Robust Multimodal Sentiment Analysis via Tag Encoding of Uncertain Missing Modalities

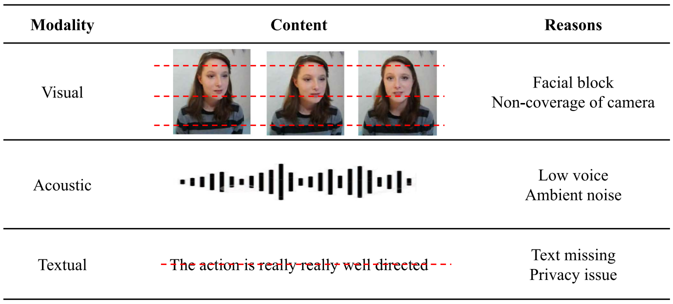
Jiandian Zeng, Jiantao Zhou, *Senior Member, IEEE*, and Tianyi Liu

***Abstract*—Multimodal sentiment analysis aims to extract emotions with multiple data sources, usually under the assumption that all modalities are available. In practice, such a strong assumption does not always hold, and most of multimodal sentiment analysis methods may fail when partial modalities are missing. Some existing works have started to address the missing modality problem; but only considered the single modality missing case, while ignoring the practically more general cases of multiple modalities missing. To this end, in this paper, we propose a Tag-Assisted Transformer Encoder (TATE) network to handle the problem of missing uncertain modalities. Specifically, we design a tag encoding module to cover both the single modality and multiple modalities missing cases, so as to guide the network's attention to those missing modalities. Besides, a new space projection pattern is adopted to align common vectors, taking into account the different importance of each modality. Afterwards, a Transformer encoder-decoder network is utilized to learn the missing modality features, and the outputs of the Transformer encoder are extracted for the final sentiment classification. Extensive experiments and analyses are conducted on CMU-MOSI, IEMOCAP, and MELD datasets, which show that the proposed method can achieve significant improvements compared with several baselines.**

***Index Terms*—Multimodal sentiment analysis, missing modality, joint representation.**

I. Introduction

In recent years, sentiment analysis has attracted intensive interest in extracting human's emotions and opinions, among which multimodal sentiment analysis is becoming an especially popular research direction with the massive amounts of online content [1], [2], [3]. Analyzing such a large scale of multimodal data is essential for online platforms to understand individual behavior or emotions. Traditional single modality analysis methods may not work well on the multimodal data. Taking the single phase "*oh, of course*!" for instance, it's hard to read the emotion without enough lexical information, and the acoustic modality may help in the emotion recognition if available. Thus, combining different modalities allows complementary features to be learned, resulting in better joint multimodal representations. Most prior works on multimodal fusion [4], [5] assumed that all modalities are always available at the training and testing stages. However, in reality, we often encounter scenarios that partial modalities could be missing. For example, as shown in Fig. 1, the visual features may be blocked due to the non-coverage of the camera; the acoustic information may be unavailable because of the enormous ambient noise; and the textual information may be absent for privacy concerns. Therefore, how to handle and recover missing modalities has becoming an important topic in the multimodal sentiment analysis.



Modality	Content	Reasons
Visual		Facial block Non-coverage of camera
Acoustic		Low voice Ambient noise
Textual	~~The action is really really well directed~~	Text missing Privacy issue

Fig. 1. An example of missing modalities. Reasons for missing modalities are also provided. The absent modality is marked with dotted red lines.

To handle the missing modality problem, several works [6], [7], [8] simply discarded missing modalities or utilized matrix completion methods to impute missing modalities, and degraded the overall performance. In [8], the visual modality was ablated when training with missing data. Zhao et al. [7] completed the kernel matrices of the modality using the common instances in different modalities. Owing to the strong learning ability of deep learning, recent works have employed deep neural networks to learn latent relationships among available modalities. Tran et al. [9] first identified the general problem of missing modality in multimodal data, and proposed a Cascaded Residual Auto-encoder (CRA) network to learn complex relationships from different modalities. More recently, Zhao et al. [10] adopted the cycle consistency learning with CRA to recover missing modalities. Yuan et al. [11] designed a Transformer-based feature reconstruction network to guide the extractor in obtaining the semantics of missing modality features. Wei et al. [12] provided a modality distillation strategy based on tensor

Manuscript received 28 April 2022; revised 30 August 2022 and 6 September 2022; accepted 12 September 2022. Date of publication 19 September 2022; date of current version 1 November 2023. This work was supported in part by Macau Science and Technology Development Fund under Grants SKLIOTSC-2021-2023 and 0072/2020/AMJ, in part by Research Committee at University of Macau under Grants MYRG2018-00029-FST, MYRG2019-00023-FST, MYRG2020-00101-FST, and MYRG2022-00152-FST, in part by the Natural Science Foundation of China under Grant 61971476, and in part by Alibaba Group through Alibaba Innovative Research Program. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Guoying Zhao. *(Corresponding author: Jiantao Zhou.)*

Jiandian Zeng and Jiantao Zhou are with the State Key Laboratory of Internet of Things for Smart City, and Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yb87470@um.edu.mo; jtzhou@umac.mo).

Tianyi Liu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: liutianyi@sjtu.edu.cn).

Digital Object Identifier 10.1109/TMM.2022.3207572

fusion network, aiming to capture the interactions from available modalities. However, most of these existing works assumed that there is only one missing modality, and ignored the practically more general cases of multiple modalities missing with uncertainties. That is, these methods required to train a new model to fit each missing modality case, which is apparently both costly and inconvenient. In practice, the pattern of missing modalities could also be uncertain, e.g., one or two modalities are randomly absent. Besides, most of the above works adopted simple feed-forward neural layers with the same parameters to align different modalities, and then directly concatenated them to acquire joint representations. Ideally, the importance of different modalities should be further considered when fusing them. To tackle the above issues, two challenges should be addressed: 1) how to handle the cases when multiple modalities are absent?, and 2) how to learn robust joint representations when the missing modalities are uncertain?

In this paper, we propose a **T**ag-**A**ssisted **T**ransformer **E**ncoder (TATE) network to learn complementary features among modalities. For the first challenge, we design a tag encoding module to mark missing modalities, aiming to direct the network's attention to absent modalities. As will be shown later, the attached tag not only can cover both the single modality and multiple modalities absent situations, but also can assist in the joint representation learning. For the second challenge, we first adopt the Transformer [13] as the extractor to capture intra-modal features. Since the importance of each modality could be different, we calculate the attention weight according to the accuracy score of each modality. Here, we emphasize adding fusion weights when concatenating multiple modalities, and apply a two-by-two projection pattern to map each modality into a common space. Then, the pre-trained network trained with full modalities is utilized to supervise the encoded vectors. Specifically, we design a forward differential loss to guide the learning process for missing modalities, a backward reconstruction loss to supervise the joint common vector reconstruction, and a tag recovery loss to allow the network to focus more on the added tag. Eventually, the outputs generated by a Transformer encoder are fed into a classifier for sentiment prediction. Our major contributions are summarized as follows:

- We propose the TATE network to handle the multiple modalities missing problem for multimodal sentiment analysis. The code is publicly available[1].
- We design a tag encoding module to cover both the single modality and multiple modalities absent situations. We also adopt a new common space projection module, which considers the importance of different modalities to learn joint representations.
- Our proposed model achieves significant improvements compared with several benchmarks on CMU-MOSI, IEMOCAP and MELD datasets, validating its superiority. Further studies also prove the effectiveness of the proposed method.

The rest of this paper is organized as follows: Section II introduces related works. Section III defines the detailed problem and describes the proposed architecture. Experimental results and further analyses are shown in Section IV. Section V concludes the paper and discusses the future directions.

Difference from conference version: Portions of the work presented in this paper have previously appeared in [14] as a conference version. We have significantly revised and clarified the paper, and improved many technique details compared with [14]. The primary improvements can be summarized as follows. First of all, since the importance of each modality could be different, we have specified the motivation and enriched our method by adding fusion weights in the common space projection module (Section III-F). Owing to this strategy, our performance on three datasets improves by 0.61% to 4.48% on M-F1 and by 0.07% to 2.43% on ACC, which could be deemed to be significant. Secondly, we have compared different pre-trained models and chosen the most efficient pre-trained network in Section III-D, showing the detailed training process with full modalities. Last but not the least, additional experiments, e.g., results on the MELD dataset (Section IV), the comparison with different word embeddings (Section IV-D4) and the case study (Section IV-E) etc., have been conducted for further statistical analysis.

II. Related Works

In this section, we first introduce the concept of multimodal sentiment analysis, and then discuss the methods for dealing with missing modalities.

A. Multimodal Sentiment Analysis

Multimodal sentiment analysis [15], [16], [17] has attracted a great deal of attention in recent years as a core branch of sentiment analysis [18], [19], [20]. Compared to a single modality case, multimodal sentiment analysis is more challenging due to the complexity of handling and analyzing data from different modalities. To learn joint representations of multimodal data, three multimodal fusion strategies are applied: 1) *early fusion* directly combines features of different modalities before the classification. Majumder et al. [21] proposed a hierarchical fusion strategy to fuse acoustic, visual and textual modalities, and proved the effectiveness of two-by-two fusion pattern; 2) *late fusion* adopts the average score of each modality as the final weights. Guo et al. [22] adopted an online early-late fusion scheme to explore complementary relationship for the sign language recognition, where the late fusion further aggregated features combined by the early fusion; and 3) *intermediate fusion* utilizes a shared layer to fuse features. Xu et al. [23] constructed the decomposition and relation networks to represent the commonality and discrepancy among modalities. Hazarika et al. [24] designed a multimodal learning framework that can learn modality-invariant and modality-specific representations by projecting each modality into two distinct sub-spaces. It should be emphasized that the above multimodal fusion models cannot handle the cases when partial modalities are missing.

B. Methods for Handling Missing Modalities

Based on the strategies for handing the missing modality problem, previous works can be generally categorized into two

[1] [Online]. Available: https://github.com/JaydenZeng/TATE

groups: 1) generative methods [9], [25], [26], [27], [28]; and 2) joint learning methods [10], [11], [29], [30].

Generative methods learn to generate new data with similar distributions to obey the distribution of the observed data. With the ability to learn latent representations, the Auto-Encoder (AE) [25] has been widely used. Vincent et al. [31] extracted features with AE based on the idea of making the learned representations robust to partial corruption of the input data. Kingma et al. [32] designed a Variational Auto-Encoder (VAE) to infer and learn features with a simple ancestral sampling. Besides, inspired by the residual connection network [33], Tran et al. [9] proposed a Cascaded Residual Auto-encoder (CRA) to impute data with missing modality, combining a series of residual AEs into a cascaded architecture to learn relationships among different modalities. Relying on Generative Adversarial Networks (GAN) [34], Shang et al. [27] treated each view as a separate domain, and identified domain-to-domain mappings via a GAN using randomly-sampled data from each view. Furthermore, the domain mapping technique is also considered to impute missing data. Along this line, Cai et al. [35] formulated the missing modality problem as a conditional image generation task, and designed a 3D encoder-decoder network to capture modality relations. They also incorporated the available category information during training to enhance the robustness of the model. Moreover, Zhao et al. [28] developed a cross partial multi-view network to model complex correlations among different views, where multiple discriminators are used to generate missing data.

On the other hand, joint learning methods try to learn joint representations based on the relations among different modalities [29], [36], [37]. Based on the idea that the cycle consistency loss can retain the maximal information from all modalities, Pham et al. [29] investigated learning robust representations via cyclic translations from source to target modalities. It was shown that translation from a source to a target modality can assist in learning joint representations. Zhao et al. [10] also applied cycle consistency learning for missing modality imputation, and the CRA-based cross-modality imagination module was designed based on paired multimodal data. They then proposed a missing modality network to handle uncertain missing cases, so as to improve the performance of emotion recognition. More recently, Yuan et al. [11] utilized the Transformer to extract intra-modal and inter-modal relations, and designed a Transformer-based feature reconstruction network to reproduce the semantics of the missing modality.

We would like to point out that most of the above works can only handle the scenarios of missing one single modality, and cannot satisfactorily deal with multiple modalities missing cases (because they need to train a new model for each modality missing case). As will be clear soon, our developed scheme differs the above works in several ways: 1) a tag encoding module is designed to cover all uncertain missing cases in a unified manner; and 2) a new mapping method considering the fusion weight of each modality is applied to learn better joint representations in the common space projection module.

III. Methodology

In this section, we first give the problem definition and associated notations. Then, we present the overall workflow of the proposed architecture and the detailed modules.

A. *Problem Definition and Notations*

Given a multimodal video segment that contains three modalities: $S = [X_v, X_a, X_t]$, where X_v, X_a and X_t denote visual, acoustic and textual modalities respectively. Without loss of generality, we use X'_m to represent the missing modality, where $m \in \{v, a, t\}$. For instance, assuming that the visual modality and acoustic modality are absent, the multimodal representation can be denoted as $[X'_v, X'_a, X_t]$. Formally, our problem is defined as follows: for the given triple (X_v, X_a, X_t), one or two modalities are randomly missing. The primary task is to classify the overall sentiment (*positive*, *neutral*, or *negative*) based on the available modalities.

B. *Overall Framework*

As can be seen in Fig. 2, the main workflow is as follows: for a given video segment, assuming that the visual modality and acoustic modality are missing, we first mask these missing modalities as 0, and then extract the remaining raw features. Afterwards, the masked multimodal representation goes through two branches: 1) one is encoded by a pre-trained model, which is trained with all full modality data; and 2) another goes through the tag encoding module and the common space projection module to acquire aligned feature vectors. Further, the updated representations are processed by a Transformer encoder, and the forward similarity loss between the pre-trained vectors and the encoder outputs is calculated. Meanwhile, the encoded outputs are fed into a classifier for the sentiment prediction. At last, we compute the backward reconstruction loss and the tag recovery loss to supervise the joint representation learning. In the following sub-sections, we present the details of each module.

C. *Multi-Head Attention*

Transformer [13] not only plays a great role in the Natural Language Processing (NLP) community, but also shows excellent representational capabilities in other areas, such as Computer Vision (CV). Rather than using an RNN-based structure to capture the sequential information, we employ the Transformer to generate the contextual representation of each modality respectively, where the key components of multi-head dot-product attention can be formalized as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

where Q, K and V are the query, the key, and the value respectively, and d is the dimension of the input.

Instead of utilizing the single attention, the multi-head attention is applied to obtain more information from different semantic spaces:

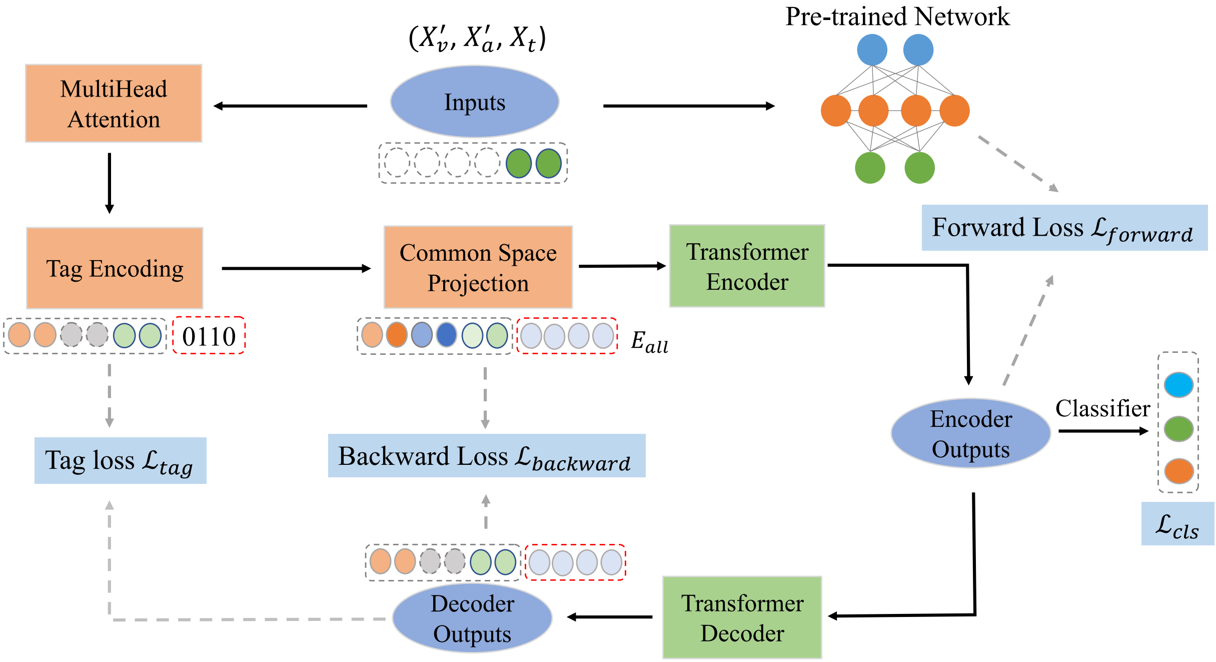


Fig. 2. Workflow of the proposed framework. The information flow goes to two branches: 1) one goes to the pre-trained network, which is trained with full modality data; and 2) another goes to the left multihead attention module for further encoding.

$$E_M = MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W^o, \qquad (2)$$

where $W^o \in \mathbb{R}^{d\times d}$ is a weight matrix and h is the head number. Given the input E, the i-th $head_i$ is calculated as follow:

$$head_i = Attention(EW_i^Q, EW_i^K, EW_i^V) \qquad (3)$$

where $W_i^Q \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$, $W_i^K \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$ and $W_i^V \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$ are the i-th weight matrices of the query, the key and the value, respectively.

Therefore, the updated modality representations can be formulated as follows:

$$\begin{aligned} E_v &= MultiHead(X'_v, X'_v, X'_v),\\ E_a &= MultiHead(X'_a, X'_a, X'_a),\\ E_t &= MultiHead(X_t, X_t, X_t). \end{aligned} \qquad (4)$$

D. Pre-Trained Network

As can be seen in Fig. 2, the pre-trained network is utilized to guide the learning process for missing modalities. We present the detailed structure of the pre-trained network in Fig. 3 for the extraction of latent vectors. Specifically, we first encode each modality by the multi-head self-attention mechanism to extract sequential relations:

$$E_k = MultiHead(X_k, X_k, X_k), k \in \{v, a, t\}. \qquad (5)$$

Then we concatenate them with a softmax activation function for the final pre-trained prediction:

$$\begin{aligned} E_{pre} &= [E_v||E_a||E_t],\\ P_{pre} &= softmax(W_{pre}E_{pre} + b_{pre}), \end{aligned} \qquad (6)$$

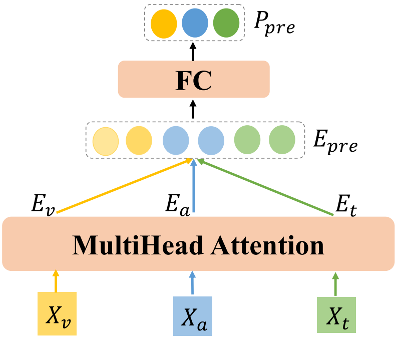


Fig. 3. Pre-trained network with full modalities.

where E_{pre} and b_{pre} are the learned weights and bias, E_{pre} is the concatenated representations, and P_{pre} is the predicted label. Noting that once the model with full modalities is well trained, we fix the pre-trained network during the whole training stage.

E. Tag Encoding

One of the key components of our network is the tag encoding module, which can cover both the single modality and multiple modalities absent situations. Also, the attached tags can indicate which modality is missing. To achieve this, we propose to employ the tag encoding technique to mark uncertain missing modalities, and direct network's attention to them. With the attached tags, the proposed model can handle various missing modality cases. In our settings, we adopt 4 digits ("0" or "1") to label missing modalities. If partial modalities of the input are missing, we set the first digit as "0"; otherwise "1". Besides, the last three digits are used to mark the corresponding visual, acoustic and textual modalities.

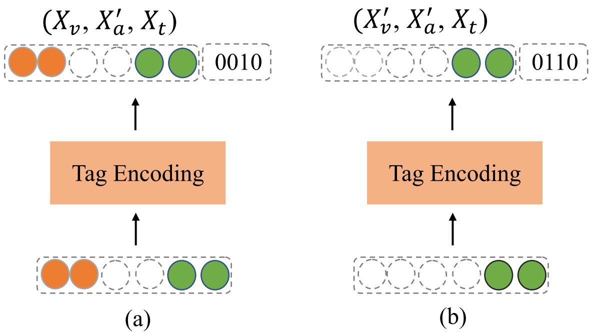


Fig. 4. Examples of modality tags. (a) modality tag with one missing modality. (b) modality tag with two missing modalities.

TABLE I
TAGS IN DIFFERENT CASES

Case	Tag	Case	Tag
(X_v, X_a, X_t)	1000	(X'_v, X'_a, X'_t)	0111
(X'_v, X_a, X_t)	0100	(X_v, X'_a, X'_t)	0011
$(X_v, X_{a'}, X_t)$	0010	(X'_v, X_a, X'_t)	0101
(X_v, X_a, X'_t)	0001	(X'_v, X'_a, X_t)	0110

As can be seen in Fig. 4, we give two examples about modality tags: in Fig. 4(a), the acoustic modality is missing, and the tag is set as "0010"; for the multiple modalities missing case (see Fig. 4(b)), we set the tag as "0110" to mark visual and acoustic modalities. The benefits are twofold: 1) the tag encoding module can cover both single and multiple modalities missing scenarios in a unified manner; and 2) the encoded tags can complementarily assist in the learning of the joint representations. Meanwhile, as will be shown later, the designed tag recovery loss can emphasize the reconstruction of missing modalities. All 8 possible tags are presented in Table I. To simplify the subsequent mathematical expression, we denote all tags as E_{tag}.

F. Common Space Projection

After the tag encoding, we now project three modalities into the common space. Previous works [23], [24] utilized simple feed-forward neural layers with the same parameters to align different modalities, and then directly concatenated them to acquire multimodal representations. However, these projection methods may fail when more than two modalities exist. In addition, the importance of different modalities should be further considered when concatenating them.

To tackle these issues, we first divide three modalities into three small subsets, then adopt a two-by-two projection strategy to align common space. Besides, we add dynamical weights according to their accuracy scores when fusing them. Fig. 5 presents the motivation for assigning modality weights, the illustration of calculating fusion weights, and the detailed process of the common space projection module. As can be seen in Fig. 5(a), the accuracy of the sentiment classification varies for each single modality on three datasets, where the textual modality achieves the best accuracy while the visual modality gets the lowest. Considering this, we calculate different weights dynamically (see Fig. 5(b)): each modality goes through a fully connected network to calculate sentiment logits L_k:

$$L_k = softmax(W_k E_k + b_k), k \in \{v, a, t\}, \tag{7}$$

where W_k is the weight matrix, and b_k is the corresponding bias. Then the maximum logit value L'_k of each modality is chosen to obtain fusion weights α:

$$\begin{aligned} \alpha &= softmax([L'_v||L'_a||L'_t]), \\ L'_k &= max(L_k), k \in \{v, a, t\}, \end{aligned} \tag{8}$$

where $||$ denotes the vertically concatenating operation.

Fig. 5(c) shows the final fusion process. For each single modality, we first obtain the self-related common space based on the following linear transformation:

$$\begin{aligned} C_v &= [W_{va} E_v || W_{vt} E_v], \\ C_a &= [W_{va} E_a || W_{ta} E_a], \\ C_t &= [W_{vt} E_t \; || W_{ta} E_t], \end{aligned} \tag{9}$$

where W_{va}, W_{vt} and W_{ta} are all weight matrices. Then, we concatenate all common vectors with the corresponding weights and the encoded tag to obtain the final common joint representations E_{all}:

$$E_{all} = [\alpha_v C_v || \alpha_a C_a || \alpha_t C_t || E_{tag}]. \tag{10}$$

G. Transformer Encoder-Decoder

To effectively model the long-term dependency of the intra-modal and the inter-modal information, we employ one sublayer in Transformer [13] to manage the information flow. As illustrated in Section III-C, the encoded outputs E_{out} can be accessed by the multi-head attention and feed-forward networks:

$$\begin{aligned} E_{out} &= MultiHead(E_{all}, E_{all}, E_{all}), \\ E_{out} &= \max(0, E_{out} W_e^1 + b_e^1) W_e^2 + b_e^2, \end{aligned} \tag{11}$$

where the query, the key, and the value are the same input E_{all}, W_e^1 and W_e^2 are two weight matrices, b_e^1 and b_e^2 are two learnable biases.

Similarly, the decoded outputs D_{out} are formulated as follows:

$$\begin{aligned} D_{out} &= MultiHead(E_{out}, E_{out}, E_{out}), \\ D_{out} &= \max(0, D_{out} W_o^1 + b_o^1) W_o^2 + b_o^2, \end{aligned} \tag{12}$$

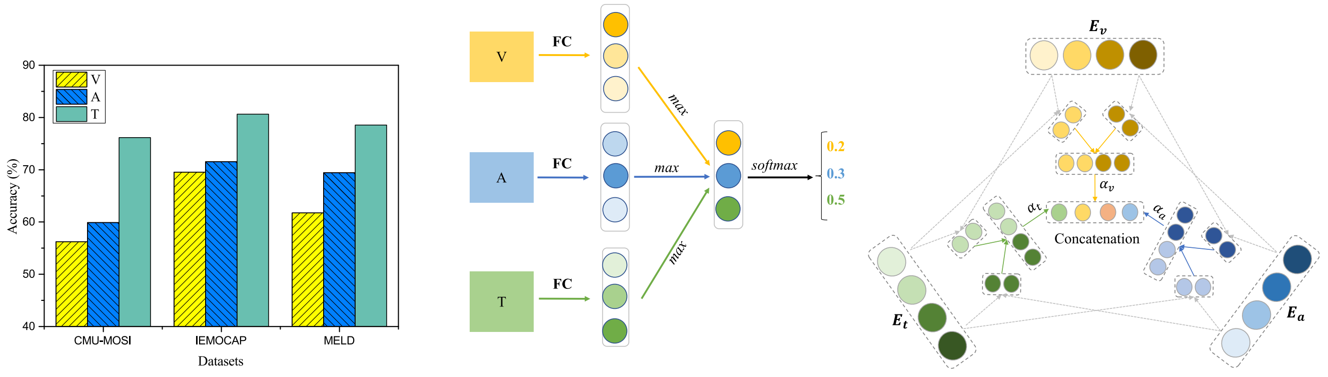
where W_o^1, W_o^2, b_o^1, and b_o^2 are parameters.

H. Training Objective

The overall training objective ($\mathcal{L}_{total}$) is expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{forward} + \lambda_2 \mathcal{L}_{backward} + \lambda_3 \mathcal{L}_{tag}, \tag{13}$$

where $\mathcal{L}_{cls}$ is the classification loss, $\mathcal{L}_{forward}$ is the forward differential loss, $\mathcal{L}_{backward}$ is the backward reconstruction loss, $\mathcal{L}_{tag}$ is the tag recovery loss, and λ_1, λ_2 and λ_3 are the corresponding weights. We now introduce these loss terms in details.



(a) (b) (c)

Fig. 5. Motivation and the detailed process of the common space projection. (a) Accuracy of the single modality on three datasets. (b) Calculating fusion weights, and (c) Illustration of the common space projection.

Algorithm 1: TATE Algorithm.

```
Input: Visual modality $X_v$, acoustic modality $X_a$,
 textual modality $X_t$, batch size $b$, learning rate $lr$,
 dropout probability $p$
Output: Predicted sentiment polarity $\hat{y}^*$
// suppose that the visual modality and
 acoustic modality are absent
$X_v' \leftarrow X_v$, $X_a' \leftarrow X_a$;
 //mask the visual modality and
 acoustic modality
$E_v \leftarrow MultiHead(X_v', X_v', X_v')$;
$E_a \leftarrow MultiHead(X_a', X_a', X_a')$;
$E_t \leftarrow MultiHead(X_t, X_t, X_t)$;
 //update modality representations
$E_{tag} \leftarrow \{X_v', X_a', X_t\}$;
 //obtain the tag based on the
 modality status
Calculate fusion weights $\alpha$ by Eqs. (8);
$C_v \leftarrow [W_{va}E_v || W_{vt}E_v]$;
$C_a \leftarrow [W_{va}E_a || W_{ta}E_a]$;
$C_t \leftarrow [W_{vt}E_t\ || W_{ta}E_t]$;
 //obtain self-related common space
 representation
$E_{all} \leftarrow [\alpha_v C_v || \alpha_a C_a || \alpha_t C_t || E_{tag}]$;
 //obtain common joint representations
 with the corresponding $\alpha$
$E_{out} \leftarrow MultiHead(E_{out}, E_{out}, E_{out})$;
 //Transformer encoder outputs
$p(\hat{y}|E_{out}, \theta) \leftarrow softmax(W_c E_{out} + b_c)$
 //obtain sentiment logits
$\hat{y}^* \leftarrow \arg\max_{\hat{y}}(p(\hat{y}|E_{out}, \theta))$;
return $\hat{y}^*$
```

1) Forward Differential Loss ($\mathcal{L}_{forward}$): As illustrated in Fig. 2, the forward loss is calculated by the difference between the pre-trained output (E_{pre}) and the Transformer encoder output (E_{out}). Since the pre-trained model is trained in the full modality settings, we employ the differential loss to guide the learning process for missing modalities. Specifically, the Kullback Leibler (KL) divergence loss is used:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot \frac{p(x_i)}{q(x_i)}, \tag{14}$$

where p and q are two probability distributions. Since KL divergence is asymmetric, we adopt the Jensen-Shannon (JS) divergence loss instead:

$$\begin{aligned}\mathcal{L}_{forward} &= JS(E_{out}||E_{pre}) \\ &= \frac{1}{2}(D_{KL}(E_{out}||E_{pre}) + D_{KL}(E_{pre}||E_{out})).\end{aligned} \tag{15}$$

2) Backward Reconstruction Loss ($\mathcal{L}_{backward}$): For the backward loss, we aim to supervise the joint common vector reconstruction. Therefore, similar to the forward differential loss, we calculate the JS divergence loss between the Transformer decoder output (D_{out}) and the updated common joint representations (E_{all}):

$$\begin{aligned}\mathcal{L}_{backward} &= JS(D_{out}||E_{all}) \\ &= \frac{1}{2}(D_{KL}(D_{out}||E_{all}) + D_{KL}(E_{all}||D_{out})).\end{aligned} \tag{16}$$

3) Tag Recovery Loss ($\mathcal{L}_{tag}$): In our settings, the tag is attached to mark missing modalities, and we expect that our network can pay more attention to them. To better guide the reconstruction of the attached tag, we design a tag recovery loss to direct the process. Specifically, we first rescale the last four digits of D_{out} with the Sigmoid function, that is:

$$D_{tag}^i = Sigmoid(D_{out}^i[-4:]), i \in [1, N], \tag{17}$$

where i corresponds to the i-th sample, and N is the total number of training samples. Then, the tag MAE (Mean Absolute Error) loss is calculated as follow:

$$\mathcal{L}_{tag} = \frac{1}{N}\sum_{i=1}^{N} |E_{tag}^i - D_{tag}^i|. \tag{18}$$

Here, the reason why we choose the MAE loss is that it is less sensitive to outliers with the absolute function.

4) Classification Loss ($\mathcal{L}_{cls}$): For the final classification module, we feed E_{out} into a fully connected network with the softmax activation function:

$$p(\hat{y}|E_{out}, \theta) = softmax(W_c E_{out} + b_c), \tag{19}$$

where W_c and b_c are the learned weights and bias. In detail, we employ the standard cross-entropy loss for this task, that is:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n=1}^{N} y_n log \hat{y}_n, \tag{20}$$

where N is the number of samples, y_n is the true label of the n-th sample, and $\hat{y}_n$ is the predicted label.

The detailed algorithm for employing TATE network for the sentiment analysis is presented in Algorithm 1.

IV. Experimental Results

All experiments are carried out on a Linux server (Ubuntu 18.04.1) with a Intel(R) Xeon(R) Gold 5120 CPU, 8 Nvidia 2080TI GPUs and 128 G RAM. Datasets and experimental settings are described as follows:

Datasets: We conduct several experiments on CMU-MOSI [38], IEMOCAP [39] and MELD [40] datasets. All datasets are multimodal benchmarks for the sentiment recognition, including visual, textual, and acoustic modalities. For the CMU-MOSI dataset, it contains 2199 segments from 93 opinion videos on YouTube. The label of each sample is annotated with a sentiment score in [-3, 3]. Following Yu et al. [41], we transform the scores into negative, neutral and positive labels. For the IEMOCAP dataset, it contains 5 sessions, each of which contains about 30 videos with at least 24 utterances. Specifically, the annotated labels are: neutral, frustration, anger, sad, happy, excited, surprise, fear, disappointing, and other. For the MELD dataset, it contains about 13,000 utterances from 1,433 dialogues with the corresponding emotion and sentiment labels. In our experiments, we report three-class (negative: [-3,0), neutral:[0], positive: (0,3]) results on CMU-MOSI, two-class (negative:[frustration, angry, sad, fear, disappointing], positive:[happy, excited]) on IEMOCAP, and three-class (positive, neutral, negative) results on MELD.

Parameters: Following a standardized procedure, we tune our model using five-fold validation and grid-searching on the training set. The learning rate lr is selected from $\{0.1, 0.001, 0.0005, 0.0001\}$, the batch size $b \in \{32, 64, 128\}$, and the hidden size $d \in \{64, 128, 300, 768\}$. Adam [42] is adopted to minimize the total loss given in Eq. (13). The epoch number is 20, the batch size is 32, the loss weight is set to 0.1, and these parameters are summarized in Table II.

Evaluation Metrics: $Accuracy$ (ACC) and $Macro - F1$ (M-F1), defined as follows, are used to measure the performance of the models.

$$ACC = \frac{T_{true}}{N},$$

$$M - F1 = \frac{1}{C}\sum_{i=1}^{C} \frac{2P_i R_i}{P_i + R_i}, \tag{21}$$

where T_{true} denotes the number of correctly predicted samples, N is the total number of samples, C is the class number, P_i is the i-th class positive predictive value, and R_i is the i-th class recall value.

TABLE II
Detailed Parameter Settings in Our Experiments

Description	Symbol	Value
Batch size	b	32
Epoch number	e	20
Dropout rate	p	0.3
Hidden size	d	300
Missing rate	η	[0, 0.5]
Learning rate	lr	0.001
Maximum textual length	n_t	25
Maximum visual length	n_v	100
Maximum acoustic length	n_a	150
Loss weights	$\lambda_1, \lambda_2, \lambda_3$	0.1

A. Feature Extraction

Visual Representations: The CMU-MOSI [38], IEMOCAP [39] and MELD [40] datasets mainly consist of human conversations, where visual features are primarily composed of human faces. Following [43], [44], we also adopt OpenFace2.0 toolkit [45] to extract facial features. Except the first to the fifth columns data that are irrelevant attributes about the frame number, the face_id, and the timestamp etc., we finally obtain 709-dimensional visual representations, where the face, the head, and the eye movement are included.

Textual Representations: For each textual utterance, the pre-trained Bert [46] is utilized to extract textual features. Eventually, we adopt the pre-trained uncased BERT-base model (12-layer, 768-hidden, 12-heads) to acquire 768-dimensional word vectors.

Acoustic Representations: As an audio analysis toolkit, Librosa [47] shows an excellent ability to extract acoustic features. For three multimodal datasets, each audio is mixed to the mono and is re-sampled to 16000 Hz. Besides, each frame is separated by 512 samples, and we choose the zero crossing rate, the Mel-Frequency Cepstral Coefficients (MFCC) and the Constant-Q Transform (CQT) features to represent audio segments. Finally, we concatenate three features to yield 33-dimensional acoustic features.

B. Baselines

To evaluate the performance of our approach, the following baselines are chosen for the comparison purpose:

AE [25]: An efficient data encoding network trained to copy its input to output. In our implementation, we employ 5 AEs with each layer of the size [512, 256, 128, 64].

CRA [9]: A missing modality reconstruction framework that employed the residual connection mechanism to approximate the difference between the input data. In our implementation,

TABLE III
RESULTS OF ALL BASELINES WITH A SINGLE MODALITY MISSING, WHERE THE BEST RESULTS ARE IN BOLD, AND THE MISSING RATE VARIES FROM 0 TO 0.5

Datasets	Models	0		0.1		0.2		0.3		0.4		0.5	
		M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
CMU-MOSI	AE	56.78	79.69	54.07	79.17	53.40	78.13	51.28	72.53	50.75	73.48	44.99	69.32
	CRA	56.85	79.73	54.37	79.38	53.57	78.24	51.67	72.84	51.02	73.79	45.38	69.45
	MCTN	57.32	79.75	55.48	79.87	53.99	77.49	52.31	71.59	51.64	73.81	45.76	68.11
	TransM	57.84	80.21	57.53	79.69	55.21	78.42	52.87	72.92	52.49	72.40	45.86	68.23
	MMIN	**60.41**	82.29	57.75	81.86	55.38	80.20	53.65	79.24	52.55	76.33	48.95	70.76
	Ours	60.18	**86.54**	**58.49**	**85.85**	**56.52**	**82.65**	**56.08**	**81.77**	**54.67**	**80.53**	**52.66**	**76.04**
IEMOCAP	AE	76.15	82.09	75.24	80.26	75.02	78.01	73.92	77.43	70.19	76.01	67.27	76.43
	CRA	77.05	82.13	75.95	80.97	75.13	78.09	74.02	78.11	70.69	76.12	67.75	76.49
	MCTN	78.57	82.27	77.74	81.02	75.37	78.27	74.69	78.52	71.75	76.29	68.17	76.63
	TransM	79.57	82.64	78.03	81.86	76.33	80.43	75.83	78.64	72.01	77.27	68.57	76.65
	MMIN	80.83	83.43	78.85	82.58	77.09	81.27	76.63	80.43	72.81	78.43	70.58	77.45
	Ours	**81.76**	**86.46**	**81.25**	**85.24**	**79.57**	**84.80**	**79.06**	**83.76**	**77.84**	**82.97**	**75.76**	**82.51**
MELD	AE	61.15	63.38	60.87	63.03	59.85	62.45	58.87	61.66	57.27	60.48	55.38	59.83
	CRA	61.78	63.77	61.21	63.21	60.46	62.72	58.95	61.89	57.39	60.68	56.16	59.95
	MCTN	62.17	64.43	62.07	64.12	61.59	63.83	60.64	62.21	59.39	61.36	58.40	60.72
	TransM	62.69	64.86	62.77	64.21	63.28	64.03	61.66	62.95	59.80	61.78	59.03	61.33
	MMIN	64.28	66.59	63.65	66.10	63.39	65.29	62.43	64.81	61.98	63.44	61.05	63.16
	Ours	**65.22**	**67.72**	**64.79**	**67.53**	**64.12**	**67.28**	**63.38**	**66.97**	**62.17**	**65.91**	**61.85**	**65.01**

we add a residual connection for the input with the same layer setting in AE [25].

MCTN[2] *[29]:* A method to learn robust joint representations by translating among modalities. It was claimed that translating from a source modality to a target one can capture joint information among modalities.

TransM [30]: An end-to-end translation based multimodal fusion method that utilized the Transformer to translate among modalities and encoded multimodal features. In our implementation, we concatenate 6 MAE losses between two modalities transformation.

MMIN[3] *[10]:* A unified multimodal emotion recognition model that adopted the cascade residual auto-encoder and cycle consistency learning to recover missing modalities.

TATE: Our proposed model.

C. Overall Results

For the single modality missing case, the experimental results are shown in Table III, where the missing rate is set from 0 to 0.5. Specifically, for each method, we report triple classification results on CMU-MOSI and MELD and two classification results on IEMOCAP, where M-F1 and ACC are used as metrics. With the increment of the missing rate, the overall results present a descending trend. Except for the M-F1 value under the full modality condition is 0.33% lower than MMIN on the CMU-MOSI dataset, our proposed method achieves the best results on all other settings, validating the effectiveness of our model. As can be seen in the table, compared to auto-encoder based methods (AE and CRA), translation-based methods (MCTN and TransM) achieve better performance, probably due to the fact that the end-to-end translation among modalities can better fuse the multimodal information. Besides, the comparative experiments suggest that the backward decoder can assist the forward encoder, so as to further improve the overall performance.

On the other hand, for multiple modalities missing cases, we also present our findings in Table IV. In these settings, one or two modalities are randomly absent. It can be seen that our proposed model still improves by 1.88% on M-F1 and by 2.69% on ACC compared to other baselines, demonstrating the robustness of the TATE network. Owing to the forward differential loss and the assistance of the tag encoding, our model can better capture semantic-relevant information.

D. Effects of Different Settings

In this subsection, we first conduct the ablation studies to better understand the effects of different modules in TATE. In addition, we further evaluate the performance of TATE by replacing several key components with alternatives.

1) Ablation Studies: To explore the effects of different modules in TATE, we evaluate our model with several settings: a) using only one modality; b) using two modalities; c) removing the tag encoding module; d) removing the common space projection module; e) removing the tag recovery loss; f) removing the forward differential loss; and g) removing the backward reconstruction loss. Note, in all above cases, that modality missing can still occur in a random manner with a certain missing rate, except the single modality experiments. Taking setting b) for example, where visual and textual modalities are considered (i.e., the case of (V, T)), either visual or textual modality could be randomly missing in a sample.

According to the results given in Table V, one interesting finding is that the performance drops sharply when the textual modality is missing, validating that textual information dominates in the multimodal sentiment analysis. A possible explanation for this phenomenon is that textual information is the manual transcription. However, similar reductions are not observed when removing the visual modality. We conjecture that it is because the visual information is not well extracted, due

[2] [Online]. Available: https://github.com/hainow/MCTN
[3] [Online]. Available: https://github.com/AIM3-RUC/MMIN/tree/master

TABLE IV
RESULTS OF ALL BASELINES WITH MULTIPLE MODALITIES MISSING, WHERE THE BEST RESULTS ARE IN BOLD, AND THE MISSING RATE VARIES FROM 0 TO 0.5

Datasets	Models	0		0.1		0.2		0.3		0.4		0.5	
		M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
CMU-MOSI	AE	56.78	79.69	52.80	75.65	50.84	74.18	46.23	69.18	44.40	69.05	40.29	66.01
	CRA	56.81	79.72	52.85	75.68	51.02	74.73	46.87	69.23	45.17	69.48	41.77	66.82
	MCTN	56.85	79.73	52.97	75.89	51.75	74.16	46.98	69.29	45.73	69.55	42.98	67.02
	TransM	57.84	80.21	53.49	77.08	51.97	74.24	48.23	70.51	47.02	70.38	43.28	67.74
	MMIN	**60.41**	82.29	55.49	80.12	52.79	76.26	48.97	73.27	47.39	74.28	44.63	68.92
	Ours	60.18	**86.54**	**57.37**	**82.81**	**55.62**	**81.67**	**53.65**	**78.12**	**53.28**	**77.98**	**51.92**	**76.60**
IEMOCAP	AE	76.15	82.09	75.07	79.84	74.20	76.91	71.55	76.07	69.73	75.16	67.15	75.22
	CRA	77.05	82.13	75.21	79.95	74.22	77.03	71.86	76.41	70.13	75.29	67.31	75.42
	MCTN	78.57	82.27	76.83	80.56	74.77	77.89	72.27	77.03	71.02	75.84	67.51	75.88
	TransM	79.57	82.64	77.21	81.13	75.87	79.01	72.36	78.15	71.38	76.88	68.02	76.04
	MMIN	80.83	83.43	78.02	82.32	76.38	79.53	73.05	79.02	71.22	77.27	69.39	77.01
	Ours	**81.76**	**86.46**	**80.67**	**85.30**	**79.12**	**83.77**	**78.99**	**84.64**	**78.44**	**82.75**	**76.97**	**82.25**
MELD	AE	61.15	63.38	59.55	62.89	59.43	62.16	58.22	61.19	56.74	60.25	54.85	59.07
	CRA	61.78	63.77	60.83	63.07	60.12	62.49	58.34	61.55	56.87	60.47	55.13	59.28
	MCTN	62.17	64.43	61.85	63.66	60.93	63.19	59.42	61.78	57.47	60.90	56.66	60.21
	TransM	62.69	64.86	62.03	63.81	62.10	63.55	60.82	62.22	58.29	61.24	57.39	60.88
	MMIN	64.28	66.59	62.99	65.15	62.37	64.46	61.33	64.05	60.41	62.86	59.82	62.10
	Ours	**65.22**	**67.72**	**63.72**	**67.17**	**63.28**	**67.09**	**62.89**	**66.74**	**61.33**	**65.42**	**60.92**	**64.67**

TABLE V
COMPARISON OF ALL MODULES IN TATE, WHERE THE MISSING RATE VARIES FROM 0 TO 0.5. THE TOP HALF OF THE TABLE SHOWS THE RESULTS OF COMBINING DIFFERENT MODALITIES; AND THE BOTTOM HALF OF THE TABLE SHOWS THE ABLATION RESULT WITH DIFFERENT MODULES. HERE, (V, A, T) MEANS THAT THREE MODALITIES ARE CONSIDERED, AND (V, A), (V, T), OR (A, T) MEANS THAT ONLY TWO MODALITIES ARE CONSIDERED

Modules	0		0.1		0.2		0.3		0.4		0.5	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
V	37.84	56.25	-	-	-	-	-	-	-	-	-	-
A	39.82	59.90	-	-	-	-	-	-	-	-	-	-
T	55.63	76.17	-	-	-	-	-	-	-	-	-	-
(V, A)	40.87	61.35	39.25	59.47	38.33	57.05	37.74	56.84	37.28	55.57	36.83	54.85
(V, T)	57.37	79.85	56.80	78.94	55.75	76.83	55.02	75.46	53.01	74.63	50.85	73.19
(A, T)	57.87	80.93	57.23	79.83	55.67	77.59	55.43	75.75	54.17	74.65	51.89	73.74
(V, A, T)	**60.18**	**86.54**	**58.49**	**85.85**	**56.52**	**82.65**	**56.08**	**81.77**	**54.67**	**80.53**	**52.66**	**76.04**
-w/o tag	58.05	80.45	57.95	80.53	55.23	80.10	54.21	80.17	52.78	76.94	49.79	72.53
-w/o tag loss	58.76	83.15	58.21	82.17	55.94	80.94	54.23	80.78	53.10	77.45	49.98	73.76
-w/o forward loss	52.77	76.80	52.12	75.84	50.76	73.87	50.21	72.86	48.95	71.53	47.73	70.83
-w/o backward loss	53.97	77.85	52.96	77.56	51.93	74.48	51.18	74.55	49.73	72.25	48.96	71.54
-w/o common space	54.03	79.76	53.20	77.59	52.97	75.99	51.23	75.01	49.83	72.37	49.05	71.85

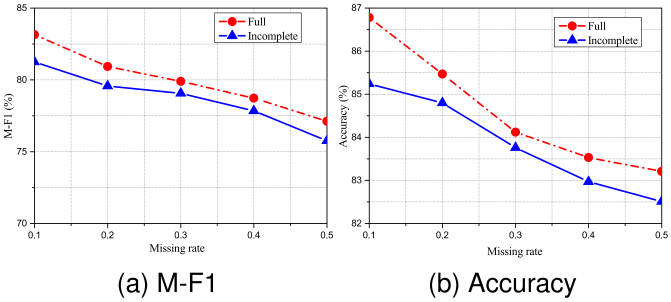
to the minor changes to the face. Besides, the top half of the table shows that the combination of two modalities provides better performance than single modality. This indicates that two modalities can learn complementary features for boosting the performance. Regarding the effects of different modules, the performance of the forward differential module decreases by 4.93% to 5.72% on M-F1 and by 5.21% to 10.01% on ACC, compared to the whole model, demonstrating the importance of the forward guidance. Since we employ full modalities to pre-train the guidance network, the forward JS divergence loss serves as a good supervision. One striking result to emerge from this table is that the tag encoding module also improves the performance as expected.

2) Effects of the Tag Encoding: To further validate the effectiveness of the tag encoding module, we incorporate it with two basic models: AE and TransM. The reason why we choose the above two models is that AE and TransM are two different kinds of encoders: AE is the auto-encoder based method, while TransM is the Transformer based one. For the above two models, we add tags after the feature extraction module. Table VI presents the detailed results on the CMU-MOSI dataset with a 30% missing rate. It can be seen that models with the tag encoding module improve by 3.50% to 4.78% on M-F1 and by 1.99% to 4.08% on ACC compared to basic models, showing the effectiveness of the tag encoding. Owing to the added tag, the network can be better guided, and can better focus on missing modalities.

TABLE VI
IMPROVEMENTS OF THE TAG ENCODING

Model	Basic		+Tag	
	M-F1	ACC	M-F1	ACC
AE	51.28	72.53	53.25 (3.69% ↑)	75.21 (3.56% ↑)
TransM	52.87	72.92	54.79 (3.50% ↑)	76.02 (4.08% ↑)
Ours	54.21	80.17	56.80 (4.78% ↑)	81.77 (1.99% ↑)

3) Effects of the Complete Modality: To see the difference between the complete and incomplete modalities of the test data,



(a) M-F1 (b) Accuracy

Fig. 6. Comparison of full modalities and incomplete modalities during the testing. (a) M-F1 values. (b) ACC values.

TABLE VII
RESULTS OF DIFFERENT WORD EMBEDDINGS, WHERE THE MISSING RATE VARIES FROM 0 TO 0.5

Datasets	Rate	Word2vec		Glove		Bert	
		M-F1	ACC	M-F1	ACC	M-F1	ACC
CMU-MOSI	0	55.46	79.73	57.52	82.60	60.18	86.54
	0.1	53.92	77.53	55.17	81.25	58.49	85.85
	0.2	51.25	74.28	53.03	77.60	56.52	82.65
	0.3	50.01	73.36	52.75	76.04	56.08	81.77
	0.4	47.75	70.18	48.83	71.35	54.67	80.53
	0.5	45.97	66.43	46.60	68.75	52.66	76.04
IEMOCAP	0	79.45	82.21	80.25	84.38	81.76	86.46
	0.1	78.32	80.59	79.87	83.36	81.25	85.24
	0.2	75.38	79.48	77.29	82.05	79.57	84.80
	0.3	74.42	79.05	76.83	81.57	79.06	83.76
	0.4	73.21	78.12	75.65	80.48	77.84	82.97
	0.5	72.16	77.41	73.39	79.73	75.76	82.51
MELD	0	62.49	64.39	64.28	66.31	65.22	67.72
	0.1	62.01	63.95	63.73	65.96	64.79	67.53
	0.2	60.88	62.21	62.49	65.08	64.12	67.28
	0.3	60.19	61.57	61.88	64.77	63.38	66.97
	0.4	58.36	60.41	60.49	62.35	62.17	65.91
	0.5	57.73	59.23	59.31	62.09	61.85	65.01

we first train the model with incomplete data, and then test the model with both full modality data and different missing rates of incomplete data. All experiments share the same parameters on the IEMOCAP dataset for a fair comparison. As can be observed in Fig. 6, the gaps between two settings on M-F1 and ACC reach the minimum when the missing rate is 0.3. As the number of missing samples in the training data increases, the correlation among modalities becomes harder to capture, resulting in weaker testing performance. Also, the gap increases when the missing rate is bigger than 0.3. One possible explanation for the above results is that the model cannot learn the joint representation well when there are too many absent samples.

4) Effects of Different Word Embeddings: We also measure the effects of the proposed model under other available word embedding methods. To this end, we choose Word2vec [48] and Glove [49] as alternative methods to the pre-trained Bert, and evaluate the respective prediction performance. Here, we set the embedding size as 128 in Word2vec and choose the cased 840 B tokens of 300 dimension in Glove. All settings share the same parameters for a fair comparison.

As presented in Table VII, the pre-trained Bert achieves the best results on three datasets, while the Word2vec model gets the worst. The reason may be that the Bert is trained from a large amount of text corpus, resulting in better word semantic correlations. However, Word2vec cannot handle the problem of polysemous words, since each word and vector correspond individually. For Word2vec and Glove models, the fusion weight becomes smaller with a lower accuracy of the single textual modality, resulting in a degradation of the overall performance. The table also shows that different word embeddings have significant effects on the overall performance. This phenomenon also implies that the textual modality may dominate the overall sentiment.

TABLE VIII
DETAILED DISTRIBUTIONS ON IEMOCAP

Category		Hap.	Ang.	Sad.	Neu.	Fru.	Exc.	Sur
4-class	Train	477	879	868	1385	-	-	-
	Test	118	224	216	323	-	-	-
7-class	Train	476	891	873	1348	1458	848	87
	Test	119	212	211	360	391	193	20

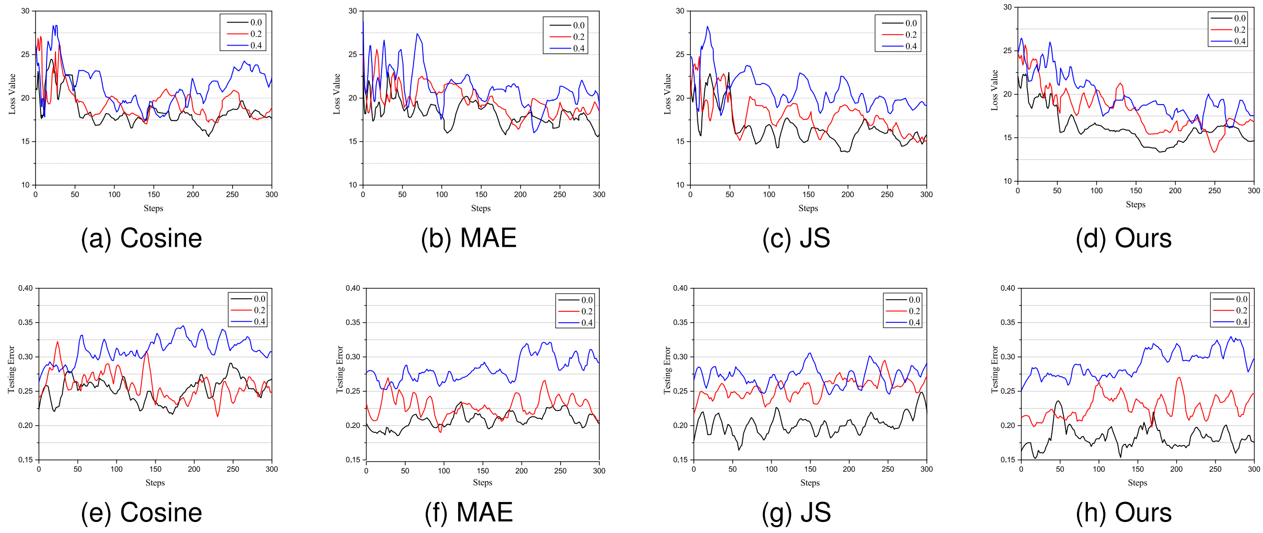
TABLE IX
RESULTS OF MULTI-CLASS ON IEMOCAP

Rate	2-class		4-class		7-class	
	M-F1	ACC	M-F1	ACC	M-F1	ACC
0	81.76	86.46	48.38	59.43	37.21	47.90
0.1	81.25	85.24	46.97	57.89	36.45	46.13
0.2	79.57	84.80	46.34	57.15	35.86	44.84
0.3	79.06	83.76	45.88	56.46	35.76	43.56
0.4	77.84	82.97	45.21	55.47	34.15	42.89
0.5	75.76	82.51	44.54	54.97	33.43	42.55

5) Multi-Class on IEMOCAP: We also explore the performance of multiple classes on the IEMOCAP dataset. Apart from the two-class results, we choose happy, angry, sad and neutral emotions as the 4-class experiment, and then choose the extra frustration, excited, and surprise emotions as the 7-class experiment. The detailed distributions and results are presented in Table VIII and Table IX, respectively. It can be seen that both M-F1 and ACC values decrease with the increment of class numbers. By comparing the results with different rates of missing modalities, the gaps among 7-class are smaller than those among 2-class and 4-class. Besides, the closer inspection of Table IX shows that the overall performance drops sharply when the class number is 7. Likely, it is caused by the confusion of multiple classes, resulting in the difficulties in the model convergence.

6) Effects of Different Losses: In our training stage, the JS divergence loss is adopted in both the forward differential loss and the backward reconstruction loss, and the MAE loss is used in the tag recovery loss. To investigate the effects of different losses, we replace them with different loss functions and observe the performance. Specifically, the cosine similarity loss, the MAE loss, and the JS divergence loss are chosen for comparison. We evaluate our model with 4 settings: a) using the cosine similarity loss for $\mathcal{L}_{forward}$, $\mathcal{L}_{backward}$ and $\mathcal{L}_{tag}$; b) using the MAE loss for $\mathcal{L}_{forward}$, $\mathcal{L}_{backward}$ and $\mathcal{L}_{tag}$; c) using the JS divergence loss for$\mathcal{L}_{forward}$, $\mathcal{L}_{backward}$ and $\mathcal{L}_{tag}$; and d) using the JS divergence loss for $\mathcal{L}_{forward}$ and $\mathcal{L}_{backward}$, and using the MAE loss for $\mathcal{L}_{tag}$ (ours).

In Fig. 7, we present both training loss and testing error curves (steps ranging from 0 to 300) on CMU-MOSI, where three missing rates of 0, 0.2, and 0.4 are considered. It can be seen that the training loss curves in our method (Fig. 7(d)) fluctuate relatively smoother than other three loss settings (Figs. 7(a)-(c)),



(a) Cosine (b) MAE (c) JS (d) Ours

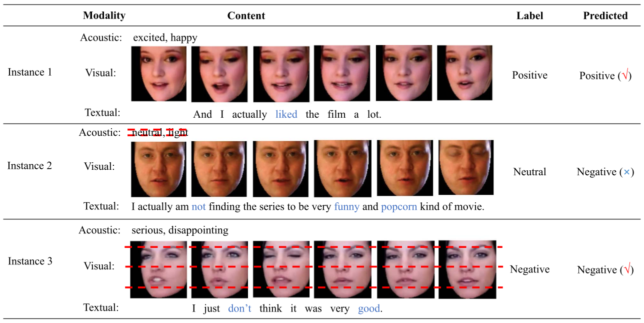
(e) Cosine (f) MAE (g) JS (h) Ours

Fig. 7. Training loss and testing error curves of different settings on CMU-MOSI, where (a)-(d) are training loss curves, and (e)-(h) are testing error curves. (0.0: full modality; 0.2: missing rate 20%; 0.4: missing rate 40%). (a) and (e) are the Cosine similarity loss for $\mathcal{L}_{forward}$, $\mathcal{L}_{backward}$ and $\mathcal{L}_{tag}$, (b) and (f) are the MAE loss for $\mathcal{L}_{forward}$, $\mathcal{L}_{backward}$ and $\mathcal{L}_{tag}$, (c) and (g) are the JS divergence loss for $\mathcal{L}_{forward}$, $\mathcal{L}_{backward}$ and $\mathcal{L}_{tag}$, and (d) and (h) are the JS divergence loss for $\mathcal{L}_{forward}$ and $\mathcal{L}_{backward}$, and MAE loss for $\mathcal{L}_{tag}$.

TABLE X
RESULTS OF DIFFERENT LOSSES, WHERE THE MISSING RATE IS SET AS 0, 0.2, AND 0.4, RESPECTIVELY

Datasets	Loss	0		0.2		0.4	
		M-F1	ACC	M-F1	ACC	M-F1	ACC
CMU-MOSI	Cosine	56.13	81.77	52.46	77.08	50.92	74.90
	MAE	56.83	82.81	52.84	77.60	51.89	75.04
	JS	58.89	84.90	53.84	79.61	52.36	75.52
	ours	60.18	86.54	56.52	82.65	54.67	80.53
IEMOCAP	Cosine	80.74	84.38	78.43	82.03	76.09	80.36
	MAE	80.93	85.03	78.82	82.81	76.24	80.88
	JS	81.08	85.28	79.12	83.57	77.09	81.45
	ours	81.76	86.46	79.57	84.80	77.84	82.97
MELD	Cosine	63.41	65.27	62.06	64.19	60.08	64.33
	MAE	63.73	65.48	62.19	64.43	60.56	64.78
	JS	65.01	67.33	63.59	66.42	61.85	65.22
	ours	65.22	67.72	64.12	67.28	62.17	65.91

validating the reasonableness of the loss setting. In addition, the training loss curves become more fluctuating with the increment of the missing rate, especially when the missing rate is 0.4. Overall speaking, our loss values are lower than those in the cosine similarity, MAE, and JS cases. As Fig. 7(d) shows, our training loss eventually fluctuates in a small interval between (approximately) 12 and 16, with the minimum value being 12.37. Similar findings can be observed in Figs. 7(e)-(h), which are consistent with the ones of the loss curves shown in Figs. 7(a)-(d). Together these results provide important insights that our loss settings are more robust under uncertain missing modalities. Meanwhile, the overall results in Table X further enhance the observation in the figure. As can be noticed, our method achieves the best performance compared to other three loss settings on three datasets. In contrast, the results of applying JS divergence achieve the secondary performance. Since the tag is composed of 4 digits ("0" or "1"), the MAE loss is more straightforward than JS divergence loss. Further analysis of the table suggests that the combination of the JS divergence loss and the MAE loss is beneficial in improving the overall performance.



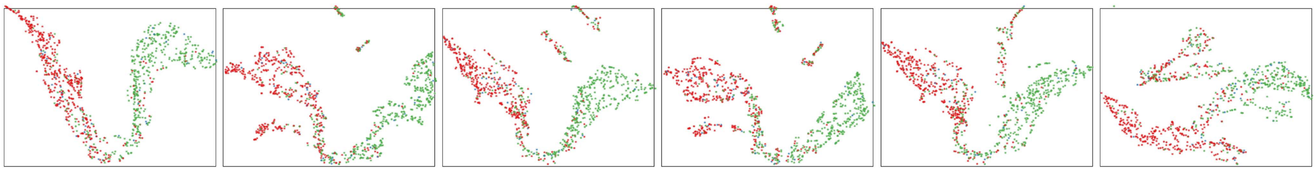
	Modality	Content	Label	Predicted
Instance 1	Acoustic:	excited, happy		
	Visual:		Positive	Positive (√)
	Textual:	And I actually liked the film a lot.		
Instance 2	Acoustic:	neutral, light		
	Visual:		Neutral	Negative (×)
	Textual:	I actually am not finding the series to be very funny and popcorn kind of movie.		
Instance 3	Acoustic:	serious, disappointing		
	Visual:		Negative	Negative (√)
	Textual:	I just don't think it was very good.		

Fig. 8. Case study. The acoustic modality is described with the emotional tone, and the visual modality consists of six facial images. Specifically, the missing modality is marked with dotted red lines, and the semantic related words in the textual modality are marked in blue.

E. Case Study

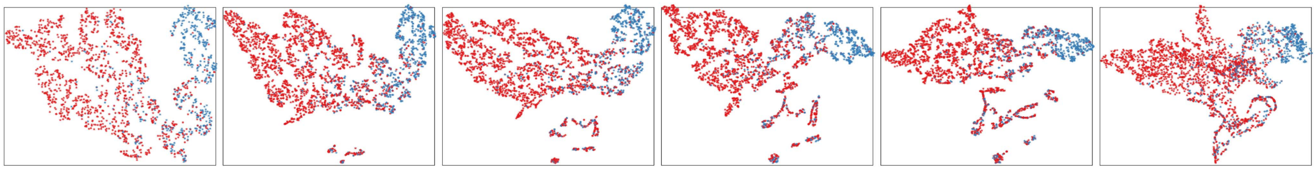
To better understand in which conditions the proposed method works, we present three samples for case analyses. In Fig. 8, the acoustic modality is described with the emotional tone for intuitive expressions, and the visual modality consists of six facial images extracted by OpenFace2.0 tookit. For the textual modality, the semantic related words are marked in blue, and the missing modality is marked with dotted red lines.

In the instance 1, the acoustic modality expresses excited and happy voice; visual features contain a set of happy faces; and the sentiment of the text is positive because of the word "*liked*". Since three modalities are all in positive semantic, our model infers it as positive correctly. In the instance 2, the textual modality expresses negative emotion; the original tone of the acoustic modality is light; and there are no major ripples in facial features. The overall sentiment is neutral when combing three modalities together. However, when the acoustic modality is missing, our model is guided by the text and misclassifies this instance into negative. As for the instance 3, three modalities all



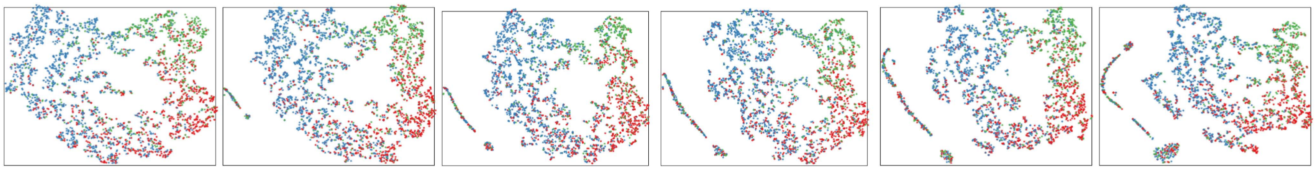
(a) $\eta = 0$ (b) $\eta = 0.1$ (c) $\eta = 0.2$ (d) $\eta = 0.3$ (e) $\eta = 0.4$ (f) $\eta = 0.5$

Fig. 9. Visualization of joint representations with different rates of missing modalities on CMU-MOSI (red: negative, blue: neutral, green: positive). (a) Full modalities, (b) missing rate 0.1, (c) missing rate 0.2, (d) missing rate 0.3, (e) missing rate 0.4, and (f) missing rate 0.5.



(a) $\eta = 0$ (b) $\eta = 0.1$ (c) $\eta = 0.2$ (d) $\eta = 0.3$ (e) $\eta = 0.4$ (f) $\eta = 0.5$

Fig. 10. Visualization of joint representations with different rates of missing modalities on IEMOCAP (red: negative, blue: positive). (a) Full modalities, (b) missing rate 0.1, (c) missing rate 0.2, (d) missing rate 0.3, (e) missing rate 0.4, and (f) missing rate 0.5.



(a) $\eta = 0$ (b) $\eta = 0.1$ (c) $\eta = 0.2$ (d) $\eta = 0.3$ (e) $\eta = 0.4$ (f) $\eta = 0.5$

Fig. 11. Visualization of joint representations with different rates of missing modalities on MELD (red: negative, blue: neutral, green: positive). (a) Full modalities, (b) missing rate 0.1, (c) missing rate 0.2, (d) missing rate 0.3, (e) missing rate 0.4, and (f) missing rate 0.5.

in the negative semantic, especially the facial features and the transcript. Though the visual modality is absent, our model still predicts the instance as negative correctly. In the above cases, our model works well when three modalities express the same emotions; but there is still room for further improvement when different modalities contain contradictory information. Despite the textual modality dominates the overall sentiment, it remains a challenge to trade off the different modalities when they express different emotions. Furthermore, what if the key modality is missing? For example, if the acoustic modality is absent in instance 2, the overall sentiment would change to negative. Thus, how to mark and recover the key modality still remains a challenging task, which needs further investigation in the future research.

F. Visualization

To demonstrate the learning ability of our model, we adopt the T-SNE toolkit to visualize the joint representations under different rates of missing modalities. Specifically, we visualize about 1000 vectors learned by the Transformer encoder on the CMU-MOSI dataset, where the red, the blue, and the green colors denote negative, neutral and positive samples, respectively. As shown in Figs. 9(a)-(e), the overall joint representations obtain the similar distribution as the full modality condition. The majority of vectors are generally divided into three categories, where neutral samples are harder to be classified because of their uncertain semantics. Besides, with the increment of the missing rate, the distributions become more discrete, especially when the missing rate is bigger than 0.3. Apart from that, as can be seen in the top of Figs. 9(b)–(e), the larger rate of missing modalities, the wider outliers. The reason is that the model cannot converge with too many absent samples. While in Fig. 9(f), the decision boundary is closer to the outliers when there are nearly half of missing samples. We suspect that absent samples dominate when training the model, resulting in a quite distinct distribution.

We further present about 3000 learned joint representations containing two categories on the IEMOCAP dataset, where the red and the blue colors denote negative and positive respectively. Due to the removal of neutral samples in the training phase, all learned joint vectors are clearly divided into two groups. Similar to Fig. 9(b)–(e), the outliers in the bottom corner of Fig. 10(b)-(e) increase with the larger number of absent samples. Besides, the decision boundary is closer to the top right-hand corner with the increment of the missing rate. There are two reasons for this phenomenon: 1) the number of negative samples is much greater than the number of positive ones, resulting in samples from the positive group appearing in the area of higher sample density in the negative group; and 2) since we mask all missing modalities into "0," more absent samples are misclassified into the negative group.

As for the visualization results on the MELD datastet, similar observations can be found in Fig. 11, where about 3000 samples are visualized. In addition, it can be seen that three clusters are not so concentrated compared to those of the IEMOCAP dataset. We suspect that visual features may not be well extracted, due to the multiple speakers participating in the dialogues.

V. Conclusion

In this paper, we propose a Tag-Assisted Transformer Encoder (TATE) network to handle the problem of missing uncertain modalities, which can be considered as a unified manner for the robust multimodal sentiment analysis. Owing to the tag encoding technique, the proposed model can cover all uncertain missing cases, and the designed tag recovery loss can in turn supervise the joint representation learning. Besides, the importance of each modality is considered when fusing three modalities, and more general aligned vectors are obtained by the common space projection module with the corresponding fusion weights. Afterwards, the aggregated features are fed into the Transformer encoder-decoder for further learning, where the forward differential loss, the backward reconstruction loss, the tag recovery loss and the classification loss are designed to supervise the learning process. All experiments and further analyses are conducted on CMU-MOSI, IEMOCAP, and MELD datasets, showing the effectiveness of the proposed method. In the future, we will explore the cases when the key modality is missing (as discussed in Section IV-E), and we believe further performance gains could be obtained if the absent key modality can be handled in a satisfactory manner.

References

[1] W. Guo et al., "LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 1785–1798, 2021.

[2] X. Guo, W.-K. A. Kong, and A. C. Kot, "Deep multimodal sequence fusion by regularized expressive representation distillation," *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2022.3142448.

[3] H. Wu, J. Zhou, J. Tian, and J. Liu, "Robust image forgery detection over online social network shared images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13440–13449.

[4] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 929–932.

[5] J. Xu, Z. Li, F. Huang, C. Li, and S. Y. Philip, "Social image sentiment analysis by exploiting multimodal content and heterogeneous relations," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2974–2982, Apr. 2021.

[6] 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.

[7] W. Shao, X. Shi, and S. Y. Philip, "Clustering on multiple incomplete datasets via collective kernel learning," in *Proc. Int. Conf. Data Mining*, 2013, pp. 1181–1186.

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

[9] 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.

[10] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. Assoc. Comput. Linguistics Int. Jt Conf. Nat. Lang. Process.*, 2021, pp. 2608–2618.

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

[12] W. Peng, X. Hong, and G. Zhao, "Adaptive modality distillation for separable multimodal sentiment analysis," *IEEE Intell. Syst.*, vol. 36, no. 3, pp. 82–89, May/Jun. 2021.

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

[14] J. Zeng, T. Liu, and J. Zhou, "Tag-assisted multimodal sentiment analysis under uncertain missing modalities," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1–10.

[15] S. Poria et al., "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov./Dec. 2018.

[16] T. Zhu et al., "Multimodal sentiment analysis with image-text interaction network," *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2022.3160060.

[17] L. Zhang et al., "Multimodal marketing intent analysis for effective targeted advertising," *IEEE Trans. Multimedia*, vol. 24, pp. 1830–1843, 2021.

[18] K. Yang et al., "Crowdtc: Crowd-powered learning for text classification," *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 1, pp. 1–23, 2021.

[19] W. Nie, R. Chang, M. Ren, Y. Su, and A. Liu, "I-GCN: Incremental graph convolution network for conversation emotion detection," *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2021.3118881.

[20] S. Ruan et al., "Color enhanced cross correlation net for image sentiment analysis," *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2021.3118208.

[21] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl. Based Syst.*, vol. 161, pp. 124–133, 2018.

[22] D. Guo, W. Zhou, H. Li, and M. Wang, "Online early-late fusion based on adaptive hmm for sign language recognition," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1, pp. 1–18, 2017.

[23] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3777–3786.

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

[25] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 37–49.

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

[27] C. Shan et al., "VIGAN: Missing view imputation with generative adversarial networks," in *Proc. IEEE Int. Conf. Big Data*, 2017, pp. 766–775.

[28] C. Zhang et al., "Deep partial multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2402–2415, May 2022.

[29] 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, vol. 33, pp. 6892–6899.

[30] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, 2020, pp. 2514–2520.

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

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–14.

[33] 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.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, and D. Warde-Farley, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

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

[36] E.-S. Kim, W. Y. Kang, K.-W. On, Y.-J. Heo, and B.-T. Zhang, "Hypergraph attention networks for multimodal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14581–14590.

[37] H. Akbari et al., “Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text,” in *Proc. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–16.

[38] 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.

[39] C. Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[40] S. Poria et al., “Meld: A multimodal multi-party dataset for emotion recognition in conversations,” in *Proc. Assoc. Comput. Linguist*., 2019, pp. 527–536.

[41] W. Yu, H. Xu, Z. Yuan, and J. Wu, “Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 12, pp. 10790–10797.

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[43] 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. Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[44] W. Yu et al., “Ch-sims: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality,” in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[45] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “OpenFace 2.0: Facial behavior analysis toolkit,” in *Proc. Int. Conf. Auto. Face Amst. Recognit.*, 2018, pp. 59–66.

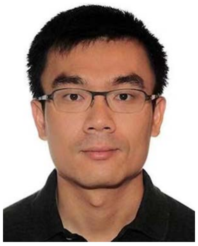
[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Tech*., 2019, pp. 4171–4186.

[47] B. McFee et al., “librosa: Audio and music signal analysis in python,” in *Proc. Python Sci. Conf.*, 2015, vol. 8, pp. 18–25.

[48] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proc. Workshop Int. Conf. Learn. Representations*, 2013, pp. 1–12.

[49] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proc. Empirical Methods Nat. Lang. Process*., 2014, pp. 1532–1543.

Jiandian Zeng received the B.S. degree in computer science and technology from Jianghan University, Wuhan, China, in 2014, and the M.S. degree in computer technology from Huaqiao University, Xiamen, China, in 2018. He is currently working towards the Ph.D. degree with Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China. His research interests include natural language processing, sentiment analysis, and knowledge engineering.



Jiantao Zhou (Senior Member, IEEE) received the B.Eng. degree from the Department of Electronic Engineering, Dalian University of Technology, Dalian, China, in 2002, the M.Phil. degree from the Department of Radio Engineering, Southeast University, Nanjing, China, in 2005, and the Ph.D. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2009. He held Various Research positions with the University of Illinois at Urbana-Champaign, Champaign, IL, USA, Hong Kong University of Science and Technology, and McMaster University, Hamilton, ON, Canada. He is currently an Associate Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China, and also the Interim Head of the newly established Centre for Artificial Intelligence and Robotics. His research interests include multimedia security and forensics, multimedia signal processing, artificial intelligence, and Big Data. He holds four granted U.S. patents and two granted Chinese patents. He has coauthored two papers that were the recipient of the Best Paper Award at the IEEE Pacific-Rim Conference on Multimedia in 2007 and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016. He is also the Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON MULTIMEDIA.

Tianyi Liu received the B.S. degree from the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the Ph.D degree from the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2022. His research interests include relation extraction, natural language processing, and knowledge engineering.