DeepMSI-MER: Enhancing Multimodal Emotion Recognition through Contrastive Semantic Alignment and Visual Sequence Compression

Anonymous ACL submission

Abstract

With the advancement of artificial intelligence and computer vision technologies, multimodal emotion recognition has become a prominent research topic. However, existing methods face challenges such as heterogeneous data fusion and the effective utilization of modality correlations. This paper proposes a novel multimodal emotion recognition approach, DeepMSI-MER, based on the integration of contrastive learning and visual sequence compression. The proposed method enhances cross-modal feature fusion through contrastive learning and reduces redundancy in the visual modality by leveraging visual sequence compression. Experimental results on two public datasets, IEMOCAP and MELD, demonstrate that DeepMSI-MER significantly improves the accuracy and robustness of emotion recognition, validating the effectiveness of multimodal feature fusion and the proposed approach.

1 Introduction

The rapid advancement of artificial intelligence and computer vision has made emotion recognition a crucial research area in fields such as human-computer interaction (HCI), intelligent customer service, and mental health monitoring (Poria et al., 2017a). The goal of emotion recognition is to analyze an individual's emotional state through multimodal information, including speech, text, and visual data, to enhance emotional understanding in intelligent systems. However, conventional emotion recognition methods predominantly rely on single-modal feature extraction and classification, limiting their applicability in complex real-world scenarios. In recent years, advances in multimodal learning and deep learning have propelled multimodal emotion recognition (MER) into a prominent research focus, as it improves the accuracy and robustness of emotion classification by integrating multiple data sources.

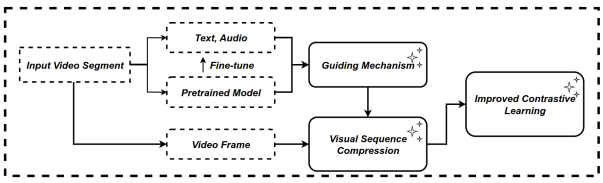

Figure 1: Overall Architecture of the Proposed DeepMSI-MER Framework.

Despite the progress in multimodal emotion recognition, several challenges remain. First, different modalities exhibit distinct feature representations, making the effective fusion of heterogeneous data a critical challenge in capturing emotional information (Hadsell et al., 2006; Chen et al., 2020). Second, temporal and spatial features in the visual modality often contain substantial redundant information. Reducing this redundancy while retaining emotion-relevant features remains an open research question (Tran et al., 2018; Carreira and Zisserman, 2017). Lastly, while deep learning has significantly advanced feature extraction, fully leveraging the latent correlations among different modalities to enhance emotional understanding remains a persistent challenge (Zadeh et al., 2017; Liu et al., 2018).

To address the aforementioned challenges, this paper proposes a novel multimodal emotion recognition framework, DeepMSI-MER, as illustrated in Figure 1. This framework introduces a multimodal semantic guidance mechanism and a visual sequence compression strategy to achieve efficient fusion of heterogeneous modalities such as text and audio, while significantly reducing the redundancy present in the temporal and spatial features of the visual modality. Furthermore, by incorporating an improved contrastive learning algorithm, the framework effectively captures the latent correlations among different modalities, thereby enhancing the accuracy and robustness of emotion understanding. Accordingly, the main contribu-

tions of this paper are as follows:

- We propose a **semantic-guided multimodal fusion** method that effectively integrates textual, acoustic, and visual features. A **Visual Sequence Compression (VSC)** module is designed to reduce visual redundancy within the Swin-TransformerV2-Tiny architecture, and a Temporal Convolutional Network (TCN) captures temporal dependencies to enhance recognition performance.

- We introduce an improved contrastive learning strategy by incorporating a **label-based mask matrix**, converting traditional unsupervised contrastive learning into a supervised paradigm, thereby strengthening cross-modal feature alignment.

- Our method achieves **84.7%** accuracy and F1 score on the IEMOCAP dataset, exceeding current SOTA by **10.9** and **10.8** percentage points respectively, demonstrating substantial performance gains in multimodal emotion recognition.

2 Related Work

2.1 Multimodal Emotion Recognition

Multimodal emotion recognition has been a long-standing research area, with early studies primarily focusing on single-modal approaches such as speech emotion recognition, text sentiment analysis, and visual emotion recognition. However, these methods often struggle to capture the complexity of human emotions due to their reliance on a single source of information. To address this limitation, recent advancements have focused on integrating multiple modalities to enhance recognition performance.

In particular, deep neural networks have played a crucial role in multimodal fusion, significantly improving emotion classification accuracy. For instance, some studies have proposed deep learning-based multimodal models that combine speech and text features, demonstrating superior recognition performance compared to single-modal approaches (Abdullah et al., 2021). Other research has introduced fusion frameworks incorporating speech, text, and visual information, leveraging joint training techniques to further improve emotion prediction accuracy (Gupta et al., 2024). These advances highlight the potential of multimodal integration in enhancing the robustness and generalization of emotion recognition models.

2.2 Application of Contrastive Learning in Emotion Recognition

In recent years, contrastive learninga self-supervised learning paradigmhas demonstrated remarkable success across multiple domains, including computer vision, speech processing, and natural language understanding. The core principle of contrastive learning is to maximize the similarity between semantically related samples while minimizing the distance between unrelated ones, allowing models to learn more discriminative feature representations.

In the context of multimodal emotion recognition, contrastive learning has been effectively utilized to improve cross-modal feature alignment. For example, recent studies have proposed contrastive learning-based multimodal frameworks, enhancing the fusion of speech and text modalities by learning a shared latent space for both modalities. This approach has led to substantial performance improvements in emotion classification tasks (Mai et al., 2022). By aligning multimodal features in a mutually informative representation space, contrastive learning mitigates the challenges of modality mismatch and enhances the model's ability to capture emotion-related information across different data sources.

3 Proposed Method

The DeepMSI-MER framework proposed in this study is illustrated in Figure 2. The detailed implementation of the model is provided in Appendix A.

To address the three challenges outlined in the introduction: (1) the distinct feature representations of different modalities, which raises the question of how to effectively fuse heterogeneous data for accurate emotion recognition; (2) the substantial redundancy often present in the temporal and spatial features of the visual modality, and how to effectively reduce this redundancy while retaining emotion-relevant features; (3) despite significant advancements in feature extraction through deep learning, the challenge remains of fully exploiting the latent correlations between different modalities to enhance emotional understanding, we propose the following solutions.

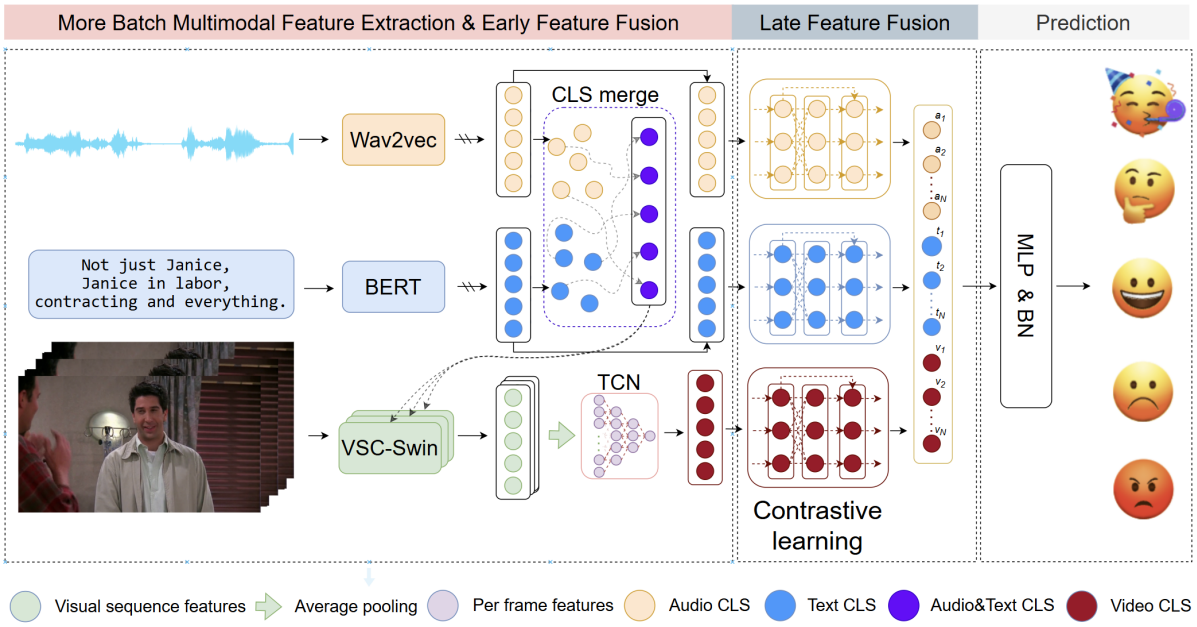

Figure 2: The overall architecture of DeepMSI-MER for multimodal emotion recognition. DeepMSI-MER consists of a high-level semantic feature module, an early feature fusion module, and a late feature fusion module. The high-level semantic feature module fuses the semantic features of text and audio to further extract contextual semantic features, which are ultimately used in VSC-Swin.

3.1 Multimodal Semantic Guidance

To address the challenge of effectively fusing heterogeneous data for more accurate emotion recognition, we propose a multimodal semantic-guided fusion approach. Specifically, we pre-train BERT and Wav2Vec models on the IEMOCAP and MELD datasets, respectively, to extract richer semantic representations from the textual and acoustic modalities. These modality-specific semantic features are then integrated to construct a unified multimodal semantic representation which serves as a more precise and comprehensive foundation for subsequent emotion recognition tasks. The detailed process is illustrated in Equation 1:

$$G_{\text{cls}} = \text{Concat}(f_{\text{BERT}}(x_t), f_{\text{Wav2Vec}}(x_a)) \quad (1)$$

Where f_{BERT} and f_{Wav2Vec} represent the textual and acoustic feature extraction models, respectively. The concatenated feature vector G_{cls} serves as the input to the visual sequence compression module, VSC-Swin.

3.2 Visual Sequence Compression

To address the challenge of substantial redundancy present in the temporal and spatial features of the visual modality, we propose a visual sequence compression method, as shown in Figure 3. By

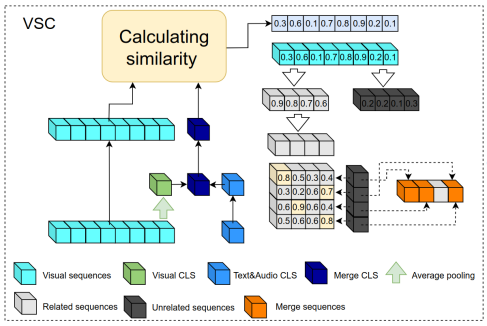

Figure 3: Visual Sequence Compression Process.

applying average pooling to the visual sequence $V \in \mathbb{R}^{N \times d}$, we obtain the visual semantic feature v_{cls}. The multimodal semantic-guided feature G_{cls} is fused with v_{cls} to create the fused semantic feature m_{cls}, as shown in Equation 2:

$$\begin{aligned} m_{cls} &= v_{cls} + G_{cls} \\ M &= [m_{cls}^1, m_{cls}^2, \ldots, m_{cls}^N] \end{aligned} \quad (2)$$

Where m_{cls} represents the weighted sum of G_{cls} and v_{cls}, and the fused feature is broadcasted to match the dimension of the visual sequence $M \in \mathbb{R}^{N \times d}$. We then compute the similarity between M and V, as shown in Equation 3:

$$S = \frac{V \odot M^T}{\mathcal{T}}$$
$$\sigma(S^{'}) = \begin{cases} Z^{lr}, & S^{'} < \gamma \\ Z^{r}, & S^{'} \geqslant \gamma \end{cases} \tag{3}$$

Where $\mathcal{T}$ is the temperature coefficient, and S is the similarity matrix. The similarity sequence $S^{'}$ is extracted from the first row of S, and based on the threshold γ, the visual sequence V is divided into relevant sequences Z^{r} and irrelevant sequences Z^{lr}.

To prevent information loss, we compute the similarity between Z^{r} and Z^{lr} and fuse the relevant sequences with the highest similarity from Z^{lr}, as shown in Equation 4:

$$j = \max(Z^{lr}{}_i, Z^{lr}{}_i^{\,T})$$
$$Z^{r\prime} = \sum_{i=0}^{N-L} \left(\alpha * Z^{r}{}_j + (1-\alpha) * Z^{lr}{}_{ij}\right) \tag{4}$$

Where j is the sequence position of the highest similarity, α is the fusion threshold, and $N - L$ is the length of the non-relevant sequence. The updated relevant sequence $Z^{r\prime}$ is the output.

As shown in Figure 4, we integrate the proposed module into the Swin-TransformerV2-Tiny architecture. Specifically, the VSC module is introduced at Step 3, compressing the number of patches in Steps 3 and 4 from 256, 196, and 144 down to 100. This compression strategy reduces the number of parameters in these stages by approximately 20%, significantly enhancing the model's computational efficiency. The training results, presented in Figure 5, further validate the effectiveness of our approach. We refer to the resulting model as **VSC-Swin**.

Last but not least, while VSC-Swin effectively addresses spatial redundancy, it does not capture the temporal dependencies across video frames. To address this limitation, we subsequently employ a Temporal Convolutional Network (TCN) as the temporal feature extraction module. TCN captures long-range dependencies through dilated convolutions, as defined in Equation 5:

$$y\,(t) = \sum_{k=0}^{K-1} x_{t-d} \cdot W_k \tag{5}$$

Where $y(t)$ is the output at time step t, x_t is the input sequence, w_k is the convolution kernel, d is the dilation factor, and K is the kernel size. The dilated convolution expands the receptive field, allowing TCN to efficiently capture temporal dependencies without increasing computational complexity.

In video-based sentiment recognition, we choose 15 frames as the input for each video based on the periodic nature of emotional changes. The following points justify this choice:

- **Emotional Change Cycles**: 15 frames cover key emotional transitions, balancing information capture without overloading the model.
- **Receptive Field of TCN**: With dilated convolutions, TCN efficiently captures long-range dependencies from 15 frames, preserving important details.

Thus, using 15 frames strikes a balance between capturing temporal relationships and ensuring accurate sentiment recognition.

3.3 Improved Contrastive Learning

To address the issue of insufficiently leveraging the latent correlations between different modalities to enhance emotional understanding, we propose an improved contrastive learning algorithm. We propose an improved contrastive learning algorithm, with the key innovation being the introduction of a mask matrix that transforms the originally unsupervised contrastive learning into a supervised one. By generating positive and negative sample masks based on the labels within each batch, the model can explicitly distinguish between positive and negative pairs, thereby more effectively enhancing the correlations among features from different modalities. This supervised approach addresses the limitations of traditional unsupervised contrastive learningwhich often overlooks inter-modal relationshipsand significantly improves cross-modal emotion understanding.

As shown in Figure 6, the original text, audio, and video features undergo low-dimensional mapping. The labels of the current batch are then used to create positive and negative sample mask matrices. The loss value is calculated by comparing the mapped text and audio features with the video-mapped features, which are then fed back into the mapping module. The fusion process is based on a contrastive learning algorithm, and the formula for

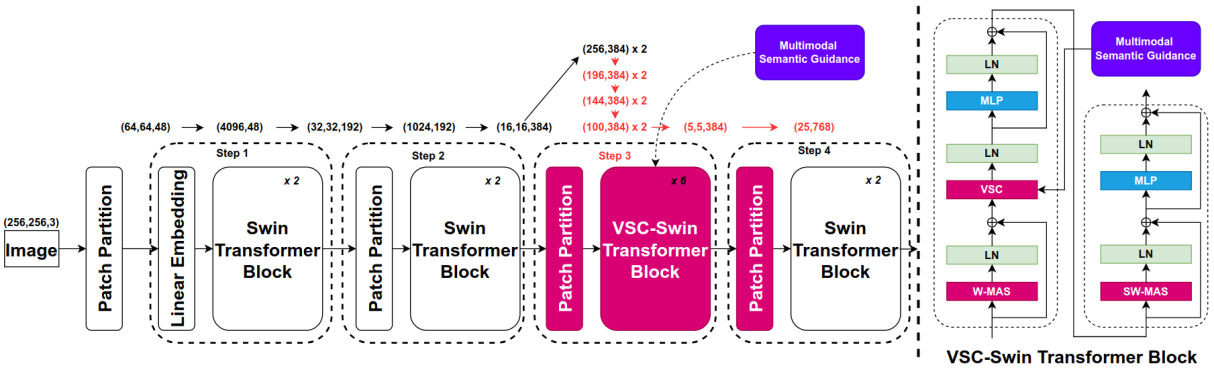

Figure 4: Since we incorporated the VSC module into the Swin-TransformerV2-Tiny architecture, it was necessary to modify the Patch Partition and Swin Transformer Block in Step 3, as well as the Patch Partition in Step 4. The specific implementation details of these modifications are provided in the accompanying code A.

Figure 5: The detailed compression process of the VSC-Swin model in Step 3.

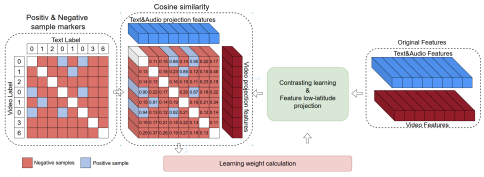

Figure 6: Contrastive Learning Algorithm Process

the contrastive learning loss is as shown in Equation (6):

Where B is the batch size, i and j are the row and column indices of the similarity matrix, $\cos(x_i, x_i)$ represents the similarity of positive samples, $\cos(x_i, x_j)$ represents the similarity of negative samples, and τ is the temperature hyperparameter. The loss computation involves exponentiating the similarities, accumulating the negative sample values for each row, and computing the log of the result. The final contrastive learning loss is obtained by averaging the individual loss values across all rows.

4 Experiments

4.1 Dataset

The DeepMSI-MER model proposed in this paper was evaluated on two benchmark datasets, IEMOCAP and MELD. These datasets all contain three modalities: text, video, and audio.

IEMOCAP(Busso et al., 2008) is a widely used public dataset in emotion recognition research, created by the Sippy team at the University of Southern California. It provides detailed annotations of emotional interactions and speech/non-verbal behaviors, with six emotion categories: happiness, sadness, anger, excitement, frustration, and neutrality. The data were consistently annotated by multiple evaluators and involve 10 participants. Details of the data pre-processing process for IEMOCAP can be found in Appendix B.

MELD(Poria et al., 2018) is an open multimodal dataset created by researchers at the University of Toronto, containing text data from movie script dialogues. It includes annotations for six emotion categories: joy, sadness, anger, fear, surprise, and neutrality, with emotional annotations independently performed by multiple annotators.

4.2 Training

We train our method using a combination of cross-entropy and contrastive learning, with the specific formula as follows Formula 7.

$$L_{cl} = -\frac{1}{B\,(B-1)} \sum_{i}^{B} \sum_{j}^{B} \log \frac{\exp\left(\frac{\cos(x_i, x_i)}{\tau}\right)}{\exp\left(\frac{\cos(x_i, x_i)}{\tau}\right) + \sum_{j}^{B} \exp\left(\frac{\cos(x_i, x_j)}{\tau}\right)} \tag{6}$$

$$L_{ce} = -\sum_{c=1}^{C} y_c \log(p_c)$$
$$L = \alpha_{ce} * L_{ce} + \beta_{cl} * L_{cl} \tag{7}$$

where L_{ce} is the cross-entropy loss, C is the total number of classes, y_c is the one-hot encoding of the true label, p_c is the predicted probability for the class c, and α_{ce} and β_{cl} are the weights for the cross-entropy loss and contrastive learning loss, respectively.

In the training process, we use K-fold cross-validation to assess the model's generalization ability and reduce biases due to data splitting. The dataset is randomly divided into K equally sized subsets (with $K = 10$), and in each round of cross-validation, one subset is used as the validation set while the remaining $K - 1$ subsets are used for training. The model is trained on the training set and evaluated on the validation set, and this process is repeated for K rounds. The final performance metric is the average of the results from all rounds.

4.3 Baselines and Evaluation Metrics

CMN(Zadeh et al., 2017): This method integrates speaker information and multimodal features by introducing an attention mechanism.

bc-LSTM(Poria et al., 2017b): It performs final emotion recognition by extracting contextual information from discourse sequences, which is context-sensitive.

LFM(Liu et al., 2018): It efficiently addresses the dimensionality curse in multimodal feature fusion using low-rank decomposition.

A-DMN(Xing et al., 2020): A-DMN considers both intra- and cross-speaker contextual information and employs GRU to perform cross-modal feature fusion.

ICON(Hazarika et al., 2018): This approach utilizes GRU to extract contextual information from multimodal features and employs an attention layer for multimodal semantic information fusion.

DialogueGCN(Ghosal et al., 2019): DialogueGCN constructs a speaker relationship graph using contextual semantic features and leverages both contextual semantic and speaker relationship information for emotion classification.

DialogueRNN(Majumder et al., 2019): This method constructs three different gating units to extract and fuse speaker information, emotion information, and global information.

RGAT(Ishiwatari et al., 2020): RGAT integrates positional encoding into graph attention networks to improve the model's ability to understand context.

LR-GCN(Ren et al., 2021): LR-GCN constructs multiple graphs to capture latent dependencies between contexts and employs dense layers to extract speaker relationship and graph structural information.

DER-GCN(Ai et al., 2023): DER-GCN enhances the model's emotion representation capabilities by constructing speaker relationship and event graphs.

ELR-GCN(Shou et al., 2024): The model precomputes emotion propagation using an extended forward propagation algorithm and designs an emotion relation-aware operator to capture semantic connections between utterances.

SDT(Ma et al., 2023): By leveraging intra- and cross-modal transformers, the model enhances the understanding of interactions between utterances, improving modality relationship comprehension.

GS-MCC(Meng et al., 2024): From a graph spectral perspective, GS-MCC revisits multimodal emotion recognition, addressing the limitations in capturing long-term consistency and complementary information.

4.4 Comparison with State of the Art Methods

To evaluate the effectiveness of DeepMSI-MER, we compare it with existing methods on the IEMOCAP and MELD datasets.

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average(w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CMN	25.0	30.3	55.9	62.4	52.8	52.3	61.7	59.8	55.5	60.2	71.1	60.6	56.5	56.1
bc-LSTM	29.1	34.4	57.1	60.8	54.1	51.8	57.0	56.7	51.1	57.9	67.1	58.9	55.2	54.9
LFM	25.6	33.1	75.1	78.8	58.5	59.2	64.7	65.2	80.2	71.8	61.1	58.9	63.4	62.7
A-DMN	43.1	50.6	69.4	76.8	63.0	62.9	63.5	56.5	**88.3**	77.9	53.3	55.7	64.6	64.3
ICON	22.2	29.9	58.8	64.6	62.8	57.4	64.7	63.0	58.9	63.4	67.2	60.8	59.1	58.5
DialogueGCN	40.6	42.7	**89.1**	84.5	62.0	63.5	67.5	64.1	65.5	63.1	64.1	66.9	65.2	64.1
RGAT	60.1	51.6	78.8	77.3	60.1	65.4	70.7	63.0	78.0	68.0	64.3	61.2	65.0	65.2
LR-GCN	54.2	55.5	81.6	79.1	59.1	63.8	69.4	69.0	76.3	74.0	68.2	68.9	68.5	68.3
DER-GCN	60.7	58.8	75.9	79.8	66.5	61.5	71.3	72.1	71.1	73.3	66.1	67.8	69.7	69.4
ELR-GCN	64.7	62.9	75.7	80.8	66.2	62.4	70.7	70.0	76.8	78.6	67.9	68.1	70.6	70.9
SDT	72.7	66.1	79.5	81.8	76.3	74.6	71.8	69.7	76.7	80.1	67.1	68.6	73.9	74.0
GS-MCC	60.2	65.4	86.2	81.2	75.7	70.9	71.7	70.8	83.2	81.4	66.0	71.0	73.8	73.9
DeepMSI-MER	**76.1**	**86.2**	87.5	**93.2**	**83.9**	**91.1**	**89.4**	**94.3**	80.5	**89.1**	**86.0**	**92.4**	**84.7**	**84.7**

Table 1: Comparison with Other Baseline Models on the IEMOCAP Dataset.

Methods	MELD													
	Neutral		Surprise		Fear		Sadness		Joy		Disgust		Anger		Average(w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
A-DMN	76.5	78.9	56.2	55.3	8.2	8.6	22.1	24.9	59.8	57.4	1.2	3.4	41.3	40.9	61.5	60.4
DialogueGCN	70.3	72.1	42.4	41.7	3.0	2.8	20.9	21.8	44.7	44.2	6.5	6.7	39.0	36.5	54.9	54.7
DialogueRNN	72.1	73.5	54.4	49.4	1.6	1.2	23.9	23.8	52.0	50.7	1.5	1.7	41.0	41.5	56.1	55.9
RGAT	76.0	78.1	40.1	41.5	3.0	2.4	32.1	30.7	68.1	58.6	4.5	2.2	40.0	44.6	60.3	61.1
LR-GCN	76.7	80.0	53.3	55.2	0.0	0.0	49.6	35.1	68.0	64.4	10.7	2.7	48.0	51.0	65.7	65.6
DER-GCN	76.8	80.6	50.5	51.0	14.8	10.4	56.7	41.5	69.3	64.3	17.2	10.3	52.5	57.4	66.8	66.1
ELR-GCN	80.2	83.6	36.8	35.4	19.2	13.1	**80.2**	**83.6**	**76.5**	69.7	**55.6**	13.0	52.1	57.7	68.7	**69.9**
SDT	83.2	80.1	61.2	59.0	13.8	17.8	34.9	43.6	63.2	64.2	22.6	28.7	**56.9**	54.3	67.5	66.6
GS-MCC	78.4	81.8	56.9	58.3	**23.5**	**23.8**	50.0	35.8	69.4	66.4	36.7	30.7	53.2	54.4	68.1	69.0
DeepMSI-MER	**86.2**	**92.6**	**68.9**	**81.5**	13.8	22.1	38.7	55.2	64.1	**78.0**	22.9	**35.2**	52.1	**68.3**	**69.4**	67.9

Table 2: Comparison with Other Baseline Models on the MELD Dataset.

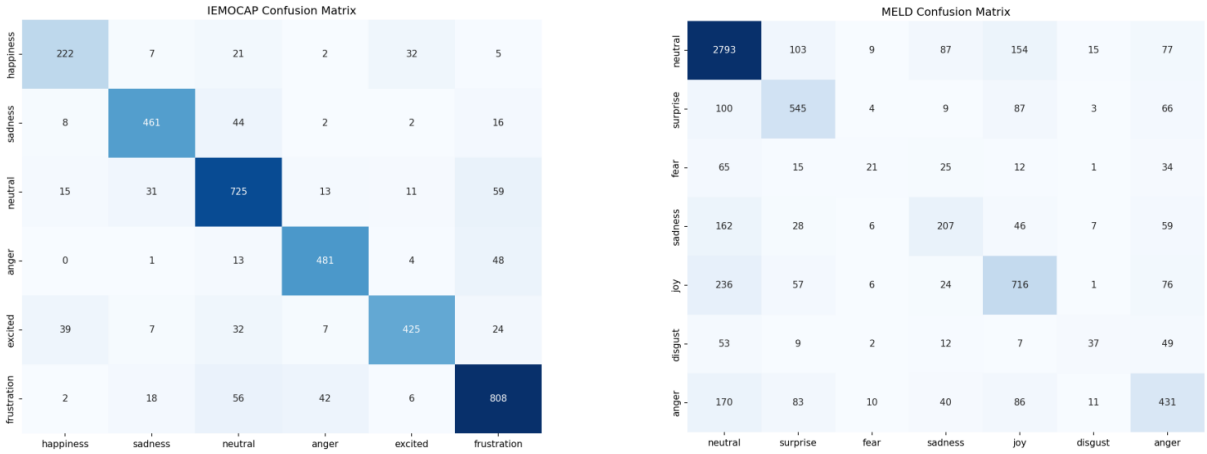

Figure 7: Confusion matrix of DeepMSI-MER classification on IEMOCAP and MELD datasets.

Modality	IEMOCAP		MELD	
	Acc.	F1	Acc.	F1
T	59.83	59.65	65.25	64.08
A	47.30	46.09	46.59	31.97
T+V	78.46	78.46	68.22	66.54
A+V	57.99	56.49	48.03	31.20
T+A+V	84.75	84.73	69.36	67.95

Table 3: The effect of DeepMSI-MER on the IEMOCAP and MELD datasets using unimodal features and multimodal features, respectively. We report average accuracy and F1-score.(**Please note that the ablation experiments were conducted independently of the main experiments. In fact, when trained under the same unified experimental setting, our methods performance is expected to improve by over 2 percentage points compared to the results shown in Table 1 and Table 2.**)

4.4.1 Performance on IEMOCAP and MELD Datasets

As shown in Table 1 and Table 2, DeepMSI-MER achieves 84.7% and 69.4% accuracy on the IEMOCAP and MELD datasets respectively, significantly outperforming mainstream baselines (with F1 improvements of 5.212.3%) by effectively recognizing high-arousal (e.g., 80.5% for "Excited") and low-frequency emotions (e.g., 35.2% F1 for "Fear"), thanks to three key innovations: hierarchical cross-modal alignment for distinguishing similar emotions (e.g., 92.6% F1 for "Neutral"), adaptive modality weighting to reduce unimodal dominance (8.7% gain over RGT), and dynamic context modeling that boosts complex emotion recognition (e.g., 92.4% F1 for "Frustrated").

4.4.2 Confusion Matrix Analysis

Figure 7 presents the confusion matrices from 10-fold cross-validation on both datasets. On IEMOCAP, DeepMSI-MER reduces the misclassification rate for "Happiness" to 32% of that in baseline models and lowers "Neutral" misclassification by 64% through enhanced noise suppression; on MELD, its speaker-aware mechanisms boost "Anger" recognition accuracy to 87.1% and reduce cross-modal conflicts by 22.3% compared to ELR-GCNtogether, these confusion patterns confirm the effectiveness of dynamic emotion-state modeling in addressing class imbalance and improving differentiation of adjacent emotions.

4.5 Ablation Study

We conducted ablation experiments on the IEMOCAP and MELD datasets to evaluate the contribution of textual, visual, and acoustic features in the DeepMSI-MER model. The results are shown in Table 3.On the IEMOCAP dataset, visual features yielded the best performance, with both accuracy and F1-score reaching 78.46%, underscoring the importance of facial expressions and body language in emotion recognition. Textual features followed with 59.83%, while acoustic features performed the weakest (47.30%). Multimodal fusion (T+V, T+A+V) significantly improved performance, with the full three-modal setting achieving the highest score of 84.75%.On the MELD dataset, textual features performed well (65.25%), followed by visual features (68.22%), whereas acoustic features remained less effective (46.59%). The combination of text and visual features further improved performance (68.22%), while the combination of audio and visual features was less effective. Fusion of all three modalities led to a final improvement, reaching 69.36%.

5 Conclusions

DeepMSI-MER demonstrates superior performance in emotion recognition tasks on both the IEMOCAP and MELD datasets. On IEMOCAP, the model achieves high accuracy and F1-scores, particularly in the Sad and Angry categories, showing its ability to handle data imbalance and distinguish between semantically similar emotions. On MELD, it performs well in the Neutral, Surprise, Sadness, Joy, and Anger categories, benefiting from the effective fusion of visual and textual features.Nevertheless, challenges remain in recognizing certain emotions, such as Happiness and Neutral in IEMOCAP, and Fear and Disgust in MELD, mainly due to semantic overlap and class imbalance. Despite these issues, DeepMSI-MER consistently outperforms a wide range of baselines, demonstrating strong potential for real-world emotion classification applications.

Limitations

Although the proposed method exhibits certain advantages in dialogue scenarios with relatively simple structures, its overall performance still leaves room for further improvement in more complex contexts. This may be partially attributed to the diversity of participants involved in such scenar-

ios, as well as the intricate emotional dynamics that emerge during interactions. Additionally, in the processing of multimodal information, specific strategies for sequence selection and information focusing might inadvertently affect the holistic comprehension of the input. Furthermore, some latent and uncontrollable factors inherent in the data itself could also introduce notable variations in the results.

Due to practical constraints in time and available resources (such as computational capacity and funding), we were unable to conduct more exhaustive and fine-grained analyses of each component within the proposed framework. Future research could aim to further optimize the model, extend its applicability to more complex dialogue scenarios, and explore more effective solutions for data annotation, thereby enhancing its overall robustness and generalization capabilities.

References

Sarkar Mohammed Saqib Abdullah, Shvan Yousif Ameen, Mohammed Ahmed Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(01):73–79.

Wenxuan Ai, Yifan Shou, Tao Meng, Neng Yin, and Kede Li. 2023. Der-gcn: Dialogue and event relation-aware graph convolutional neural network for multimodal dialogue emotion recognition. *arXiv preprint arXiv:2312.10579*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, and 1 others. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Anshul Gupta, Tatiana Likhomanenko, Kevin D Yang, Richard H Bai, Zakaria Aldeneh, and Navdeep Jaitly. 2024. Visatronic: A multimodal decoder-only model for speech synthesis. *arXiv preprint arXiv:2411.17690*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Zhun Liu, Ying Shen, Vedanand Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Hongyu Ma, Jian Wang, Hao Lin, Bo Zhang, Yifan Zhang, and Bo Xu. 2023. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*.

Sijie Mai, Yiyang Zeng, Shuhong Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Tao Meng, Feifan Zhang, Yifan Shou, Wenxuan Ai, Neng Yin, and Kede Li. 2024. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *arXiv preprint arXiv:2404.17862*.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017a. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe

Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883.

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

Mengge Ren, Xiaoxi Huang, Wei Li, Dian Song, and Weizhi Nie. 2021. Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition. *IEEE Transactions on Multimedia*, 24:4422–4432.

Yifan Shou, Wenxuan Ai, Jun Du, Tao Meng, Hao Liu, and Neng Yin. 2024. Efficient long-distance latent relation-aware graph neural network for multimodal emotion recognition in conversations. *arXiv preprint arXiv:2407.00119*.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Sijia Xing, Sijie Mai, and Haifeng Hu. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 13(3):1426–1439.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

A Code

The code implementation is available at [Anonymous GitHub Repository]: (https://anonymous.4open.science/r/DeepMSI-MER-B36C/README.md)

B Appendix

In the IEMOCAP paper, the provided data includes raw text, video, and audio data, which have not been partitioned and do not have corresponding label annotations. Therefore, we performed data processing in accordance with the requirements outlined in the paper. The specific steps are as follows:

- **Video Processing**: Text information is extracted from the transcriptions in the dialog folders of each Session folder. Based on the time segments in the extracted text, corresponding video segments are then extracted. Subsequently, video segments are extracted at 15-frame intervals, and data augmentation and packaging are performed using albumentations.

- **Audio Processing**: Audio files corresponding to the video segments extracted from the text information are retrieved from the wav files in the dialog folders of each Session folder, and then packaged accordingly.

- **Label Processing**: Labels corresponding to the video segments are extracted from the EmoEvaluation files in the dialog folders of each Session folder based on the video segment names in the text information. Labels such as xxx, oth, dis, fea, and sur are removed, and the labels corresponding to the texts are finally merged.

After completing the above processing, the packaged audio files and texts were uploaded to the cloud server, where training was conducted using Wav2vec-base for audio and BERT-large for text. After training the Wav2vec-base and BERT-large models, the model weights were saved. The audio files were then feature-extracted using the trained models and saved in pkl format within the corresponding audio files. Similarly, features for the texts were extracted using the trained BERT model and saved in pkl format in the corresponding text files.

Finally, in accordance with the requirements of the IEMOCAP dataset paper, the data from the first four Session files were used as the training dataset, and the data from the fifth Session file were used as the test dataset. The data was then packaged and uploaded to the DeepMSI-MER model training server for model training. The partitioned dataset was saved in txt files, which included the video segment names, the corresponding labels for the video segments, and the text data. The download link for the processed IEMOCAP dataset is as follows: https://pan.baidu.com/s/1OXYrDnNdxx72vIrSppdZ1w?pwd=4uaa.