Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24716

State of the Art in the Application of Multimodal Affective Methods for Comparative Analysis of Modal Deficits

Yanhaotian Zhao

School of Information Science & Engineering, Lanzhou University, Lanzhou, China
zhaoyht21@lzu.edu.cn

Abstract. The advent of multimedia technology has precipitated a paradigm shift in the realm of human-computer interaction and affective computing, thus rendering multimodal emotion recognition a pivotal domain. However, the issue of modal absence, resulting from equipment failure or environmental interference in practical applications, significantly impacts the accuracy of emotion recognition. The objective of this paper is to analyse multimodal emotion recognition methods oriented to modal absence. The focus is on comparing and analysing the advantages and disadvantages of techniques such as generative class and joint representation class. Experimental findings demonstrate the efficacy of these methods in surpassing the conventional baseline on diverse datasets, including IEMOCAP, CMU-MOSI, and others. Notably, CIF-MMIN enhances the mean accuracy by 0.92% in missing conditions while concurrently reducing the UniMF parameter by 30%, thus preserving the SOTA performance. Key challenges currently being faced by researchers in the field of multimodal emotion recognition for modal absence include cross-modal dependencies and semantic consistency, model generalisation ability, and dynamic scene adaptation. These challenges may be addressed in the future through the development of a lightweight solution that does not require full-modal pre-training, and by combining comparative learning with generative modelling to enhance semantic fidelity. The present paper provides both theoretical support and practical guidance for the development of a highly robust and efficient emotion recognition system.

Keywords: Multimodal emotion recognition, modal absence, robustness, cross-modal imagery

1. Introduction

In the contemporary era of rapid advancements in multimedia technology, multimodal emotion recognition has emerged as a pivotal research trajectory within the domains of human-computer interaction and affective computing. However, in practical applications, multimodal data is often subject to the problem of modal missing due to equipment failures, environmental interference and other factors. This poses a serious challenge to the accuracy of emotion recognition.

In recent years, scholars have proposed two main types of solutions around the modality missing problem: generative class methods and joint representation class methods. For example, the Missing Modality Imagination Network (MMIN) model proposed by Zhao et al. improves the performance

DOI: 10.54254/2755-2721/2025.PO24716

of emotion recognition under uncertain missing modality conditions through cross-modality imagery and recurrent consistency learning [1]. Multimodality-Diffused Emotion Recognition (IMDer) effectively solves the problem of emotion recognition under incomplete multimodal data using a score-based diffusion model and a conditionally guided recovery process [2]. Huan et al. proposed a unified cross-modal framework Missing Modality Imagination Network (UniMF) for sentiment analysis of missing modality and aligned multimodal sequences, which can efficiently handle both missing modality and aligned multimodal sequences and outperforms existing methods in terms of the number of parameters [3]. In addition, Liu et al. proposed a modality-invariant feature acquisition method based on contrast learning and applied it to Missing Modality Imagination Network Missing Modality Imagination Network (CIF-MMIN) Despite the progress made by these methods, the existing researches still face the bottlenecks such as insufficient modeling of cross-modal dependencies, computationally inefficient, dynamic missing bottlenecks such as poor real-time performance [4-6].

The purpose of this paper is to systematically review the latest advances in multimodal emotion recognition for modal absence, analyze the core challenges and solutions, and provide directional suggestions for future research. The full paper is structured as follows: Section 2 compares and analyzes the generative classes (MMIN, IMDer, Spectral Domain Reconstruction Graph Neural Network (SDR-GNN), Multimodal Cascaded Framework, Text-guided Reconstruction Network (TgRN)), joint representation classes (CIF-MMIN, UniMF, Unimodal Label Generation and Modality Decomposition (ULMD)) and an innovative multimodal sentiment analysis method (Tag-assisted Transformer Encoder). Assisted Transformer Encoder (TATE)) and an innovative multi-task co-optimization type approach; Section 3 deconstructs the technical challenges from the dimensions of cross-modal dependency, semantic consistency, and dynamic missing; Section 4 looks ahead to the potential research directions such as real-time processing of dynamic missing modalities, cross-modal semantic consistency enhancement, and multi-task co-optimization. By combing the literature evolution paths and performance boundaries, this paper attempts to provide theoretical support and practical insights for building highly robust and efficient emotion recognition systems.

2. Comparison and analysis of typical techniques

In addressing the issue of missing modality in multimodal emotion recognition, the extant methods can be principally categorised into three distinct groups: generative, joint representation, and other innovative strategies. Specifically, this phenomenon is illustrated in Table 1.

Table 1: Multimodal emotion recognition missing modality processing methods classification

Category	Core features	Thesis methodology
Generation Classes	Solve the missing modality problem by generating or recovering data for missing modalities.	MMIN, IMDer, SDR-GNN, Multimodal Cascaded Framework, TgRN
Joint Representation Classes	Learning joint representations across modalities, directly utilizing available modal information for sentiment analysis without explicitly generating missing data.	CIF, UniMF, ULMD
Other Classes	Innovative strategies using non-generative or non-joint representations, e.g., labeling assistance, module-specific design, etc.	TATE

Table 1 provides a systematic summary of the three major classes of methods and their typical algorithms for dealing with missing modalities in multimodal emotion recognition. The generative

DOI: 10.54254/2755-2721/2025.PO24716

class of methods is concerned with the recovery of missing modalities through the utilisation of neurogenerative techniques. For instance, MMIN employs cross-modal imagery, while IMDer utilises a diffusion model. The joint representation class is dedicated to the construction of robust joint multimodal representations. For example, CIF employs contrast learning, and UniMF employs Transformer unified modelling. The remaining classes adopt different technical routes. One such example is TATE, which enhances semantic consistency through label encoding. The three categories outlined here essentially encompass the prevailing technical approaches to addressing the missing modality problem in the contemporary mainstream.

The generative class of methods has been shown to restore missing modalities through data reconstruction mechanisms. Among these mechanisms, MMIN is an innovative multimodal sentiment recognition method. MMIN employs a cross-modal feature generation approach through CRA and cyclic consistency learning, addressing the challenge posed by uncertain missing modalities. Its unified model design and robust performance position it as a leading method in the field of multimodal sentiment recognition. This method is not only applicable to benchmark datasets such as IEMOCAP and CMU-MOSI, but also performs well in dealing with uncertain missing modalities, thus bringing new ideas and methods to the field of multimodal sentiment analysis [1].

IMDer is a state-of-the-art multimodal sentiment recognition method designed to address missing modal data. The method is predicated on a fractional diffusion model that gradually recovers missing data by mapping random noise to the distribution space of the missing modalities.IMDer utilizes the available modalities as conditions to guide the diffusion process, ensuring that the recovered modalities remain consistent with the original data in terms of distribution and semantics. This approach has been demonstrated to achieve state-of-the-art sentiment recognition accuracy on the CMU-MOSI and CMU-MOSEI datasets. Furthermore, it has been shown to significantly improve the performance of sentiment analysis in the presence of missing modalities.IMDer is not only technically innovative, but it also demonstrates strong robustness and generalization ability in practical applications. As such, it provides strong support to the field of multimodal sentiment recognition [2].

SDR-GNN approach is an innovative graph neural network strategy that has been developed for the classification of sentiment in incomplete multimodal data. The method effectively captures complex sentiment dependencies by combining multi-frequency feature fusion through spectral domain reconstruction techniques. Specifically, the SDR-GNN constructs a graph structure containing speaker and contextual relations, and utilizes a neighborhood relation-aware layer and a hyperedge relation-aware layer to extract high-level semantic information. Additionally, the model performs multi-frequency aggregation in the spectral domain, a process that preserves high-frequency detail information and enhances the model's expressive power. This approach has been demonstrated to outperform existing sentiment classification methods on multiple datasets, thereby showing its superior performance in handling missing or incomplete multimodal data.The SDR-GNN approach is not only technically innovative, but it also provides new perspectives and solutions for multimodal sentiment analysis [7].

TgRN is a cutting-edge multimodal sentiment analysis method that addresses the issue of missing modalities. The method utilizes the information of textual modalities to recover missing modal features through a text-guided reconstruction module, thereby ensuring effective sentiment recognition despite incomplete data.TgRN contains a text-guided extraction module (TEM) and a reconstruction module (RM), which can efficiently capture both cross-modal features and intra-modal features. Furthermore, TgRN introduces a text-guided fusion module (TFM) to explore the dynamic correlation between nonverbal and verbal modalities, thereby enhancing the robustness and

DOI: 10.54254/2755-2721/2025.PO24716

accuracy of the model. The experimental findings on the CMU-MOSI and CH-SIMS datasets demonstrate that TgRN exhibits optimal performance at elevated missing rates, accompanied by minimal performance degradation. This performance is found to be significantly superior to that of existing multimodal sentiment analysis methods [8,9].

Multimodal Cascaded Framework is a recently developed multimodal classification framework that was designed to address the problem of missing modalities in multimodal data. The proposed framework enhances the discriminative ability of the feature space by generating complete multimodal data and employs a variety of novel loss functions to improve the classification accuracy. The framework is comprised of three distinct steps: feature extraction, feature generation, and classification. It employs multiple novel loss designs (e.g., missing modality union, center-of-mass union, and latent a priori loss) to enhance the classifiability and ensure that similar classes of data remain close in the latent space in the presence of missing modalities. The experimental results demonstrate that the framework exhibits a high degree of proficiency in managing incomplete multimodal data, thereby leading to a substantial enhancement in classification accuracy [10].

The joint Representation Class Approach is an innovative approach to cross-modal representation learning. UniMF is a multimodal sentiment analysis framework that has been designed to efficiently handle missing modalities and unaligned multimodal sequences. The framework under consideration consists of two modules: a translation module and a prediction module. The translation module employs the Multimodal Generation Transformer (MGT) and Multimodal Generation Mask (MGM) to recover the missing modalities through the information of the existing modalities to ensure the integrity of the data. The prediction module employs the Multimodal Understanding Transformer (MUT) and the Multimodal Understanding Mask (MUM), in conjunction with the distinctive sequence MultiModalSequence (MMSeq), to effectively integrate multimodal data and yield the final sentiment analysis outcomes. The UniMF design not only reduces the complexity of multimodal processing but also attains performance that is comparable to or even superior to state-of-the-art methods across multiple datasets. The UniMF design not only simplifies the complexity of multimodal processing but also achieves comparable or even superior performance to state-of-the-art methods on multiple datasets, while significantly reducing the number of model parameters, thereby demonstrating its dual advantages in computational efficiency and accuracy [3].

CIF-MMIN is an innovative multimodal emotion recognition method that combines the advantages of contrast learning and cross-modal imaginative networks. The proposed methodology first extracts modality-invariance features through contrast learning to ensure feature consistency across different modalities. This, in turn, serves to mitigate the challenges posed by modality gaps. Subsequently, CIF-MMIN employs the cross-modal imagination and cyclic consistency learning of MMIN (Missing Modality Imagination Network) to effectively reconstruct the information of missing modalities. This dual mechanism enables CIF-MMIN to demonstrate a high degree of proficiency in the handling of missing modality data, thereby significantly enhancing the robustness and accuracy of emotion recognition. A comprehensive evaluation of experimental results on numerous benchmark datasets has been conducted, revealing the efficacy of CIF-MMIN in outperforming existing methods under various conditions. This outcome is indicative of its strong adaptability and exceptional performance in complex multimodal environments [4].

ULMD is an innovative multimodal sentiment analysis method that aims to optimize the joint representation of multimodal data through unimodal label generation and modal decomposition. The proposed method employs a multi-task learning framework that decomposes the multimodal sentiment analysis task into a multimodal task and three unimodal tasks. This approach enables the utilization of the independent information inherent in each modality.ULMD introduces a modal

DOI: 10.54254/2755-2721/2025.PO24716

representation separator that decomposes the modal features into modality-specific and -invariance parts. These parts are employed in the multimodal task and the unimodal task, respectively, to ensure that the model captures both the common and individual features among the modalities. Consequently, ULMD enhances not only the model's capacity for generalization but also its performance in sentiment analysis across diverse datasets. The ULMD system's modular design and multi-task learning strategy facilitate its proficiency in managing intricate multimodal data, signifying a substantial technological advancement within the domain of multimodal sentiment analysis [9].

TATE is an innovative multimodal sentiment analysis method designed to effectively address the problem of missing modalities. The proposed method has been shown to enhance the performance of sentiment classification in both unimodal and missing multimodal scenarios through the integration of label-assisted techniques and the Transformer encoder. The Transformer encoder's label-encoding module is central to the TATE approach, as it addresses both unimodal and missing multimodal situations by aligning common vectors through a novel spatial projection pattern. This approach enables the network to prioritize the analysis of missing modal features. Furthermore, TATE employs a Transformer encoder-decoder network to learn the features of missing modalities, thereby enhancing the model's expressive power. The experimental results demonstrate that TATE attains substantial enhancements over alternative baseline methods on the CMU-MOSI and IEMOCAP datasets, thereby substantiating its preeminence and efficacy in addressing uncertain missing modalities [5,6].

For the above-proposed methods, the performance advantages and limitations of these methods are specifically shown in Table 2.

DOI: 10.54254/2755-2721/2025.PO24716

Table 2: Comparative analysis of multimodal emotion recognition methods

Category	Method Name	Performance Advantages	Disadvantages
Generative class	MMIN	Significantly improves sentiment recognition performance on benchmark dataset	Higher computational complexity, may affect real-time performance
	IMDer	Achieves state-of-the-art sentiment recognition accuracy, distributional consistency and semantic disambiguation	Requires large computational resources, long training time
	SDR-GNN	Outperforms existing methods, captures complex sentiment dependencies	Model is complex and difficult to interpret
	TgRN	Stable at high missing rate, small performance degradation	High text dependency, lack of multimodal fusion
	Multimodal Cascaded Framework	High text dependency, lack of multimodal fusion	Complex generation process, high computational effort
Joint Representation Class	UniMF	with few parameters, performance is comparable or better than SOTA	Limited ability to deal with complex modal relationships
	CIF-MMIN	Outperforms existing methods in all missing conditions, with significantly improved robustness	Depends on the quality of comparative learning and may be limited by the dataset
	ULMD	Optimized joint representation through modal separation and multi-task learning, improved model generalization ability	Complex model structure, difficult to train
Other classes	TATE	Significantly improves classification performance in missing scenarios	Strong label dependence, lack of autonomous learning ability

Table 2 systematically compares the performance of different methods in multimodal emotion recognition. Generative class methods (e.g., MMIN, IMDer) enhance the effect by completing the missing modalities; however, they generally suffer from computational complexity and high resource consumption. Joint representation class methods (e.g., UniMF, CIF-MMIN) focus on robust representation learning but are sensitive to data quality. Other classes of methods (e.g., TATE) rely on the assistance of external labels and perform outstandingly in specific scenarios. With respect to computational efficiency, UniMF exhibits the smallest number of parameters, while the deployment feasibility of SDR-GNN and ULMD is significantly influenced by their model complexity. These methods provide tradeoffs for technology selection in different application scenarios.

3. Challenges and prospects

3.1. Challenge analysis

The research on missing modality recovery presents several significant challenges, including cross-modal dependencies and semantic consistency, model generalization capability, and dynamic scene adaptation.

First, concerning cross-modal semantic consistency and noise control, the recovery of missing modalities in multimodal emotion recognition must ensure that the generated modalities are semantically and distributionally consistent with the existing modalities. For instance, when an IMDer generates missing modalities through a diffusion model, the diffusion process must be conditioned by the available modalities to avoid introducing semantic ambiguities (e.g., the

DOI: 10.54254/2755-2721/2025.PO24716

recovered visual expressions conflict with the textual descriptions). Nonetheless, the intricacy of cross-modal dependencies may result in noise amplification issues. Even though TgRN aligns multimodal semantics through a text-guided attention mechanism (TCA module), low-quality inputs from non-textual modalities (e.g., audio) may still contaminate the generation results. Furthermore, CIF-MMIN extracts modality- invariant features through contrast learning; however, its efficacy may be hindered when the inter-modal distribution difference is too substantial (e.g., textual discrete symbols vs. visual continuous pixels).

Secondly, about the model's capacity for generalization, extant methodologies demonstrate efficacy in specific datasets or scenarios involving missing modalities; however, their generalizability remains constrained. For instance, MMIN demonstrates consistent performance under fixed missing protocols (e.g., only visual modalities are absent), yet its accuracy declines by 8% when 50% of the modalities are randomly missing, suggesting that it is not adequately adaptive to dynamic missing combinations.CIF-MMIN utilizes full-modal pre-training to extract invariant features, and if the modal missing modes of the test scenario differ significantly from the training (e.g., It has been determined that the training phase lacks audio, while the test phase is missing both audio and visual components. UniMF has been demonstrated to enhance generalization by reducing the complexity of the translation module. However, its performance is 2% lower than the optimal method when linguistic modalities are absent, thereby illustrating the impact of critical modal absence on the model.

Furthermore, the types, proportions, and timing of missing modalities in actual scenarios may undergo dynamic changes, and existing methods are confronted with two significant challenges: first, uncertain missing processing, such as transient loss of audio/video in videoconferencing, where MMIN necessitates predefined missing modalities, and computationally inefficient, although the hypergraph structure of the SDR-GNN can model complex dependencies. Secondly, the computational efficiency and performance trade-off is addressed by the iterative generation of IMDer's diffusion model, which results in increased inference latency. UniMF reduces the parameters by unifying the framework, but this may compromise part of the performance.

Finally, experimental data demonstrate that public datasets (e.g., CMU-MOSI, CH-SIMS) exhibit significant category imbalance (less than 20% of neutral samples), which may result in the model's suboptimal performance in minority class emotion recognition. Furthermore, the presence of noise (e.g., sensor failure, insufficient light) in real-world scenarios can exacerbate the degradation of modal quality and augment the complexity of recovery.

3.2. Outlook

Future research can focus on the following areas: For real-time processing of missing modes, the SDR-GNN spectral domain reconstruction technique can be optimized and combined with a dynamic complementation strategy to improve processing efficiency for real-time data streams. At the same time, the IMDer dynamic recovery mechanism based on the fractional diffusion model can be explored to enhance the model's adaptability. In terms of semantic consistency between intermodal modes, modal invariant features can be extracted using CIF-MMIN's contrastive learning framework. These features can then be integrated with MMIN's cross-modal imagery network to create a joint representation learning system with cyclic consistency constraints. In low-resource scenarios, a lightweight multimodal fusion architecture can be developed by combining CIF-MMIN's contrastive dimensionality reduction with SDR-GNN's graph structure compression strategy. To enhance robustness, TATE's interference-resistant projection mode and CRA's cascading residual structure must be integrated to strengthen the model's tolerance to noise and missing data.

DOI: 10.54254/2755-2721/2025.PO24716

Additionally, the single-modal labeling mechanism of ULMD needs to be improved and combined with the text-guided reconstruction module of TgRN to achieve unimodal and missing mode decomposition under uncertainty. Furthermore, the unimodal labeling mechanism of ULMD must be improved and combined with TgRN's text-guided reconstruction module to refine decomposition under uncertain missing modalities. This optimization is achieved through the multimodal cascade framework's multiple loss function, forming a comprehensive low-resource multimodal sentiment analysis solution. See Table 3 for details.

Table 3: Overview of future research directions and methods in multimodal emotion recognition

Research Direction	Involved Papers	Method Description	Specific Techniques
Dynamic missing modality real-time processing	SDR-GNN、IMDer	Development of lightweight models to process real-time data streams, combined with dynamic complementation strategies	Spectral Domain Reconfiguration Graph Neural Network (SDR-GNN), Fraction-based Diffusion Model (IMDer)
Cross-modal semantic consistency enhancement	CIF-MMIN、MMIN	Combining contrastive learning and generative modeling to improve semantic fidelity	Modal invariant feature extraction (CIF), cross-modal imagery network (MMIN)
Multi-task joint optimization	UniMF	Extended unified framework to support multi-task joint training	Joint architecture of translation module + prediction module (UniMF)
Low Resource Adaptation Research	CIF-MMIN、SDR-GNN	Designing lightweight solutions without full modal pre-training	Graph Structure Compression (SDR-GNN), Comparative Learning Dimensionality Reduction (CIF)
Robust feature extraction	TATE、CRA	Constructing joint representation learning methods that are resistant to interference Label-assisted	Transformer (TATE), Cascaded Residual Auto-Encoder (CRA)
Unimodal Label Generation and Modal Decomposition	ULMD	Unimodal Label Generation and Modal Decomposition using Multi-Task Learning	Framework Modal Representation Separator, Text Centered Fusion Module, Assisted Unimodal Label Generation Tasks
Uncertain Missing Modal Processing	TgRN	Uncertain missing modal processing using text-guided extraction and reconstruction modules	Text-guided extraction module (TEM), reconstruction module (RM), text-guided fusion module (TFM)
Multimodal Cascaded Processing	Multimodal Cascaded Framework	Generate complete multimodal data through a cascaded framework and use a loss function to improve classification accuracy Missing Modal	Joint Losses, Center of Mass Joint Losses, Potential A Priori Losses

4. Conclusion

Significant progress has been made in studying the robustness of multimodal sentiment recognition under modal absence conditions. The generative and joint representation classes of approaches have their advantages. Generative techniques achieve modal complementation through data reconstruction but must balance computational overhead with real-time performance. Joint representation methods directly model cross-modal associations with lightweight architectures but experience performance bottlenecks with complex modal relationships. Future research should focus on three areas: first, developing real-time, dynamic, missing-processing mechanisms (e.g., IMDer

DOI: 10.54254/2755-2721/2025.PO24716

diffusion model acceleration) combined with SDR-GNN graph structure optimization to improve efficiency; second, enhancing cross-modal semantic consistency by integrating comparative learning from CIF-MMIN and cyclic constraints from MMIN and reducing generative noise; and third, constructing multi-task, collaborative frameworks (e.g., combining UniMF with TATE) to efficiently deploy in low-resource scenarios through labeling assistance and parameter sharing. Additionally, it is necessary to pay attention to data quality and category imbalance, promote the deep integration of theory and practice, and ultimately develop an emotional computing system that can adapt to complex real-world scenarios.

References

[1] Zhao, Jinming, Ruichen Li, and Qin Jin. "Missing modality imagination network for emotion recognition with uncertain missing modalities." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.

[2] Wang, Yuanzhi, Yong Li, and Zhen Cui. "Incomplete multimodality-diffused emotion recognition." Advances in Neural Information Processing Systems 36 (2023): 17117-17128.

[3] Huan, Ruohong, et al. "Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences." IEEE Transactions on Multimedia 26 (2023): 5753-5768.

[4] Liu, Rui, et al. "Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities." IEEE Transactions on Affective Computing 15.4 (2024): 1856-1873.

[5] Zeng, Jiandian, Tianyi Liu, and Jiantao Zhou. "Tag-assisted multimodal sentiment analysis under uncertain missing modalities." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.

[6] Zeng, Jiandian, Jiantao Zhou, and Tianyi Liu. "Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities." IEEE Transactions on Multimedia 25 (2022): 6301-6314.

[7] Fu, Fangze, et al. "SDR-GNN: Spectral Domain Reconstruction Graph Neural Network for incomplete multimodal learning in conversational emotion recognition." Knowledge-Based Systems 309 (2025): 112825.

[8] Shi, Piao, et al. "Text-guided Reconstruction Network for Sentiment Analysis with Uncertain Missing Modalities." IEEE Transactions on Affective Computing (2025).

[9] Zhu, Linan, et al. "Multimodal sentiment analysis with unimodal label generation and modality decomposition." Information Fusion 116 (2025): 102787.

[10] John, Vijay, and Yasutomo Kawanishi. "Multimodal Cascaded Framework with Multimodal Latent Loss Functions Robust to Missing Modalities." ACM Transactions on Multimedia Computing, Communications and Applications (2025).