# GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, *Member, IEEE*, and Jianhua Tao, *Senior Member, IEEE*

***Abstract*—Conversations have become a critical data format on social media platforms. Understanding conversation from emotion, content and other aspects also attracts increasing attention from researchers due to its widespread application in human-computer interaction. In real-world environments, we often encounter the problem of incomplete modalities, which has become a core issue of conversation understanding. To address this problem, researchers propose various methods. However, existing approaches are mainly designed for individual utterances rather than conversational data, which cannot fully exploit temporal and speaker information in conversations. To this end, we propose a novel framework for incomplete multimodal learning in conversations, called "Graph Complete Network (GCNet)," filling the gap of existing works. Our GCNet contains two well-designed graph neural network-based modules, "Speaker GNN" and "Temporal GNN," to capture temporal and speaker dependencies. To make full use of complete and incomplete data, we jointly optimize classification and reconstruction tasks in an end-to-end manner. To verify the effectiveness of our method, we conduct experiments on three benchmark conversational datasets. Experimental results demonstrate that our GCNet is superior to existing state-of-the-art approaches in incomplete multimodal learning.**

***Index Terms*—Conversational data, graph complete network (GCNet), incomplete multimodal learning, speaker-sensitive modeling, temporal-sensitive modeling.**

## I. Introduction

Conversation understanding has become an active research area due to its widespread applications in many tasks, including dialogue systems [1], [2] and recommender systems [3], [4]. To understand conversations from different aspects (such as emotion and content), researchers collect a large amount of conversational data through various approaches, especially from social media platforms [5], [6]. However, in real-world environments, many factors may lead to missing modalities. For example, the speech is probably missing due to background noise or sensor failure; the text is perhaps unavailable due to automatic speech recognition errors or unknown words; the faces may not be detected due to lighting, motion, or occlusion. The problem of incomplete modalities increases the difficulty of understanding conversations accurately [7], [8], [9].

To this end, researchers propose various methods to deal with incomplete modalities. However, existing approaches are mainly designed for individual utterances or medical images rather than conversational data [10], [11]. For example, Pham et al. [10] took modality-incomplete utterances as the input. They learned utterance-level representations via cyclic translations to ensure robustness to missing modalities. To diagnose Alzheimer's disease from incomplete medical images, Suo et al. [12] jointly optimized image imputation and metric learning. Liu et al. [13] predicted missing parts by distilling knowledge from complete data to incomplete data, resulting in better performance on medical images with incomplete modalities. Unlike individual utterances or medical images, conversations contain rich temporal and speaker information [14], [15]. On the one hand, adjacent utterances are usually semantically related in a conversation. On the other hand, each speaker has their means of expression, generally consistent in a conversation. However, existing works usually fail to exploit them [16], [17], thus limiting their performance in conversational data.

In this paper, we propose a novel framework for incomplete multimodal learning in conversations called "Graph Complete Network (GCNet)". We aim to take full advantage of temporal and speaker information in conversations to deal with incomplete modalities. Fig. 1 shows the overall structure of our method. Specifically, we first randomly discard multimodal features to mimic real-world missing patterns [16], [18]. To capture speaker and temporal dependencies in conversations, we propose two graph neural network-based modules, "Speaker GNN (SGNN)" and "Temporal GNN (TGNN)". These two modules share the same edges but different edge types. Finally, we jointly optimize classification and reconstruction in an end-to-end manner. To verify the effectiveness of our method, we conduct experiments on three benchmark conversational datasets. Through quantitative and qualitative analysis, we prove that our GCNet outperforms currently advanced approaches in

Manuscript received 4 March 2022; revised 20 September 2022; accepted 29 December 2022. Date of publication 1 June 2023; date of current version 5 June 2023. This work was supported in part by the National Key Research and Development Plan of China under Grant 2020AAA0140003, in part by the National Natural Science Foundation of China (NSFC) under Grants 62201572, 61831022, 62276259, and U21B2010, and in part by the Open Research Projects of Zhejiang Lab under Grant 2021KH0AB06. Recommended for acceptance by L. Li. *(Corresponding authors: Jianhua Tao; Bin Liu.)*

Zheng Lian, Lan Chen, and Bin Liu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lianzheng2016@ia.ac.cn; chenlan2016@ia.ac.cn; liubin@nlpr.ia.ac.cn).

Licai Sun is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: sunlicai2019@ia.ac.cn).

Jianhua Tao is with the Department of Automation, Tsinghua University, Bejing 100084, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jhtao@tsinghua.edu.cn).

Code is available at https://github.com/zeroQiaoba/GCNet.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPAMI.2023.3234553.

Digital Object Identifier 10.1109/TPAMI.2023.3234553

0162-8828 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

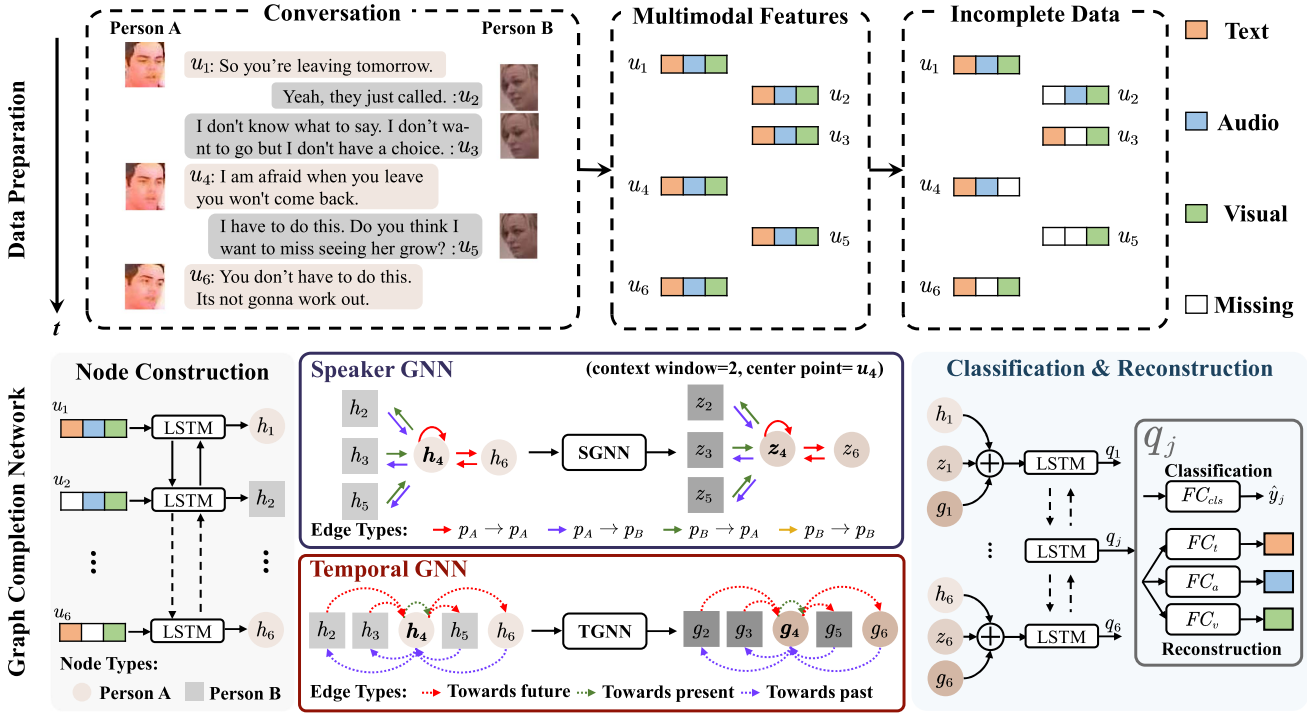



Fig. 1. The overall structure of Graph Complete Network (GCNet) with the trimodal setting ($M = 3$). This model focuses on the classification task based on incomplete conversational data.

both classification and imputation. The main contribution of this paper can be summarized as follows:

- Unlike existing works that mainly focus on medical images or individual utterances, we study the problem of incomplete modalities on conversational data, filling the gap of current works.
- We design a novel framework, GCNet, to deal with incomplete conversational data. This model leverages graph neural networks to capture temporal and speaker information in conversations.
- Experimental results on three benchmark datasets verify the effectiveness of our method. GCNet is superior to existing state-of-the-art approaches in incomplete multimodal learning.

The remainder of this paper is organized as follows: In Section II, we briefly review some recent works on incomplete multimodal learning. In Section III, we propose a novel framework for conversational data with incomplete modalities. In Section IV, we introduce experimental datasets and setup in detail. In Section V, we conduct experiments to verify the effectiveness of our method. Finally, we conclude this paper and discuss future work in Section VI.

## II. Related Works

Learning from incomplete multimodal data is a fundamental research area in machine learning. A straightforward approach is to conduct data imputation and then utilize existing classification methods on the imputed data. In addition to imputation methods, there are also some strategies that can directly conduct learning without imputation.

### A. Imputation Methods

Imputation methods attempt to estimate missing data from partially observed input. We review previous works and roughly divide them into three groups: zero/average imputation, low-rank imputation and DNN-based imputation.

*Zero/Average Imputation:* Padding missing modalities with zero vectors or average values are widely utilized for data imputation [16], [19], [20]. For example, Parthasarathy et al. [19] filled missing frames of videos with zero vectors. Zhang et al. [16] padded missing modalities with average values based on the available samples within the same class. Zero/average imputation can achieve competitive performance in incomplete multimodal learning. However, since no supervision information is utilized, there is still a gap between filled values and original data, thus degrading the performance of downstream tasks.

*Low-Rank Imputation:* Complete multimodal data exhibits correlations between different modalities and leads to the low-rank data matrix. However, incomplete data breaks these correlations and increases tensor rank [21], [22]. To capture multimodal correlations, previous works project data into a common space by using low-rankness. These approaches are usually based on nuclear norm minimization, such as singular value thresholding (SVT) [23] and Soft-Impture [24]. Besides nuclear norm, Fan et al. [25] also minimized tensor tubal rank to deal with various missing patterns. Furthermore, Liang et al. [22] combined the strength of non-linear functions to learn complex correlations in

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

tensor rank minimization. However, these methods are usually computationally expensive for Big Data [26].

*DNN-Based Imputation:* Due to the generative ability of DNNs, several DNN-based models have emerged to estimate missing data from partially observed input, e.g., autoencoder [27], [28], GAN [29], [30], VAE [31], [32] and Transformer [33], [34]. Among these approaches, autoencoder and its variants are widely utilized due to their promising results in incomplete multimodal learning [17]. For example, Duan et al. [35] leveraged autoencoders to impute missing data. To improve the modeling ability of autoencoders, Tran et al. [28] proposed the cascaded residual autoencoder (CRA). It combined a series of residual autoencoders [36] into a cascaded architecture for data imputation. Furthermore, Zhao et al. [17] incorporated CRA with cycle consistency loss for cross-modal imputation, which achieved superior performance over existing methods.

### B. Non-Imputation Methods

Existing non-imputation methods can be roughly divided into grouping strategies, correlation maximization and encoderless models.

*Grouping Strategy:* Complete data is easier to deal with than incomplete data. The grouping strategy directly partitions incomplete data into multiple complete subgroups, and then feature learning is carried out independently for each subgroup [37], [38], [39]. Despite its effectiveness, the number of subgroups grows exponentially with the number of modalities. Therefore, this strategy cannot work well for data with a large number of modalities or limited samples.

*Correlation Maximization:* To deal with the problem of incomplete data, an efficient approach is to maximize correlations between different modalities. In this way, we can constrain different modalities of the same sample to have related low-dimensional representations. Recently, several works based on correlation maximization have been proposed, including canonical correlation maximization [40], [41], HGR correlation maximization [42], mutual information maximization [43] and likelihood maximization [20]. Among these approaches, canonical correlation and its variants are widely utilized due to their promising results in incomplete multimodal learning. For example, Hotelling et al. [40] proposed CCA that learned relationships between multi-modalities by linearly mapping them into a low-dimensional common space with maximal canonical correlations. Different from CCA that focused on linear mappings, Andrew et al. [41] proposed DCCA that leveraged deep neural networks to learn more complex nonlinear combinations between multi-modalities. Wang et al. [44] further combined canonical correlations with reconstruction errors of autoencoders, trading off the structure information of each modality and the relationship between multi-modalities.

*Encoderless Model:* Unlike previous works that rely on encoders, encoderless models can learn latent representations without encoders. They directly optimize latent representations to reconstruct modality-incomplete data regardless of missing patterns [45], [46]. Typically, Zhang et al. [16] proposed CPM-Net, a robust encoderless model for incomplete multimodal learning. It combined the encoderless model with a clustering-like classification loss to learn well structured features, which has validated its effectiveness on multimodal data with missing modalities.

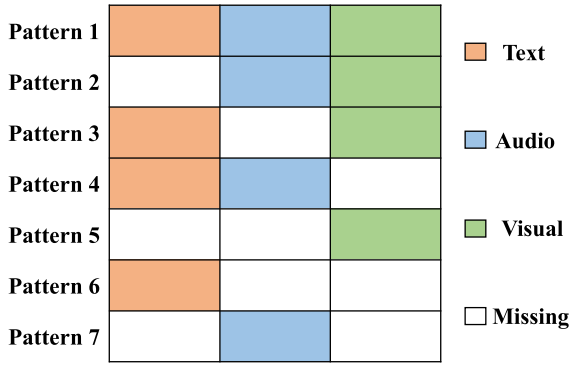



Fig. 2. Seven missing patterns for a trimodal dataset.

## III. Methodology

In this paper, we focus on the classification task based on the conversational data with missing modalities. Each conversation contains multiple continuous utterances. For each utterance, we first extract multimodal features with random missing patterns. Then we propose a novel graph neural network-based framework, GCNet, to deal with incomplete conversational data. Fig. 1 shows the overall structure of our proposed method.

### A. Data Preparation

Let us define a conversation $C = \{(u_i, y_i)\}_{i=1}^{L}$, where $L$ is the number of utterances in the conversation, $u_i$ is the $i^{th}$ utterance in $C$ and $y_i$ is the true label of $u_i$. Here, $y_i \in \{1, 2, \ldots, c\}$ and $c$ is the total number of labels. Each utterance $u_i$ is uttered by the speaker $p_{s(u_i)}$, where the function $s(\cdot)$ maps the index of utterance into its corresponding speaker. For each utterance $u_i$, we extract multimodal features $x_i = \{x_i^m\}_{m\in\{a,l,v\}}$. Here, $x_i^a \in \mathbb{R}^{d_a}$, $x_i^l \in \mathbb{R}^{d_l}$ and $x_i^v \in \mathbb{R}^{d_v}$ are the utterance-level features of acoustic, lexical and visual modalities, respectively. And $\{d_m\}_{m\in\{a,l,v\}}$ is the feature dimension of each modality.

To mimic real-world missing scenarios, we randomly discard some modalities by guaranteeing at least one modality is available for each sample, in line with previous works [16], [18]. Therefore, an incomplete $M$-modal dataset has $(2^M - 1)$ different missing patterns. Fig. 2 illustrates a trimodal ($M = 3$) dataset with seven missing patterns. Suppose $\sigma_i$ is the missing pattern of $u_i$ and $\phi(\cdot)$ is a function that maps each missing pattern to its available modalities. The incomplete representation of $u_i$ is denoted as $\widetilde{x}_i = \{\lambda_i^m x_i^m\}_{m\in\{a,l,v\}}$, where $\lambda_i^m$ is defined as follows:

$$\lambda_i^m = \begin{cases} 1, m \in \phi(\sigma_i) \\ 0, m \notin \phi(\sigma_i) \end{cases}. \tag{1}$$

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

### B. Graph Completion Network

Recent works have verified the importance of temporal and speaker information in conversations [47], [48]. To this end, we leverage this information to improve classification performance on incomplete conversational data. Our method consists of three key modules: Node Construction, Speaker & Temporal GNN and Classification & Reconstruction.

*1) Node Construction:* We take the conversational data with missing modalities as the input. Each conversation contains multiple utterances, and we represent each utterance $u_i$ as a node $v_i$. To extract initial representations of $v_i$, we first concatenate incomplete multimodal features $\widetilde{x}_i = \{\lambda_i^m x_i^m\}_{m\in\{a,l,v\}}$, followed with bi-directional long short-term memory (Bi-LSTM) to capture contextual information. Bi-LSTM consists of gating mechanisms to control the flow of information, which is capable of modeling long-term contextual dependencies in both forward and backward directions

$$f_i = Concat\left(\lambda_i^a x_i^a, \lambda_i^l x_i^l, \lambda_i^v x_i^v\right), \tag{2}$$

$$H = BiLSTM\left(F, \theta_h\right), \tag{3}$$

where $F = \{f_i\}_{i=1}^L \in \mathbb{R}^{L\times(d_a+d_l+d_v)}$ and $H = \{h_i\}_{i=1}^L \in \mathbb{R}^{L\times d}$ are the input and the output of Bi-LSTM, respectively. Here, $d$ is the feature dimension of output features. And $\theta_h$ is the trainable parameters. Finally, $H = \{h_i\}_{i=1}^L$ is utilized as the initial node representations.

*2) Speaker & Temporal GNN:* Speaker GNN (SGNN) and Temporal GNN (TGNN) are key modules for capturing speaker and temporal dependencies in conversations. In these modules, edges measure the importance of connections between nodes. Edge types define different aggregation approaches in node interactions [49], [50]. SGNN and TGNN share the same edges but distinct edge types to capture different dependencies. In this section, we describe these two modules in detail.

*Edges:* For each node, its interaction with the context nodes should be considered. If each node $v_i$ interacts with all context nodes $\{v_j\}_{j\in[1,L]}$, the designed graph will contain a large number of edges, making it difficult to store and optimize. Inspired by previous works that most utterances focus on their local context [51], we fix the size of the context window to $w$, and each node $v_i$ only interacts with the nodes within the context window $\{v_j\}_{j\in[\max(i-w,1),\min(i+w,L)]}$. In our implementation, we choose $w$ from $\{1, 2, 3, 4\}$. And we use $e_{ij}$ to represent the edge from $v_i$ to $v_j$.

*Speaker Types:* SGNN leverages speaker types to capture speaker-sensitive dependencies in conversations. Specifically, we assign each edge $e_{ij}$ with a speaker type identifier $\alpha_{ij} \in \boldsymbol{\alpha}$. Here, $\boldsymbol{\alpha}$ denotes the set of speaker types, and $|\boldsymbol{\alpha}|$ is the number of $\boldsymbol{\alpha}$. For each edge $e_{ij}$, $\alpha_{ij}$ is set to $(p_{s(u_i)} \rightarrow p_{s(u_j)})$, where $p_{s(u_i)}$ and $p_{s(u_j)}$ represent the speaker identity of $u_i$ and $u_j$, respectively. If there are $S$ distinct speakers in a conversation, there can be a maximum of $|\boldsymbol{\alpha}| = S^2$ distinct speaker types.

*Temporal Types:* TGNN leverages temporal types to capture temporal-sensitive dependencies in conversations. Specifically, we assign each edge $e_{ij}$ with a temporal type identifier $\beta_{ij} \in \boldsymbol{\beta}$. Here, $\boldsymbol{\beta}$ denotes the set of temporal types, and $|\boldsymbol{\beta}|$ is the number of $\boldsymbol{\beta}$. Depending on the relative position of occurrence of $v_i$ and $v_j$ in a conversation, we determine the value of $\beta_{ij}$ to be either of {past, present, future}, resulting $|\boldsymbol{\beta}| = 3$.

*Graph Learning:* Recent works have verified the effectiveness of relation graph convolutional networks (R-GCN) in relation modeling problems (such as link prediction and entity classification) [52]. Inspired by its success, we leverage R-GCN to aggregate the neighborhood information in the graph. For SGNN and TGNN, we pass information via edges with parameters dependent on edge types. The calculation formula is shown as follows:

$$z_i = \sigma\left(\sum_{r\in\boldsymbol{\alpha}}\sum_{j\in N_i^r}\frac{1}{|N_i^r|}W_r^s h_j\right), \tag{4}$$

$$g_i = \sigma\left(\sum_{r\in\boldsymbol{\beta}}\sum_{j\in N_i^r}\frac{1}{|N_i^r|}W_r^t h_j\right), \tag{5}$$

where $z_i \in \mathbb{R}^h$ and $g_i \in \mathbb{R}^h$ are the outputs of SGNN and TGNN, respectively. Here, $N_i^r$ represents the set of neighboring indexes of the node $v_i$ under relation $r$, and $|N_i^r|$ is the number of $N_i^r$. $\sigma(\cdot)$ is the activation function, and we choose $ReLU(\cdot)$ is our implementation. $W_r^s$ and $W_r^t$ are the trainable parameters for SGNN and TGNN under relation $r$, respectively.

*3) Classification & Reconstruction:* For each node $v_i$, we extract its initial representation $h_i$, a representation $z_i$ considering speaker information, and a representation $g_i$ considering temporal information. To aggregate these representations, we concatenate them together and then utilize Bi-LSTM for context-sensitive modeling

$$\hat{h}_i = Concat\left(h_i, z_i, g_i\right), \tag{6}$$

$$Q = BiLSTM\left(\hat{H}, \theta_q\right), \tag{7}$$

where $\hat{H} = \{\hat{h}_i\}_{i=1}^L \in \mathbb{R}^{L\times(3h)}$ and $Q = \{q_i\}_{i=1}^L \in \mathbb{R}^{L\times h}$ are the input and the output of Bi-LSTM, respectively. Here, $\theta_q$ is the trainable parameters.

*Classification:* To learn more discriminative features for conversation understanding, we feed the latent representations $Q = \{q_i\}_{i=1}^L$ into a fully-connected layer, followed by a softmax layer to estimate classification probabilities:

$$\hat{Y} = softmax(QW_c + b_c), \tag{8}$$

where $\hat{Y} = \{\hat{y}_i\}_{i=1}^L \in \mathbb{R}^{L\times c}$ is the estimated probabilities, where $c$ is the number of discrete labels in the corpus. Here, $W_c \in \mathbb{R}^{d\times c}$ and $b_c \in \mathbb{R}^c$ are the trainable parameters.

*Reconstruction:* Reconstructing complete data from the latent space can guide the model to learn the semantics of missing parts [34]. Therefore, we propose a modality-specific reconstruction module. For each modality $m \in \{a, l, v\}$, we perform a linear transformation mapping the extracted features into the input space

$$\hat{X}^m = QW_m + b_m, m \in \{a, l, v\}, \tag{9}$$

where $\hat{X}^m = \{\hat{x}_i\}_{i=1}^L \in \mathbb{R}^{L\times d_m}$ is the estimated complete data. $W_m \in \mathbb{R}^{d\times d_m}$ and $b_c \in \mathbb{R}^{d_m}$ are the trainable parameters, where $d_m$ is the feature dimension for each modality.

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

*4) Joint Optimization:* To leverage both complete and incomplete data, we jointly optimize classification and imputation in an end-to-end manner. Therefore, our loss function consists of two parts: the classification loss $L_{cls}$ and the reconstruction loss $L_{rec}$.

Minimizing the classification loss ensures learning more discriminative features, making the margins between different classes more explicit. During training, we choose the cross-entropy loss as the classification loss. The calculation formula is shown as follows:

$$L_{cls} = -\frac{1}{L}\sum_{i=1}^{L} y_i \log(\hat{y}_i), \tag{10}$$

where $y_i \in \mathbb{R}^c$ is the true one-hot label.

To better estimate missing data from partially observed input, we calculate the reconstruction loss between the original and filled features at the missing positions

$$L_{rec} = \sum_{m\in\{a,l,v\}} \frac{1}{d_m L}\sum_{i=1}^{L} \|(1-\lambda_i^m)(\hat{x}_i^m - x_i^m)\|^2, \tag{11}$$

where $\lambda_i^m$ reveals available modalities for each utterance $u_i$, as defined in (1). Therefore, $(1-\lambda_i^m)$ is the mask revealing missing positions.

Finally, we combine these two loss functions into a joint objective function. This joint loss is utilized to optimize all trainable parameters in an end-to-end manner.

$$L = L_{cls} + L_{rec}. \tag{12}$$

This paper assumes that modality-complete data is available during training, consistent with previous works [17], [28]. In practice, we first collect some modality-complete data using specific sensors. The inability to collect any such data rarely happens, which is left for our future work. Then, we can obtain a robust classifier for missing conditions by jointly optimizing the classification and reconstruction modules. During inference, we only need to leverage the pretrained classification module for conversation understanding. Therefore, although the raw information of missing modalities is unavailable in real-world scenarios, it does not affect our inference process.

## IV. Experimental Databases and Setup

In this section, we first describe three benchmark conversational datasets in our experiments. Following that, we illustrate evaluation metrics, implementation details and multimodal features in detail. Finally, we introduce various currently advanced baselines in incomplete multimodal learning for comparison.

### A. Corpus Description

To verify the effectiveness of GCNet, we conduct experiments on three benchmark conversational datasets, including IEMOCAP [53], CMU-MOSI [5] and CMU-MOSEI [6].

*IEMOCAP* contains conversations where two actors perform improvised or scripted scenarios, especially chosen to evoke emotional expressions. These conversations are divided into five sessions. Each conversation is segmented into multiple utterances and each utterance is annotated with categorical labels. For a fair comparison, we follow two popular label process manners, resulting in four-class [47], [54] and six-class [48], [55] datasets. Since no predefined data split manner is provided, we perform five-fold cross-validation using the leave-one-session-out strategy, in line with previous works [56], [57]. We calculate the statistics of each session in Table I.

TABLE I
Statistical Information on IEMOCAP

| Dataset | | # conversations | # utterances |
|---|---|---|---|
| IEMOCAP (four-class) | Session1 | 28 | 1085 |
| | Session2 | 30 | 1023 |
| | Session3 | 32 | 1151 |
| | Session4 | 30 | 1031 |
| | Session5 | 31 | 1241 |
| IEMOCAP (six-class) | Session1 | 28 | 1373 |
| | Session2 | 30 | 1356 |
| | Session3 | 32 | 1569 |
| | Session4 | 30 | 1512 |
| | Session5 | 31 | 1623 |

TABLE II
Statistical Information on CMU-MOSI and CMU-MOSEI

| Dataset | # conversations | | | # utterances | | |
|---|---|---|---|---|---|---|
| | train | val | test | train | val | test |
| CMU-MOSI | 52 | 10 | 31 | 1284 | 229 | 686 |
| CMU-MOSEI | 2249 | 300 | 676 | 16326 | 1871 | 4659 |

*CMU-MOSI* is a collection of movie review videos from online websites. To reflect sentiment intensity, each utterance is annotated with a sentiment score from $-3$ (extremely negative sentiment) to +3 (extremely positive sentiment).

*CMU-MOSEI* is created by extending CMU-MOSI with more utterances and a greater variety of topics. Following the annotation manner in CMU-MOSI, each utterance is annotated with a sentiment score between $[-3, 3]$. Since train/validation/test splits are provided in CMU-MOSI and CMU-MOSEI, we utilize the official dataset split manner for a fair comparison. The statistics are shown in Table II.

### B. Evaluation Metrics

To evaluate GCNet against previous methods, we choose the following evaluation metrics for a fair comparison.

*IEMOCAP* is labeled in categorical labels. Due to the natural imbalance across different classes [53], [58], we use weighted average F1-score (WAF) as our evaluation metric, in line with previous works [47], [48].

*CMU-MOSI and CMU-MOSEI* are annotated in continuous sentiment scores between $[-3, 3]$. In this paper, we focus on the negative/positive classification task. Positive and negative classes are assigned for $< 0$ and $> 0$ scores, respectively. We also utilize WAF as our evaluation metric, in line with previous works [59], [60].

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

### C. Implementation Details

We investigate the performance of different methods on multimodal datasets with varying missing rates. The missing rate is defined as $\eta = 1 - \frac{\sum_{i=1}^{L} m_i}{L \times M}$, where $m_i$ indicates the number of available modalities for the $i^{th}$ sample. Here, $L$ denotes the total number of samples and $M$ represents the total number of modalities. For each sample with $M$ modalities, we randomly select the missing modalities with the probability $\eta$. We also guarantee that at least one modality is available for each sample, resulting $m_i \geqslant 1$ and $\eta \leqslant \frac{M-1}{M}$. Therefore, if the number of modalities is set to $M = 3$, we choose the missing rate $\eta$ from $[0.0, 0.1, \ldots, 0.7]$, where 0.7 is an approximation of $\frac{M-1}{M}$ with the same effect. We keep the same missing rate during training, validation and testing periods, in line with previous works [16], [34].

Since train/validation/test splits are provided in the CMU-MOSI and CMU-MOSEI datasets, we select the best model on the validation set and report its performance on the test set. For the IEMOCAP dataset, we turn all parameters with five-fold cross-validation. There are two user-specific parameters in our GCNet, i.e., the dimension of latent representations $h$ and the context window size $w$. We select $h$ from $\{50, 100, 200\}$ and $w$ from $\{1, 2, 3, 4\}$ for all datasets. During training, we use the Adam optimization scheme with a learning rate of 0.001 and a weight decay of 0.00001. To alleviate the over-fitting problem, Dropout [61] is also utilized with a rate of $p = 0.5$. Since our GCNet takes conversation-level features as the input, we pad conversations of the same mini-batch to have the same number of utterances. Bit masking is also implemented to eliminate the effect of padded parts. To evaluate the performance of different methods, we run each experiment ten times and report the average values on the test set.

### D. Multimodal Feature Extraction

For each utterance, we extract multimodal features from acoustic, lexical and visual modalities. The multimodal feature extraction process is described as follows:

*Acoustic Modality:* Pre-trained wav2vec [62] is used as the acoustic feature extractor. This model is a multi-layer convolutional neural network trained on a large amount of unlabeled data. Inspired by its success in many downstream tasks [63], we leverage the pre-trained *wav2vec-large*[1] to extract 512-dimensional acoustics features for each utterance.

*Lexical Modality:* Pre-trained DeBERTa [64] is exploited as the lexical feature extractor. This model improves BERT [65] and RoBERTa [66] using the disentangled attention mechanism and the enhanced mask decoder. Inspired by its success in both natural language understand and natural language generation [64], we leverage the pre-trained *DeBERTa-large*[2] to encode word sequences into 1,024-dimensional lexical features for each utterance.

*Visual Modality:* Pre-trained MA-Net [67] is utilized as the visual feature extractor. This model leverages global multi-scale and local attention to address occlusions and non-frontal poses. Inspired by its success in facial emotion recognition, we use the MTCNN face detection algorithm to extract aligned faces [68] and then employ the pre-trained MA-Net[3] for facial feature extraction. Finally, we compress frame-level facial features into 1024-dimensional utterance-level features by average encoding.

### E. Baselines

To evaluate the performance of our proposed GCNet, we implement the following state-of-the-art incomplete multimodal learning methods as the baselines.

*CCA* [40] is a strong benchmark model. To deal with the problem of incomplete data, it learns relationships across different modalities by linearly mapping them into low-dimensional common space with maximum correlations.

*DCCA* [41] is an extension of CCA. CCA focuses on linear combinations of different modalities. Differently, DCCA leverages deep neural networks to learn more complex nonlinear combinations between multi-modalities.

*DCCAE* [44] is another extension of CCA. It employs autoencoders to learn hidden features for each modality, and then optimizes both reconstruction errors of autoencoders and canonical correlations. This objective function trades off the structure information of each modality and the inherent association between multiple modalities.

*AE* [69] is widely utilized in incomplete multimodal learning [35], [70]. It leverages autoencoders to impute missing data from partially observed input. Followed with previous works [17], [34], we jointly optimize the reconstruction loss of autoencoders and the classification loss of downstream tasks in our implementation.

*CRA* [28] is an extension of AE. It combines a series of residual autoencoders into a cascaded architecture for data imputation. In our implementation, we jointly optimize imputation and downstream tasks end-to-end.

*MMIN* [17] is another extension of AE. It combines CRA with cycle consistency learning to predict latent representations of missing modalities. This model is a strong benchmark model, which outperforms other approaches under varying missing conditions.

*CPM-Net* [16] jointly considers completeness and structure to learn discriminative latent representations. As for completeness, it designs an encoderless model that projects all samples into a common space regardless of missing patterns. As for structuring, it learns well-structured features by equipping with a clustering-like classification loss.

## V. Results and Discussion

In this paper, we propose a novel framework, GCNet, for incomplete multimodal learning in conversations. To verify the effectiveness of our method, we first conduct comparative

[1][Online]. Available: https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

[2][Online]. Available: https://huggingface.co/microsoft/deberta-large

[3][Online]. Available: https://github.com/zengqunzhao/MA-Net

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPARISON OF CLASSIFICATION PERFORMANCE WITH DIFFERENT MISSING RATES

| Dataset | Method | Missing Rate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | Average |
| IEMOCAP (four-class) | CCA [40] | 64.52 | 65.19 | 62.60 | 59.35 | 55.25 | 51.38 | 45.73 | 30.61 | 54.33 |
| | DCCA [41] | 60.03 | 57.25 | 51.74 | 42.53 | 36.54 | 34.82 | 33.65 | 41.09 | 44.71 |
| | DCCAE [44] | 63.42 | 61.66 | 57.67 | 54.95 | 51.08 | 45.71 | 39.07 | 41.42 | 51.87 |
| | CPM-Net [16] | 58.00 | 55.29 | 53.65 | 52.52 | 51.01 | 49.09 | 47.38 | 44.76 | 51.46 |
| | AE [69] | 74.82 | 71.36 | 67.40 | 62.02 | 57.24 | 50.56 | 43.04 | 39.86 | 58.29 |
| | CRA [28] | 76.26† | 71.28 | 67.34 | 62.24 | 57.04 | 49.86 | 43.22 | 38.56 | 58.23 |
| | MMIN [17] | 74.94 | 71.84† | 69.36† | 66.34† | 63.30† | 60.54† | 57.52† | 55.44† | 64.91† |
| | **GCNet** | **78.36** | **77.48** | **77.34** | **76.22** | **75.14** | **73.80** | **71.88** | **71.38** | **75.20** |
| | $\Delta_{SOTA}$ | ↑2.10 | ↑5.64 | ↑7.98 | ↑9.88 | ↑11.84 | ↑13.26 | ↑14.36 | ↑15.94 | ↑10.29 |
| IEMOCAP (six-class) | CCA [40] | 43.04 | 46.06 | 43.86 | 41.66 | 37.13 | 34.94 | 32.06 | 21.80 | 37.57 |
| | DCCA [41] | 42.18 | 39.15 | 34.47 | 27.65 | 23.69 | 22.86 | 22.71 | 27.38 | 30.01 |
| | DCCAE [44] | 46.19 | 43.77 | 41.28 | 37.98 | 34.58 | 30.02 | 26.78 | 27.66 | 36.03 |
| | CPM-Net [16] | 41.05 | 37.33 | 36.22 | 35.73 | 35.11 | 33.64 | 32.26 | 31.25 | 35.32 |
| | AE [69] | 56.76 | 52.82 | 48.66 | 42.26 | 35.18 | 29.12 | 25.08 | 23.18 | 39.13 |
| | CRA [28] | **58.68** | 53.50 | 49.76 | 45.88 | 39.94 | 32.88 | 28.08 | 26.16 | 41.86 |
| | MMIN [17] | 56.96 | 53.94† | 51.46† | 48.42† | 45.60† | 42.82† | 40.18† | 37.84† | 47.15† |
| | **GCNet** | 58.64† | **58.50** | **57.64** | **57.08** | **56.12** | **54.40** | **53.60** | **53.46** | **56.18** |
| | $\Delta_{SOTA}$ | ↓0.04 | ↑4.56 | ↑6.18 | ↑8.66 | ↑10.52 | ↑11.58 | ↑13.42 | ↑15.62 | ↑9.03 |
| CMU-MOSI | CCA [40] | 74.74 | 71.60 | 68.84 | 65.71 | 62.13 | 60.37 | 59.92 | 53.26 | 64.57 |
| | DCCA [41] | 67.60 | 65.67 | 61.46 | 58.81 | 54.43 | 52.77 | 50.13 | 53.29 | 58.02 |
| | DCCAE [44] | 66.76 | 64.21 | 62.61 | 59.23 | 59.61 | 53.56 | 51.87 | 53.49 | 58.92 |
| | CPM-Net [16] | 71.90 | 68.91 | 71.12 | 70.59 | 67.95 | 65.88 | 64.02 | 61.79† | 67.77 |
| | AE [69] | **85.86** | 82.27 | 78.43 | 74.00 | 69.83 | 66.62 | 60.22 | 55.64 | 71.61 |
| | CRA [28] | 85.68† | **82.61** | 78.53† | 75.12† | 70.20† | 67.40 | 62.40 | 59.40 | 72.67 |
| | MMIN [17] | 85.20 | 81.91 | 78.22 | 74.60 | 70.14 | 67.72† | 64.04† | 61.53 | 72.92† |
| | **GCNet** | 85.01 | 82.54† | **80.17** | **78.54** | **76.48** | **73.45** | **69.46** | **68.35** | **76.75** |
| | $\Delta_{SOTA}$ | ↓0.85 | ↓0.07 | ↑1.64 | ↑3.42 | ↑6.28 | ↑5.73 | ↑5.42 | ↑6.56 | ↑3.83 |
| CMU-MOSEI | CCA [40] | 81.23 | 79.06 | 78.09 | 76.25 | 74.62 | 73.22 | 70.57 | 56.29 | 73.67 |
| | DCCA [41] | 74.06 | 70.11 | 64.69 | 61.31 | 61.11 | 60.47 | 58.27 | 59.76 | 63.72 |
| | DCCAE [44] | 75.70 | 74.67 | 73.16 | 71.82 | 70.26 | 67.86 | 64.15 | 62.75 | 70.05 |
| | CPM-Net [16] | 78.47 | 74.79 | 74.48 | 73.81 | 72.39 | 70.43 | 68.73 | 67.07 | 72.52 |
| | AE [69] | 86.66† | 84.37† | 82.58† | 80.57† | 78.80† | 76.43† | 74.26† | 72.81† | 79.56† |
| | CRA [28] | 86.48 | 84.19 | 82.25 | 80.12 | 78.55 | 75.85 | 74.07 | 72.46 | 79.25 |
| | MMIN [17] | 85.78 | 83.77 | 81.85 | 79.77 | 77.63 | 75.36 | 72.95 | 71.18 | 78.54 |
| | **GCNet** | **87.12** | **86.50** | **85.50** | **84.53** | **83.55** | **82.44** | **80.27** | **80.20** | **83.76** |
| | $\Delta_{SOTA}$ | ↑0.46 | ↑2.13 | ↑2.92 | ↑3.96 | ↑4.75 | ↑6.01 | ↑6.01 | ↑7.39 | ↑4.20 |

We report waf scores(%), and higher waf indicates better performance. the best performance is highlighted in bold, and the second-highest result is labeled by y. the row with $\Delta_{sota}$ means the improvements or reductions of gcnet compared to existing state-of-the-art systems.

experiments with currently advanced approaches, investigating the classification and imputation performance under different missing rates. Then, we show the importance of incomplete data in multimodal learning. Next, we reveal the necessity of each component in GCNet, including SGNN for speaker-sensitive modeling and TGNN for temporal-sensitive modeling. After that, we reveal the impact of different hyper-parameters and study the convergence property. To qualitatively analyze the improvement of our proposed method, we further visualize latent representations and conduct case studies.

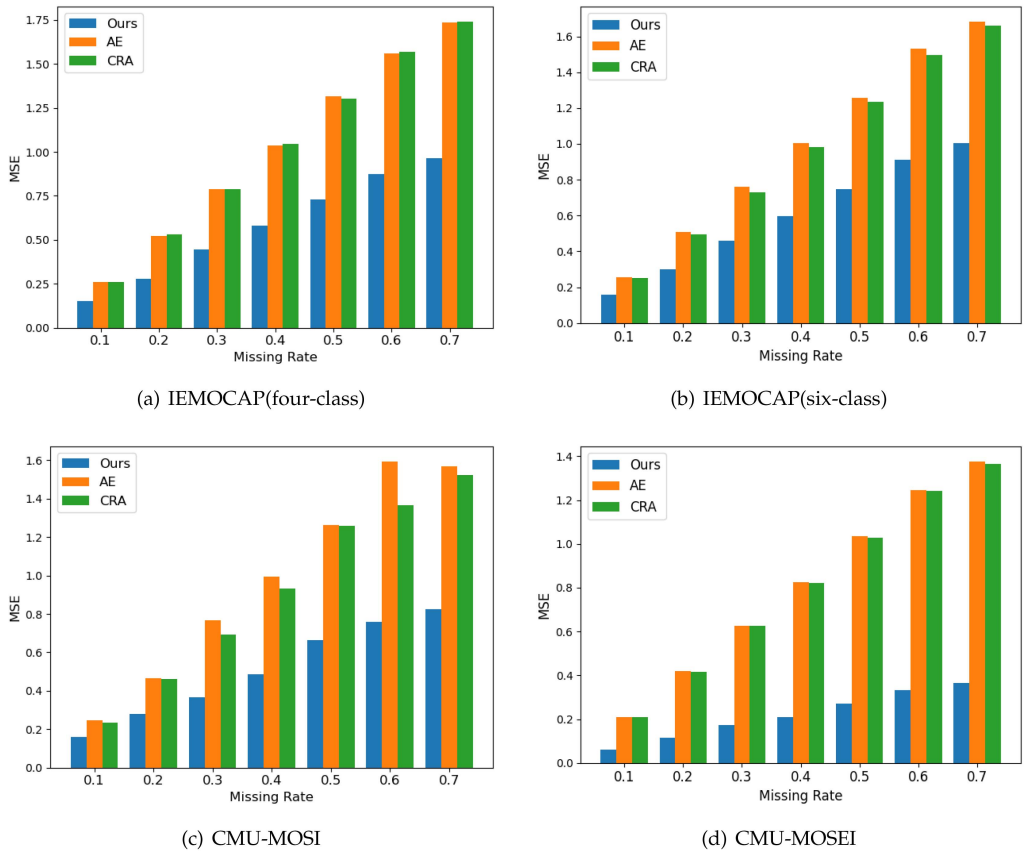
### A. Classification Performance

Table III presents the classification performance of different approaches under varying missing rates. From these experimental results, we have the following observations:

1) On average, our method achieves the best performance on all datasets. For the IEMOCAP(four-class) dataset, GCNet succeeds over currently advanced approaches by 10.29%. For the IEMOCAP(six-class) dataset, GCNet achieves the new state-of-the-art record 56.18%, which shows an absolute improvement of 9.03%. We also observe a similar phenomenon on the CMU-MOSI and CMU-MOSEI datasets. These quantitative results verify the effectiveness of our method in incomplete multimodal learning. We argue that these baselines do not explicitly model temporal and speaker dependencies in conversations, which are essential for conversation understanding. Differently, the proposed method leverages graph neural networks to capture these dependencies, resulting in better classification performance.

2) Experimental results in Table III show that the performance degradation of our method is much smaller than that of baselines as the missing rate increases. Taking the results on the

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.




(a) IEMOCAP(four-class) (b) IEMOCAP(six-class)

(c) CMU-MOSI (d) CMU-MOSEI

Fig. 3. Comparison of imputation performance with different missing rates. Lower MSE indicates better imputation performance.

IEMOCAP(four-class) dataset as an example, as the missing rate increases from 0.0 to 0.7, the performance of baselines declines 13.24%∼37.70% while our GCNet only declines 6.98%. More notably, the performance gap between GCNet and baselines becomes more significant as the missing rate increases. Taking the results on the IEMOCAP(four-class) dataset as an example, the performance gap increases from 2.10% to 15.94% as the missing rate increases from 0.0 to 0.7. These results verify that GCNet improves classification performance, especially in severely missing cases.

3) Experimental results in Table III demonstrate that our method also exhibits competitive performance on complete multimodal data ($\eta = 0.0$). For the IEMOCAP(four-class) dataset, our GCNet shows an absolute improvement of 2.10% over currently advanced approaches. For the CMU-MOSEI dataset, our proposed method succeeds over state-of-the-art baselines by 0.46%. Although our method achieves slightly worse performance than baselines on the CMU-MOSI and IEMOCAP(six-class) datasets, the difference is not significant. These results validate the effectiveness of our method on modality-complete data.

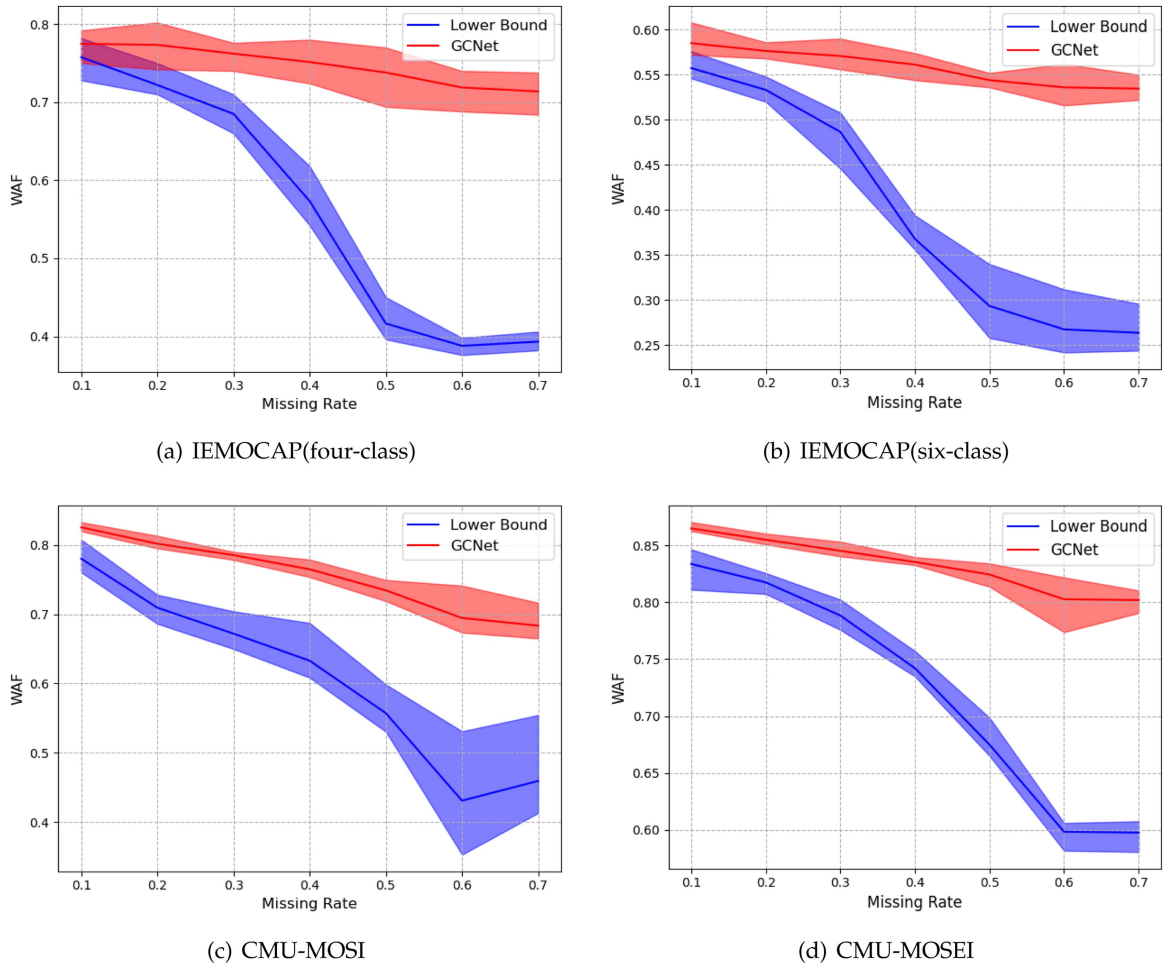
### B. Imputation Performance

GCNet leverages graph neural networks to impute missing data from partially observed input. To verify the effectiveness of our method, we compare the imputation performance of GCNet with two currently advanced imputation methods, *AE* [69] and *CRA* [28]. To evaluate the imputation performance of different methods, we calculate the mean square error (MSE) between original features and estimated features at the missing position.

Fig. 3 presents the imputation results of different methods under varying missing rates. Experimental results demonstrate that our GCNet consistently outperforms all baselines under all missing rates on all datasets. Speaker and temporal dependencies are crucial in data imputation. First, adjacent utterances in a conversation are usually semantically related. Second, each speaker has their means of expression, which is generally consistent in a conversation. But most baselines ignore these dependencies, thus degrading their imputation performance. Differently, our GCNet makes full use of speaker and temporal information via graph neural networks and achieves better imputation performance. These results verify the importance of speaker and temporal information in data imputation and the effectiveness of our method in incomplete multimodal learning.

Meanwhile, experimental results in Fig. 3 demonstrate that the imputation performance declines as the missing rate increases. It is reasonable because higher missing rates can result in fewer observed data, increasing the difficulty of data imputation. Meanwhile, the performance gap between GCNet and baselines becomes more significant as the missing rate increases. These results validate that our model is more robust to missing data than baselines.

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.




(a) IEMOCAP(four-class) (b) IEMOCAP(six-class)

(c) CMU-MOSI (d) CMU-MOSEI

Fig. 4. Classification performance comparison of GCNet and Lower Bound under different missing rates.

### C. Importance of Incomplete Data

Besides modality-complete data, GCNet also makes full use of modality-incomplete data during training. To investigate the importance of modality-incomplete data, we implement one comparison system. In Fig. 4, we compare the performance of different methods under varying missing rates.

- *GCNet:* It is our proposed method that leverages both modality-complete and modality-incomplete data in the learning process.
- *Lower Bound:* It comes from GCNet, but abandons modality-incomplete data and only focuses on modality-complete data. It is one of the most straightforward strategies to handle modality missing and is often regarded as the lower bound [20].

As shown in Fig. 4, GCNet consistently outperforms the comparison system with all missing rates on all datasets. It is reasonable because the comparison system does not use extra information in modality-incomplete data, resulting in a loss of available information.

Meanwhile, the performance gap between GCNet and the comparison system becomes more significant as the missing rate increases. The reason lies in that the comparison system will delete a large number of training samples, especially in severe missing cases. It is challenging to learn discriminative classifiers with limited data. Differently, our GCNet incorporates both modality-complete data and modality-incomplete data in the learning process. With more available training samples, we can use the complementary information in these two types of data and learn more discriminative representations for classification.

### D. Role of SGNN and TGNN

In GCNet, we capture speaker and temporal dependencies via SGNN and TGNN. To verify the necessity of these components, we implement three comparison systems. Experimental results are listed in Table IV.

- *GCNet:* It is our method that captures speaker and temporal dependencies via SGNN and TGNN.
- *GCNet-T:* It comes from GCNet but omits the SGNN component. Therefore, this model ignores speaker information in conversations.
- *GCNet-S:* It comes from GCNet but omits the TGNN component. Therefore, this model ignores temporal information in conversations.
- *GCNet-ST:* It comes from DialogueGCN [71], a currently advanced graphic model for conversation understanding.

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
ABLATION RESULTS FOR SGNN AND TGNN ON THE IEMOCAP DATASET

| $\eta$ | Model | IEMOCAP (four-class) | IEMOCAP (six-class) |
|---|---|---|---|
| 0.0 | GCNet | **78.36** | **58.64** |
| 0.0 | GCNet-T | 76.90 | 57.43 |
| 0.0 | GCNet-S | 77.65 | 58.03 |
| 0.0 | GCNet-ST | 77.86 | 58.60 |
| 0.1 | GCNet | **77.48** | **58.50** |
| 0.1 | GCNet-T | 75.60 | 57.06 |
| 0.1 | GCNet-S | 76.80 | 57.37 |
| 0.1 | GCNet-ST | 77.04 | 57.80 |
| 0.2 | GCNet | **77.34** | **57.64** |
| 0.2 | GCNet-T | 74.55 | 56.06 |
| 0.2 | GCNet-S | 75.80 | 56.94 |
| 0.2 | GCNet-ST | 76.70 | 56.64 |
| 0.3 | GCNet | **76.22** | **57.08** |
| 0.3 | GCNet-T | 74.05 | 55.14 |
| 0.3 | GCNet-S | 75.35 | 55.54 |
| 0.3 | GCNet-ST | 75.44 | 55.84 |
| 0.4 | GCNet | 75.14 | **56.12** |
| 0.4 | GCNet-T | 72.05 | 55.06 |
| 0.4 | GCNet-S | 74.25 | 55.20 |
| 0.4 | GCNet-ST | **75.20** | 55.40 |
| 0.5 | GCNet | **73.80** | 54.40 |
| 0.5 | GCNet-T | 72.10 | 54.09 |
| 0.5 | GCNet-S | 72.85 | 54.14 |
| 0.5 | GCNet-ST | 73.68 | **54.54** |
| 0.6 | GCNet | **71.88** | 53.60 |
| 0.6 | GCNet-T | 71.40 | 52.40 |
| 0.6 | GCNet-S | 71.60 | **53.74** |
| 0.6 | GCNet-ST | 71.80 | 52.94 |
| 0.7 | GCNet | **71.38** | **53.46** |
| 0.7 | GCNet-T | 71.20 | 51.11 |
| 0.7 | GCNet-S | 71.30 | 51.69 |
| 0.7 | GCNet-ST | 71.00 | 50.36 |

We report the classification performance of different models under varying missing rates $\eta$. the bold front represents the best performance.

Unlike our approach that captures speaker and temporal dependencies through two separate graphs, this comparison system captures "speaker-temporal" dependencies simultaneously in one graph. This coupling strategy will increase relation types that need to be optimized.

Experimental results in Table IV demonstrate that GCNet exhibits performance improvement over *GCNet-S* in most cases. The difference between these two models lies in whether TGNN is utilized for temporal-sensitive modeling. Our method can leverage temporal information to learn more discriminative features, achieving better classification performance. These results verify the importance of temporal information in incomplete multimodal learning and the effectiveness of our method in temporal-sensitive modeling.

Meanwhile, our GCNet outperforms *GCNet-T* on all missing rates. Taking the results on the IEMOCAP(six-class) dataset as an example, GCNet shows an absolute improvement of 0.31%∼2.35% over *GCNet-T*. Compared with this comparison system, we further exploit SGNN to capture speaker information in conversations. Since each speaker has its means of expression,

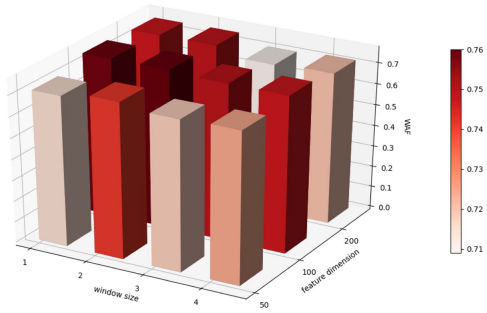
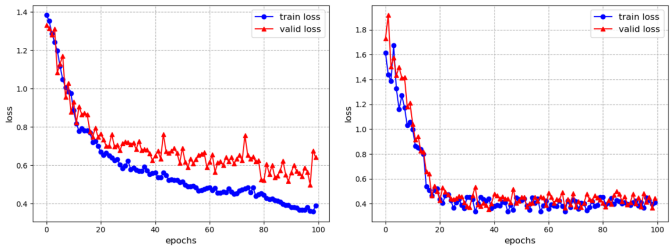



Fig. 5. Parameter tuning on the IEMOCAP(four-class) dataset with the missing rate $\eta = 0.3$.




(a) classification loss (b) imputation loss

Fig. 6. Trends in loss functions on the IEMOCAP(four-class) dataset with the missing rate $\eta = 0.3$.

we can well reconstruct conversational data with the help of speaker information and achieve better classification performance.

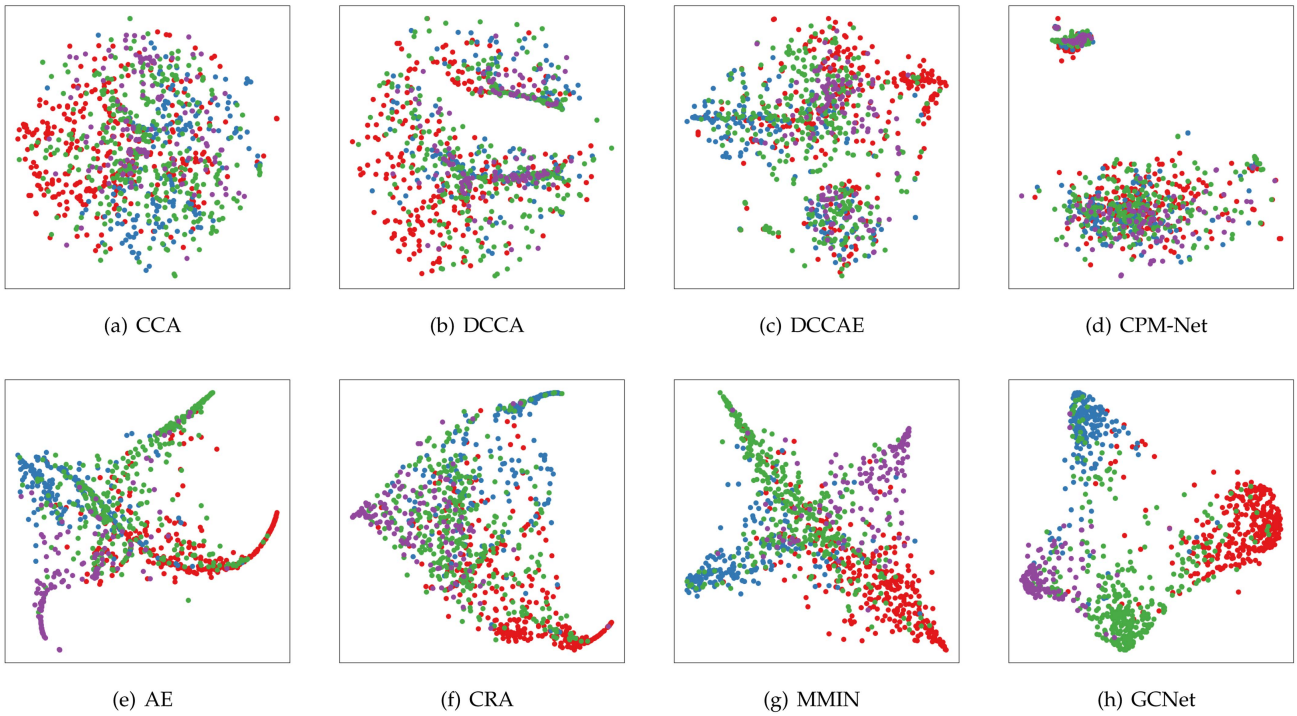
From Table IV, we also observe that GCNet outperforms *GCNet-ST* in most cases. Although we can easily collect conversational data from social media platforms, the annotation process still requires a lot of manual effort, resulting in insufficient amounts of labeled data. Compared with GCNet, *GCNet-ST* needs to optimize more relation types with limited labeled data. It increases the difficulty of model optimization, and each relation type is hard to be fully learned. Our GCNet captures speaker and temporal dependencies through two separate models, which reduces the number of relation types, alleviates the difficulty of model optimization, and achieves better performance.

### E. Parameter Tuning

Our GCNet contains two user-specified parameters: the context window size $w$ and the dimension of latent representations $h$. To evaluate the influence of these parameters, we visualize the parameter tuning process on the IEMOCAP(four-class) dataset with the missing rate $\eta = 0.3$. We choose $w$ from $\{1, 2, 3, 4\}$ and $h$ from $\{50, 100, 200\}$. Experimental results are presented in Fig. 5.

As $w$ increases, the classification performance improves first and then degrades. The reasons are probably twofold. First, as the window size increases, the designed graph will contain a large number of edges, which increases the difficulty of model optimization. Second, most utterances focus on their local context [51]. A larger window size will include more irrelevant

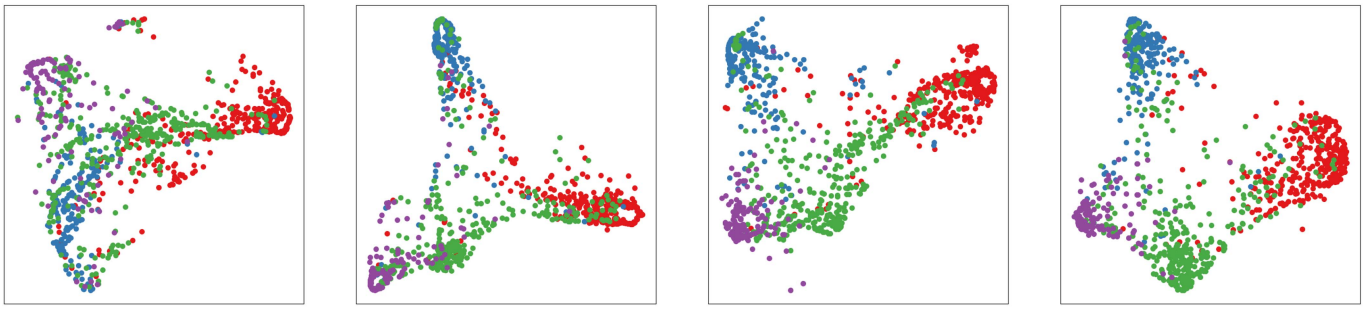
Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.



(a) CCA (b) DCCA (c) DCCAE (d) CPM-Net

(e) AE (f) CRA (g) MMIN (h) GCNet

Fig. 7. Visualization the representations of different methods on the IEMOCAP(four-class) test set with the missing rate $\eta = 0.3$. In these figures, we use red, blue, green and purple to represent happiness, sadness, neutral and anger, respectively.



(a) 10 epoch (b) 20 epoch (c) 50 epoch (d) 78 epoch

Fig. 8. T-SNE visualization results on the IEMOCAP(four-class) test set with increasing training iterations (the missing rate $\eta = 0.3$). In these figures, we use red, blue, green and purple to represent happiness, sadness, neutral and anger, respectively.

contextual information, which may affect the prediction results of the target utterance. Therefore, unlike previous works [72] that exploit fully connected graphs for conversation understanding, we fix the context window size to limit the number of edges.

Meanwhile, as $h$ gradually increases, the classification performance improves first and then degrades. When the feature dimension increases, our GCNet will contain more trainable parameters, increasing the risk of over-fitting. Therefore, a good choice of parameters can remarkably improve the performance of conversation understanding.

### *F. Convergence Analysis*

In this section, we investigate the convergence property on the IEMOCAP(four-class) dataset with the missing rate $\eta = 0.3$. Figs. 6(a) and (b) display the loss curves of classification and imputation, respectively. Despite the slight fluctuation, the loss curves maintain a descending trend and finally converge. For each loss function, the loss curves on training and validation sets converge at close epochs. These results prove the generalization ability of our GCNet on unseen data. For each data subset, the classification and imputation loss curves also converge at close epochs. These results reveal the correlation between classification and imputation. Our GCNet can achieve better classification performance when it can well reconstruct multimodal data.

### *G. Visualization of Embedding Space*

To qualitatively analyze the improvements of GCNet, we visualize the latent representation of different methods on the

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

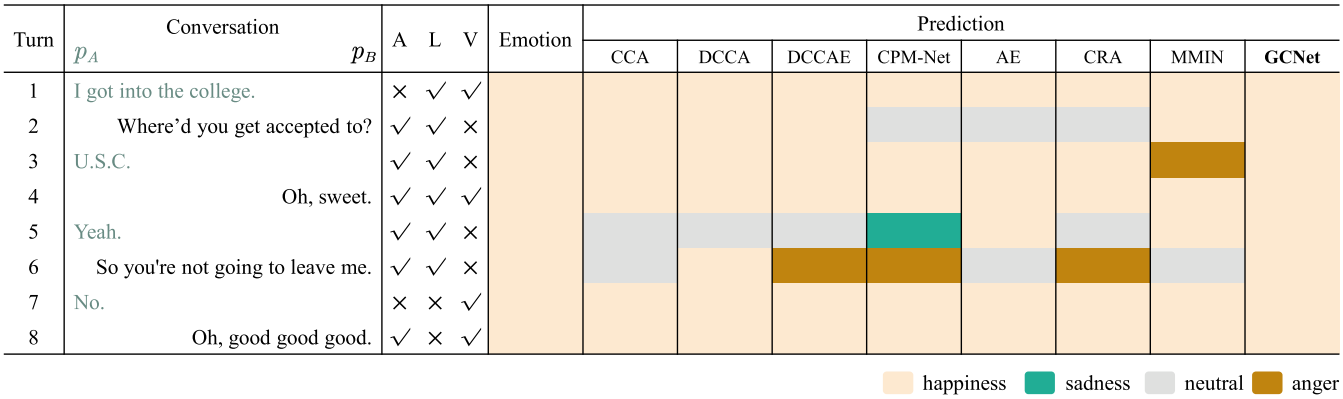



Fig. 9. Prediction results on incomplete conversational data. "✓" denotes the available modalities and "×" denotes the missing modalities.

IEMOCAP(four-class) test set. In the following experiments, we fix the missing rate $\eta$ to 0.3.

Figs. 7(a)–(g) present the t-SNE [73] visualization results of baselines. Fig. 7(h) presents the visualization result of our GCNet. As can be observed, our proposed method can effectively disentangle the latent representation, making the margin between different classes clearer. Meanwhile, we visualize the latent representation of our method with increasing training iterations. Figs. 8(a)–(d) present the t-SNE results of the 10th, 20th, 50th and 78th epochs, respectively. As can be observed, with the increase of the epoch, the latent representation reveals the underlying class distribution much better.

### H. Case Study

In this section, we compare the prediction results of different methods on incomplete conversational data. Fig. 9 provides a representative example from the IEMOCAP(four-class) test set. In this example, $p_A$ is accepted to a college and shares his happiness with $p_B$. We observe that GCNet generates more accurate results than other methods. Indeed, it is hard to make correct predictions on incomplete conversational data without incorporating temporal and speaker information. These qualitative results verify the effectiveness of our GCNet in incomplete multimodal learning.

## VI. Conclusion

In this paper, we propose a novel framework, GCNet, for incomplete multimodal learning in conversations. It takes full advantage of speaker and temporal information in conversations, aiming to learn discriminative representations from modality-incomplete conversational data. Unlike existing works that mainly focus on medical images or individual utterances, we study incomplete multimodal learning on conversational data, filling the gap of current works. Experimental results on three benchmark datasets demonstrate the effectiveness of our method. Through quantitative and qualitative analysis, we first verify that GCNet consistently outperforms currently advanced approaches under varying missing rates, achieving the best classification and imputation performance. Then, we show the importance of incomplete data in feature learning and prove the necessity of each component in GCNet. After that, we visualize the parameter tuning process and reveal the impact of different hyper-parameters. Through convergence analysis, we also present the correlation between classification and imputation. Our GCNet can achieve better classification performance when it can well reconstruct modality-complete data.

In the future, we will extend the applications of our proposed method. Besides conversational emotion recognition, we will leverage GCNet to deal with modality missing problems in other types of conversation understanding tasks. Furthermore, in addition to fixing the context window size to limit the number of edges, we will explore some dynamic edge construction strategies based on correlations between utterances in our future work.

## References

[1] Y. Liang, F. Meng, Y. Zhang, Y. Chen, J. Xu, and J. Zhou, "Emotional conversation generation with heterogeneous graph neural network," *Artif. Intell.*, vol. 308, 2022, Art. no. 103714.
[2] T. Fu, S. Gao, X. Zhao, J.-R. Wen, and R. Yan, "Learning towards conversational AI: A survey," *AI Open*, vol. 3, pp. 14–28, 2022.
[3] L. Nie, W. Wang, R. Hong, M. Wang, and Q. Tian, "Multimodal dialog system: Generating responses via adaptive decoders," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1098–1106.
[4] C. Gao, W. Lei, X. He, M. de Rijke, and T.-S. Chua, "Advances and challenges in conversational recommender systems: A survey," *AI Open*, vol. 2, pp. 100–126, 2021.
[5] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.
[6] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[7] Y. Yang, D.-C. Zhan, X.-R. Sheng, and Y. Jiang, "Semi-supervised multi-modal learning with incomplete modalities," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2998–3004.
[8] Z. Xue, J. Du, D. Du, W. Ren, and S. Lyu, "Deep correlated predictive subspace learning for incomplete multi-view semi-supervised classification," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 4026–4032.
[9] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2302–2310.
[10] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

[11] L. Zhang, Y. Zhao, Z. Zhu, D. Shen, and S. Ji, "Multi-view missing data completion," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1296–1309, Jul. 2018.

[12] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang, "Metric learning on healthcare data with incomplete modalities," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 3534–3540.

[13] Y. Liu et al., "Incomplete multi-modal representation learning for alzheimer's disease diagnosis," *Med. Image Anal.*, vol. 69, pp. 1–11, 2021.

[14] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1936–1940.

[15] Z. Lian, B. Liu, and J. Tao, "DECN: Dialogical emotion correction network for conversational emotion recognition," *Neurocomputing*, vol. 454, pp. 483–495, 2021.

[16] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 05, pp. 2402–2415, May 2022.

[17] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 2608–2618.

[18] J. Chen and A. Zhang, "HGMF: Heterogeneous graph-based fusion for multimodal data with incompleteness," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 1295–1305.

[19] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Proc. Companion Publication Int. Conf. Multimodal Interaction*, 2020, pp. 400–404.

[20] F. Ma, X. Xu, S.-L. Huang, and L. Zhang, "Maximum likelihood estimation for multimodal learning with missing modality," 2021, *arXiv:2108.10513*.

[21] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. Lee Giles, "Learning to extract semantic structure from documents using multimodal fully convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5315–5324.

[22] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1569–1576.

[23] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[24] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, 2010.

[25] H. Fan, Y. Chen, Y. Guo, H. Zhang, and G. Kuang, "Hyperspectral image restoration using low-rank tensor recovery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 10, pp. 4589–4604, Oct. 2017.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[27] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[28] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1405–1414.

[29] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu, "Partial multi-view clustering via consistent GAN," in *Proc. IEEE Int. Conf. Data Mining*, 2018, pp. 1290–1295.

[30] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1158–1166.

[31] O. Ivanov, M. Figurnov, and D. Vetrov, "Variational autoencoder with arbitrary conditioning," in *Proc. 7th Int. Conf. Learn. Representations*, 2019, pp. 1–25.

[32] C. Du et al., "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 108–116.

[33] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[34] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4400–4407.

[35] Y. Duan, Y. Lv, W. Kang, and Y. Zhao, "A deep learning based approach for traffic data imputation," in *Proc. IEEE 17th Int. Conf. Intell. Transp. Syst.*, 2014, pp. 912–917.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[37] L. Yuan et al., "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.

[38] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Multi-source learning with block-wise missing data for alzheimer's disease prediction," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2013, pp. 185–193.

[39] Y. Li, T. Yang, J. Zhou, and J. Ye, "Multi-task learning based survival analysis for predicting alzheimer's disease progression with multi-source block-wise missing data," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 288–296.

[40] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. Berlin, Germany: Springer, 1992, pp. 162–190.

[41] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[42] F. Ma, S.-L. Huang, and L. Zhang, "An efficient approach for audio-visual emotion recognition with missing labels and missing modalities," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[43] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "Completer: Incomplete multi-view clustering via contrastive prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11174–11183.

[44] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[45] A. Zadeh, Y.-C. Lim, P. P. Liang, and L.-P. Morency, "Variational auto-decoder: A method for neural generative modeling from incomplete data," 2019, *arXiv:1903.00840*.

[46] A. Zadeh, S. Benoit, and L.-P. Morency, "Relay variational inference: A method for accelerated encoderless VI," 2021, *arXiv:2110.13422*.

[47] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[48] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6818–6825.

[49] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7370–7377.

[50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. 6th Int. Conf. Learn. Representations*, 2018, pp. 1–12.

[51] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 985–1000, Jan. 2021 .

[52] M. Schlichtkrull et al., "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.*, 2018, pp. 593–607.

[53] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.

[54] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 2122–2132.

[55] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 164–172.

[56] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech emotion recognition via contrastive loss under siamese networks," in *Proc. Joint Workshop 4th Workshop Affect. Social Multimedia Comput. 1st Multi-Modal Affect. Comput. Large-Scale Multimedia Data*, 2018, pp. 21–26.

[57] Z. Zhao et al., "Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition," *Neural Netw.*, vol. 141, pp. 52–60, 2021.

[58] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "Box of lies: Multimodal deception detection in dialogues," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 1768–1777.

[59] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[60] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.

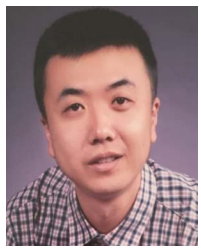

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.

[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[62] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3465–3469.
[63] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," 2020, *arXiv:2012.06185*.
[64] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, pp. 1–21.
[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[66] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[67] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Trans. Image Process.*, vol. 30, pp. 6544–6556, Jul. 2021.
[68] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[69] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.
[70] L. Z. Wong, H. Chen, S. Lin, and D. C. Chen, "Imputing missing values in sensor networks using sparse data representations," in *Proc. 17th ACM Int. Conf. Model., Anal. Simul. Wireless Mobile Syst.*, 2014, pp. 227–230.
[71] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2019, pp. 154–164.
[72] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5666–5675.
[73] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

**Zheng Lian** received the BS degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021. He is currently an assistant professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing, deep learning and multimodal emotion recognition.

**Lan Chen** received the BS degree from the China University of Petroleum, Beijing, China, in 2016, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2022. Her current research interests include computer graphics and image processing.

**Licai Sun** received the BS degree from Beijing Forestry University, Beijing, China, in 2016, and the MS degree from the University of Chinese Academy of Sciences, Beijing, China, in 2019. He is currently working toward the PhD degree with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing, deep learning, and multimodal representation learning.



**Bin Liu** (Member, IEEE) received the BS and MS degrees from the Beijing Institute of Technology, Beijing, China, in 2007 and 2009 respectively, and the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an associate professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing and audio signal processing.

**Jianhua Tao** (Senior Member, IEEE) received the MS degree from Nanjing University, Nanjing, China, in 1996, and the PhD degree from Tsinghua University, Beijing, China, in 2001. He is currently a professor with the Department of Automation, Tsinghua University, Beijing, China. He has authored or coauthored more than eighty papers on major journals and proceedings. His current research interests include speech recognition, speech synthesis and coding methods, human–computer interaction, multimedia information processing, and pattern recognition. He is the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, NCMMSC, etc. He is also the steering committee member for *IEEE Transactions on Affective Computing*, an associate editor for *Journal on Multimodal User Interface* and *International Journal on Synthetic Emotions*, and the deputy editor-in-chief for *Chinese Journal of Phonetics*. He was the recipient of several awards from the important conferences, such as Eurospeech, NCMMSC, etc.

Authorized licensed use limited to: Monash University. Downloaded on August 12,2023 at 03:28:49 UTC from IEEE Xplore. Restrictions apply.