RAMer: Reconstruction-based Adversarial Model for Multi-party Multi-modal Multi-label Emotion Recognition

Xudong Yang[1], **Yizhang Zhu**[1], **Hanfeng Liu**[1], **Zeyi Wen**[1,2], **Nan Tang**[1,2*] and **Yuyu Luo**[1,2]
[1]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2]The Hong Kong University of Science and Technology, Hong Kong SAR, China
sootungyoung@gmail.com, {yzhu305, hliu174}@connect.hkust-gz.edu.cn, {wenzeyi, nantang, yuyluo}@hkust-gz.edu.cn

Abstract

Conventional Multi-modal multi-label emotion recognition (MMER) assumes complete access to visual, textual, and acoustic modalities. However, real-world *multi-party* settings often violate this assumption, as non-speakers frequently lack acoustic and textual inputs, leading to a significant degradation in model performance. Existing approaches also tend to unify heterogeneous modalities into a single representation, overlooking each modality's unique characteristics. To address these challenges, we propose **RAMer** (**R**econstruction-based **A**dversarial **M**odel for **E**motion **R**ecognition), which refines multi-modal representations by not only exploring modality commonality and specificity but crucially by leveraging reconstructed features, enhanced by contrastive learning, to overcome data incompleteness and enrich feature quality. **RAMer** also introduces a personality auxiliary task to complement missing modalities using modality-level attention, improving emotion reasoning. To further strengthen the model's ability to capture label and modality interdependency, we propose a stack shuffle strategy to enrich correlations between labels and modality-specific features. Experiments on three benchmarks, i.e., MEmoR, CMU-MOSEI, and M^3ED, demonstrate that **RAMer** achieves state-of-the-art performance in dyadic and multi-party MMER scenarios.

1 Introduction

Emotion recognition from videos is crucial for advancing human-computer interaction and social intelligence. Multi-modal multi-label emotion recognition (MMER) leverages visual, textual, and acoustic signals to identify multiple emotions (e.g., happy, sad) simultaneously [Zhang *et al.*, 2020; Zhang *et al.*, 2021a]. Conventional MMER methods, as shown in Figure 1(a), typically focus on monologue or dyadic settings, assuming all modalities are fully available. However, real-world conversations often involve multiple participants (i.e., *multi-party*) with incomplete modality data for non-speakers who always lack acoustic and textual signals.

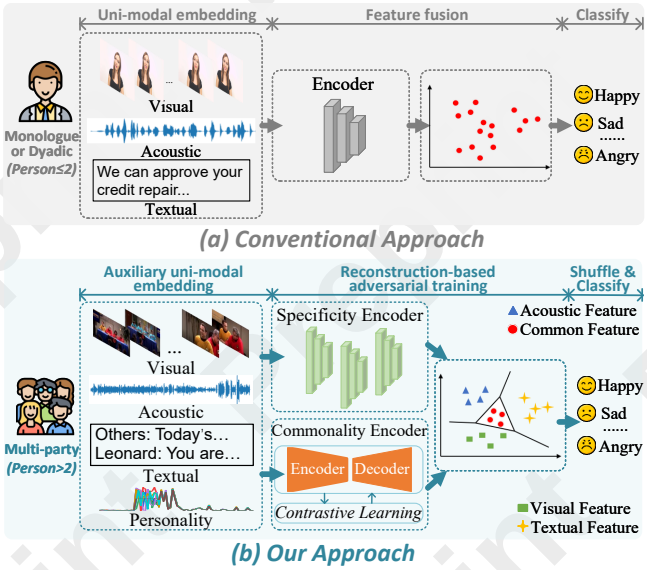


Figure 1: (a) shows the conventional approach for MMER in monologue and dyadic conversations with complete modalities in a uniform representation; (b) depicts our approach for multi-party conversations with incomplete modalities, reconstructing and projecting them into both specificity and commonality representations.

***Multi-party* MMER**, a more complex and practical setting, introduces three key challenges. Firstly, handling incomplete modalities is a significant challenge, which requires robust methods to reconstruct or infer missing information. Most existing works [Zhang *et al.*, 2021a; Zhang *et al.*, 2022; Ge *et al.*, 2023] assume complete modality access and encode each modality independently, overlooking missing data. While some methods [Ghosal *et al.*, 2019; Hu *et al.*, 2021] leverage speaker-aware context modeling, their performance degrades in multi-party settings where non-speakers often lack critical modalities. Secondly, representing diverse modalities effectively remains challenging, often requiring techniques that can not only integrate disparate information but also reconstruct rich, complete representations from potential modalities. Current fusion strategies, such as aggregation-based methods (e.g., concatenation, averaging)[Shen *et al.*, 2020] and hybrid approaches [Manzoor *et al.*, 2023], project modalities into a shared subspace,

* Corresponding author.

often neglecting their unique characteristics and reducing discriminative ability. Recent methods [Zhang *et al.*, 2022] attempt to separate modality-specific and shared features but often suffer from information loss due to inadequate handling of inter-modal correlations. Similarly, methods preserving modality-specific information [Peng *et al.*, 2023] may overlook cross-modal commonalities, limiting their ability to fully capture inter-modal relationships. Finally, multi-label learning presents challenges in modeling robust label correlations and capturing complex interdependency between modalities and labels. Existing approaches [Cisse *et al.*, 2013; Ma *et al.*, 2021] often fail to fully exploit collaborative label relationships. Moreover, emotions vary across modalities, and different emotions rely on distinct modality features, further complicating the task.

To address these issues, we propose **RAMer**, a novel framework designed to tackle the challenges of the ***Multi-party*** **MMER** problem. RAMer integrates multimodal representation learning with multi-label modeling to effectively handle incomplete modalities in multi-party settings.

As illustrated in Figure 1(b), RAMer addresses the challenge of multi-party MMER by following techniques. To address the challenge of incomplete modalities, we propose an auxiliary task that incorporates external knowledge, such as personality traits, to complement the existing modalities. Leveraging this, we employ modality-level attention mechanisms to capture both inter- and intra-personal features. A reconstruction-based network is utilized to recover and enrich the features of any modality by leveraging information from the other modalities.

To represent diverse modalities effectively and capture discriminative features, we design an adversarial network that extracts commonality across modalities while amplifying the specificity inherent to each one. This helps ensure minimal information loss during the fusion process.

Additionally, to model robust interconnections between modalities and labels, we propose a novel modality shuffle strategy. This strategy enriches the feature space by shuffling both samples and modalities, based on the commonality and specificity of the modalities, improving the model's ability to capture label correlations and modality-to-label relationships.

In summary, the contributions of this work are:

- *A Novel Model for the Multi-party MMER Problem.* We present **RAMer**, a new framework that centrally integrates feature reconstruction within an adversarial learning paradigm. **RAMer** adeptly captures both commonality and specificity across modalities, crucially utilizing robustly reconstructed features to significantly improve emotion recognition, especially even with incomplete modality data.
- *Optimization Techniques.* To enhance the robustness of multi-party emotion recognition, **RAMer** employs contrastive learning to enrich reconstructed features and integrates a personality auxiliary task to capture modality-level attention. We also propose a stack shuffle strategy, enhancing the modeling of label correlations and modality-to-label relationships by leveraging the commonality and specificity of different modalities.
- *Extensive Experiments.* We conduct comprehensive experiments on three benchmarks, i.e., MEmoR, CMU-MOSEI, and M^3ED, across various conversation scenarios. Results show that **RAMer** surpasses existing approaches and achieves state-of-the-art performance in both dyadic and multi-party MMER problems.

2 Related Work

Multi-modal Representation Learning. Emotion recognition has progressed from uni-modal approaches [Huang *et al.*, 2021; Saha *et al.*, 2020] to multi-modal methods [Lv *et al.*, 2021; Zhang *et al.*, 2022] that exploit complementary features across modalities. While uni-modal approaches often face recognition biases [Huang *et al.*, 2021], multi-modal learning has gained significant attention, with a key challenge being the effective integration of heterogeneous modalities. Early fusion methods, such as concatenation [Ngiam *et al.*, 2011a], tensor fusion [Liu *et al.*, 2018a], and averaging [Hazirbas *et al.*, 2017], struggle with modality gaps that hinder effective feature alignment. To address this, attention-based methods [Ge *et al.*, 2023; Tsai *et al.*, 2019] leverage cross-attention mechanisms to dynamically align features in the latent space, while contrastive learning [Chen *et al.*, 2020; Peng *et al.*, 2023] further improves robustness. However, most attention-based methods merge modalities into a joint embedding, often overlooking modality-specific features.

Multi-label Emotion Recognition in Videos. MMER involves assigning multiple emotion labels to a target person in the video. Early methods treated multi-label classification as independent binary tasks [Boutell *et al.*, 2004], but advancements explore label correlations using techniques like Adjacency-based Similarity Graph Embedding(ASGE) [You *et al.*, 2020], Graph Convolutional Network (GCN) [Chen *et al.*, 2019], and multi-task pattern [Tsai and Lee, 2020] to explore label correlations. Some noteworthy strategies [Zhang *et al.*, 2021b; Zhang *et al.*, 2022] focus on modeling label-feature correlations through label-specific representations enabled by visual attention [Chen *et al.*, 2019] and transformers [Zhang *et al.*, 2022]. Beyond monologue settings, MMER in conversations has gained interest, using GCN [Ghosal *et al.*, 2019] and memory networks [Hazarika *et al.*, 2018] to model dynamic speaker interactions. However, multi-party MMER is especially challenging, as it requires recognizing emotions for multiple speakers with incomplete modalities. Additionally, the influence of individual personalities on label-feature correlations remains underexplored.

Adversarial Training. Adversarial training (AT) [Goodfellow *et al.*, 2014], involves two models: a discriminator that estimates the probability of samples, and a generator that creates samples indistinguishable from actual data. This setup forms a minimax two-player game, enhancing the robustness of the model. The technique has since been adapted for CV and NLP applications [Wang *et al.*, 2017]. For instance, Miyato et al. [Miyato *et al.*, 2016] extended AT to text categorization by introducing perturbations to word embeddings. Wu et al. [Wu *et al.*, 2017] applied it within a multi-label learning framework to facilitate relationship extraction. Additionally, AT has been used to learn joint distributions between

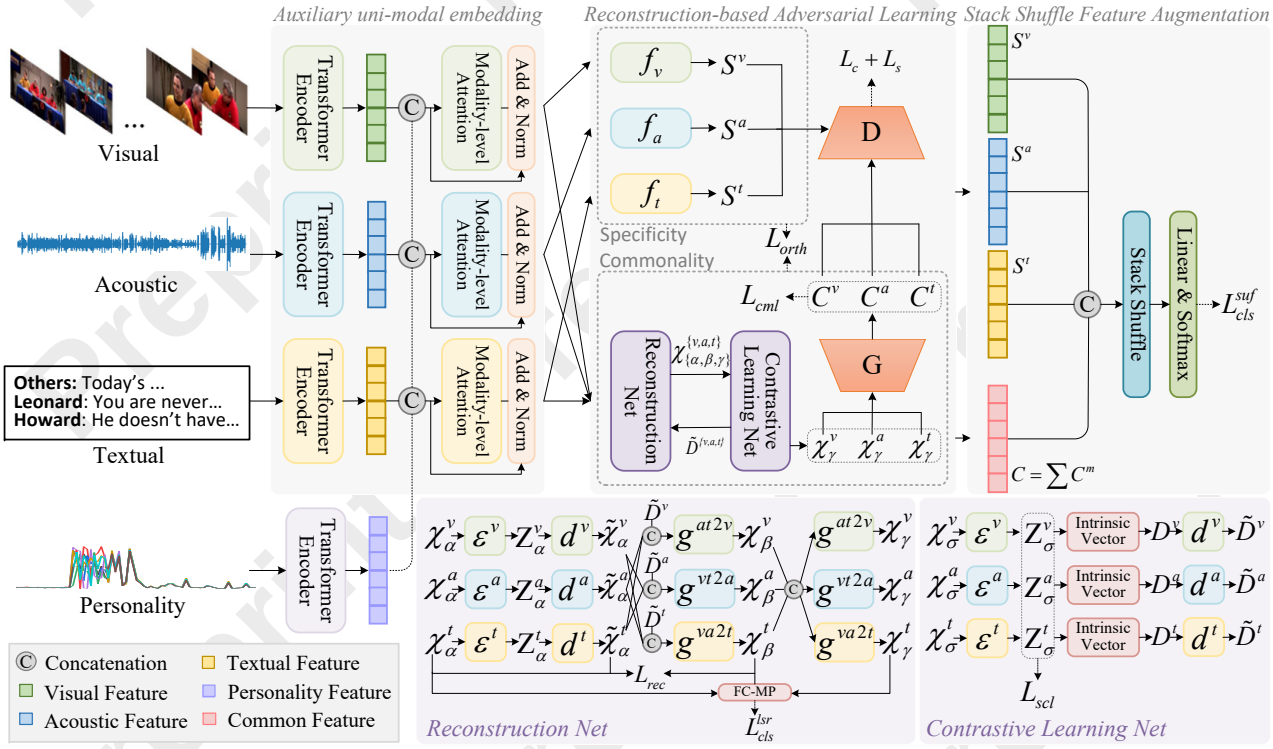


Figure 2: The framework of RAMer. Given incomplete multi-modal inputs, RAMer first encodes each individual modality through an auxiliary task, then feeds the features into a reconstruction-based adversarial network to extract specificity and commonality. Finally, a stack shuffle strategy is employed to learn enhanced representations.

multi-modal [Tsai *et al.*, 2018]. More recently, Ge et al. [Ge *et al.*, 2023] applied AT to reduce modal and data biases in MMER tasks. However, Zhang et al. [Zhang *et al.*, 2022] implemented AT to extract multi-modal commonality and diversity, but suffered a significant loss of modality information due to inadequate cross-modal information fusion.

3 Problem Formulation

In this section, we introduce the notations used and formally define the *Multi-party* Multi-modal Multi-label Emotion Recognition (*Multi-party* MMER) problem.

Notations. We use lowercase letters for scalars (e.g., v), uppercase letters for vectors (e.g., Y), and boldface for matrices (e.g., $\boldsymbol{X}$). A data sample is represented by the tuple $(V, P_t, S_r, E_{t,r})$, where:

- $V = \left(\{P_i\}_{i=1}^{T}, \{S_j\}_{j=1}^{R}\right)$ is a video clip containing T persons and R semantic segments.
- $\{P_i\}_{i=1}^{T}$ refers to the set of target persons, and $\{S_j\}_{j=1}^{R}$ represents the target segments, each annotated with an emotion moment.
- $E_{t,r}$ denotes the labeled emotion for person P_t in S_r.

Each sample involves modalities such as visual (v), acoustic (a), textual (t), and personality traits (p).

For each modality $m \in \{v, a, t, p\}$, the corresponding features are represented as $\left(\boldsymbol{\mathcal{X}}^1, \boldsymbol{\mathcal{X}}^2, \cdots, \boldsymbol{\mathcal{X}}^m\right)$, where $\boldsymbol{\mathcal{X}}^k \in \mathbb{R}^{l_k \times d_k}$ represents the feature space of the k-th modality. Here: l_k denotes the sequence length, and d_k denotes the dimension of the modality. Let $\mathcal{Y} = \{y_1, y_2, \cdots, y_\zeta\}$ represent a label space with ζ possible emotion labels.

Multi-party MMER Problem. Given a training dataset $\mathcal{D} = \left\{\boldsymbol{X}_\tau^{\{1,2,\cdots,m\}}, Y_\tau\right\}_{\tau=1}^{N}$ with N data samples, where: (1) $\boldsymbol{X}_\tau^m \in \boldsymbol{\mathcal{X}}^m$ represents the features for each modality m in sample τ, and (2) $Y_\tau = \{0,1\}^\zeta$ is a multi-hot vector indicating the presence (1) or absence (0) of emotion labels, where $Y_\tau^v = 1$ indicates that sample τ belongs to class v, and $Y_\tau^v = 0$ otherwise. The **goal** of the Multi-party MMER problem is to learn a function $\mathcal{F} : \boldsymbol{\mathcal{X}}^1 \times \boldsymbol{\mathcal{X}}^2 \times \cdots \times \boldsymbol{\mathcal{X}}^m \mapsto \mathcal{Y}$ that predicts the target emotion $E_{t,r}$ for person P_t in segment S_r, leveraging contextual information from multiple modalities.

Discussion. It is important to note that the target person P_t may have incomplete modality information, meaning they may not simultaneously possess visual, textual, or acoustic representations. This introduces uncertainty in the modality of the target segment S_r, making the prediction task more challenging.

4 Methodology

Figure 2 shows the framework of RAMer, which consists of three components: auxiliary uni-modal embedding, reconstruction-based adversarial Learning, and stack shuffle feature augmentation.

4.1 Auxiliary Uni-modal Embedding

To extract contextual information from each modality, we employ four independent transformer encoders [Vaswani *et al.*, 2017], each dedicated to a specific modality m.Each encoder consists of n_m identical layers to ensure consistent and deep representation. For multi-party conversation videos with T participants with incomplete modalities, we introduce an optional auxiliary task leveraging personality to complement missing modalities. Specifically, we concatenate personality embedding $\boldsymbol{\mathcal{X}}^p$ with each modality $\boldsymbol{\mathcal{X}}^m \in \{v, t, a\}$ to enrich the feature space. We then apply the scaled dot-product attention to compute inter-person attention across the person dimension within each segment, and intra-person attention along the segment dimension for each individual [Shen *et al.*, 2020]. This modality-level attention mechanism is designed to enhance the model's emotion reasoning ability by effectively capturing both interpersonal dynamics and temporal patterns within the data. In this way, we obtain personality enhanced representation $\boldsymbol{\mathcal{X}}_\alpha^m \in \mathbb{R}^{l\times d}$.

4.2 Reconstruction-based Adversarial Learning

The second component leverages multiple modalities by capturing inter-modal commonalities while preserving modality-specific features. To address the limitations of adversarial networks [Goodfellow *et al.*, 2014; Zhang *et al.*, 2022], which can result in information loss and difficulty in learning modality-label dependencies, we introduce a reconstruction-based approach. It employs contrastive learning to learn modality-independent but label-relevant representations while reconstructing missing modalities during training to enhance robustness.

Adversarial Training. To balance modality specificity and commonality, we adopt adversarial training to extract discriminative features. The uni-modal embeddings $\boldsymbol{\mathcal{X}}_\alpha^m$ are fed to three fully connected networks f_m to extract specificity $\boldsymbol{S}^m, m \in \{v, a, t\}$. In parallel, $\boldsymbol{\mathcal{X}}_\alpha^m$ are also passed through a reconstruction network, which is coupled with a contrastive learning network, followed by a generator $G(\cdot;\theta_G)$ to derive the commonality $\boldsymbol{C}^m$. Both specificity and commonality are then passed through linear layers with softmax activation in the discriminator $D(\cdot;\theta_D)$ that is designed to distinguish which modality the inputs come from. The generator captures commonality C^m by projecting different reconstructed embedding $\boldsymbol{\mathcal{X}}_\gamma^m$ into a shared latent subspace, ensuring distributional alignment across modalities. Consequently, this architecture encourages the generator $G(\cdot;\theta_G)$ to produce outputs that challenge the discriminator $D(\cdot;\theta_D)$ by obscuring the source modality of $\boldsymbol{C}^m$. The generator and discriminator are jointly trained in a game-theoretic setup to enhance feature robustness against modality-specific biases. Both the commonality adversarial loss $\mathcal{L}_C$ and the specificity adversarial loss $\mathcal{L}_S$ are calculated by cross-entropy loss as,

$$\mathcal{L}_C = -\frac{1}{N}\sum_{m\in\{v,t,a\}}\sum_{\tau=\overline{1}}^{N}\Big(U^m \log\left(D(C_\tau^m;\theta_D)\right)\Big), \quad (1)$$

$$\mathcal{L}_S = -\frac{1}{N}\sum_{m\in\{v,t,a\}}\sum_{\tau=1}^{N}\Big(U^m \log\left(D(S_\tau^m;\theta_D)\right)\Big), \quad (2)$$

where $U \in \{U^v, U^t, U^a\}$ represents the ground truth label corresponding to the discriminator's input.

In the shared subspace, it is advantageous to employ a unified representation of various modalities to facilitate multi-label classification. This representation is designed to eliminate redundant information and extract the elements common to the different modalities, thereby introducing a common semantic loss defined as,

$$\mathcal{L}_{cml} = -\sum_{m\in\{v,t,a\}}\sum_{\tau=1}^{N}\sum_{v=1}^{\zeta} y_\tau^v \log \hat{y}_\tau^{v,m} + (1-y_\tau^v)\log(1-\hat{y}_\tau^{v,m}), \quad (3)$$

where $\hat{y}_\tau^{v,m}$ is predicted with $\boldsymbol{C}_m$ and y_τ^v is the ground-truth label. To encode diverse aspects of multi-modal data, we introduce an orthogonal loss $\mathcal{L}_{orth}$ that encourages the commonality $\boldsymbol{C}^m$ and specificity $\boldsymbol{S}^m$ subspaces to remain distinct by minimizing their overlap.

$$\mathcal{L}_{orth} = -\sum_{m\in\{v,t,a\}}\sum_{\tau=1}^{N}\left\|(\boldsymbol{C}_\tau^m)^T\boldsymbol{S}_\tau^m\right\|_F^2, \quad (4)$$

where $\|\cdot\|_F$ is Frobenius norm. Hereby, the objective of adversarial training $\mathcal{L}_{adv}$ is

$$\mathcal{L}_{adv} = \lambda_a\left(\mathcal{L}_C + \mathcal{L}_S\right) + \lambda_o\mathcal{L}_{orth} + \lambda_c\mathcal{L}_{cml}, \quad (5)$$

where λ_a, λ_o and λ_c are trade-off parameters.

Multi-modal Feature Reconstruction. To reconstruct the features of any modality by leveraging information from the other modalities. We employ a reconstruction network that is composed of modality-specific encoders ε^m, decoders d^m, and a two-level reconstruction process utilizing multi-layer linear networks $g(\cdot)$. Given input $\boldsymbol{\mathcal{X}}_\alpha^m$ from different modality, three encoders ε^m that consist of MLPs are utilized to project $\boldsymbol{\mathcal{X}}_\alpha^m$ into latent embedding $\boldsymbol{Z}_\alpha^m$ within the latent space $\mathcal{S}^z$. Subsequently, three corresponding decoders d^m transform these latent vectors into the decoded vectors $\widetilde{\boldsymbol{\mathcal{X}}}_\alpha^m$. At the first level of reconstruction network, the intrinsic vector $\widetilde{\boldsymbol{D}}^m$ that derived from contrastive learning network and semantic features $\widetilde{\mathcal{X}}_\alpha^{\{v,t,a\}\setminus m}$ are concatenated to form the input, which is processed to produce $\boldsymbol{\mathcal{X}}_\beta^m$ used for the second-level reconstruction network. Hereby, the reconstruction network can be formulated as,

$$\boldsymbol{\mathcal{X}}_\gamma^m = g\left(g\left(d^m(\varepsilon^m(\boldsymbol{\mathcal{X}}_\alpha^m;\theta_m)), \widetilde{\boldsymbol{D}}^m\right)\right). \quad (6)$$

The obtained three embedding $\boldsymbol{\mathcal{X}}_\alpha^m$, $\boldsymbol{\mathcal{X}}_\beta^m$, and $\boldsymbol{\mathcal{X}}_\gamma^m$ from three distinct feature spaces are fed into fully connected network followed by max pooling. We can formulate the reconstruction loss $\mathcal{L}_{rec}$ and classification loss $\mathcal{L}_{cls}^{lsr}$ as,

$$\mathcal{L}_{rec} = \sum_{m=1}^{M}\left(\left\|\boldsymbol{\mathcal{X}}_\alpha^m - \widetilde{\boldsymbol{\mathcal{X}}}_\alpha^m\right\|_F + \left\|\boldsymbol{\mathcal{X}}_\alpha^m - \boldsymbol{\mathcal{X}}_\beta^m\right\|_F\right), \quad (7)$$

$$\mathcal{L}_{cls}^{lsr} = \lambda_\alpha\mathcal{L}_B\left(\mathcal{S}^\alpha, Y\right) + \lambda_\beta\mathcal{L}_B\left(\mathcal{S}^\beta, Y\right) + \lambda_\gamma\mathcal{L}_B\left(\mathcal{S}^\gamma, Y\right), \quad (8)$$

where $\|\cdot\|_F$ is the Frobenius norm, $\lambda_{\alpha,\beta,\gamma}$ are trade-off parameters, $\mathcal{L}_B$ is the binary cross entropy (BCE) loss.

To capture the feature distributions of different modalities and use them to guide the restoration of incomplete modalities, intrinsic vectors $\widetilde{\boldsymbol{D}}^m$ obtained through a supervised contrastive learning network [Khosla *et al.*, 2020] are incorporated into the reconstruction network. The encoders ε^m project input $\boldsymbol{\mathcal{X}}_\sigma^m$ to contrastive embeddings $\boldsymbol{Z}_\sigma^m$,

$\sigma \in \{\alpha, \beta, \gamma\}$. Given a contrastive embedding set $\boldsymbol{\mathcal{Z}} = \left\{ \boldsymbol{Z}_{\sigma}^{\{v,t,a\}} \right\}$, an anchor vector $z_i \in \mathcal{Z}$ and assuming the prototype vector updated during the training process based on the moving average is $\mu_{j,k}^m$, where modality $m \in \{v, t, a\}$, label category $j \in [\zeta]$, label polarity $k \in \{pos, neg\}$, then the intrinsic vector $\widetilde{D}^m$ can be derived from:

$$\delta_j^m = \sum_{k}^{\{pos,neg\}} o_{j,k}^m \cdot u_{j,k}^m, \quad o_{j,k}^m = \frac{\exp(z_i \cdot u_{j,k}^m)}{\sum_{k'}^{\{pos,neg\}} \exp(z_i \cdot u_{j,k'}^m)} \tag{9}$$

$$\widetilde{D}^m = d^m \left(\left[\delta_1^m, \cdots, \delta_\zeta^m \right]; \theta_m \right). \tag{10}$$

The loss of contrastive learning network is defined as,

$$\mathcal{L}_{scl}\left(i, \mathcal{Z}\right) = \sum_{i \in \mathcal{Z}} -\frac{1}{|\boldsymbol{P}(i)|} \sum_{p \in \boldsymbol{P}(i)} \log \frac{\exp(z_i \cdot z_p / \eta)}{\sum_{r \in \boldsymbol{A}(i)} \exp(z_i \cdot z_r / \eta)}, \tag{11}$$

where $\boldsymbol{P}(i)$ is the positive set, $\eta \in \mathbb{R}^+$ is a temperature parameter, and $\boldsymbol{A}(i) = \boldsymbol{\mathcal{Z}} \setminus \{i\}$.

4.3 Stack Shuffle for Feature Augmentation

To construct more robust correlations among labels and model the complex interconnections between modalities and labels, we propose a multi-modal feature augmentation strategy that incorporates a stack shuffle mechanism. After obtaining the commonality and specificity representations, we perform sample-wise and modality-wise shuffling processes sequentially on a batch of samples. To strengthen the correlations between labels, we first apply a sample-wise shuffle. The features derived from $\boldsymbol{C}^m$ and $\boldsymbol{S}^m$ are split into k stacks along the sample dimension, with the top elements of each stack cyclically popped and appended to form new vectors. Next, a modality-wise shuffle is introduced to help the model capture and integrate information across different modalities. For each sample, features are divided into stacks along the modality dimension, and iterative pop-and-append operations are applied. Finally, the shuffled samples $\mathbf{V}$ are used to fine-tune the classifier c_ζ with the binary cross-entropy (BCE) loss.

$$\mathcal{L}_{cls}^{suf} = -\frac{1}{N} \sum_{m \in \{v,t,a\}} \sum_{\tau=1}^{N} \left(Y^m \log \left(c_\zeta(\mathbf{V}^m) \right) \right), \tag{12}$$

Combing the Eq.(5), Eq.(7) $\sim$ Eq.(11) and Eq.(12), the final objective function $\mathcal{L}$ is formulated as,

$$\mathcal{L} = \mathcal{L}_{cls}^{suf} + \mathcal{L}_{cls}^{lsr} + \lambda_r \mathcal{L}_{rec} + \lambda_s \mathcal{L}_{scl} + \mathcal{L}_{adv} \tag{13}$$

where λ_r, λ_s are trade-off parameters.

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. We evaluated RAMer on three benchmark datasets: MEmoR [Shen *et al.*, 2020], a multi-party conversation dataset that includes personality, and CMU-MOSEI [Zadeh *et al.*, 2018] and M^3ED [Zhao *et al.*, 2022], which are dyadic conversation datasets that do not include personality information. The evaluation is conducted under the protocols of these datasets. For CMU-MOSEI and M^3ED, we employed four commonly used evaluation metrics: Accuracy (Acc), Micro-F1, Precision (P), and Recall (R). For MEmoR, we followed the benchmark's protocol and used Micro-F1, Macro-F1, and Weighted-F1 metrics.

Baselines. For the MEmoR dataset, we compare RAMer with multi-party conversation baselines, including MDL, MDAE [Ngiam *et al.*, 2011b], BiLSTM+TFN [Zadeh *et al.*, 2017], BiLSTM+LMF [Liu *et al.*, 2018b], DialogueGCN [Ghosal *et al.*, 2019], DialogueCRN [Hu *et al.*, 2021], and AMER [Shen *et al.*, 2020]. We further assess its robustness against recent dyadic models, including CARAT [Peng *et al.*, 2023] and TAILOR [Zhang *et al.*, 2022]. For the CMU-MOSEI and M^3ED datasets, we test three categories of methods. 1) Classic methods. CC [Read *et al.*, 2011], which concatenates all available modalities as input for binary classifiers. 2) Deep-based methods. ML-GCN [Chen *et al.*, 2019], using Graph Convolutional Networks to map label representations and capture label correlations. 3) Multi-modal multi-label methods. These include MulT [Tsai *et al.*, 2019] for cross-modal interactions, MISA [Hazarika *et al.*, 2020] for learning modality-invariant and modality-specific features, and methods like MMS2S [Zhang *et al.*, 2020], HHMPN [Zhang *et al.*, 2021a], TAILOR [Zhang *et al.*, 2022], AMP [Ge *et al.*, 2023], and CARAT [Peng *et al.*, 2023].

5.2 Comparison with the State-of-the-Art

We present the performance comparisons of RAMer on the MEmoR, CMU-MOSEI, and M3ED datasets in Table 1, Table 2, and Table 3, respectively, with following observations.

1) On the MEmoR dataset, RAMer outperforms all baselines by a significant margin. While TAILOR achieves a high weighted-F1 score in the fine-grained setting, its overall performance is weaker due to biases toward frequent and easier-to-recognize classes. RAMer consistently delivers strong results across all settings, demonstrating its ability to learn more effective representations. 2) On the CMU-MOSEI and M3ED datasets, RAMer surpasses state-of-the-art methods on all metrics except recall, which is less critical compared to accuracy and Micro-F1 in these contexts. 3) Deep-based methods outperform classical ones, highlighting the importance of capturing label correlations for improved classification performance. 4) Multimodal methods like HHMPN and AMP significantly outperform the unimodal ML-GCN, emphasizing the necessity of multimodal interactions. 5) Models optimized for dyadic conversations, such as CARAT, experience a notable performance drop in multi-party settings with incomplete modalities. In contrast, RAMer excels in both scenarios, achieving substantial improvements in Micro-F1 and Macro-F1 scores on the MEmoR dataset.

5.3 Ablation Study

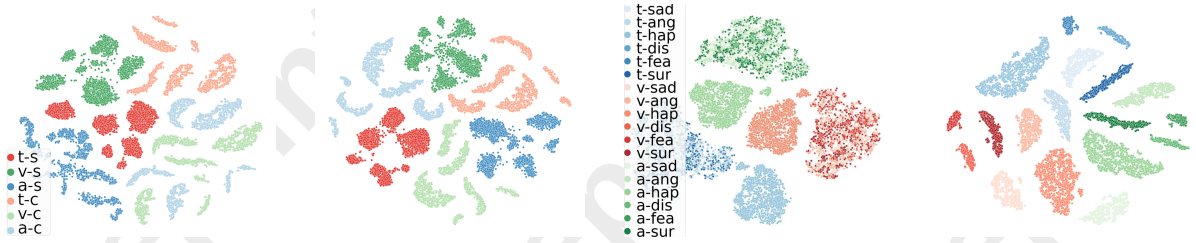
To better understand the importance of each component of RAMer, we compared various ablated variants.

As shown in Table 4, we make the following observations:

- The specificity and commonality enhance MMER performance. Variants (1), (2), and (3) show lower Micro-

Methods	Modality	Primary			Fine-grained		
		Micro-F1	Macro-F1	Weighted-F1	Micro-F1	Macro-F1	Weighted-F1
MDL with Personality	v, a, t, p	0.429	0.317	0.423	0.363	0.217	0.345
MDAE	v, a, t, p	0.421	0.303	0.410	0.363	0.219	0.341
BiLSTM+TFN	v, a, t, p	0.470	0.310	0.454	0.366	0.207	0.350
BiLSTM+LMF	v, a, t, p	0.449	0.294	0.432	0.364	0.198	0.351
DialogueGCN	v, a, t, p	0.441	0.310	0.425	0.373	0.229	0.373
AMER w/o Personality	v, a, t	0.446	0.339	0.440	0.401	0.246	0.379
AMER	v, a, t, p	0.477	0.353	0.465	0.419	0.262	0.400
DialogueCRN	v, a, t, p	0.441	0.310	0.425	0.373	0.229	0.373
TAILOR	v, a, t, p	0.341	0.287	0.326	0.303	0.069	**0.490**
CARAT	v, a, t, p	0.399	0.224	0.422	0.346	0.090	0.483
RAMer	v, a, t, p	**0.499**	**0.402**	**0.503**	**0.431**	**0.299**	0.404

Table 1: Performance comparison on the MEmoR dataset under primary and fine-grained settings. With various modality combinations (visual(v), acoustic(a), textual(t), personality(p)).



(a) w/o Adversarial Training (b) w/ Adversarial Training (c) w/o RN and CLN (d) w/ RN and CLN

Figure 3: t-SNE visualizations of modality embeddings. (a)(b): Specificity and commonality features without/with adversarial training. Color indicates modality (textual, visual, acoustic); saturation differentiates specificity (dark) and commonality (light) components. (c)(d): Reconstruction embeddings without/with RN and CLN, where hue denotes modality and saturation encodes emotion.

Methods	Aligned				Unaligned			
	Acc	P	R	Micro-F1	Acc	P	R	Micro-F1
CC	0.225	0.306	0.523	0.386	0.235	0.320	0.550	0.404
ML-GCN	0.411	0.546	0.476	0.509	0.437	0.573	0.482	0.524
MulT	0.445	0.619	0.465	0.531	0.423	0.636	0.445	0.523
MISA	0.43	0.453	**0.582**	0.509	0.398	0.371	**0.571**	0.45
MMS2S	0.475	0.629	0.504	0.56	0.447	0.619	0.462	0.529
HHMPN	0.459	0.602	0.496	0.556	0.434	0.591	0.476	0.528
TAILOR	0.488	0.641	0.512	0.569	0.46	0.639	0.452	0.529
AMP	0.484	0.643	0.511	0.569	0.462	0.642	0.459	0.535
CARAT	0.494	0.661	0.518	0.581	0.466	0.652	0.466	0.544
RAMer	**0.505**	**0.668**	0.551	**0.604**	**0.469**	**0.660**	0.486	**0.560**

Table 2: Performance Comparison on CMU-MOSEI dataset.

Methods	Acc	P	R	Micro-F1
MMS2S	0.645	0.813	0.737	0.773
HHMPN	0.648	0.816	0.743	0.778
TAILOR	0.647	0.814	0.739	0.775
AMP	0.654	0.819	0.748	0.782
CARAT	0.664	0.824	0.755	0.788
RAMer	**0.665**	**0.826**	**0.759**	**0.791**

Table 3: Performance Comparison on the M^{3}ED dataset

F1 than variant (11). This indicates that jointly learning specificity and commonality yields superior performance, underscoring the importance of capturing both modality-specific specificity and shared commonality.

- Contrastive learning benefits the MMER. The inclusion of loss functions $\mathcal{L}_{scl}$ in adversarial training leads to progressive performance improvements, as evidenced by the superior results of (4).
- Feature reconstruction net benefits MMER. Variants (5), (6), (7) are worse than (11), and (8) shows an 0.045 decrease in Micro-F1, which indicates that feature reconstruction can improve model performance. When the entire reconstruction process is omitted, the performance of (8) declines even more compared to (6) and (7), confirming the effectiveness of multi-level feature reconstruction in achieving multi-modal fusion.
- Changing the fusion order leads to poor performance, variants (9) and (10) perform worse than (11). It validates the rationality and optimality of feature fusion.

5.4 Qualitative Analysis

Visualization of Learned Modality Representations

To evaluate the effectiveness of reconstruction-based adversarial training, we use t-SNE to visualize the commonality and specificity representations learned in the aligned CMU-MOSEI dataset. In figure 3(a), without adversarial training, specificity and commonality are loosely separated, but their distributions overlap in certain areas, such as the lower-right corner. In contrast, Figure 3(b) shows clearer separation of commonality and specificity, forming distinct boundaries and effectively distinguishing emotions across modalities. Figure 3(c) demonstrates that without the reconstruction net (RN) and contrastive learning net (CLN), modality-wise embeddings are distinguishable, but emotion labels within the same

Approaches	Acc	P	R	Micro-F1
(1) w/o L_{sc}	0.474	0.610	0.517	0.573
(2) w/o $\boldsymbol{C}^{\{v,a,t\}}$	0.467	0.612	0.501	0.552
(3) w/o $\boldsymbol{S}^{\{v,a,t\}}$	0.460	0.599	0.491	0.552
(4) w/o L_{scl}	0.492	0.651	0.540	0.588
(5) w/o ε^m, d^m	0.480	0.633	0.524	0.580
(6) w/o $\boldsymbol{\mathcal{X}}_{\beta}^{m}$	0.481	0.641	0.538	0.590
(7) w/o $\boldsymbol{\mathcal{X}}_{\gamma}^{m}$	0.485	0.620	0.523	0.586
(8) w/o $\boldsymbol{\mathcal{X}}_{\beta}^{m}$ + $\boldsymbol{\mathcal{X}}_{\gamma}^{m}$	0.477	0.603	0.490	0.557
(9) $\Lambda\{v,t,a,\boldsymbol{C}\}$	0.489	0.603	0.514	0.564
(10) $\Lambda\{t,a,v,\boldsymbol{C}\}$	0.494	0.650	0.525	0.582
(11) **RAMer**	**0.502**	**0.672**	**0.545**	**0.602**

Table 4: Ablation study on the aligned CMU-MOSEI dataset. Λ refers to the fusion order, and L_{sc} represents the specific and common loss. "w/o ε^m, d^m" denotes the removal of the encoding and decoding processes.

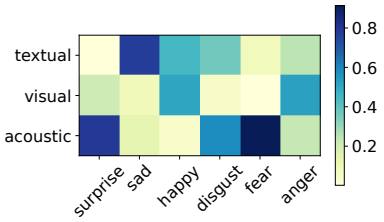


Figure 4: The correlation of modality-to-label dependencies.

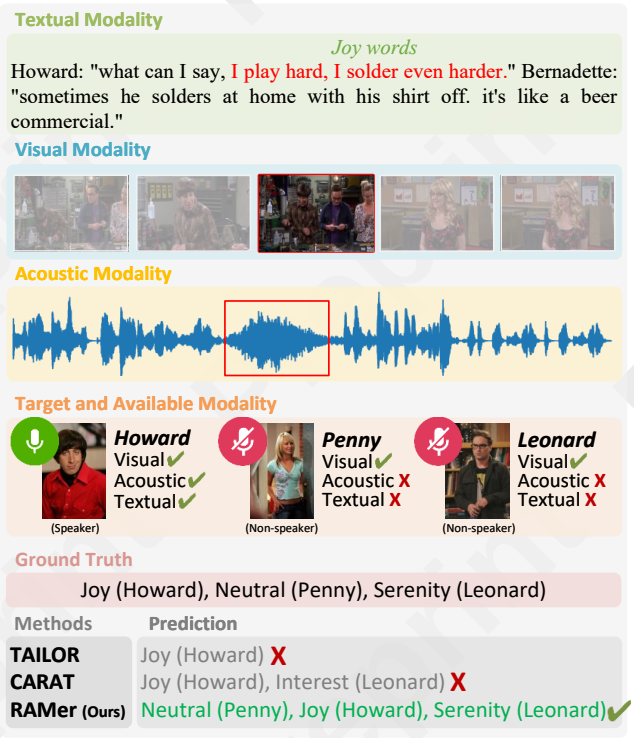


Figure 5: An example of the case study results.

modality remain intermixed. In contrast, Figure 3(d) shows clear separation across both modalities and emotions, indicating that the reconstruction-based module improves representation distinctiveness. Overall, RAMer accurately captures both the commonality and specificity of different modalities.

Visualization of Modality-to-Label Correlations

To explore the relationship between modalities and labels, we visualized the correlation of labels with their most relevant modalities. As shown in Figure 4, regardless of the presence of adversarial training, different emotion label is influenced by different modalities. For instance, surprise is predominantly correlated with the acoustic modality, while anger is primarily associated with the visual modality. This indicates that each modality captures the distinguishable semantic information of the labels from distinct perspectives.

Case Study

To demonstrate RAMer's robustness in complex scenarios, Figure 5 shows an example of MMER on the MEmoR dataset where specific target persons have incomplete modality signals. The top three rows display different modalities from a video clip, segmented semantically with aligned multi-modal signals. Key observations include: 1) The target moment requires recognizing emotions for both the speaker (e.g., Howard) and non-speakers (e.g., Penny and Leonard). While the speaker typically has complete multi-modal signals, non-speakers often lack certain modalities. TAILOR, limited by missing modalities, yields partial predictions as its self-attention mechanism struggles to align labels with missing features. 2) Limitations of single or incomplete modalities. A single modality, such as text, is often insufficient for accurate emotion inference (e.g., only Howard's Joy is detectable from text alone). Although CARAT attempts to reconstruct missing information, it fails to capture cross-modal commonality, leading to incorrect predictions. 3)Inter-person interactions and external knowledge (e.g., personality traits) play important roles. Inter-person attention helps compensate for missing data, while personality-aware reasoning improves emotion inference across participants, highlighting the synergy between user profiling and emotion recognition. Experimental results demonstrate that RAMer achieves superior robustness and effectiveness in complex real-world scenarios.

6 Conclusion

In this paper, we proposed RAMer, a framework that refines multi-modal representations using reconstruction-based adversarial learning to address the Multi-party Multi-modal Multi-label Emotion Recognition problem. RAMer captures both the commonality and specificity across modalities using an adversarial learning module, with reconstruction and contrastive learning enhancing its ability to differentiate emotion labels, even with missing data. We also introduce a personality auxiliary task to complement incomplete modalities, improving emotion reasoning through modality-level attention. Furthermore, the stack shuffle strategy enriches the feature space and strengthens correlations between labels and modalities. Extensive experiments on three datasets demonstrate that RAMer consistently outperforms state-of-the-art methods in both dyadic and multi-party MMER scenarios.

Acknowledgments

We thank anonymous reviewers and scholars for their insightful comments and suggestions which helped us significantly improve an earlier draft of this paper. This paper is supported by NSF of China (62402409), Guangdong provincial project (2023CX10X008), Guangdong Basic and Applied Basic Research Foundation (2023A1515110545), Guangzhou Basic and Applied Basic Research Foundation (2025A04J3935), and Guangzhou-HKUST(GZ) Joint Funding Program (2025A03J3714).

References

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[Chen *et al.*, 2019] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Cisse *et al.*, 2013] Moustapha M Cisse, Nicolas Usunier, Thierry Artieres, and Patrick Gallinari. Robust bloom filters for large multilabel classification tasks. *Advances in neural information processing systems*, 26, 2013.

[Ge *et al.*, 2023] Shiping Ge, Zhiwei Jiang, Zifeng Cheng, Cong Wang, Yafeng Yin, and Qing Gu. Learning robust multi-modal representation for multi-label emotion recognition via adversarial masking and perturbation. In *Proceedings of the ACM Web Conference 2023*, pages 1510–1518, 2023.

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

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Hazarika *et al.*, 2018] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access, 2018.

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

[Hazirbas *et al.*, 2017] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, pages 213–228. Springer, 2017.

[Hu *et al.*, 2021] Dou Hu, Lingwei Wei, and Xiaoyong Huai. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. *arXiv e-prints*, page arXiv:2106.01978, June 2021.

[Huang *et al.*, 2021] Zhiqi Huang, Fenglin Liu, Xian Wu, Shen Ge, Helin Wang, Wei Fan, and Yuexian Zou. Audio-oriented multimodal machine comprehension via dynamic inter-and intra-modality attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13098–13106, 2021.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. *arXiv e-prints*, page arXiv:2004.11362, April 2020.

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

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

[Lv *et al.*, 2021] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562, 2021.

[Ma *et al.*, 2021] Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3855–3864, 2021.

[Manzoor *et al.*, 2023] Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34, 2023.

[Miyato *et al.*, 2016] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for

semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[Ngiam *et al.*, 2011a] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[Ngiam *et al.*, 2011b] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[Peng *et al.*, 2023] Cheng Peng, Ke Chen, Lidan Shou, and Gang Chen. CARAT: Contrastive Feature Reconstruction and Aggregation for Multi-Modal Multi-Label Emotion Recognition. *arXiv e-prints*, page arXiv:2312.10201, December 2023.

[Read *et al.*, 2011] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

[Saha *et al.*, 2020] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, 2020.

[Shen *et al.*, 2020] Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on multimedia*, pages 493–502, 2020.

[Tsai and Lee, 2020] Che-Ping Tsai and Hung-Yi Lee. Order-free learning alleviating exposure bias in multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6038–6045, 2020.

[Tsai *et al.*, 2018] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2017] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, and Fei-Yue Wang. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, 2017.

[Wu *et al.*, 2017] Yi Wu, David Bamman, and Stuart Russell. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783, 2017.

[You *et al.*, 2020] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12709–12716, 2020.

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

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

[Zhang *et al.*, 2020] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3584–3593, 2020.

[Zhang *et al.*, 2021a] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14338–14346, 2021.

[Zhang *et al.*, 2021b] Min-Ling Zhang, Jun-Peng Fang, and Yi-Bo Wang. Bilabel-specific features for multi-label classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(1):1–23, 2021.

[Zhang *et al.*, 2022] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. Tailor versatile multi-modal learning for multi-label emotion recognition. *arXiv e-prints*, page arXiv:2201.05834, January 2022.

[Zhao *et al.*, 2022] Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*, 2022.