Disentangled Representation Learning for Multimodal Emotion Recognition

Dingkang Yang
Academy for Engineering and Technology, Fudan University

Shuai Huang
Academy for Engineering and Technology, Fudan University

Haopeng Kuang
Academy for Engineering and Technology, Fudan University

Yangtao Du
Academy for Engineering and Technology, Fudan University
Engineering Research Center of AI and Robotics, Ministry of Education
Artifical Intelligence and Unmanned Systems Engineering Research Center of Jilin Province

Lihua Zhang*
Academy for Engineering and Technology, Fudan University
Jilin Provincial Key Laboratory of Intelligence Science and Engineering
Ji Hua Laboratory
lihuazhang@fudan.edu.cn

ABSTRACT

Multimodal emotion recognition aims to identify human emotions from text, audio, and visual modalities. Previous methods either explore correlations between different modalities or design sophisticated fusion strategies. However, the serious problem is that the distribution gap and information redundancy often exist across heterogeneous modalities, resulting in learned multimodal representations that may be unrefined. Motivated by these observations, we propose a Feature-Disentangled Multimodal Emotion Recognition (FDMER) method, which learns the common and private feature representations for each modality. Specifically, we design the common and private encoders to project each modality into modality-invariant and modality-specific subspaces, respectively. The modality-invariant subspace aims to explore the commonality among different modalities and reduce the distribution gap sufficiently. The modality-specific subspaces attempt to enhance the diversity and capture the unique characteristics of each modality. After that, a modality discriminator is introduced to guide the parameter learning of the common and private encoders in an adversarial manner. We achieve the modality consistency and disparity constraints by designing tailored losses for the above subspaces. Furthermore, we present a cross-modal attention fusion module to learn adaptive weights for obtaining effective multimodal representations. The final representation is used for different downstream tasks. Experimental results show that the FDMER outperforms the state-of-the-art methods on two multimodal emotion recognition benchmarks. Moreover, we further verify the effectiveness of our model via experiments on the multimodal humor detection task.

* indicates corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547754

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Neural networks*; • **Information systems** → *Multimedia streaming*.

KEYWORDS

disentangled representation learning, emotion recognition, adversarial learning, multimodal fusion

ACM Reference Format:
Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547754

1 INTRODUCTION

Emotion plays a role in human communication. Recently, Multimodal Emotion Recognition (MER) has become an active research area with essential applications in various fields, such as human-computer interaction [5], intelligent healthcare [13], and robotics [32]. Human expressions of emotions are usually a mixture of natural language, facial gestures, and acoustic behaviors. Different modalities can provide rich information to understand human emotions and intents. Benefiting from the excellent performance of deep learning technologies in processing diverse signals [8, 21, 26–28, 46, 54], various models have been developed to extract emotion-related information from multimodal sequences, such as convolution neural networks [20], recurrent neural networks [22, 41], transformers [1], and their variants [30, 42]. The mainstream research focuses on two aspects: 1) learning correlations between elements from different modalities to obtain refined modality semantics [25, 30, 43], and 2) designing sophisticated fusion strategies to produce effective representations [44, 48, 51, 56]. Nevertheless, the inherent heterogeneity across modalities often introduces information redundancy and distribution gap, increasing the difficulty of multimodal representation learning and feature fusion. In this case, most previous methods treat the representation of each modality in a holistic learning manner, causing the learned multimodal representations that may be unrefined and redundant.

Recent works have provided some first insights into learning distinct multimodal representations. Liang *et al.* [25] attempt to learn reliable cross-modal interactions over modality-invariant subspace where the distribution is bridged. However, their method neglects the fact that different modalities reveal the unique characteristic of emotions from different perspectives. Hazarika *et al.* [11] use similarity loss and difference loss to explore consistency and complementarity between multiple modalities. It is a sub-optimal solution because utilizing only the simple constraints fail to guarantee that the learned representations are perfectly projected to the desired subspaces. Wu *et al.* [50] propose a text-centric framework for learning shared and private semantics in acoustic and visual modalities. Unfortunately, their framework is trained in stages and is not generalizable, *i.e.*, it depends on the specific modality.

Motivated by the above observations, we propose a Feature-Disentangled Multimodal Emotion Recognition (FDMER) method to deal with modality heterogeneity by learning two distinct representations for each modality. The first is the common representation, which aims to project all modalities into a modality-invariant shared subspace with aligned distributions. Our FDMER can capture the commonality among modalities regarding suggested emotions and reduce the modality gap in this subspace. The second is the private representation, which aims to provide a modality-specific subspace for each modality. In these subspaces, our FDMER can learn the unique characteristics of different modalities and eliminate redundant information. We design the common and private encoders to achieve the feature disentanglement described above. In addition, the proposed consistency and disparity constraints are utilized to guarantee consistency in the common representations and diversity in the private representations, respectively. To further guarantee that the different representations are projected perfectly into the corresponding subspaces, a modality discriminator is introduced to guide the parameter learning of the common and private encoders. For alleviating the modality heterogeneity challenge, we employ a spherical modality discriminative loss to enhance the intra-class compactness and inter-class discrepancy for the hidden representations and parameters of the modality discriminator in a hyper-sphere. After that, we propose a cross-modal attention fusion module based on adaptive attention weights to effectively fuse the distinct representations. The refined multimodal representation eventually serves downstream tasks.

The main contributions can be summarized as follows:

- We propose FDMER, a novel multimodal emotion recognition method based on feature disentanglement. The FDMER tackles heterogeneity gap by learning the common and private representations across multiple modalities in the modality-invariant and -specific subspaces, respectively.
- We present a Cross-Modal Attention Fusion (CMAF) module to fuse multimodal representations effectively. The CMAF module adaptively assigns weights to different representations to highlight the stronger ones and suppress the weaker ones based on their importance.
- Our FDMER outperforms previous state-of-the-art methods on three standard multimodal benchmarks. Comprehensive experiments demonstrate that our method can clearly capture distinct multimodal representations and depict the commonality and diversity among multiple modalities.

2 RELATED WORK

2.1 Multimodal Emotion Recognition

Emotion recognition is a research hotspot that has attracted widespread attention in the multimedia community. Unlike conventional works [2, 39] that use only isolated modality (*e.g.*, text or audio), multimodal emotion recognition aims to combine information from multiple sources to improve the understanding and perception of human emotions. Previous multimodal methods have contributed to leveraging the complementary information across modalities [29, 31, 42, 44, 48, 51, 53]. For instance, Zadeh *et al.* [51] propose a tensor fusion to explicitly capture uni-modal, bi-modal, and tri-modal interactions. Liu *et al.* [29] utilize the low-rank tensors to accelerate the fusion process. Mai *et al.* [31] propose a graph fusion network to model the interactions between different modalities. The aforementioned methods are aggregation-based fusion paradigms, and the modality gap heavily hurts multimodal fusion. To bridge the modality gap, some recent works [25, 30, 43] attempt to achieve potential adaption from one modality to another based on the cross-modal attention. However, they tend to perform fusion into a joint embedding space, which neglects the diversity of each modality.

2.2 Disentangled Representation Learning

Early work on disentangled representation learning is mostly based on auto-encoders [4] and generative adversarial networks [35]. Chen *et al.* [6] introduce joint bayesian formulation to decompose a face representation into three parts, including intrinsic difference, transformation difference, and noise. Besides, the FactorVAE [24] is proposed to disentangle by encouraging the representation to be factorial and independent across the dimensions. Recently, disentangled representation learning has been increasingly applied in cross-modal tasks. For example, Wu *et al.* [49] propose a disentangled variational representation method for heterogeneous face matching by aligning the correlation between different modality variations. Guo *et al.* [17] present a cross-modal retrieval method based on deep mutual information estimation that disentangles the exclusive modality information from the shared representations. In comparison, we extend the one-to-one disentangled paradigm to the multimodal pattern in an adversarial manner.

3 APPROACH

3.1 Model Overview

In this section, we describe the details of the proposed Feature-Disentangled Multimodal Emotion Recognition (FDMER) method. The overall structure of the FDMER is illustrated in Figure 1. We consider three primary modalities that express emotion: text, audio, and visual modalities. Their corresponding sequences are represented as $\boldsymbol{X}_t \in \mathbb{R}^{L_t \times d_t}$, $\boldsymbol{X}_a \in \mathbb{R}^{L_a \times d_a}$, and $\boldsymbol{X}_v \in \mathbb{R}^{L_v \times d_v}$, respectively, where $L_{(\cdot)}$ is the sequence length and $d_{(\cdot)}$ is the embedding dimension. Our goal is to extract distinct multimodal representations from heterogeneous modalities for performing effective multimodal fusion.

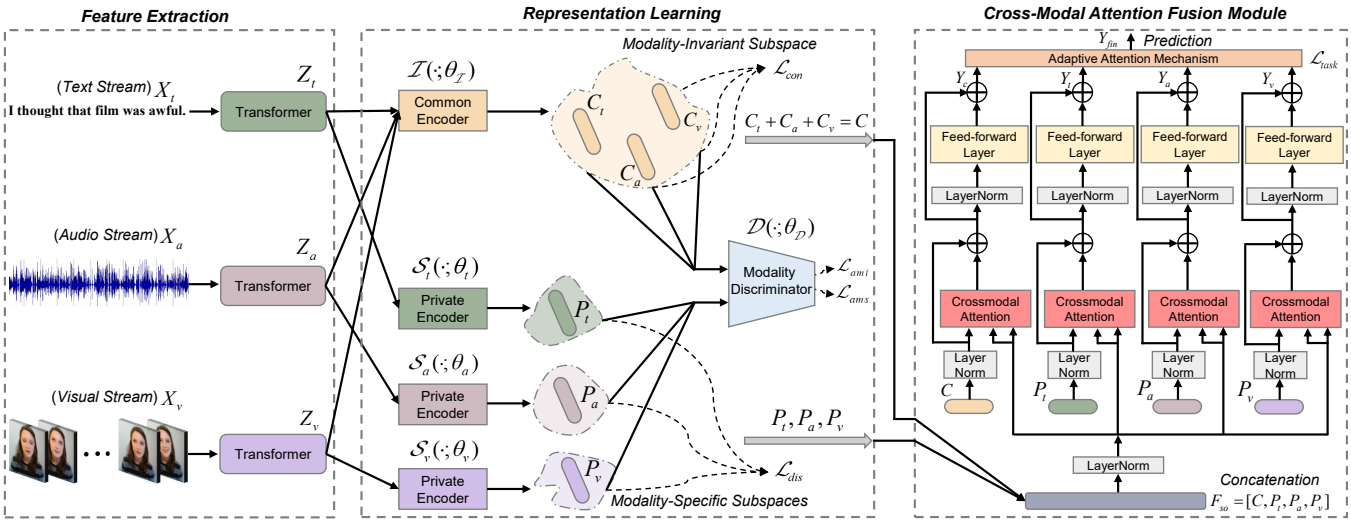

Figure 1: The overall structure of the proposed FDMER. For the multimodal sequences X_t, X_a and X_v, we first progressively enrich the low-level features through the respective transformers to obtain reinforced representations Z_t, Z_a, and Z_v. Immediately, a common encoder and three private encoders are employed to extract the common representations $C_{\{t,a,v\}}$ and the private representations $P_{\{t,a,v\}}$ of different modalities, respectively. On the one hand, we propose the consistency and disparity losses to constrain the projections of modality-invariant and -specific subspaces, respectively. On the other hand, a modality discriminator is presented to supervise and guide the learning of the distinct representations. Eventually, we introduce a cross-modal attention fusion module to fuse multiple representations and make predictions via the fully connected layers.

The core idea is to learn the common and private feature representations by projecting each modality into modality-invariant and modality-specific subspaces. To this end, we design the consistency and disparity constraints to enhance the commonality across the common representations and reduce the redundancy among the private representations, respectively (introduced in Section 3.3). Moreover, a modality discriminator is introduced to supervise the learning of the common and private representations explicitly, *i.e.*, reducing the distribution gap among modalities in the modality-invariant subspace while learning the unique characteristics of each modality in the modality-specific subspaces. After that, we propose a cross-modal attention fusion module to achieve information interaction and fusion between different representations (introduced in Section 3.4). Eventually, the fused multimodal representation is used to perform downstream tasks.

3.2 Feature Extraction

To explore long-range contextual information, we utilize n_t-layer, n_a-layer, n_v-layer transformer encoder [45] to enrich the text features, audio features, and visual features separately. The transformer encoder consists of a multi-head self-attention module and a position-wise feed-forward layer, where residual connections are adopted, followed by layer normalization. Please refer to [45] for more details. The extracted features are denoted as Z_m:

$$Z_m = \text{Transformer}(X_m; \theta_m^{trans}) \in \mathbb{R}^{L_m \times d_m}, \tag{1}$$

where $m \in \{t, a, v\}$ and θ_m^{trans} are the learnable parameters. Subsequently, we obtain the refined features of each modality in a fixed dimension as $Z_m \in \mathbb{R}^{d_k}$ through the fully connected layers.

3.3 Representation Learning

Common and Private Representations. Although the feature extractors based on temporal models can capture the long-range contextual dependencies of multimodal sequences, they cannot address feature redundancy due to the modality gap [56]. In addition, the divide-and-conquer processing pattern suffers from the heterogeneous nature among different modalities. Motivated by the above observations, we propose the common and private encoders to embed the pre-extracted features from each modality into modality-invariant and modality-specific subspaces, respectively. The nature of this disentangled representation learning is to leverage the common and private representations to capture the consistency and specificity of heterogeneous modalities, respectively. More formally, both the common encoder $\mathcal{I}(\cdot;\theta_{\mathcal{I}})$ and the private encoders $\mathcal{S}_m(\cdot;\theta_m)$ are implemented as two-layer perceptrons with the activation function of GeLU [19], where $m \in \{t, a, v\}$. The common and private representations can be formulated as:

$$\boldsymbol{C}_t = \mathcal{I}(Z_t;\theta_{\mathcal{I}}), \boldsymbol{C}_a = \mathcal{I}(Z_a;\theta_{\mathcal{I}}), \boldsymbol{C}_v = \mathcal{I}(Z_v;\theta_{\mathcal{I}}), \tag{2}$$

$$\boldsymbol{P}_t = \mathcal{S}_t(Z_t;\theta_t), \boldsymbol{P}_a = \mathcal{S}_a(Z_a;\theta_a), \boldsymbol{P}_v = \mathcal{S}_v(Z_v;\theta_v), \tag{3}$$

where $\boldsymbol{C}_{\{t,a,v\}}$, $\boldsymbol{P}_{\{t,a,v\}} \in \mathbb{R}^{d_k}$. The common encoder $\mathcal{I}(\cdot;\theta_{\mathcal{I}})$ shares the parameters $\theta_{\mathcal{I}}$ across all modalities, and the private encoders $\mathcal{S}_m(\cdot;\theta_m)$ learn the parameters θ_m for each modality.

Consistency Constraint. Inspired by [47], we apply the tailored constraints to disentangled representation learning for multiple modalities. For the common representations, we introduce a consistency constraint to strengthen the commonality among different modalities. Concretely, we use L_2-normalization to normalize the

representations C_m of the common encoder output as C_m^{nor}. The normalized matrices can be utilized to depict the similarity as S_m:

$$S_m = C_m \cdot C_m^{nor^T}. \tag{4}$$

The consistency means that the two similarity matrices should be similar, which results in the following constraint:

$$\mathcal{L}_{con} = \frac{1}{3} \sum_{(m_1,m_2)} \| S_{m_1} - S_{m_2} \|_F^2, \tag{5}$$

where $(m_1, m_2) \in \{(t,a),(t,v),(a,v)\}$, and $\| \cdot \|_F^2$ is the squared Frobenius norm.

Disparity Constraint. To ensure that the private representations model the different aspects of multimodal data and reduce information redundancy across different modalities, we employ the Hilbert-Schmidt Independence Criterion (HSIC) [40] to effectively measure the independence between the private representations. If the independence between the two representations is high, their discrepancy is significant. Benefiting from its fast convergence, the HSIC has been applied to several machine learning tasks [16, 34]. More formally, the HSIC constraint between any two private representations is defined as:

$$\text{HSIC}(P_{m_1}, P_{m_2}) = (n-1)^{-2} Tr(UK_{m_1}UK_{m_2}), \tag{6}$$

where K_{m_1} and K_{m_1} are the Gram matrices with $k_{m_1,ij} = k_{m_1}(p_{m_1}^i, p_{m_1}^j)$ and $k_{m_2,ij} = k_{m_2}(p_{m_2}^i, p_{m_2}^j)$. $U = I - (1/n)ee^T$, where I is an identity matrix and e is an all-one column vector. In practice, we use the inner product kernel function [47] for K_{m_1} and K_{m_2}. In this case, we perform the HSIC constraint between the private representations of each pair of modalities, and the overall disparity loss is expressed as:

$$\mathcal{L}_{dis} = \frac{1}{3} \sum_{(m_1,m_2)} \text{HSIC}(P_{m_1}, P_{m_2}), \tag{7}$$

where $(m_1, m_2) \in \{(t,a),(t,v),(a,v)\}$.

Adversarial Learning. Although the consistency and disparity loss can encourage the common and private encoders to produce different representations, they do not guarantee that the common representations belong to a latent subspace shared across modalities and that the private representations explicitly reflect the unique characteristics of each modality. As an example of learning modality-invariant subspace, previous state-of-the-art work [11] focuses on using the Central Moment Discrepancy (CMD) metric to align common cross-modal features on a shared subspace. However, such the simple constraint is supervised only by a task-specific pattern, resulting in the produced representations that are potentially impure and the projected subspace that is potentially mixed.

Inspired by generative adversarial network [15], we design a modality discriminator to identify modality labels and guide parameter learning of common and private encoders in an adversarial learning manner. To guarantee the purity of the common and private representations, the modality discriminator $\mathcal{D}(\cdot;\theta_{\mathcal{D}})$ maps the input into a probability distribution and estimates the modality from which the representation comes. The formula is defined as:

$$\mathcal{D}(h;\theta_{\mathcal{D}}) = softmax(W_{\mathcal{D}}^T \cdot \mathcal{F}(h)), \tag{8}$$

where $W_{\mathcal{D}} \in \mathbb{R}^{d\times 3}$ and $h \in \mathbb{R}^{d_k}$ is the input representation of the modality discriminator, which can be either the output C_m of the common encoder or the output P_m of the private encoders. $\mathcal{F}(h) = W_{\mathcal{F}} \cdot h + b_{\mathcal{F}}$, where $W_{\mathcal{F}} \in \mathbb{R}^{d\times d_k}$ and $b_{\mathcal{F}} \in \mathbb{R}^{d\times 1}$.

As a multi-class classifier, it is straightforward to learn the modality discriminator with cross-entropy loss. However, the modality heterogeneity leads to the conventional cross-entropy that fails to explicitly optimize intra-class similarity and inter-class diversity, limiting the discriminative power of the modality discriminator. To address this issue, we firstly normalize the hidden representations $\hat{h} = \mathcal{F}(h) \in \mathbb{R}^d$ and each column of $W_{\mathcal{D}}$ with L_2-normalization. The normalization step on the representations and weights makes the predictions only depend on the angle between the representation and the weight. The learned representations are distributed on a hyper-sphere. In this case, we introduce an additive angular margin loss [10] to enhance the intra-class compactness and inter-class discrepancy for the modality discriminator:

$$\mathcal{L}_{am} = -log \frac{e^{\alpha cos(\theta_{y_m}+\tau)}}{e^{\alpha cos(\theta_{y_m}+\tau)} + \sum_{m=1,m\neq y_m}^{M} e^{\alpha cos(\theta_m)}}, \tag{9}$$

where $\theta_{y_m} = arccos(W_{y_m}^T \cdot \hat{h})$ and $\theta_m = arccos(W_m^T \cdot \hat{h})$. y_m denotes the ground-truth modality label. $W_{y_m} \in \mathbb{R}^d$ denotes the y_m-th column of the weight matrix $W_{\mathcal{D}}$. Similarly, $W_m \in \mathbb{R}^d$ denotes the m-th column of the weight matrix $W_{\mathcal{D}}$. α is a scale factor and τ is a margin factor. Based on the above improvement, the common representations C_m are encoded in a modality-invariant subspace, which tends to be in the same distribution. The common adversarial loss is expressed as:

$$\mathcal{L}_{ami} = \frac{1}{n} \sum_{i=1}^{n} \sum_{m\in\{t,a,v\}} \mathcal{L}_{am}(C_m, y_m), \tag{10}$$

where $\mathcal{L}_{ami}$ is trained with gradient reversal layer [14] that keeps the input fixed during forward propagation and multiply the gradient by -1 during the backpropagation. Furthermore, the private representations P_m are embedded in the modality-specific subspaces, which tend to be in different distributions. Therefore, the modality discriminator is encouraged to distinguish the source of the modality. The private adversarial loss is expressed as:

$$\mathcal{L}_{ams} = \frac{1}{n} \sum_{i=1}^{n} \sum_{m\in\{t,a,v\}} \mathcal{L}_{am}(P_m, y_m). \tag{11}$$

3.4 Cross-Modal Attention Fusion Module

Different aspects of the representations have different importance for the final prediction. Simply concatenating them ignores modality interactions, which might introduce redundant information and lead to a sub-optimal problem [55]. As shown in the right half of Figure 1, to fully use consistent and complementary information in the refined common and private representations, we propose a novel Cross-Modal Attention Fusion (CMAF) module to achieve information interaction and knowledge exchange between different representations comprehensively. The CMAF module consists of a cross-modal interaction phase and an adaptive attention phase. We describe the details of each below.

Cross-Modal Interaction Phase (Phase 1). Assuming that there is already a strong commonality between the refined common representations, we merge C_m from different modalities into the overall representation as $C = \sum_{m\in\{t,a,v\}} C_m \in \mathbb{R}^{d_k}$. Then, we concatenate all the representations as $F_{so} = [C, P_t, P_a, P_v] \in \mathbb{R}^{4d_k}$. The core strategy of this phase is to explore the potential adaptation process from the integrated source representation F_{so} to the target representations $F_{ta} \in \{C, P_t, P_a, P_v\}$ via cross-modal attention. Each target representation is effectively reinforced and improved in sufficient cross-modal interaction. Inspired by the self-attention mechanism [45], we first embed F_{ta} into a space denoted as $\mathcal{G}_{ta} = LN(F_{ta})W_{\mathcal{G}_{ta}}$, while embedding F_{so} into two spaces denoted as $\mathcal{Q}_{so} = LN(F_{so})W_{\mathcal{Q}_{so}}$ and $\mathcal{U}_{so} = LN(F_{so})W_{\mathcal{U}_{so}}$, respectively. $W_{\mathcal{G}_{ta}} \in \mathbb{R}^{d_k\times d_k}$, $\{W_{\mathcal{Q}_{so}}, W_{\mathcal{U}_{so}}\} \in \mathbb{R}^{4d_k\times 4d_k}$ are embedding weights, and LN means layer normalization. Then, the cross-modal interaction is defined as follows:

$$F_{so\to ta} = softmax(\mathcal{G}_{ta}\mathcal{Q}_{so}^T)\mathcal{U}_{so} \in \mathbb{R}^{d_k}. \tag{12}$$

Immediately, the forward computation is expressed as:

$$Y_{ta} = LN(F_{ta}) + F_{so\to ta}, \tag{13}$$

$$Y_{ta} = f_\delta(LN(Y_{ta})) + Y_{ta}, \tag{14}$$

where $f_\delta(\cdot)$ is the feed-forward layers parametrized by δ, and $Y_{ta} \in \{Y_c, Y_t, Y_a, Y_v\} \in \mathbb{R}^{d_k}$.

Adaptive Attention Phase (Phase 2). This phase introduces an adaptive attention mechanism to assign dynamic weights for each reinforced representation based on its importance. More formally, we use one shared attention vector $q \in \mathbb{R}^{d_k\times 1}$ to get the attention values μ_{ta} as follows:

$$\mu_{ta} = q^T \cdot tanh(W_{ta} \cdot Y_{ta} + b_{ta}), \tag{15}$$

where $W_{ta} \in \mathbb{R}^{d_k\times d_k}$ and $b_{ta} \in \mathbb{R}^{d_k\times 1}$. Then the attention values μ_{ta} are normalized with softmax function to obtain the final weights:

$$\psi_{ta} = \frac{exp(\mu_{ta})}{\sum_{ta\in\{c,t,a,v\}} exp(\mu_{ta})}. \tag{16}$$

A large ψ_{ta} implies the corresponding representation is important. We obtain the final representation $Y_{fin} \in \mathbb{R}^{d_k}$ by weighted summation:

$$Y_{fin} = \sum_{ta\in\{c,t,a,v\}} \psi_{ta} \odot Y_{ta}. \tag{17}$$

Eventually, the Y_{fin} passes through the fully connected layers to perform downstream tasks.

3.5 Objective Optimization

For the classification task, we employ the standard cross-entropy loss as $\mathcal{L}_{task} = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot log\hat{y}_i$. For the regression task, we use the mean squared error loss as $\mathcal{L}_{task} = \frac{1}{n}\sum_{i=1}^{n} \parallel y_i - \hat{y}_i \parallel_2^2$, where y_i is the ground truth and n is the number of samples in a batch. Combining the task loss $\mathcal{L}_{task}$, the constraint losses $\mathcal{L}_{con}$, $\mathcal{L}_{dis}$, the common adversarial loss $\mathcal{L}_{ami}$, and the private adversarial loss $\mathcal{L}_{ams}$, the final objective function is computed as:

$$\mathcal{L}_{all} = \mathcal{L}_{task} + \beta(\mathcal{L}_{ami} + \mathcal{L}_{ams}) + \gamma(\mathcal{L}_{con} + \mathcal{L}_{dis}), \tag{18}$$

where β and γ are the trade-off parameters.

4 EXPERIMENTS

4.1 Benchmarks and Evaluation Metrics

We conduct comprehensive experiments on two standard multimodal emotion recognition benchmarks and a multimodal humor detection benchmark. These benchmarks provide word-aligned multimodal signals for each sample.

CMU-MOSI. CMU-MOSI [53] is a human multimodal dataset containing 2,199 short monologue video clips. The standard partitioning of the dataset is 1,284 samples in the training set, 229 in the validation set, and 686 in the testing set. The acoustic and visual features are extracted at a sampling rate of 12.5 and 15 Hz, respectively. Each multimodal sample has a sentiment score that ranges from -3 to 3 . As in the previous works [38, 43, 44], the model performance is evaluated by the 7-class accuracy (Acc_7), the binary accuracy (Acc_2), mean absolute error (MAE), the correlation of the model's prediction with human ($Corr$), and the $F1$ score.

CMU-MOSEI. CMU-MOSEI [52] is a dataset that contains 22,856 samples of movie review video clips from YouTube. Its predetermined data split includes 16,326 training samples, 1,871 validation samples, and 4,659 testing samples. The acoustic and visual features are extracted at a sampling rate of 20 and 15 Hz, respectively. The samples are also labeled with the sentiment scores ranging from -3 to 3. The same metrics are employed as in the above setting.

UR_FUNNY. UR_FUNNY [18] is a dataset that contains 16,514 samples of multimodal utterances from TED talks. The standard partitioning of the dataset is 10,598 samples in the training set, 2,626 in the validation set, and 3,290 in the testing set. Each target utterance is labeled with a binary label for humor/non-humor instance. It also provides related context for each target utterance. For the binary classification, we report the binary accuracy (Acc_2).

4.2 Implementation Details

Feature Embedding. Most previous methods [25, 30, 44] have obtained the textual features through Glove embedding [37]. However, benefiting from the excellent performance of the BERT model [12], several recent works [11, 42] have been BERT-based. We provide the results using Glove and BERT for a comprehensive and fair comparison. Specifically, we convert the transcripts of video into pre-trained Glove word embedding with a 300-dimensional vector while using the BERT-base-uncased pre-trained model to obtain a 768-dimensional hidden state. For the MOSI & MOSEI, we use Facet [23] to indicate 35 facial action units, recording facial muscle movement to represent emotions. For the UR_FUNNY, we use the OpenFace [3] to extract 75-dimensional features related to the facial expressions. Moreover, the COVAREP toolkit [9] is used to extract low-level acoustic features, where the dimension on the MOSI & MOSEI is 74 and on the UR_FUNNY is 81. The features include 12 Mel-frequency cepstral coefficients (MFCCs), voiced/unvoiced segmenting features, glottal source parameters, etc.

Experimental Setup. All models are built on the Pytorch toolbox [36] with four Nvidia Tesla V100 GPUs. The number of transformer encoder layers for text, audio and visual are $\{n_t = 5, n_a = 4, n_v = 4\}$. For the MOSI, MOSEI and UR_FUNNY benchmarks, the

Table 1: Comparison on the CMU-MOSI benchmark.

Model	$Acc_7 \uparrow$	$Acc_2 \uparrow$	$F1 \uparrow$	$MAE \downarrow$	$Corr \uparrow$
Glove-based					
TFN [51]	32.1	73.9	73.4	0.970	0.633
LMF [29]	32.8	76.4	75.7	0.912	0.668
RAVEN [48]	33.2	78.0	76.6	0.915	0.691
MCTN [38]	35.6	79.3	79.1	0.909	0.676
MFM [44]	36.2	78.1	78.1	0.951	0.662
MulT [43]	40.0	83.0	82.8	0.871	0.698
TCSP [50]	-	80.9	81.0	0.908	0.710
PMR [30]	40.6	83.6	83.4	-	-
FDMER (ours)	**42.1**	**84.2**	**83.9**	**0.845**	**0.732**
BERT-based					
TFN [51]	34.9	80.8	80.7	0.901	0.698
LMF [29]	33.2	82.5	82.4	0.917	0.695
MFM [44]	35.4	81.7	81.6	0.877	0.706
ICCN [42]	39.0	83.0	83.0	0.860	0.710
MISA [11]	42.3	83.4	83.6	0.783	0.761
FDMER (ours)	**44.1**	**84.6**	**84.7**	**0.724**	**0.788**

Table 2: Comparison on the CMU-MOSEI benchmark.

Model	$Acc_7 \uparrow$	$Acc_2 \uparrow$	$F1 \uparrow$	$MAE \downarrow$	$Corr \uparrow$
Glove-based					
Graph-MFN [53]	45.0	76.9	77.0	0.710	0.540
RAVEN [48]	50.0	79.1	79.5	0.614	0.662
MCTN [38]	49.6	79.8	80.6	0.609	0.670
MulT [43]	51.8	82.5	82.3	0.580	0.703
TCSP [50]	-	82.8	82.7	0.576	0.715
PMR [30]	52.5	83.3	82.6	-	-
FDMER (ours)	**53.8**	**83.9**	**83.8**	**0.568**	**0.736**
BERT-based					
TFN [51]	50.2	82.5	82.1	0.593	0.700
LMF [29]	48.0	82.0	82.1	0.623	0.677
MFM [44]	51.3	84.4	84.3	0.568	0.717
ICCN [42]	51.6	84.2	84.2	0.565	0.713
MISA [11]	52.2	85.5	85.3	0.555	0.756
FDMER (ours)	**54.1**	**86.1**	**85.8**	**0.536**	**0.773**

batch sizes and epochs are $\{64, 16, 32\}$ and $\{120, 100, 60\}$, respectively. The Adam optimizer is adopted for network optimization with an initial learning rate of $\{1e^{-3}, 1e^{-3}, 2e^{-3}\}$. The trade-off parameters β and γ are set to $\{2e^{-2}, 3e^{-2}, 2e^{-2}\}$ and $\{2e^{-2}, 4e^{-2}, 3e^{-2}\}$, respectively. The hidden dimension d_k is set to 256 and the output dimension d of $\hat{\boldsymbol{h}}$ is 128. The margin factor τ and scale factor α of the angular margin loss in Eq. (9) are set to 0.5 and 72. In practice, all the hyper-parameters are determined via the validation set.

4.3 Comparison with State-of-the-Art Methods

Model Zoo. We compare our model with the state-of-the-art (SOTA) works, including the Glove-based mehtods: TFN [51], LMF [29], Graph-MFN [53], RAVEN [48], MCTN [38], MFM [44], MulT [43], TCSP [50], PMR [30], and the BERT-based methods: ICCN [42], MISA [11]. Note that the results of some works [29, 44, 51] based on BERT come from [42].

Table 3: Comparison on the UR_FUNNY benchmark.

Model	Context	Target	$Acc_2 \uparrow$
Glove-based			
C-MFN [18]	✓		58.45
C-MFN [18]		✓	64.47
TFN [51]		✓	64.71
LMF [29]		✓	65.16
C-MFN [18]	✓	✓	65.23
MISA [11]		✓	68.60
FDMER (ours)		✓	**70.55**
BERT-based			
LMF [29]		✓	67.53
TFN [51]		✓	68.57
MISA [11]		✓	70.61
FDMER (ours)		✓	**71.87**

Multimodal Emotion Recognition. Comparative results on the CMU-MOSI and CMU-MOSEI benchmarks are reported in Tables 1 and 2, respectively. We have the following observations. Our method significantly outperforms the previous SOTA methods on all metrics for both benchmarks. Compared to the recent PMR [30], which is based on complex cross-modal interactions, the proposed FDMER learns effective multimodal representations with a simple structure in the perspective of feature disentanglement. Compared to the MISA [11], which also learns different subspace representations for multiple modalities, our method demonstrates the rationality and effectiveness of using an adversarial manner to learn modality-invariant and modality-specific subspaces with superior results. Furthermore, we observe a significant performance improvement of the FDMER using BERT to extract low-level features compared to Glove-based embedding. This observation indicates the advantages of using BERT to extract textual features.

Multimodal Humor Detection. To further verify the applicability of the FDMER, comparative experiments are conducted on the UR_FUNNY benchmark. Since humor detection is sensitive to heterogeneous representations of different modalities [18], the best results presented by our method in Table 3 show the superiority of the proposed multimodal framework in learning distinct representations. It is worth noting that our Glove-based variant achieves comparable performance to the BERT-based MISA [11].

4.4 Ablation Studies

We perform thorough ablation studies on all benchmarks to understand the necessity of the different components in the FDMER. Table 4 shows the results with the following observations.

Importance of Modality. Firstly, we remove a modality separately to explore the performance of the bi-modal FDMER. There is a significant drop in the model's performance when the text modality is removed, indicating that the text modality dominates the multimodal emotion recognition (MER) and multimodal humor detection (MHD) tasks. A reasonable explanation is that the acoustic and visual features contain more noisy and redundant information than the textual features, limiting the model's performance [7]. Furthermore, the consistently worse performance of the

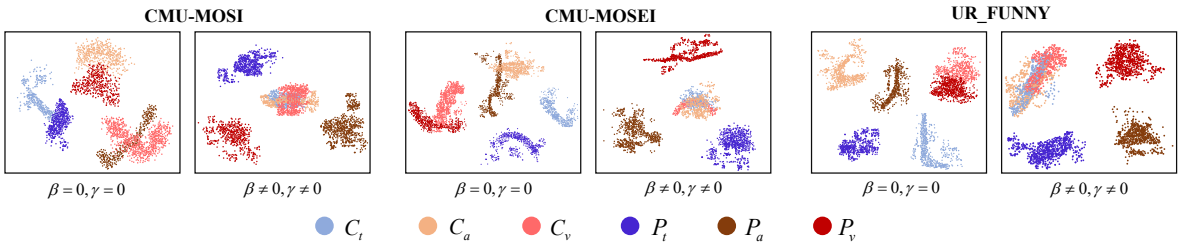

Figure 2: Visualization of the common and private representations in the testing set on three benchmarks. $\beta = 0, \gamma = 0$ denotes without adversarial and constraint losses, and vice versa. The blue, brown, and red colors represent text, audio, and visual modalities, respectively. The light colors correspond to common parts, while dark colors correspond to private parts.

Table 4: Results of ablation studies on three benchmarks.

Model	CMU-MOSI		CMU-MOSEI		UR_FUNNY
	MAE ↓	*Corr* ↑	*MAE* ↓	*Corr* ↑	Acc_2 ↑
FDMER	**0.845**	**0.732**	**0.568**	**0.736**	**70.55**
Importance of Modality					
w/o Text	1.275	0.345	0.896	0.378	57.58
w/o Audio	0.883	0.726	0.604	0.727	70.43
w/o Visual	0.925	0.714	0.639	0.713	69.81
Importance of Regularization					
w/o $\mathcal{L}_{ami} + \mathcal{L}_{ams}$	0.872	0.718	0.590	0.715	68.36
w/o $\mathcal{L}_{con}$	0.867	0.723	0.577	0.728	69.87
w/o $\mathcal{L}_{dis}$	0.854	0.728	0.571	0.731	70.28
Cross-Entropy Loss	0.852	0.725	0.582	0.724	68.95
Importance of Representations					
w/o Common	0.868	0.720	0.575	0.724	68.31
w/o Private	0.885	0.714	0.582	0.712	69.27
Non-Disentangled	0.862	0.725	0.571	0.727	69.68
Different Fusion Strategies					
w/o CMAF	0.871	0.712	0.606	0.710	67.33
w/o Phase 1	0.864	0.724	0.582	0.726	68.41
w/o Phase 2	0.855	0.728	0.570	0.729	69.75
Addition	0.869	0.715	0.607	0.718	67.85
Multiplication	0.854	0.727	0.579	0.726	68.96

bi-modal FDMER compared to the tri-modal FDMER suggests that each modality provides an indispensable contribution.

Importance of Regularization. We remove the proposed losses separately to verify the role played by the different regularizations. When there is no adversarial loss ($\mathcal{L}_{ami} + \mathcal{L}_{ams}$), the model learns distinct multimodal representations that rely on the constraint losses and do not involve the modality discriminator. The worst performance demonstrates the importance of the adversarial manner in disentangled representation learning. Meanwhile, we observe that both the consistency constraint $\mathcal{L}_{con}$ and the disparity constraint $\mathcal{L}_{dis}$ improve the model's performance. Moreover, we replace the adversarial loss with the Cross-Entropy loss (CE-loss) to explore its effectiveness. As shown in Table 4, the CE-loss causes a significant performance drop on each benchmark. The result is inevitable because the traditional CE-loss cannot overcome modality heterogeneity and capture the inter-class discrepancy in this case.

Importance of Representations. To prove the effectiveness of the different representations, we first remove the common or private representations separately to experiments. Specifically, we keep the representation learning process but use only partial representations in the fusion and prediction phases. The decreased results reveal that both representations learned are essential and meaningful. Interestingly, the private representations are more expressive in the MER task. Conversely, the common representations provide a significant contribution to the MHD task. In addition, we conduct a version that does not learn the distinct subspaces, *i.e.*, the extracted features from the transformers are used directly for fusion. We show that the model without feature disentanglement is slightly better than those containing only common or private representations, revealing the learning limitations in a partial subspace.

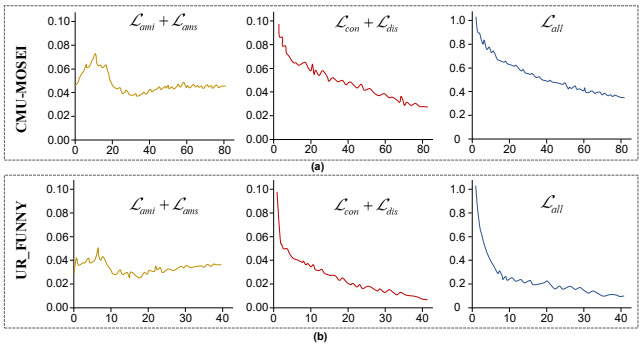

Figure 3: Visualization of the adversarial loss, constraint loss and overall loss during the training process. Similar trends are also observed on the CMU-MOSI benchmark.

Different Fusion Strategies. Finally, we explore the effect of different fusion strategies. The feature fusion is performed with the simple concatenation when the CMAF module is removed. In this case, the poor results prove that our fusion strategy is indispensable. When one phase of the CMAF module is removed separately, both decreased results suggest that it is beneficial to consider cross-modal interactions and dynamic weights in multimodal fusion. Furthermore, our strategy remains competitive when compared to additive fusion and advanced multiplicative fusion [33].

4.5 Visualization Results

Visualization of Adversarial Representations. Understanding the subspace distributions of the different representations is

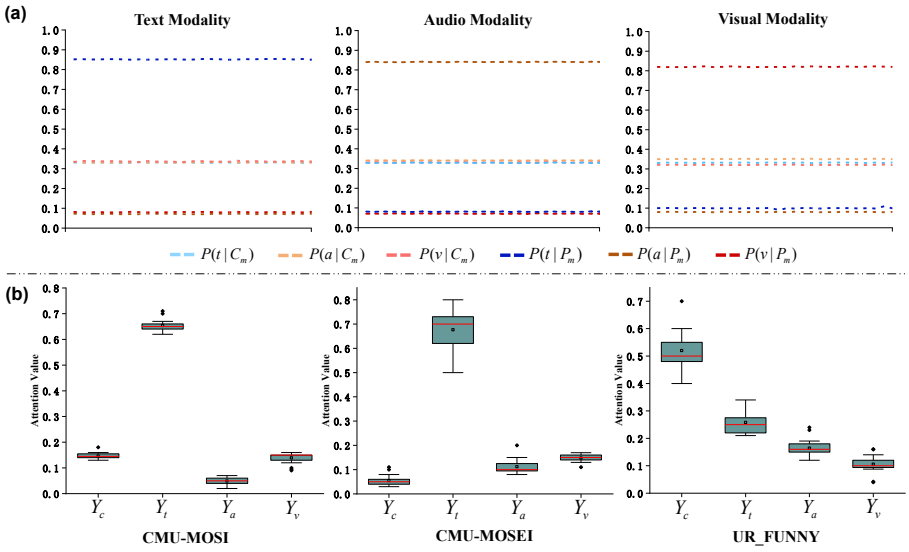

Figure 4: (a) Distributions of the common and private representations from different modalities on the CMU-MOSI benchmark. For modality $m \in \{t, a, v\}$, $[P(t|\boldsymbol{C}_m), P(a|\boldsymbol{C}_m), P(v|\boldsymbol{C}_m)] = \mathcal{D}(\boldsymbol{C}_m; \theta_{\mathcal{D}})$, $[P(t|\boldsymbol{P}_m), P(a|\boldsymbol{P}_m), P(v|\boldsymbol{P}_m)] = \mathcal{D}(\boldsymbol{P}_m; \theta_{\mathcal{D}})$. (b) The attention distributions from the cross-modal attention fusion module in the testing set on the three benchmarks.

essential. To this end, we visualize the common and private representations $\boldsymbol{C}_{\{t,a,v\}}$ and $\boldsymbol{P}_{\{t,a,v\}}$ learned without or with the adversarial training and constraint losses on all benchmarks in Figure 2. When $\beta = 0, \gamma = 0$, the distributions of $\boldsymbol{C}_{\{t,a,v\}}$ and $\boldsymbol{P}_{\{t,a,v\}}$ sometimes overlap, and the common representations are not learned. Contrarily, when $\beta \neq 0, \gamma \neq 0$, the distributions of $\boldsymbol{C}_{\{t,a,v\}}$ are blended together and gradually blurred, where adversarial training effectively aligns distributions of different modalities and minimizes the modality gap. Meanwhile, each modality-specific subspace is separable, where the disparity constraint punishes redundant latent representations. The above observations prove that our method captures the commonality and specificity of different modalities.

Regularization Trends. The losses $\{\mathcal{L}_{ami}, \mathcal{L}_{ams}, \mathcal{L}_{con}, \mathcal{L}_{dis}\}$ act as measures to evaluate the ability of the model to learn the distinct representations. In Figure 3, the constraint loss ($\mathcal{L}_{con} + \mathcal{L}_{dis}$) and overall loss $\mathcal{L}_{all}$ decrease almost monotonously and converge smoothly, while the adversarial loss ($\mathcal{L}_{ami} + \mathcal{L}_{ams}$) gradually stabilizes after the vibration. The above observations prove that the model is indeed learning the representations as designed.

Probability Distributions from Modality Discriminator. To further understand the effect of adversarial learning, we visualize the probabilities generated by the modality discriminator during the adversarial process in Figure 4.(a). For each modality m, the probabilities of the common representations $P(t|\boldsymbol{C}_m), P(a|\boldsymbol{C}_m), P(v|\boldsymbol{C}_m)$ are centered around 0.33, which is hard to differentiate the source of common modalities. Contrarily, taking text modality for example, $P(t|\boldsymbol{P}_m)$ is higher than $P(a|\boldsymbol{P}_m)$ and $P(v|\boldsymbol{P}_m)$ by a large margin, leading to increasingly the private representations.

Analysis of Attention Distributions. In Figure 4.(b), we conduct the box plots to analyze the attention distributions in the testing set on all benchmarks. For CMU-MOSI and CMU-MOSEI benchmarks, we note that the attention values of the private representations are usually larger than the values of the common representations, and the private representation of text modality has the largest weights. This implies that the information in the modality-specific subspace is more important than in the modality-invariant subspace in the MER task. Conversely, for the UR_FUNNY benchmark, we find that the attention values of the common representations are generally larger than those of the private representations. This means that the information in the modality-invariant subspace facilitates a better understanding of humor. The above observations demonstrate that our CMAF module can adaptively assign larger attention values to meaningful representations.

5 CONCLUSION

In this paper, we propose the Feature-Disentangled Multimodal Emotion Recognition (FDMER) method to tackle modality heterogeneity by projecting each modality into modality-invariant and modality-specific subspaces. Our FDMER elegantly refines the common and private representations of feature disentanglement via the tailored constraints and adversarial learning strategy. Furthermore, the novel cross-modal attention fusion module provides new insight into fusing multimodal representations. Experimental results demonstrate the superiority of our method. It is worth noting that the FDMER can be expanded to various multimodal application scenarios to facilitate the development of the communities.

ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China (2021ZD0113503), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103) and National Natural Science Foundation of China under Grant (82090052).

REFERENCES

[1] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178* (2021).

[2] Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*. Springer, 196–205.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems* 29 (2016).

[5] Scott Brave and Cliff Nass. 2007. Emotion in human-computer interaction. In *The human-computer interaction handbook*. CRC Press, 103–118.

[6] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. 2012. Bayesian face revisited: A joint formulation. In *European conference on computer vision*. Springer, 566–579.

[7] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 163–171.

[8] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. 2022. Towards Practical Certifiable Patch Defense with Vision Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15148–15158.

[9] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[11] Hazarika Devamanyu, Zimmermann Roger, and Poria Soujanya. 2020. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis.. In *Proceedings of the 28th ACM International Conference on Multimedia*, Vol. 34. 1122–1131.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Faiyaz Doctor, Charalampos Karyotis, Rahat Iqbal, and Anne James. 2016. An intelligent framework for emotion aware e-healthcare support systems. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–8.

[14] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

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

[16] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

[17] Weikuo Guo, Huaibo Huang, Xiangwei Kong, and Ran He. 2019. Learning disentangled representation for cross-modal retrieval with deep mutual information estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1712–1720.

[18] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618* (2019).

[19] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[20] Haiping Huang, Zhenchao Hu, Wenming Wang, and Min Wu. 2019. Multimodal emotion recognition based on ensemble convolutional neural network. *IEEE Access* 8 (2019), 3265–3271.

[21] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. 2022. CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 1 (Jun. 2022), 989–997. https://doi.org/10.1609/aaai.v36i1.19982

[22] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Zhengqi Wen, Minghao Yang, and Jiangyan Yi. 2017. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 11–18.

[23] iMotions. 2017. *Facial expression analysis*.

[24] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *International Conference on Machine Learning*. PMLR, 2649–2658.

[25] 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*. 8148–8156.

[26] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. 2022. Efficient Universal Shuffle Attack for Visual Object Tracking. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2739–2743. https://doi.org/10.1109/ICASSP43922.2022.9747773

[27] Yang Liu, Jing Liu, Mengyang Zhao, Shuang Li, and Liang Song. 2022. Collaborative Normality Learning Framework for Weakly Supervised Video Anomaly Detection. *IEEE Transactions on Circuits and Systems II: Express Briefs* 69, 5 (2022), 2508–2512. https://doi.org/10.1109/TCSII.2022.3161061

[28] Yang Liu, Jing Liu, Xiaoguang Zhu, Donglai Wei, Xiaohong Huang, and Liang Song. 2022. Learning Task-Specific Representation for Video Anomaly Detection with Spatial-Temporal Attention. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2190–2194. https://doi.org/10.1109/ICASSP43922.2022.9746822

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

[30] 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*. 2554–2562.

[31] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 164–172.

[32] François Michaud, Paolo Pirjanian, Jonathan Audet, and Dominic Létourneau. 2000. Artificial emotion and social robotics. In *Distributed autonomous robotic systems 4*. Springer, 121–130.

[33] Trisha Mittal, Aniket Bera, and Dinesh Manocha. 2021. Multimodal and Context-Aware Emotion Perception Model With Multiplicative Fusion. *IEEE MultiMedia* 28, 2 (2021), 67–75.

[34] Donglin Niu, Jennifer G Dy, and Michael I Jordan. 2010. Multiple non-redundant spectral clustering views. In *ICML*.

[35] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*. PMLR, 2642–2651.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[38] 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*, Vol. 33. 6892–6899.

[39] Amir Shirian and Tanaya Guha. 2021. Compact graph architecture for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6284–6288.

[40] Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. 2007. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*. 823–830.

[41] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. 27–34.

[42] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8992–8999.

[43] 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 conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

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

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[46] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. 2021. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4902–4910.

[47] Xiao Wang, Meiqi Zhu, Deyu Bo, Peng Cui, Chuan Shi, and Jian Pei. 2020. Am-gcn: Adaptive multi-channel graph convolutional networks. In *Proceedings of the 26th ACM SIGKDD International conference on knowledge discovery & data mining*. 1243–1253.

[48] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7216–7223.

[49] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. 2019. Disentangled variational representation for heterogeneous face recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9005–9012.

[50] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4730–4738.

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

[52] Amir Zadeh and Paul Pu. 2018. 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 (Long Papers)*.

[53] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[54] Peng Zhai, Jie Luo, Zhiyan Dong, Lihua Zhang, Shunli Wang, and Dingkang Yang. 2022. Robust Adversarial Reinforcement Learning with Dissipation Inequation Constraint. (2022).

[55] Changqing Zhang, Ziwei Yu, Qinghua Hu, Pengfei Zhu, Xinwang Liu, and Xiaobo Wang. 2018. Latent semantic aware multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[56] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor Versatile Multi-modal Learning for Multi-label Emotion Recognition. *arXiv preprint arXiv:2201.05834* (2022).