Integrating Representation Subspace Mapping with Unimodal Auxiliary Loss for Attention-based Multimodal Emotion Recognition

Xulong Du [1,2], Xingnan Zhang [1], Dandan Wang [1], Yingying Xu [1],
Zhiyuan Wu [3], Shiqing Zhang[1] [*], Xiaoming Zhao[1] [†], Jun Yu[2],Liangliang Lou[1]
[1] Taizhou University, Taizhou, China
[2] Hangzhou Dianzi University, Hangzhou, China
[3] Broward College, Bonaventure Blvd, Weston USA
duxulong23@163.com, tzczsq@163.com, tzxyzxm@163.com

Abstract

Multimodal emotion recognition (MER) aims to identify emotions by utilizing affective information from multiple modalities. Due to the inherent disparities among these heterogeneous modalities, there is a large modality gap in their representations, leading to the challenge of fusing multiple modalities for MER. To address this issue, this work proposes a novel attention-based MER framework associated with audio and text by integrating representation subspace mapping with unimodal auxiliary loss for enhancing multimodal fusion capabilities. Initially, a representation subspace mapping module is proposed to map each modality into two distinct subspaces. One is modality-public, enabling the acquisition of common representations and reducing the discrepancies across modalities. The other is modality-unique, retaining the unique characteristics of each modality while eliminating redundant inter-modal attributes. Then, a cross-modality attention is leveraged to bridge the modality gap in unique representations and facilitate modality adaptation. Additionally, our method designs an unimodal auxiliary loss to remove the redundancy unrelated to emotion classification, resulting in robust and meaningful representations for MER. Comprehensive experiments are conducted on the IEMOCAP and MSP-Improv datasets, and experiment results show that our method achieves superior performance to state-of-the-art MER methods.

Keywords: Multimodal emotion recognition, representation subspace mapping, cross-modality attention, unimodal auxiliary loss, fusion

1. Introduction

Multimodal emotion recognition (MER) plays a crucial role in various domains, including facilitating natural human-machine interaction (Slovak et al., 2023), enhancing intelligent educational tutoring (Sabaritha et al., 2023), contributing to mental health diagnoses (Wang et al., 2022a), and so on. Human beings commonly express their emotions through a combination of verbal and non-verbal cues, such as audio, visual and text modalities (Zhang et al., 2023c). Previous works primarily focus on unimodal emotion recognition in specific modalities, including textual contents (Zhao et al., 2022a; Li et al., 2022; Zhao et al., 2022b), facial expressions (Guo et al., 2023), and audio signals (Zhang et al., 2017, 2019). However, unimodal emotion recognition usually obtains definitely limited performance. To mitigate this problem, there is a growing interest in adopting multimodal approaches for emotion recognition (Lin et al., 2022; Li et al., 2022; Wu et al., 2023). Most MER methods have concentrated on developing advanced multimodal fusion techniques, such as tensor-based fusion methods, and attention-based fusion methods (Zhang et al., 2023c). These fusion techniques still face challenges due to the inherent modality gap among heterogeneous modalities. To address this issue, this work aims to integrate complementary modalities to reduce redundancy and capture comprehensive representations for MER. To this end, we propose a novel attention-based MER framework associated with audio and text, which integrates representation subspace mapping with unimodal auxiliary loss for enhancing multimodal fusion capabilities.

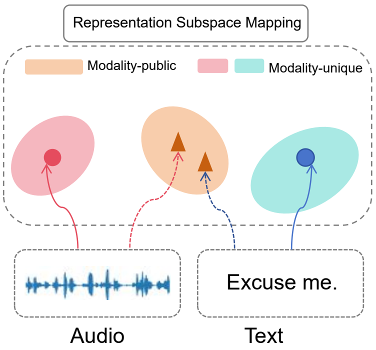

Figure 1: An illustration of multimodal learning of modality-public and modality-unique representations.

[*] Corresponding author
[†] Corresponding author

LREC-COLING 2024, pages 9120–9130
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Figure 1 presents an illustration of multimodal learning of modality-public and -unique representations from audio and text modalities. As depicted in Figure 1, our goal is to learn two distinct representations from audio and text modalities for MER. The first one, referred to as the 'modality-public' subspace, serves as the purpose of bridging the modality gap. To achieve this, we map all modalities of an utterance into a common subspace, thereby capturing the shared representations and reducing the discrepancies across modalities. The second one, known as the 'modality-unique' subspace, aims at retaining the unique representation of each modality while eliminating redundant inter-modal attributes.These unique representations often exhibit little correlation with other modalities and are sometimes considered as redundancy. However, they present a complementarity to the shared representations, thereby providing a more complete and comprehensive representation when combined with the shared representations. To further mitigate the modality gap in these unique representations, a cross-modality attention is employed to bridge the gap and facilitate modality adaptation. This is beneficial for enhancing the fusion of multiple modalities. In addition, we introduce an unimodal auxiliary loss in our model to enhance the representations by applying two unimodal classification cross-entropy constraints. This unimodal auxiliary loss is to eliminate extraneous redundancy while preserving the effectiveness and robustness of unimodal representations.

Our main contributions can be summarized as follows:

- We present a novel attention-based framework integrating audio and text modalities for MER. Our framework incorporates representation subspace mapping and unimodal auxiliary loss to enhance the capability of multimodal fusion. This approach offers a holistic perspective on MER tasks, acquiring both modality-public and modality-unique representations simultaneously.
- The unimodal auxiliary loss is designed to eliminate the redundancy unrelated to emotion classification while preserving efficient unimodal representations. Integrating representation subspace mapping with the designed unimodal auxiliary loss, results in robust and meaningful representations for promoting the overall performance on MER tasks.
- Extensive experiments were conducted on two audio-text emotional datasets, such as IEMOCAP and MSP-Improv, and the results demonstrate that our proposed method outperforms state-of-the-art MER methods.

2. Related Work

Multimodal emotion recognition (MER) aims to distinguish human emotions by integrating multiple modalities like audio, text, visual data, and so on. The diversity among these heterogeneous modalities provides varying levels of affective information for MER. To integrate these diverse modalities, numerous studies have focused on devising various multimodal fusion strategies, such as attention-based methods (Tsai et al., 2019; Lv et al., 2021) and tensor fusion-based methods (Zadeh et al., 2017; Sahay et al., 2018; Wang et al., 2022b). More specially, extensive interests have been attracted to crossmodal attention-based methods (Zhang et al., 2023b), which enable the acquisition of strengthened modality representations by learning cross-modal correlations. The representative MER method is multimodal Transformer (MulT) (Tsai et al., 2019) incorporating a cross-modal attention mechanism. Lv et al. (Lv et al., 2021) developed a Progressive Modality Reinforcement (PMR) technique based on the crossmodal Transformer for MER. PMR provides a message hub to perform information exchanging with each modality.

In contrast, tensor fusion-based methods focus on modeling inter-modality dynamics representations with a tensor fusion strategy. Zadeh et al., (Zadeh et al., 2017) proposed a Tensor Fusion Network (TFN) to capture both the intra-modality and inter-modality dynamics in an end-to-end manner. Wang et al., (Wang et al., 2022b) provided a Deep Tensor Evidence Fusion (DTEF) network, in which a common view evaluation network combining a long short-term memory (LSTM) network and a tensor-based neural network was designed to capture rich intermodal and intramodal features for MER. However, the inherent heterogeneity and redundancy of multimodal representations make these methods face challenges for multimdoal fusion.

To alleviate the above problem, some endeavors are dedicated to investigating the effect of the specific and shared aspects of multimodal representations on multimodal fusion tasks through feature decoupling (Zheng et al., 2021; Zhang et al., 2022b; Li et al., 2023). Li et al., (Li et al., 2023) presented a decoupled multimodal distillation method for MER, in which they employed a flexible and adaptive crossmodal knowledge distillation scheme to enhance the discriminating power of each modality representation. Nevertheless, these methods overlook the potential redundancy hidden in the modality representation that is unrelated to emotion classification.

To address the above issue, this work introduces a unimodal auxiliary loss to eliminate the irrelevant redundancy hidden in the modality representation, thereby enhancing the effectiveness of feature decoupling. Then, a representation subspace mapping strategy is designed to decouple original modality representations into two distinct subspaces for the corresponding multimodal learning of modality-public and modality-unique representations. This gives rise to our proposed attention-based MER framework associated with audio and text to enhance multimodal fusion capabilities.

3. Method

3.1. Problem Definition

A set of dialogues associated with audio (a) and text (t) modalities, denoted as $S = \{D_1, D_2, \ldots, D_N\}$, comprises N dialogues. Each dialogue $D = \{\mathcal{U}_1, E_1, \mathcal{U}_2, E_2, \cdots, \mathcal{U}_M, E_M\}$ consists of M utterances, where $\mathcal{U}_i$ represents the i_{th} utterance, and E_i represents the corresponding emotion label $E_i \in \{Happy, Sad, Neutral, Angry\}$. The main objective of MER is to develop a model that can accurately detect and identify emotion categories in a multi-sensory context from labeled dialogues S.

3.2. Proposed Method

The overall architecture of our proposed method is illustrated in Figure 2. Our method consists of four key components: **(1) unimodal feature extraction** (Sec.3.3), **(2) unimodal auxiliary loss for removing redundancy**(Sec.3.4), **(3) representation subspace mapping** (Sec.3.5) and **(4) modality adaptation with cross-modality attention** (Sec.3.6). In the followings, these four components are described in detail.

3.3. Unimodal Feature Extraction

Utterance-level features are extracted from raw audio and text samples in terms of the used emotional datasets, as described below.

Audio Features: For the IEMOCAP dataset, following in (Wu et al., 2023) we extract the typical 'ComParE' set consisting of 6,373 features for each utterance. Then, we employ a fully connected (FC) layer to reduce the audio features to 100. For the MSP-Improv dataset, following in (Ye et al., 2023), we derive Mel-Frequency Cepstral Coefficients (MFCCs) for each utterance. We set the mean signal length to 96,000 and the embedding length as 100, resulting in 100-dimensional audio features.

Text Features: For text modality, we extract contextual word embeddings by using a pretrained BERT-base model (Kenton and Toutanova, 2019) including 12 transformer layers and 110 million parameters. Then, we employ a mean pooling of all token representations to obtain a 768-dimensional text vector. Finally, we reduce the text vector to 100 through a FC layer.

To further learn rich contextual and high-level information within dialogues, we apply a Bidirectional Gated Recurrent Unit (Bi-GRU) to encode all utterances. Then, a self-attention mechanism (Poria et al., 2017b; Zhang et al., 2023a) is utilized to capture relationships and dependencies between elements in the extracted feature vectors. The output from the used self-attention module for each modality is divided into two branches. One branch is dedicated to emotion recognition tasks, while the other serves as an input for representation subspace mapping.

3.4. Unimodal Auxiliary Loss for Removing Redundancy

The obtained representations of text and audio usually contain redundancy that is irrelevant to emotion classification. To remove redundancy hidden in the obtained representations, we introduce the unimodal auxiliary loss to retain robust unimodal representations as the input of representation subspace mapping. Specifically, we utilize the unimodal representations learned from the self-attention module to calculate the cross-entropy losses, denoted by $\mathcal{L}_a$ and $\mathcal{L}_t$ respectively. Then, we combine these constraints to constitute the unimodal auxiliary loss ($\mathcal{L}_{ua}$). Formally, the unimodal auxiliary loss can be formulated as:

$$\mathcal{L}_a = \sum_{D \in S}^{N} \sum_{i=1}^{M} - \log P(\hat{y}_i^a = y_i) \tag{1}$$

$$\mathcal{L}_t = \sum_{D \in S}^{N} \sum_{i=1}^{M} - \log P(\hat{y}_i^t = y_i) \tag{2}$$

$$\mathcal{L}_{ua} = \mathcal{L}_a + \mathcal{L}_t \tag{3}$$

For each modality the used cross-entropy loss enables the model to retain helpful unimodal representations for emotion classification while eliminating redundant representations unrelated to emotion classification. In this sense, our model is able to preserve the key and discriminating unimodal characteristics of specific emotions in the subsequent representation subspace mapping.

3.5. Representation Subspace Mapping

To map the multimodal representations into two parts, namely the modality-public representation $\mathcal{U}_m^{pub}$ and the modality-unique representation $\mathcal{U}_m^{uni}$, where m$\in \{a, t\}$ indicates a modality, we employ a public multimodal encoder $\mathcal{E}^{pub}$ and two unique multimodal encoders $\mathcal{E}_m^{uni}$ to learn the mapped representations, as defined below:

$$\mathcal{U}_m^{pub} = \mathcal{E}^{pub}(\mathcal{U}_m), \quad \mathcal{U}_m^{uni} = \mathcal{E}_m^{uni}(\mathcal{U}_m) \tag{4}$$

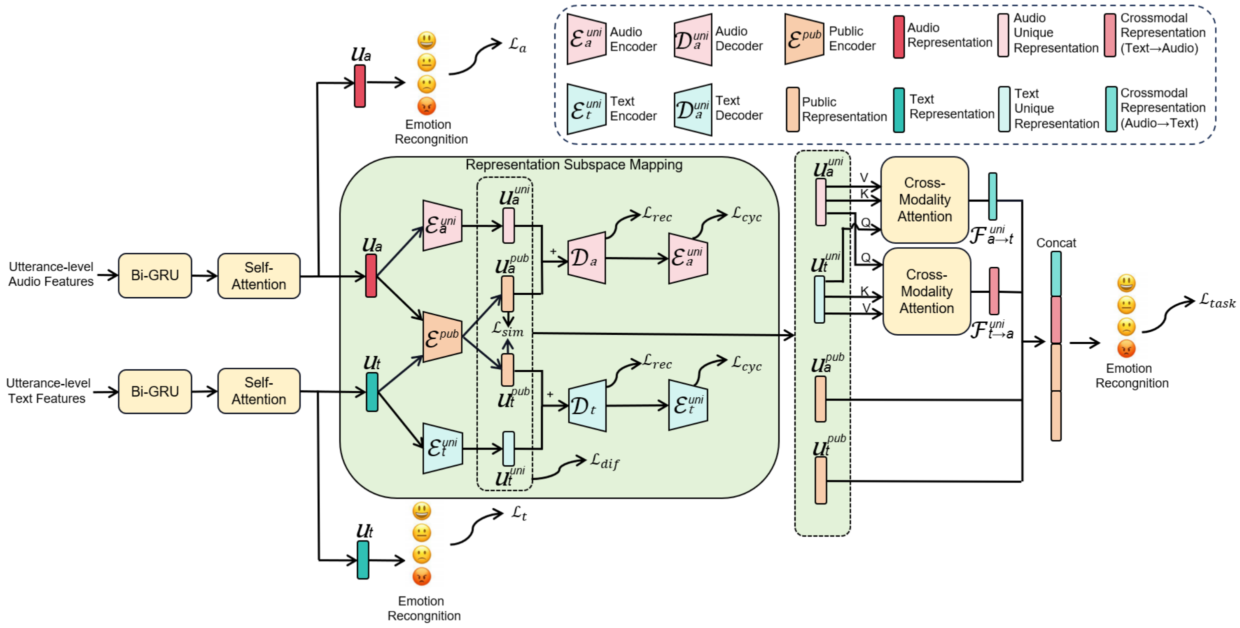

Figure 2: Overview of our proposed method integrating representation subspace mapping with unimodal auxiliary loss for MER.

To bridge the gap between text public representation $\mathcal{U}_t^{pub}$ and audio public representation $\mathcal{U}_a^{pub}$, we employ the Central Moment Discrepancy (CMD) metric (Zellinger et al., 2017) to measure the difference $\mathcal{L}_{sim}$ between the distributions of these two representations. This is because CMD explicitly matches higher-order moments without the need for computationally expensive distance and kernel matrix computations. To this end, in modality-public representation subspace we aim to minimize the following function:

$$\begin{aligned}\mathcal{L}_{sim} &= \|E(\mathcal{U}_t^{pub}) - E(\mathcal{U}_a^{pub})\|_2 \\ &\quad + \sum_{k=2}^{K} \|C_k(\mathcal{U}_t^{pub}) - C_k(\mathcal{U}_a^{pub})\|_2\end{aligned} \tag{5}$$

where $E(\mathcal{U}_m^{pub})$ is the empirical expectation vector of input sample $\mathcal{U}_m^{pub}$, and $C_k(\mathcal{U}_m^{pub}) = E((\mathcal{U}_m^{pub} - E(\mathcal{U}_m^{pub}))^k)$ is the vector of all k^{th} order sample central moments of the coordinates of $\mathcal{U}_m^{pub}$. $\|\cdot\|_2$ is the squared Frobenius norm.

In modality-unique representation subspace, we leverage difference constraint $\mathcal{L}_{dif}$ to constrain unique representations $\mathcal{U}_m^{uni}$, retaining specific representations and removing redundant representations, as defined below:

$$\begin{aligned}\mathcal{L}_{dif} &= \|\mathcal{U}_t^{uni^T}\mathcal{U}_a^{uni}\|_2 + \|\mathcal{U}_a^{uni^T}\mathcal{U}_a^{pub}\|_2 \\ &\quad + \|\mathcal{U}_t^{uni^T}\mathcal{U}_t^{pub}\|_2\end{aligned} \tag{6}$$

where T represents the transpose operation of a matrix.

By enforcing $\mathcal{L}_{dif}$, there remains a risk of learning trivial representations by the modality-specific encoders. Trivial cases can arise if the encoder function approximates an orthogonal but unrepresentative vector of the modality. To mitigate this issue, we introduce a modality reconstruction loss to ensure that the hidden representations capture the details of their respective modalities. In particular, we sum $\mathcal{U}_m^{pub}$ and $\mathcal{U}_m^{uni}$ from each modality as the input to the unique decoder $\mathcal{D}_m$ to reconstruct the corresponding representations, *i.e.*, $\mathcal{D}_m(\mathcal{U}_m^{pub} + \mathcal{U}_m^{uni})$. Then, the discrepancy between the raw and reconstructed multimodal representations can be formulated as:

$$\mathcal{L}_{rec} = \sum_{m\in\{a,t\}} \frac{\|\mathcal{U}_m - \mathcal{D}_m(\mathcal{U}_m^{pub} + \mathcal{U}_m^{uni})\|_2^2}{d_h} \tag{7}$$

where, $\|\cdot\|_2^2$ is the squared L_2-norm and d_h represents the size of $\mathcal{U}_m$.

Drawing inspiration from the field of generative workers (Zhu et al., 2017), we incorporate a cycle consistency loss ($\mathcal{L}_{cyc}$) to further promote the process of representation reconstruction. The reconstructed representations will be re-encoded via the unique encoders $\mathcal{E}_m^{uni}$ to regress the unique representations $\mathcal{U}_m^{uni}$. In this sense, the discrepancy between the regressed and unique representations can be formulated as:

$$\begin{aligned}\mathcal{L}_{cyc} &= \frac{\|\ \mathcal{U}_a^{uni} - \mathcal{E}_a^{uni}(\mathcal{D}_a(\mathcal{U}_a^{pub} + \mathcal{U}_a^{uni}))\|\ _2^2}{d_h} \\ &\quad + \frac{\|\ \mathcal{U}_t^{uni} - \mathcal{E}_t^{uni}(\mathcal{D}_t(\mathcal{U}_t^{pub} + \mathcal{U}_t^{uni}))\|\ _2^2}{d_h}\end{aligned} \tag{8}$$

Finally, we combine the above constraints to form the whole subspace loss, as defined below:

$$\mathcal{L}_{sp} = \mathcal{L}_{rec} + \mathcal{L}_{cyc} + \alpha(\mathcal{L}_{sim} + \mathcal{L}_{dif}) \quad (9)$$

where α is the balance factor.

3.6. Modality Adaptation with Cross-modality attention

The unique representations emphasize the diversity and distinct characteristics of each modality, thus manifesting a substantial modality gap. To bridge the modality gap, we employ the multimodal Transformer strategy (Tsai et al., 2019) to perform modality adaptation. More specially, when taking the audio modality $\mathcal{U}_a^{uni}$ as the source and the text modality $\mathcal{U}_t^{uni}$ as the target, the cross-modality attention can be defined as: $Q_t = \mathcal{U}_t^{uni} P_q$, $K_a = \mathcal{U}_a^{uni} P_k$, and $V_a = \mathcal{U}_a^{uni} P_v$. Here, P_q,P_k,P_v are the learnable parameters. Q_t, K_a and V_a separately represent $Query$, Key, and $Value$. The individual head of cross-modality attention can be expressed as:

$$\mathcal{F}_{a \to t}^{uni} = softmax(\frac{Q_t K_a^T}{\sqrt{d}})V_a \quad (10)$$

Here, $\mathcal{F}_{a \to t}^{uni}$ denotes the representation enhancement from audio to text modality. d is the dimension of Q_t and K_a.

Similarly, the individual head of cross-modality attention, which takes the text modality $\mathcal{U}_t^{uni}$ as the source and the audio modality $\mathcal{U}_a^{uni}$ as the target, can be expressed as:

$$\mathcal{F}_{t \to a}^{uni} = softmax(\frac{Q_a K_t^T}{\sqrt{d}})V_t \quad (11)$$

Then, we concatenate all public representations and enhanced cross-modal representations from the source to target modalities as the final fused features, *i.e.*,$[\mathcal{F}_{a \to t}^{uni}, \mathcal{F}_{t \to a}^{uni}, \mathcal{U}_a^{pub}, \mathcal{U}_t^{pub}]$. The notation $[\cdot]$ means feature concatenation. Next, the fused features are used to conduct MER tasks in terms of the following cross-entropy loss $\mathcal{L}_{task}$:

$$\mathcal{L}_{task} = \sum_{D \in S}^{N} \sum_{i=1}^{M} -\log P(\hat{y}_i = y_i) \quad (12)$$

Finally, we merge the above-mentioned loss functions into a total target loss, as defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{ua} + \beta \mathcal{L}_{sp} \quad (13)$$

where β is the balance factor.

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset

We evaluate our proposed method on two benchmark datasets: Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2008) and MSP-Improv (Busso et al., 2016). The statistics of these two datasets are listed in Table 1.

Dataset	Happy	Angry	Sad	Neutral	Total
IEMOCAP	1636	1103	1084	1708	5531
MSP-Improv	999	460	627	1733	3819

Table 1: Data statistics of datasets.

IEMOCAP This dataset records 9 emotions from 10 actors: happy, angry, sad, disgust, neutral, excited, surprised, frustrated and fearful. Following in (Zhang et al., 2023b), we use 4 emotion categories for experiments, in which happy and excited are merged into the happy category. The final dataset contains 5,531 utterances in total.

MSP-Improv This dataset records 4 emotions from 12 actors: happy, angry, sad and neutral. As in (Zhang et al., 2022a) , our initial step involves the removal of audio clips with a duration shorter than 1 second. Then, we meticulously select audio clips along with their corresponding textual transcripts from the 'Other-improvised' group. These recordings are specifically derived from improvisational scenarios and includes a total of 3,819 utterances.

4.1.2. Implementation Details

Description	Symbol	Value
Batch size	b	20
Epoch number	e	150
Dropout rate	p	0.2
Hidden size	d	100
Self- and cross-modality attention heads	h	4
L2 regularization weight	w	0.00001
Loss balance factor	α, β	0.1

Table 2: Parameter settings in experiments.

All experiments were conducted using the PyTorch deep learning toolkit on the Nvidia RTX 3090 GPU with 24GB of memory. The experiment configurations are detailed in Table 2. The hidden size is 100, which is equal to the feature dimension of modality-unique and cross-modality attention embeddings. The dropout rate is 0.2, and is employed after self-attention and cross-modality attention layers. The epoch number is 150, and the batch size is 20. The L2 regularization weight value is 0.00001. On the IEMOCAP dataset, the first 4 sessions were used for training and the fifth session was employed for testing. On the MSP-Improv dataset, we leveraged a 12-fold speaker-independent cross-validation strategy for experiments. Specially, 10 speakers' utterances were selected for training and the remaining 2 speakers were adopted for testing. The obtained average results in the 12-fold cross-validation experiments were used as the final reported results.

Dataset	Modality	Happy(%)	Sad(%)	Neutral(%)	Angry(%)	Average(%)
IEMOCAP	A	57.11	64.49	66.15	**91.18**	69.73
	T	90.74	88.16	72.39	71.18	80.62
	A+T (w/o $\mathcal{L}_{ua}$)	86.00	87.76	66.15	86.47	81.10
	A+T	**94.01**	**88.30**	**73.40**	88.00	**85.92**
MSP-Improv	A	84.55	72.73	67.21	51.79	69.07
	T	87.27	69.32	72.13	**90.18**	82.00
	A+T (w/o $\mathcal{L}_{ua}$)	87.27	77.27	**83.61**	83.04	82.80
	A+T	**95.01**	**84.86**	80.01	87.12	**86.75**

Table 3: Performance comparisons of our method for unimodal and multimodal emotion recognition on the IEMOCAP and MSP-Improv datasets. A and T refer to the audio and text modality, respectively. (w/o $\mathcal{L}_{ua}$) means to remove $\mathcal{L}_{ua}$.

For evaluation metric, on the IEMOCAP dataset we select Weighted Accuracy (WA) for a comparison. On the MSP-Improv dataset, we choose F1-score as the evaluation metric due to the imbalance of emotion categories.

4.2. Experimental Results and Analysis

4.2.1. Unimodal vs. Multimodal

To present a performance evaluation of unimodal and multimodal emotion recognition, we employ the achieved results on the IEMOCAP and MSP-Improv dataset for a comparison. Table 3 shows the performance comparisons of unimodal and multimodal emotion recognition on the IEMOCAP and MSP-Improv dataset. According to the results presented in Table 3, the text modality exhibited superior performance compared to the audio modality, achieving an improvement of 10.89$\%$ and 12.93$\%$ on the IEMOCAP and MSP-Improvs datasets, respectively. This indicates that the textual modality contains much richer emotional clues for identifying emotion than the audio modality. In addition, integrating both audio and text modalities clearly outperforms their unimodal counterparts. In particular, when fusing these two modalities, the reported accuracy is 85.92$\%$ and 86.75$\%$ on two datasets, which is much higher than the unimodal audio and text modality. This demonstrates the complementarity of audio and text modalities. Besides, Table 3 presents a performance comparison of our method when integrating both audio and text modalities whether or not with the unimodal auxiliary loss. The results in Table 3 show that the designed unimodal auxiliary loss yields an improvement of 4.82$\%$ and 3.95$\%$ on two datasets when it is leveraged for fusing both audio and text modalities. This indicates the effectiveness of the designed unimodal auxiliary loss for removing redundancy, before performing representation subspace mapping and feature fusion.

Approaches	WA(%)
bc-LSTM(Poria et al., 2017a)	75.60
CATF-LSTM(Poria et al., 2017b)	80.10
Zheng.(Lian et al., 2019)	78.02
DANN (Lian et al., 2020)	82.68
CONSK-GCN(Fu et al., 2021)	84.79
Wen. (Wu et al., 2021)	83.08
Soumya. (Dutta and Ganapathy, 2022)	83.80
MER-HAN (Zhang et al., 2023b)	73.33
Bubai. (Maji et al., 2023)	83.57
Our Method (A)	69.73
Our Method (T)	80.62
Our Method (A+T)	**85.92**

Table 4: Performance comparisons of different methods on the IEMOCAP dataset. A and T refer to the audio and text modality, respectively.

Approaches	Metric	A	T	A+T
MCTN(Pham et al., 2019)	F1(%)	32.85	50.50	56.11
MMIN(Zhao et al., 2021)	F1(%)	46.47	55.73	61.88
Bi-LSTM	F1(%)	44.06‡	60.04‡	63.57‡
GRU	F1(%)	43.49‡	73.92‡	73.18‡
Bi-GRU	F1(%)	49.08‡	82.27‡	82.70‡
Our Method	F1(%)	**55.50**	**82.68**	**85.26**

Table 5: Performance comparisons of different methods on the MSP-Improv dataset. ‡ indicates the obtained results of reproducing the corresponding methods. A and T refer to the audio and text modality, respectively.

4.2.2. Our Method vs. Previous Works

To show the advantages of our proposed method, we provide a comparison of our method and state-of-the-art approaches. Table 4 and 5 separately show the performance comparisons of our method as well as other comparing approaches on the IEMOCAP and MSP-Improv datasets. In particular, on these two datasets our method performs best among all used comparing approaches. This highlights the advantages of our method over other used approaches. Moreover, on the MSP-Improv dataset our method obtains much higher F1-score than other comparing approaches when performing unimodal and multimodal emotion classification. This is attributed to integrating representation subspace mapping with the unimodal auxiliary loss for enhancing multimodal fusion capabilities.

To show the recognition accuracy of each emotion, Figure 3 displays the confusion matrices of recognition results obtained by our method on two datasets. As depicted in Figure 3, we can see that our method yields excellent performance on these two datasets. Specially, on both datasets, happy, sad, and angry can be identified well with an accu-

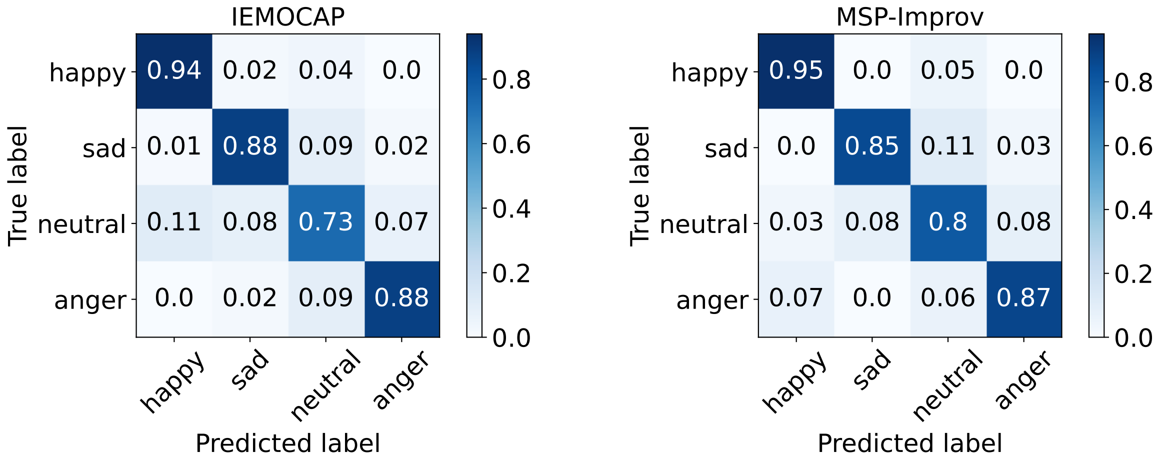

Figure 3: The confusion matrices of recognition results obtained by our method on two datasets: (left) IEMOCAP, (right) MSP-Improv.

Dataset	$\mathcal{L}_{ua}$	$\mathcal{L}_{sp}$	Cross-modality Attention	WA(%)	F1(%)
IEMOCAP	✓	✓	✓	**85.92**	**84.48**
	✓	✓	✗	85.23(↓)	83.96(↓)
	✓	✗	✗	84.79(↓)	83.02(↓)
	✗	✗	✗	80.85(↓)	78.91(↓)
MSP-Improv	✓	✓	✓	**86.75**	**85.26**
	✓	✓	✗	86.13(↓)	84.83(↓)
	✓	✗	✗	84.64(↓)	84.29(↓)
	✗	✗	✗	79.36(↓)	78.33(↓)

Table 6: The effect of key components in our method.

Dataset	Methods	w/o $\mathcal{L}_{ua}$	w/ $\mathcal{L}_{ua}$
		WA (%) / F1 (%)	WA (%) / F1 (%)
IEMOCAP	Bi-LSTM	71.36/69.23	73.21/71.28
	Bi-GRU	83.14/82.01	84.64/83.63
	Our Method	81.84/80.51	**85.92/84.48**
MSP-Improv	Bi-LSTM	69.54/67.63	71.26/69.28
	Bi-GRU	83.35/82.64	84.93/84.05
	Our Method	82.80/81.34	**86.75/85.26**

Table 7: The effect of unimodal auxiliary loss ($\mathcal{L}_{ua}$) on the IEMOCAP and MSP-Improv datasets.

racy of over 85%. By contrast, neutral is classified relatively poorly with an accuracy of 73% on the IEMOCAP dataset, and 80% on the MSP-Improv dataset, respectively. This further emphasizes the effectiveness and robustness of our method.

4.2.3. The Effect of Representation Subspace Mapping

In order to efficiently reduce the modality gap while retaining the unique representations of each modality simultaneously, we design a representation subspace mapping module to project the representations of each modality into both a public subspace and a unique subspace. To intuitively investigate the effect of representation subspace mapping, we provide the visualization results of mapped representations as well as the quantitative analysis of the regularization losses, as described below.

Visualizing mapped representations. We present the t-SNE visualization results of public representations $\mathcal{U}_m^{pub}$ and unique representations $\mathcal{U}_m^{uni}$ on two datasets, as shown in Figure 4. Here, the balance factor α in the whole subspace loss $\mathcal{L}_{sp}$, is used to control the effect of $\mathcal{L}_{sim}$ and $\mathcal{L}_{dif}$ regularization. Specially, $\alpha = 0$ indicates the absence of constraints imposed by $\mathcal{L}_{sim}$ and $\mathcal{L}_{dif}$ regularization. By contrast, $\alpha \neq 0$ signifies the presence of these constraints. When using $\mathcal{L}_{sim}$ and $\mathcal{L}_{dif}$ regularization, the results in Figure 4 clearly show that public representations tend to overlap, while unique representations have more distinct clusters from each other. This indicates that our method has successfully learned both public and unique representations.

Quantitative analysis. To quantify whether our model can learn public and unique representations or not through the designed representation subspace mapping, we use the subspace loss functions $\mathcal{L}_{sp}$ and $\mathcal{L}_{sim}$ to measure the ability of learning representations. Figure 5 depicts the trends in the regularization loss during training. From Figure 5, we can see that with the increase of training epochs, $\mathcal{L}_{sp}$ can converge rapidly. This rapid convergence indicates that our model can capture meaningful representations through the representation subspace mapping. Meanwhile, the rapid convergence of $\mathcal{L}_{sim}$ with increasing epochs also demonstrates our model's ability of learning modality-public representations.

4.2.4. Ablation Study

In this section, we present a comprehensive analysis of several key components in our method, includ-

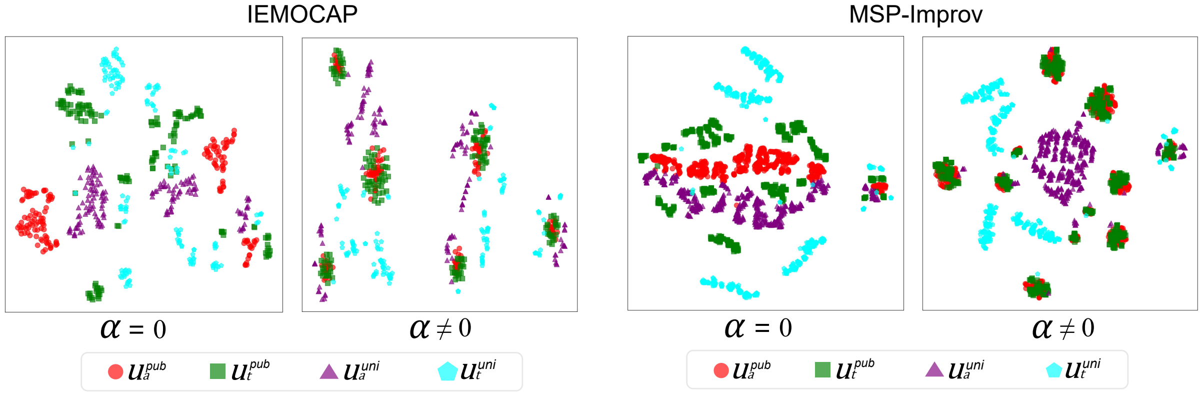

Figure 4: t-SNE visualizations of mapped features: (left) IEMOCAP, (right) MSP-Improv.

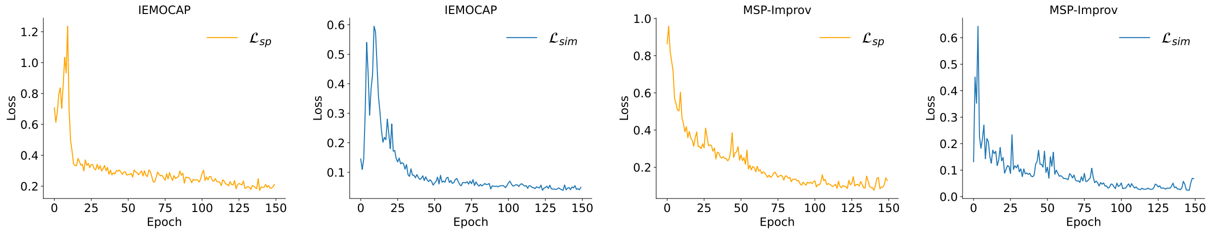

Figure 5: Trends in the regularization loss during training on the IEMOCAP and MSP-Improv datasets.

ing the unimodal auxiliary loss, the subspace loss, and cross-modality attention unit. The results of various evaluation are presented in Table 6. From Table 6, our findings can be summarized as follows:

(1) The used unimodal auxiliary loss significantly enhances MER performance. This is attributed to the unimodal auxiliary loss which is able to filter out the representations that are unrelated to emotion classification, and yield more meaningful representations for downstream tasks.

(2) Combining the unimodal auxiliary loss with the subspace loss yields the best performance. The representation subspace mapping leverages the subspace constraints to decompose raw representations into both public and unique representations. The public representations can substantially reduce the discrepancies across modalities. The unique representations can eliminates redundant inter-modal attributes while preserving modality-unique characteristics.

(3) The cross-modality attention can effectively bridge the modality gap in unique representations and facilitate modality adaptation. In this sense, our model can diminish the distribution differences among unique representations from different modalities, and enhance modality adaptation simultaneously.

To further validate the effectiveness of the proposed unimodal auxiliary loss, we also provide a comparative analysis when whether using the unimodal auxiliary loss or not on the IEMOCAP and MSP-Improv dataset. In this study, we compared our method with classic models such as Bi-LSTM and Bi-GRU. The comparing results are presented in Table 7. Here, "Bi-LSTM (w/ $\mathcal{L}_{ua}$)" and "Bi-GRU (w/ $\mathcal{L}_{ua}$)" indicate that we integrate the unimodal auxiliary loss with Bi-LSTM and Bi-GRU to eliminate emotion-irrelevant representations. The results in Table 7 demonstrate that the used models incorporating $\mathcal{L}_{ua}$ clearly exhibit an improvement compared to those without $\mathcal{L}_{ua}$. This observation highlights the positive impact of the introduced unimodal auxiliary loss on both fusion and representation subspace mapping. Moreover, the key difference between our model and Bi-GRU is whether representation subspace mapping is incorporated or not. From Table 7, we can see that representation subspace mapping in our model contributes significantly to the overall performance improvement with the aid of the unimodal auxiliary loss.

5. Conclusion

This work presents a novel attention-based audio-text emotion recognition framework which integrates the proposed representation subspace mapping and unimodal auxiliary loss. The designed representation subspace mapping can project raw representations into modality-public and modality-unique subspaces. The unimodal auxiliary loss can filter out the irrelevant redundancy, ensuring robust and meaningful representations for MER. Experimental results on the IEMOCAP and MSP-Improv datasets show that our method outperforms state-of-the-art approaches on MER tasks. In future, it is interesting to explore alternative multimodal fusion methods, and integrate more modalities such as visual and physiological signals related to emotion expression for MER (Zhao et al., 2022a; Can et al., 2023).

6. Acknowledgments

This work was supported by the grant from the National Natural Science Foundation of China (NSFC) project (No. 62276180,61976149,62101375). It was also supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ24F020014,LQ21F020001.

7. References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Yekta Said Can, Bhargavi Mahesh, and Elisabeth André. 2023. Approaches, applications, and challenges in physiological emotion recognition—a tutorial overview. *Proceedings of the IEEE*, 111(10):1287–1313.

Soumya Dutta and Sriram Ganapathy. 2022. Multimodal transformer with learnable frontend and self attention for emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6917–6921.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2021. Consk-gcn: conversational semantic-and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Wenping Guo, Xiaoming Zhao, Shiqing Zhang, and Xianzhang Pan. 2023. Learning inter-class optical flow difference using generative adversarial networks for facial expression recognition. *Multimedia Tools and Applications*, 82(7):10099–10116.

Jiaxuan He, Sijie Mai, and Haifeng Hu. 2021. A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis. *IEEE Signal Processing Letters*, 28:992–996.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640.

Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. 2022. AMOA: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7136–7146, Gyeongju, Republic of Korea.

Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang. 2019. Conversational Emotion Analysis via Attention Mechanisms. In *Proc. Interspeech 2019*, pages 1936–1940.

Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Context-dependent domain adversarial neural network for multimodal emotion recognition. In *Interspeech*, pages 394–398.

Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. 2021. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8148–8156.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7124–7135, Gyeongju, Republic of Korea.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. 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.

Bubai Maji, Monorama Swain, Rajlakshmi Guha, and Aurobinda Routray. 2023. Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038. IEEE.

A Sabaritha, Palati Sinduja, and R Priyadharshini. 2023. Knowledge awareness, and practice on emotional intelligence and academic performance among lecturers. In *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, pages 1–7.

Saurav Sahay, Shachi H Kumar, Rui Xia, Jonathan Huang, and Lama Nachman. 2018. Multimodal relational tensor network for sentiment and emotion classification. *ACL 2018*, pages 20–27.

Petr Slovak, Alissa Antle, Nikki Theofanopoulou, Claudia Daudén Roquet, James Gross, and Katherine Isbister. 2023. Designing for emotion regulation interventions: an agenda for hci theory and research. *ACM Transactions on Computer-Human Interaction*, 30(1):1–51.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy.

Shu-Lin Wang, Alex Kuo, and Jing-Ya Lin. 2022a. Mobile emotion healthcare system applying sentiment analysis. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2814–2820.

Zhongyang Wang, Guoxia Xu, Xiaokang Zhou, Jung Yoon Kim, Hu Zhu, and Lizhen Deng. 2022b. Deep tensor evidence fusion network for sentiment classification. *IEEE Transactions on Computational Social Systems*, pages 1–9.

Wen Wu, Chao Zhang, and Philip C Woodland. 2021. Emotion recognition by fusing time synchronous and time asynchronous representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6269–6273. IEEE.

Zhen Wu, Yizhe Lu, and Xinyu Dai. 2023. An empirical study and improvement for speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. 2023. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. In *5th International Conference on Learning Representations, ICLR 2017*.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2023a. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9154–9167.

Shiqing Zhang, Ruixin Liu, Yijiao Yang, Xiaoming Zhao, and Jun Yu. 2022a. Unsupervised domain adaptation integrating transformer and mutual information for cross-corpus speech emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 120–129.

Shiqing Zhang, Yijiao Yang, Chen Chen, Ruixin Liu, Xin Tao, Wenping Guo, Yicheng Xu, and Xiaoming Zhao. 2023b. Multimodal emotion recognition based on audio and text by using hybrid attention networks. *Biomedical Signal Processing and Control*, 85:105052.

Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. 2023c. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, page 121692.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. 2017. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590.

Shiqing Zhang, Xiaoming Zhao, and Qi Tian. 2019. Spontaneous speech emotion recognition using multiscale deep convolutional lstm. *IEEE Transactions on Affective Computing*, 13(2):680–688.

Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022b. Tailor versatile multimodal learning for multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9100–9108.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. 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 2608–2618.

Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. 2022a. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4703–4707. IEEE.

Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2022b. MuCDN: Mutual conversational detachment network for emotion recognition in multi-party conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7020–7030, Gyeongju, Republic of Korea.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*, pages 2980–2991.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.