ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446166

GUIDED CIRCULAR DECOMPOSITION AND CROSS-MODAL RECOMBINATION FOR MULTIMODAL SENTIMENT ANALYSIS

Haijian Liang[1], *Weicheng Xie*[1,3,*], *Xilin He*[1], *Siyang Song*[4], *Linlin Shen*[1,2,3]

[1]Computer Vision Institue, School of Computer Science & Software Engineering, Shenzhen University
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[3]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University
[4]School of Computing and Mathematical Sciences, University of Leicester

ABSTRACT

Multimodal Sentiment Analysis is a burgeoning research area, leveraging various modalities to predict the sentiment score. Nevertheless, previous studies have disregarded the impact of noise interference on specific modal sentiments during video recording, thereby compromising the accuracy of sentiment prediction. In this paper, we propose the Guided Circular Decomposition and Cross-Modal Recombination (GCD-CMR) model, which aims to eliminate contaminated sentiment features in a fine-grained way. To achieve this, we utilize tailored global information specific to each modality to guide the circular decomposing process in the GCD module, to produce a set of sentiment prototypes. Subsequently, in the CMR module, we align cross-modal sentiment prototypes and remove the contaminated prototypes for recombination. Experimental results on two publicly available datasets demonstrate that our model surpasses state-of-the-art models, confirming the effectiveness of our proposed method. We release the code at: `https://github.com/nianhua20/GCD-CMR`.

***Index Terms*—** multimodal sentiment analysis, modality decomposition, reduction of contaminated sentiment

1. INTRODUCTION

Multimodal Sentiment Analysis (MSA) is a research hotspot that integrates heterogeneous modal information to mine and analyse sentiments and perspectives of individuals in video [1, 2, 3]. It commonly encompasses multimodal information including text, audio and vision. Diverse modalities can supply abundant information, mutually complementing one another to enable a more comprehensive and precise sentiment analysis. However, as shown in Fig. 1, real-world video recordings are susceptible to environmental interference, which can compromise modal emotional features and introduce potential inaccuracies in sentiment analysis.

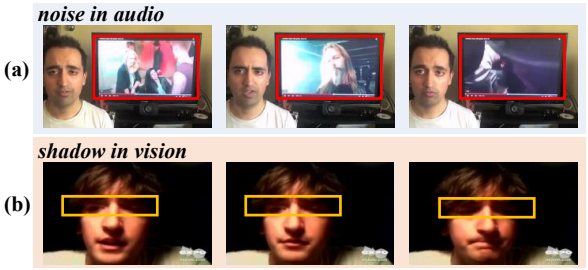


Fig. 1. **Examples of disturbances during video recording.** (a) For the audio modality, the speaker's angry voice with high-frequency pitch features is interfered with the computer playing video; (b) For the vision modality, shadows on the eye area obstruct the identification of eye features, compromising the analysis of the speaker's negative sentiment.

Most prior studies have concentrated on the development of sophisticated fusion networks for effectively harnessing diverse modal information, typically classified as early fusion [4, 5, 2] and late fusion [6, 7, 8]. Nevertheless, these fusion strategies frequently encounter huge distribution gaps among various modal representations. To bridge modality gaps, certain MSA models [9, 10, 11] decouple each modal representations to modality-invariant and modality-specific features to extract the consistency and distinctiveness of heterogeneous modalities. These methods employ a shared encoder to project diverse modalities into a modality-irrelevant space, and the resulting transformed representations are used as holistic features. However, none of these methods eliminates perturbed information within each modal commonality.

To this end, we propose Guided Circular Decomposition and Cross-Modal Recombination (GCD-CMR) network to reduce the contamination of modal sentiment features, inspired by [12, 13] which decouples expression features into a set of facial latent ones. Our contributions can be summarized as:

- We propose the GCD-CMR model, a robust model that

*: Corresponding author.

The work was supported by the Natural Science Foundation of China under grants no. 62276170, 82261138629, the Science and Technology Project of Guangdong Province under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20220531101412030.

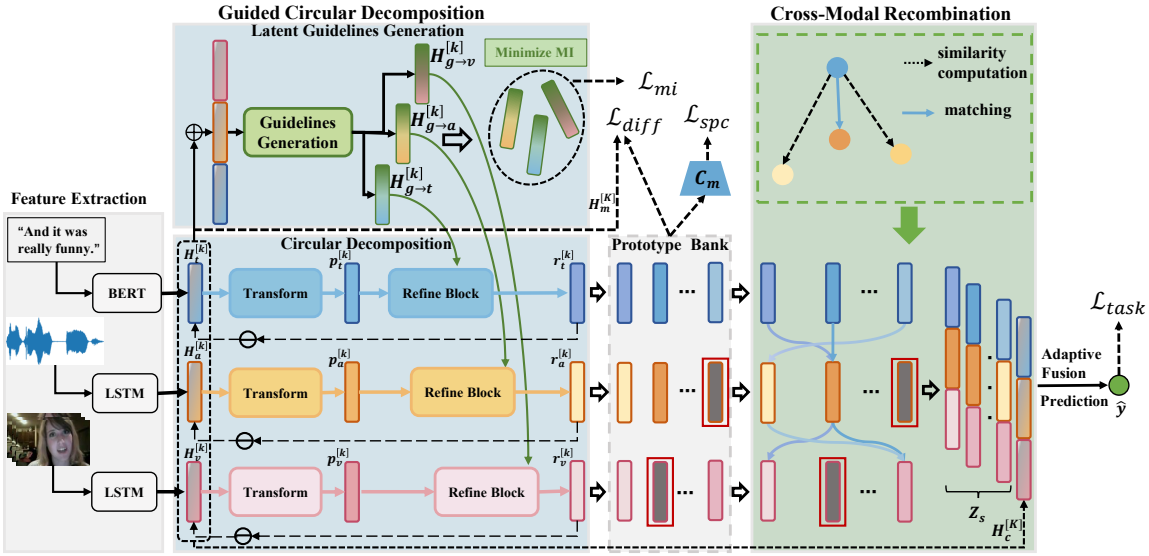


Fig. 2. The overall architecture of proposed GCD-CMR. $\oplus$ and $\ominus$ represent concatenation and subtraction operations. The contaminated sentiment prototypes are highlighted in the red boxes and will be excluded from the feature recombination.

effectively removes contaminated modal features in a fine-grained manner.

- We design a Guided Circular Decomposition (GCD) module, which can integrate multimodal information to precisely guide unimodal decomposition to derive a set of intra-modal sentiment prototypes. To enhance the robustness of the recombination representation, we eliminate contaminated prototypes and align the inter-modal pure prototypes in the Cross-Modal Recombination (CMR) module.
- Extensive experiments are conducted on public datasets and the state-of-the-art performances are achieved, validating the effectiveness and superiority of our method.

2. PROPOSED METHOD

The task of MSA is to predict a sentiment intensity score $y \in \mathcal{R}$ by leveraging multimodal signals, including text (t), audio (a) and vision (v). To encode each modal features, we utilize pre-trained BERT and two LSTMs, resulting in the initialized text features H_t, audio features H_a and visual features H_v, with the modality dimensions of d_t, d_a and d_v, respectively. The overview of our method is shown in Fig. 2.

2.1. Guided Circular Decomposition

In this part, we introduce our approach that generates tailored global cues to guide the decomposition of each modality feature into a set of sentiment prototypes, which is different from previous methods that employ a shared encoder for different modalities to derive a holistic modality-invariant feature.

Circular Decomposition. In order to decompose each modal feature into a series of sentiment prototype features, we circularly decompose the three modal features for K times. During the k-th iteration, we initially derive intra-modal sentiment prototype features $p_m^{[k]}$, which are then subtracted from the basic features $H_m^{[k]}$ for the subsequent decomposition. In this procedure, we propose to utilize the global-local guidance feature denoted as $H_{g\to m}^{[k]}$ for each modality (further detailed in Eq. (2)) to direct the decomposition and refine the resulting sentiment prototypes in the Refine block $f_m(\cdot;\theta_f^m)$. The k-th decomposition can be formulated as:

$$\begin{aligned} p_m^{[k]} &= ReLU(W_m^T H_m^{[k]}) \\ r_m^{[k]} &= f_m([H_{g\to m}^{[k]}, p_m^{[k]}];\theta_f^m) \cdot p_m^{[k]} \\ H_m^{[k+1]} &= H_m^{[k]} - r_m^{[k]} \end{aligned} \tag{1}$$

where $W_m \in R^{d_m \times d_m}$, $m \in \{t, a, v\}$ and ReLU represents the ReLU activation function. $[\cdot,\cdot]$ represents the concatenation operation on feature dim and θ_f^m are the learnable parameters of the Fully-Connected (FC) layer f_m. 1

Latent Guidelines Generation. When employing multimodal information for sentiment analysis, certain sentiment prototypes influenced by interference become important within the current modality, but they may be inconsistent or irrelevant for the overall sentiment representation. To address this issue, we design the Guidelines Generation block to generate global information and guide the decomposition of each modality in a top-down fashion, which takes the concatenated features of all three modalities as input and generates their corresponding guiding features. Specifically, the k-th decomposition of this block is formulated as:

$$\begin{aligned} H_g^{[k]} &= E([H_t^{[k]}, H_a^{[k]}, H_v^{[k]}];\theta_E) \\ H_{g\to m}^{[k]} &= D_m(H_c^{[k]};\theta_D^m) \end{aligned} \tag{2}$$

Models	CMU-MOSI					CMU-MOSEI				
	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
MISA [9]$_{\text{MM'20}}$	43.5	81.8/83.5	81.7/83.5	0.752	0.784	52.2	81.6/84.3	82.0/84.3	0.550	0.758
MAG-BERT [3]$_{\text{ACL'20}}$	45.1	82.4/84.6	82.2/84.6	0.730	0.789	52.8	81.9/85.1	82.3/85.1	0.558	0.761
Self-MM [14]$_{\text{AAAI'21}}$	45.8	82.7/84.9	82.6/84.8	0.731	0.785	53.0	82.6/85.2	82.8/85.2	0.540	0.763
MMIM [1]$_{\text{EMNLP'21}}$	45.0	83.0/85.1	82.9/85.0	0.738	0.781	53.1	81.9/85.1	82.3/85.0	0.547	0.752
FDMER [10]$_{\text{MM'22}}$	44.1	-/84.6	-/84.7	0.724	0.788	54.1	-/86.1	-/85.8	0.536	0.773
DMD [11]$_{\text{CVPR'23}}$	45.6	-/86.0	-/86.0	-	-	54.5	-/86.6	-/86.6	-	-
EMT [15]$_{\text{TAC'23}}$	47.4	83.3/85.0	83.2/85.0	0.705	0.798	54.5	83.4/86.0	83.7/86.0	0.527	**0.774**
GCD-CMR(ours)	**47.7**	**83.8/86.3**	**83.5/86.1**	**0.697**	**0.802**	**54.8**	**84.0/86.7**	**84.3/86.6**	**0.525**	0.772

Table 1. Performances of the state of the arts and ours on CMU-MOSI and CMU-MOSEI. The best results are labeled in bold.

where $\{\theta_E, \theta_D^m\}$ are learnable parameters. The universal encoder $E(\cdot;\theta_E)$ is used to learn the shared global-view feature and the modality-specific decoder $D_m(\cdot;\theta_D^m)$ is used to decode the global representation into tailored features for each modality.

To provide global guidance for the disentanglement specific to each modality, we propose to minimize the mutual information (MI) to effectively distinguish among them. Concretely, vCLUB [16] is used to compute the upper bound of MI, and the following MI minimization regularizer is employed to decrease correlation among the global-local view features:

$$\mathcal{L}_{mi} = \sum_{k=1}^{K} \sum_{(m_1,m_2)} MI_{vCLUB}(H_{m_1}^{[k]}, H_{m_2}^{[k]}) \tag{3}$$

where $(m_1, m_2) \in \{(t,a),(t,v),(a,v)\}$.

Prototype Bank. After K iterations in Eq. (1), the basic modality features H_m are circularly decomposed into different sentiment prototypes and stored in the prototype bank $\mathcal{B}_m = \{r_m^{[1]}, r_m^{[2]}, ..., r_m^{[K]}\}$, $m \in \{t,a,v\}$. The residual modality features $H_m^{[K]}$ exclusively contains the unique characteristics specific to each modality.

To enhance the discriminative quality of learned sentiment prototypes, we define a classifier $C_m(\cdot;\theta_C^m)$, which takes the prototypes stored in the bank of each modality $Q_m = \sum_{k=1}^{K} r_m^{[k]}$ as input, and the specific loss is defined as:

$$\mathcal{L}_{spc} = \sum_{m\in\{t,a,v\}} |C_m(Q_m;\theta_C^m) - y| \tag{4}$$

where C_m maps R^{d_m} to R^1 and θ_C^m is the learnable parameter. In order to achieve a separation of sentiment prototypes and reduce information redundancy between sentiment prototypes and modality-specific representation, the orthogonality constraint is used and formulated as:

$$\mathcal{L}_{diff} = \sum_{m\in\{t,a,v\}} ||Q_m H_m^{[K]}||_F^2 \tag{5}$$

Here, $||\cdot||_F^2$ is the Frobenius norm.

2.2. Cross-Modal Recombination

In addition to reducing the influence of contaminated sentiment prototype features to enhance the model's robustness, we further resort to the realignment of cross-modal prototypes to obtain more discriminative fusion features. Since previous studies [2, 17] have demonstrated the more importance of the textual modality compared to the other two modalities, we opt to employ the textual sentiment prototype as an anchor. This anchor is aligned with the most relevant sentiment prototypes from the audio bank $\mathcal{B}_a$ and the visual bank $\mathcal{B}_v$ for recombination. The matching process for each textual sentiment prototype $r_t^{[k]}$ in the prototype bank is formulated as:

$$r_n^{[k]} = \underset{r_n^{[i]}\in\mathcal{B}_n}{\arg\max} \frac{r_t^{[k]} \cdot r_n^{[i]}}{||r_t^{[k]}||||r_n^{[i]}||} \tag{6}$$

where $n \in \{a,v\}$. Then we concatenate mutually matched cross-modal prototypes of the k-th textual sentiment prototype $Z_c^{[k]} = [r_t^{[k]}, r_a^{[k]}, r_v^{[k]}]$ for the fusion representation. The contaminated prototypes are not matched in this process, and thus do not participate in the features fusion.

Adaptive Feature fusion. Following [10, 11], we employ a dynamic fusion strategy, i.e., assigning adaptive weights to modality-specific representation $H_c^{[K]} = [H_t^{[K]}, H_a^{[K]}, H_v^{[K]}]$ and each sentiment feature $Z_s = \{Z_c^{[1]}, Z_c^{[2]}, \ldots, Z_c^{[K]}\}$ to generate the fusion feature, and subsequently pass it through the FC layer to produce the sentiment prediction $\hat{y}$. To train our model, we utilize L1 loss as the prediction loss, i.e.,

$$\mathcal{L}_{task} = \frac{1}{N}\sum_{j=1}^{N} |y^{(j)} - \hat{y}^{(j)}| \tag{7}$$

where $y^{(j)}$ represents the label of the j-th sample and N represents the number of samples in a batch. The overall learning of our model is performed by minimizing:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_1\mathcal{L}_{mi} + \lambda_2\mathcal{L}_{spc} + \lambda_3\mathcal{L}_{diff} \tag{8}$$

where λ_1, λ_2 and λ_3 are the trade-off parameters.

3. EXPERIMENTS

3.1. Experimental Setup

In this paper, we use two of the most commonly used public datasets for the evaluation, i.e. CMU-MOSI [18] and CMU-MOSEI [19]. In detail, CMU-MOSI consists of 2,199 opinion segments, with annotations capturing sentiment on a scale ranging from negative to positive (-3 to 3). CMU-MOSEI comprises 23,453 annotated video clips from 1,000 speakers, each with a sentiment score interval of [-3,3].

We evaluate the performance of our algorithm on the tasks of classification and regression. For classification, we use Acc7 (%), Acc2 (%) and F1 scores (%) in negative/positive (zero excluded) and negative/non-negative (zero included) settings. For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). For the benchmarks of MOSI and MOSEI, we employ the Adam optimizer with initial learning rates of 5e-5 and 5e-6, respectively. The trade-off parameters λ_1, λ_2, and λ_3 are set as $\{0.01,0.1,0.1\}$ and $\{0.02,0.01,0.1\}$, respectively. We set the number of decomposing cycles, i.e. K as 5 and 3, respectively. All the experiments are conducted on four Nvidia Tesla P100 GPUs, using the PyTorch framework.

3.2. Experimental Results and Analysis

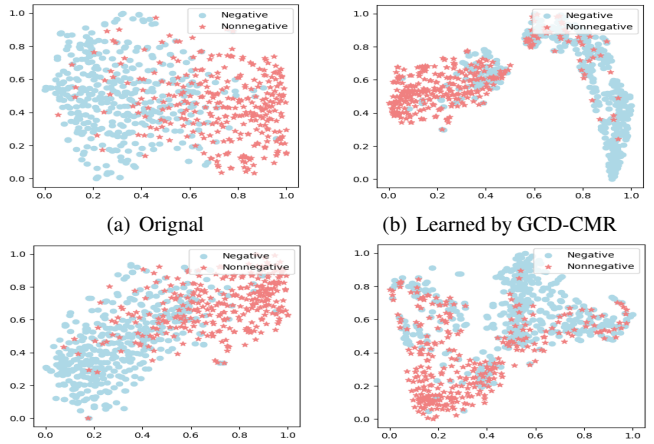
Comparison with the state of the arts. As presented in Table 1, our proposed model, GCD-CMR, consistently achieves superior performance across the majority of the evaluation metrics. In comparison to feature disentanglement models [9, 10, 11], our GCD-CMR model outperforms them across all metrics and surpasses the best-performing model, i.e. DMD by margins of 2.1% in terms of Acc7 and 0.3% in terms of Acc2 on CMU-MOSI. This can be attributed to the capability of our model to eliminate the perturbed sentiment features. Compared to the recent EMT [15], which focuses on modal features disturbed by the environment, our model achieves improvements of 1.3% and 0.7% in terms of Acc2, and 1.1% and 0.6% in terms of F1 score on CMU-MOSI and CMU-MOSEI, respectively.

Visualization. To assess the robustness of our model, we applied T-SNE [20] to visualize the distributions of multimodal features in Fig. 3. It shows that our GCD-CMR model can achieve clearer boundaries and better separation compared with the benchmark in both the cases of original features and those with Gaussian noise inserted into the visual modality.

Ablation study. To study the performance of each module, we conduct an ablation study in Table 2. From the results, we can conclude that all the designed losses contribute to improving the model. In particular, the proposed $\mathcal{L}_{mi}$ shows significant improvements, demonstrating the necessity of specialized global features tailored to each modality. Furthermore, after removing the guidelines generation module, Acc2 decreases from 86.7 to 85.0, and F1 decreases from 86.6 to 85.1, highlighting the importance of global information for refining sentiment prototypes. Additionally, we verify the effects of recombining pure sentiment prototypes. The results demonstrate that directly combining all sentiment prototypes to produce the final fused feature shows to be mediocre, leading to a decrease in Acc2 and F1 by approximately 1%.

Description	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
W/O $\mathcal{L}_{mi}$	53.9	82.6/85.9	83.0/85.8	0.531	0.753
W/O $\mathcal{L}_{spc}$	54.5	82.7/86.5	83.1/86.5	0.528	0.731
W/O $\mathcal{L}_{diff}$	54.4	82.9/86.1	83.2/86.1	0.528	0.771
W/O Guidelines	53.9	80.7/85.0	81.4/85.1	0.533	0.769
W/O Recombination	54.4	82.3/85.8	82.7/85.8	0.532	0.772
GCD-CMR(ours)	**54.8**	**84.0/86.7**	**84.3/86.6**	**0.525**	**0.772**

Table 2. Ablation study of GCD-CMR on CMU-MOSEI. 'W/O' represents removing the mentioned component.



(a) Orignal (b) Learned by GCD-CMR

(c) Orignal, visual modality with 20% Gaussian noise (d) Learned by GCD-CMR, visual modality with 20% Gaussian noise

Fig. 3. T-SNE visualization of multimodal representation on the testing set of CMU-MOSI.

4. CONCLUSION

In this paper, to mitigate the impact of noise on modality features during video recording, we introduce a novel framework, i.e. Guided Circular Decomposition and Cross-Modal Recombination. Specifically, we leverage global information to create local guidance features for modality-specific sentiment prototypes within the Guided Circular Decomposition module. Additionally, we preserve the pure sentiment prototypes and recombine them with the most relevant prototypes from different modalities in the Cross-Modal Recombination module. This process has been shown to be helpful to produce robust fusion representation by the extensive experiments on two datasets, and the comparison between seven state-of-the-arts published in the past three years and ours in terms of five metrics verifies the effectiveness of our entire framework.

5. REFERENCES

[1] Wei Han, Hui Chen, Poria, and et al., "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9180–9192.

[2] Yao-Hung Hubert Tsai, Shaojie Bai, and et al., "Multimodal transformer for unaligned multimodal language sequences," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, pp. 6558–6569, 2019.

[3] Wasifur Rahman, M. Hasan, and et al., "Integrating multimodal information in large pretrained transformers," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020, pp. 2359–2369, 2020.

[4] Jennifer Williams, Kleinegesse, and et al., "Recognizing emotions in video using multimodal dnn feature fusion," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 2018, pp. 11–19.

[5] Amir Zadeh, Liang, and et al., "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[6] Amir Zadeh, Minghai Chen, and et al., "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Sept. 2017, pp. 1103–1114.

[7] Zhun Liu, Ying Shen, and et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Annual Meeting of the Association for Computational Linguistics*, 2018.

[8] Soujanya Poria, Cambria, and et al., "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)*, 2017, pp. 873–883.

[9] Devamanyu Hazarika, Zimmermann, and et al., "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[10] Dingkang Yang, Shuai Huang, and et al., "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1642–1651.

[11] Yong Li, Yuanzhi Wang, and et al., "Decoupled multimodal distilling for emotion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6631–6640.

[12] Xinyi Zou, Yan Yan, and et al., "Learn-to-decompose: cascaded decomposition network for cross-domain few-shot facial expression recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 683–700.

[13] Delian Ruan, Yan Yan, and et al., "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7660–7669.

[14] Wenmeng Yu, Hua Xu, and et al., "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 10790–10797.

[15] Licai Sun, Zheng Lian, and et al., "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2023.

[16] Pengyu Cheng, Weituo Hao, and et al., "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*. PMLR, 2020, pp. 1779–1788.

[17] Odysseas S Chlapanis, Georgios Paraskevopoulos, and et al., "Adapted multimodal bert with layer-wise fusion for sentiment analysis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] Amir Zadeh, Rowan Zellers, and et al., "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[19] AmirAli Bagher Zadeh, Paul Pu Liang, and et al., "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[20] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.