin{aligned} H_V^{(\text{MSR})} &= H_V^{(\text{GIA})_S} + H_V^{(\text{GIA})_T} \in \mathbb{R}^{k \times d}, \\ H_S^{(\text{MSR})} &= H_S^{(\text{GIA})_V} + H_S^{(\text{GIA})_T} \in \mathbb{R}^{m \times d}, \\ H_T^{(\text{MSR})} &= H_T^{(\text{GIA})_V} + H_T^{(\text{GIA})_S} \in \mathbb{R}^{n \times d}. \end{aligned} \quad (6)$$

The overall MSR is obtained by concatenating the MSR of each modality:

$$H^{(\text{MSR})} = [H_V^{(\text{MSR})}, H_S^{(\text{MSR})}, H_T^{(\text{MSR})}] \in \mathbb{R}^{(k+m+n) \times d}, \quad (7)$$

which ensures that each modality retains its distinct emotional features while benefiting from cross-modal interactions.

2.3.2. Modality-invariant Representations (MIR) Module

The MIR module aims to extract shared emotional features across different modalities, ensuring that the learned representations capture modality-agnostic emotional cues. Unlike the MSR module, which emphasizes modality-dependent features, the MIR module focuses on learning a common latent space where multimodal features are aligned. The MIR module consists of three modality-invariant generator (MIG) blocks, each is responsible for refining the representation of a specific modality by incorporating cross-modal interactions.

MIG block. To establish a shared interaction space, we concatenate the preliminary embeddings from all three modalities to form the query Q :

$$H_{VST} = [H_V, H_S, H_T] \in \mathbb{R}^{(k+m+n) \times d}. \quad (8)$$

For each modality $M \in \{V, S, T\}$, we use the corresponding $H_M^{(\text{MSR})}$ as the key K and value V . Cross-modal attention is then applied to compute inter-modal interactions:

$$H_M^{(\text{share})} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \in \mathbb{R}^{(k+m+n) \times d}. \quad (9)$$

To enhance the feature’s modality invariance, a parallel convolutional network is employed to learn a mask that filters out modality-specific information:

$$\begin{aligned} H_M^{(b)} &= H_M^{(\text{share})} \otimes \sigma(\text{Conv}_{1d}([H_M^{(\text{MSR})}, H_{VST}])) \\ &\in \mathbb{R}^{(k+m+n) \times d}, \quad M \in \{V, S, T\}, \end{aligned} \quad (10)$$

where “ \otimes ” indicates element-wise multiplication and “Conv_{1d}” denotes 1×1 convolution followed by PReLU activation [18]. To further refine the modality-invariant representation, we introduce an additional 1D convolutional layer with a stride of 2 that enhances local feature extraction and interaction, helping to capture fine-grained dependencies across modalities. Additionally, we incorporate a residual connection to preserve the original multimodal information:

$$\begin{aligned} H_M^{(\text{MIG})} &= \text{Norm}\left(H_{VST} + \text{Conv}_{1d}\left(H_M^{(b)}\right)\right) \\ &\in \mathbb{R}^{(k+m+n) \times d}, \quad M \in \{V, S, T\}, \end{aligned} \quad (11)$$

where “Norm” represents layer normalization [19]. The overall MIR is obtained by concatenating the MIR of each modality:

$$H^{(\text{MIR})} = [H_V^{(\text{MIG})}, H_S^{(\text{MIG})}, H_T^{(\text{MIG})}] \in \mathbb{R}^{3(k+m+n) \times d}, \quad (12)$$

Finally, the MSR and MIR are concatenated together to obtain the final multimodal interactive representation $H_{VST}^{(\text{fus})}$:

$$H_{VST}^{(\text{fus})} = H^{(\text{MSR})} + H^{(\text{MIR})} \in \mathbb{R}^{4(k+m+n) \times d}, \quad (13)$$

Modality-Invariant Learning Constraints (MIC). To enforce constraints on the consistency of modality-invariant representations across different modalities, we utilize the symmetric KL divergence (SKL) to obtain a more stable and bidirectional similarity measure. The SKL divergence between two modality-invariant representations $H_M^{(\text{MIR})}$ and $H_N^{(\text{MIR})}$ is computed as follows:

$$\begin{aligned} D_{\text{SKL}}(H_M^{(\text{MIR})}, H_N^{(\text{MIR})}) &= \frac{1}{2} \left(D_{\text{KL}}(H_M^{(\text{MIR})} \parallel H_N^{(\text{MIR})}) \right. \\ &\quad \left. + D_{\text{KL}}(H_N^{(\text{MIR})} \parallel H_M^{(\text{MIR})}) \right), \end{aligned} \quad (14)$$

where $M, N \in \{V, S, T\}$ represent the three modalities, and $D_{\text{KL}}(P \parallel Q)$ is defined as:

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \quad (15)$$

To enforce cross-modal alignment, we compute the SKL divergence for all modality pairs:

$$\begin{aligned} \mathcal{L}_{\text{MIR}} &= D_{\text{SKL}}(H_V^{(\text{MIR})}, H_S^{(\text{MIR})}) + \\ &D_{\text{SKL}}(H_S^{(\text{MIR})}, H_T^{(\text{MIR})}) + D_{\text{SKL}}(H_T^{(\text{MIR})}, H_V^{(\text{MIR})}). \end{aligned} \quad (16)$$

This loss function encourages the MIR distributions of different modalities to be similar, thereby improving the consistency of cross-modal representations. By minimizing \mathcal{L}_{MIR} , we ensure that the learned MIR representations reside in a shared latent space, facilitating more effective multimodal fusion and enhancing the overall performance of emotion recognition.

2.4. Emotion Classification Module

The emotion classification is performed by applying the temporal average pooling layer on the concatenated MSR and MIR features $H_{VST}^{(\text{fus})}$ followed by a linear layer and a Softmax activation function.

$$P(\hat{l}|H_{VST}^{(\text{fus})}) = \text{Softmax}((W_c(\text{AvgPooling}(H_{VST}^{(\text{fus})})) + b_c)), \quad (17)$$

where W_c and b_c are learnable parameters and \hat{l} is the predicted emotion classification. The corresponding loss function can be defined as

$$\mathcal{L}_{\text{ER}} = - \sum \log(P(\hat{l}|H_{VST}^{(\text{fus})})). \quad (18)$$

Joint Training. During the training stage, the two loss functions are linearly combined as the overall training objective:

$$\mathcal{L} = \mathcal{L}_{\text{ER}} + \gamma \mathcal{L}_{\text{MIR}}, \quad (19)$$

where γ is the hyperparameter for balancing the weight of two loss functions.

3. Experiments and Results

3.1. Experimental Conditions

Experiment Settings. Our method was implemented with Python 3.10.0 and Pytorch 1.11.0 and was trained on a system with an Intel Xeon Gold 6248 CPU, 32GB RAM, and an NVIDIA Tesla V100 GPU. The visual, speech, and text encoders were initialized using CLIP¹, WavLM², and RoBERTa³,

¹<https://huggingface.co/openai/clip-vit-large-patch14>

²<https://huggingface.co/microsoft/wavlm-large>

³<https://huggingface.co/FacebookAI/roberta-base>

producing feature representations of 768, 1,024, and 768 dimensions, respectively. A linear layer mapped speech features to 768 dimensions. During training, CLIP and RoBERTa were kept frozen, while WavLM was fine-tuned. The speech modality input consisted of 6 seconds of audio at a 16 kHz sampling rate. Given the presence of two speakers in the video, speaker separation was required. Ultimately, the visual modality input comprised 180 individual images. The feature extractor comprised a single-layer transformer with a hidden size of 768. We optimized using Adam [20] with a batch size of 32, a fixed learning rate of $1e^{-5}$, and γ set to 0.1 in the loss function.

Dataset. To demonstrate the effectiveness of our proposed method, we conducted experiments on the IEMOCAP dataset [21]. This dataset comprised roughly 12 hours of audio, video, transcriptions, and motion-capture information from ten speakers in five scripted sessions. Following prior work, we employed 5,531 utterances from four emotion categories: “neutral”, “angry”, “happy”, and “sad”, with “excited” merged into the “happy” category. We conducted experiments with 5-fold leave-one-session-out cross-validation.

3.2. Results and Analysis

Comparisons of the SOTA Methods. Table 1 compares the performance of the GIA-MIC method with recent multimodal SER approaches on IEMOCAP. Our proposed GIA-MIC achieves a weighted accuracy (WA) of 80.7% and an unweighted accuracy (UA) of 81.1% on IEMOCAP, outperforming other SOTA methods. We used Whisper [22] to generate ASR transcripts, achieving a word error rate (WER) of 20.48% on the IEMOCAP dataset. Even with ASR transcripts, GIA-MIC outperforms other methods, further demonstrating its effectiveness.

Table 1: *Performance Comparison of SOTA Models on IEMOCAP (%)*. “S”, “T” and “V” represent speech, text and visual modalities, respectively. “ASR” and “GT” denote ASR and ground truth transcripts, respectively. The best and second-best results are marked in **bold** and underlined, respectively.

Method	Year	Modality	WA	UA
RMSER-AEA [23]	2023	S+T(ASR)	76.4	76.9
MGAT [24]	2023	S+T(GT)	78.5	79.3
IMISA [25]	2024	S+T(GT)	77.4	77.9
MFLA [26]	2024	S+T(ASR)	-	77.7
MF-AED-AEC [27]	2024	S+T(ASR)	78.1	79.3
MAF-DCT [28]	2024	S+T(GT)	78.5	79.3
FDRL [29]	2024	S+T(GT)	78.3	79.4
CAT-BC [30]	2025	S+T(GT)	<u>79.5</u>	<u>80.3</u>
MCWSA-CMHA [31]	2022	V+S	78.9	-
GCNet [32]	2023	V+S	78.4	-
Foal-Net [33]	2024	V+S	79.5	80.1
S2MER-CMDM [34]	2020	V+S+T(GT)	75.6	74.5
AMED [35]	2022	V+S+T(GT)	-	77.6
FM-MER [36]	2023	V+S+T(GT)	-	78.9
GIA-MIC (Ours)	2025	V+S+T(ASR)	79.6	80.3
GIA-MIC (Ours)	2025	V+S+T(GT)	80.7	81.3

Ablation Studies. We conduct ablation studies to evaluate the contributions of the MSR, MIR, and MIC modules in our GIA-MIC framework. Results in Table 2 show that removing any component causes a performance drop, emphasizing their role in enhancing MER. Among the three, MSR has the greatest impact, with its removal resulting in a 1.1% WA and 0.9% UA decrease, highlighting the importance of MSR information for emotion recognition. The MIR module also proves vital, with

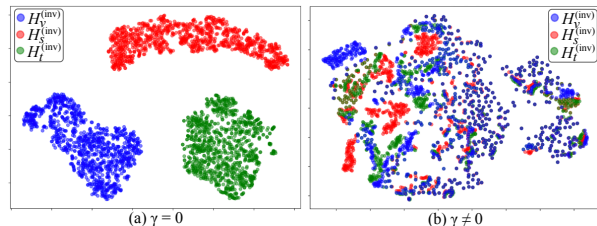


Figure 2: *t-SNE visualizations of the distribution of modality-invariant representations before and after introducing the modality-invariant learning constraints.*

a 0.6% WA drop when removed, suggesting its role in cross-modal alignment. Although MIC has the smallest effect, removing it leads to a 0.5% WA decrease, indicating its contribution to refining learned representations through consistency across modalities. Overall, the full GIA-MIC model outperforms the “Baseline” approach, confirming that combining MSR, MIR, and MIC is crucial for effective multimodal fusion.

Table 2: *The results of ablation experiments on IEMOCAP (%)*.

Modality	Method	WA	UA
V	CLIP	46.6	47.9
S	WavLM	71.7	72.4
T(GT)	RoBERTa	68.4	69.4
V+S	Baseline	72.3	73.1
S+T(GT)	Baseline	76.4	77.2
V+T(GT)	Baseline	69.0	69.9
V+S	GIA-MIC (ours)	79.4	80.1
S+T(GT)	GIA-MIC (ours)	79.0	80.4
V+T(GT)	GIA-MIC (ours)	71.3	72.0
V+S+T(GT)	GIA-MIC (ours)	80.7	81.3
w/o MSR	GIA-MIC (ours)	79.6	80.4
w/o MIR	GIA-MIC (ours)	80.1	80.7
w/o MIC	GIA-MIC (ours)	80.2	80.9

Visualization Analysis. We employ t-SNE [37] to visualize the multimodal modality-invariant representations, as shown in Fig. 2. When $\gamma = 0$, meaning no modality-invariant constraints are applied, the representations of different modalities remain distinct. In contrast, with temporal modality-invariant constraints ($\gamma \neq 0$), the three modalities exhibit greater overlap, indicating increased shared information. This suggests that the constraints effectively enhance modality alignment, leading to more similar representations after training.

4. Conclusion

In this paper, we propose GIA-MIC, a novel multimodal emotion recognition framework that effectively integrates modality-invariant and modality-specific representations while enforcing cross-modal consistency through modality-invariant constraints. Our approach addresses the challenges of modality heterogeneity and misalignment by learning both shared and distinct representations, enabling a more robust fusion of visual, speech, and text modalities. Extensive experiments on the IEMOCAP dataset demonstrate that GIA-MIC significantly outperforms baseline methods, achieving SOTA performance.

5. Acknowledgements

This work was partly supported by JST AIP Acceleration Research JPMJCR25U5 and JSPS KAKENHI Grant Number 21H05054, Japan.

6. References

- [1] J. Tian, D. Hu, X. Shi, J. He, X. Li, Y. Gao, T. Toda, X. Xu, and X. Hu, "Semi-supervised multimodal emotion recognition with consensus decision-making and label correction," in *Proc. MRAC*, 2023, pp. 67–73.
- [2] P. Liu, K. Li, and H. Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," in *Proc. Interspeech*, 2020, pp. 379–383.
- [3] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Systems with Applications*, vol. 173, p. 114683, 2021.
- [4] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *Proc. ICASSP*, 2021, pp. 6304–6308.
- [5] J. Mi, X. Shi, D. Ma, J. He, T. Fujimura, and T. Toda, "Two-stage framework for robust speech emotion recognition using target speaker extraction in human speech noise conditions," in *Proc. APSIPA ASC*, 2024.
- [6] Y. Wu, P. Yue, L. Qu, T. Li, and Y.-P. Ruan, "Multi-modal emotion recognition using multiple acoustic features and dual cross-modal transformer," in *Proc. ICASSP*, 2024, pp. 10 496–10 500.
- [7] P. Waligora, M. H. Aslam, M. O. Zeeshan, S. Belharbi, A. L. Koerich, M. Pedersoli, S. Bacon, and E. Granger, "Joint multi-modal transformer for emotion recognition in the wild," in *Proc. CVPRW*, 2024, pp. 4625–4635.
- [8] Z. Yang, J. He, and T. Toda, "Multi-modal video summarization based on two-stage fusion of audio, visual, and recognized text information," in *Proc. APSIPA ASC*, 2024, pp. 1–6.
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.
- [10] J. He and T. Toda, "2DP-2MRC: 2-dimensional pointer-based machine reading comprehension method for multimodal moment retrieval," in *Proc. Interspeech*, 2024, pp. 5073–5077.
- [11] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," in *Proc. ICLR*, 2024.
- [12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [13] 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, M. Zeng, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.
- [14] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," in *Proc. ICLR*, 2021.
- [15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: a robustly optimized bert pretraining approach," in *Proc. ICLR*, 2020.
- [16] Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao, "Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework," *Speech Communication*, vol. 139, pp. 1–9, 2022.
- [17] R. G. Praveen and J. Alam, "Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition," in *Proc. CVPRW*, 2024, pp. 4803–4813.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.
- [19] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.
- [20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 7–9.
- [21] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, 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.
- [22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.
- [23] B. Lin and L. Wang, "Robust multi-modal speech emotion recognition with ASR error adaptation," in *Proc. ICASSP*, 2023, pp. 1–5.
- [24] W. Fan, X. Xing, B. Cai, and X. Xu, "MGAT: Multi-granularity attention based transformers for multi-modal emotion recognition," in *Proc. ICASSP*, 2023, pp. 1–5.
- [25] Y. Wang, D. Li, and J. Shen, "Inter-modality and intra-sample alignment for multi-modal emotion recognition," in *Proc. ICASSP*, 2024, pp. 8301–8305.
- [26] X. Shi, Y. Gao, J. He, J. Mi, X. Li, and T. Toda, "A study on multimodal fusion and layer adapter in emotion recognition," in *Proc. APSIPA ASC*, 2024, pp. 1–6.
- [27] J. He, X. Shi, X. Li, and T. Toda, "MF-AED-AEC: Speech emotion recognition by leveraging multimodal fusion, ASR error detection, and ASR error correction," in *Proc. ICASSP*, 2024, pp. 11 066–11 070.
- [28] Y. Wu, P. Yue, L. Qu, T. Li, and Y.-P. Ruan, "Multi-modal emotion recognition using multiple acoustic features and dual cross-modal transformer," in *Proc. ICASSP*, 2024, pp. 10 496–10 500.
- [29] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," in *Proc. ICASSP*, 2024, pp. 11 051–11 055.
- [30] W. Fan, X. Xu, G. Zhou, X. Deng, and X. Xing, "Coordination attention based transformers with bidirectional contrastive loss for multimodal speech emotion recognition," *Speech Communication*, pp. 103 198–103 207, 2025.
- [31] J. Zheng, S. Zhang, Z. Wang, X. Wang, and Z. Zeng, "Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 2213–2225, 2022.
- [32] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "GCNet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 45, no. 7, pp. 8419–8432, 2023.
- [33] Q. Li, Y. Gao, Y. Wen, C. Wang, and Y. Li, "Enhancing modal fusion by alignment and label matching for multimodal emotion recognition," in *Proc. Interspeech*, 2024, pp. 4663–4667.
- [34] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proc. ACM-MM*, 2020, pp. 2852–2861.
- [35] J. Heredia, E. Lopes-Silva, Y. Cardinale, J. Diaz-Amado, I. Dongo, W. Graterol, and A. Aguilera, "Adaptive multimodal emotion detection architecture for social robots," *IEEE Access*, vol. 10, pp. 20 727–20 744, 2022.
- [36] D. Peña, A. Aguilera, I. Dongo, J. Heredia, and Y. Cardinale, "A framework to evaluate fusion methods for multimodal emotion recognition," *IEEE Access*, vol. 11, pp. 10 218–10 237, 2023.
- [37] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.