DRKF: Decoupled Representations with Knowledge Fusion for Multimodal Emotion Recognition

Peiyuan Jiang
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu, Sichuan, China
darcy981020@gmail.com

Yao Liu*
School of Information and Software Engineering, University of Electronic Science and Technology of China
Chengdu, Sichuan, China
liuyao@uestc.edu.cn

Qiao Liu
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu, Sichuan, China
qliu@uestc.edu.cn

Zongshun Zhang
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu, Sichuan, China
202421081411@std.uestc.edu.cn

Jiaye Yang
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu, Sichuan, China
202411081710@std.uestc.edu.cn

Lu Liu
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu, Sichuan, China
202522080827@std.uestc.edu.cn

Daibing Yao
Yizhou Prison, Sichuan Province
Chengdu, Sichuan, China
357497551@qq.com

Abstract

Multimodal emotion recognition (MER) aims to identify emotional states by integrating and analyzing information from multiple modalities. However, inherent modality heterogeneity and inconsistencies in emotional cues remain key challenges that hinder performance. To address these issues, we propose a Decoupled Representations with Knowledge Fusion (DRKF) method for MER. DRKF consists of two main modules: an Optimized Representation Learning (ORL) Module and a Knowledge Fusion (KF) Module. ORL employs a contrastive mutual information estimation method with progressive modality augmentation to decouple task-relevant shared representations and modality-specific features while mitigating modality heterogeneity. KF includes a lightweight self-attention-based Fusion Encoder (FE) that identifies the dominant modality and integrates emotional information from other modalities to enhance the fused representation. To handle potential errors from incorrect dominant modality selection under emotionally inconsistent conditions, we introduce an Emotion Discrimination Submodule (ED), which enforces the fused representation to retain discriminative cues of emotional inconsistency. This ensures that even if the FE selects an inappropriate dominant modality, the Emotion Classification Submodule (EC) can still make accurate predictions by leveraging preserved inconsistency information. Experiments show that DRKF achieves state-of-the-art (SOTA) performance on IEMOCAP, MELD, and M3ED. The source code is publicly available at https://github.com/PANPANKK/DRKF.

CCS Concepts

• **Computing methodologies** → Probabilistic reasoning; **Neural networks**; *Learning latent representations.*

Keywords

Multimodal emotion recognition, Contrastive learning, Mutual information, Multimodal fusion

ACM Reference Format:
Peiyuan Jiang, Yao Liu, Qiao Liu, Zongshun Zhang, Jiaye Yang, Lu Liu, and Daibing Yao. 2025. DRKF: Decoupled Representations with Knowledge Fusion for Multimodal Emotion Recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754758

1 Introduction

Multimodal emotion recognition based on speech and text is essential for human-computer interaction (HCI) [1]. The MER's fundamental concept is to acquire modality representations and subsequently fuse them [2, 3]. In representation learning, contrastive learning-based methods have been widely applied to various multimodal tasks [4–6]. These methods rely on the multi-view redundancy assumption, which states that the shared information across different modalities can sufficiently capture the critical features required for downstream tasks [7, 8].

Although multimodal emotion representation learning has made significant progress under this assumption, it does not always hold

*Corresponding author: Yao Liu.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754758

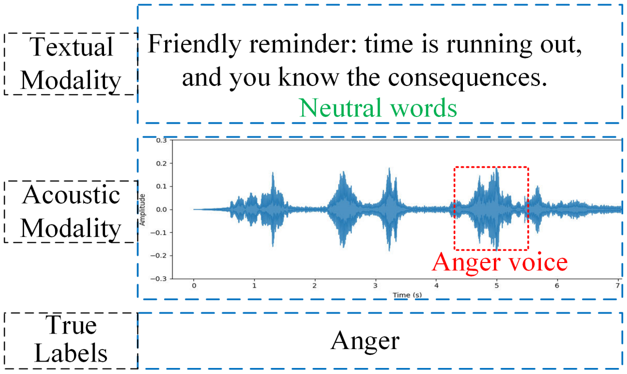


Figure 1: Illustration of Modality-specific Emotions and True Labels in a Multi-view Non-redundant Scenario.

in broader real-world multimodal scenarios. In multi-view non-redundant scenarios, the information contained in each modality is not necessarily relevant to the downstream task. To address this issue, existing studies [9–12] leverage techniques such as adversarial learning, parameter sharing, and subspace learning to decouple modality-specific and shared features.

The aforementioned methods are capable of extracting modality-specific and shared features. However, they are unable to ensure that the learned representations are pertinent to the tasks. [13] adopts an information-theoretic perspective to describe the modality-specific and shared information pertaining to a given task. A Contrastive Mutual Information Estimation (CMIE) method has been introduced to optimize modality-specific and shared representations that are applicable to the task [14]. In such methods, neural networks (NNs) are typically employed to determine mutual information scores between modalities [15, 16]. However, the efficacy of representation learning can be compromised by the instability of these approaches when there is a substantial distributional disparity between task labels and input modalities.

In multimodal representation fusion, early methods primarily relied on tensor fusion and simple feature concatenation [17, 18]. Recent advances have introduced cross-attention mechanisms to better model semantic dependencies and task-oriented alignment across modalities. However, these mechanisms can still suffer from the introduction of redundant noise and increased modeling ambiguity, particularly when emotion-related information is inconsistently conveyed across different modalities [19, 20].

In this work, we aim to address two key challenges in MER: the difficulty of extracting and aligning task-relevant information across heterogeneous modalities—stemming from their inherent representational differences—and the inconsistencies in emotional cues conveyed by different modalities, as illustrated in Fig. 1. To tackle these challenges, we propose a Decoupled Representations with Knowledge Fusion Method (DRKF) for MER. Our model consists of two modules: *the Optimized Representation Learning Module (ORL)*, inspired by [21], indirectly aligns the input modalities with the label distribution through progressive modality augmentation learning, thereby overcoming the challenge of mutual information estimation in the CMIE method caused by modality-label distribution discrepancies. Following the ORL, the Knowledge Fusion Module (KF) consists of a Fusion Encoder (FE), an Emotion Classification Submodule (EC), and an Emotion Discrimination Submodule (ED). The FE, based on self-attention mechanism, identifies the dominant modality of the current sample and integrates complementary emotional information from other modalities to enhance the fused representation. Under emotionally inconsistent conditions, the ED further constrains the fused representation to retain discriminative cues regarding intermodal emotional discrepancies, thereby mitigating potential errors caused by incorrect dominant modality selection. Finally, the EC takes the fused representation as input to perform the emotion classification task. Through their collaborative design, these three components enable more robust and adaptable multimodal emotion recognition.

- We introduce an optimized representation learning module, which learns an optimal enhanced modality to guide the alignment of distributions across modalities as well as between modalities and labels, thereby facilitating more effective representation decoupling.
- We introduce a Knowledge Fusion Module that leverages collaborative learning to integrate fusion encoding, emotion consistency discrimination, and emotion classification, ensuring reliable emotion recognition.
- Extensive experiments on three benchmark datasets demonstrated that the proposed DRKF framework surpasses state-of-the-art methods.

2 Related Work

2.1 Representation Learning

Bengio [22] emphasized that the performance of machine learning models heavily depends on the selection of input features. Different feature representations can entangle varying underlying explanatory factors, potentially obscuring their distinct contributions to the learning process. Multimodal emotion representations can be categorized into feature engineering-based representations and deep neural network-based representations. The former relies on expert knowledge, including acoustic features extracted using openSMILE [23] and textual features from emotion lexicons [24] and syntactic structures [25], while the latter leverages deep learning to automatically extract high-level features. Representative models include BERT [26], RoBERTa [27], wav2vec 2.0 [28], and WavLM [29].

Building on deep representations, multimodal emotion representation learning can be further categorized into single-stream and multi-stream models. Single-stream models use a shared encoder to learn joint representations within a unified latent space, whereas multi-stream models maintain separate pathways for each modality and integrate them later to capture cross-modal interactions [30]. While multi-stream architectures effectively preserve modality-specific information and enhance cross-modal interactions, they can also introduce irrelevant noise, hindering optimal fusion performance. Decoupled representation serves as a key mechanism to filter out irrelevant information and improve fusion quality [9].

To enhance the effectiveness of decoupled representations, recent studies have incorporated contrastive learning to reduce intermodal distributional discrepancies and achieve decoupling of shared

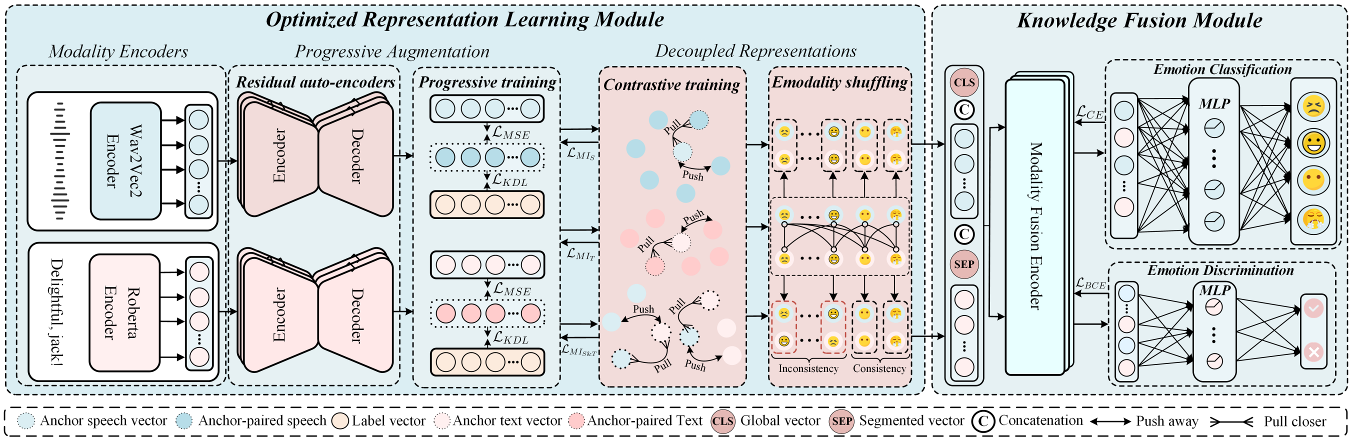


Figure 2: Overview of the Proposed DRKF Framework. It consists of two components: the ORL Module, which improves task-relevant modality mutual information estimation, and the KF Module, which models modality interactions for final emotion classification.

representations across modalities [31, 32]. Additionally, some approaches employ subspace mapping techniques, adversarial learning, and orthogonality constraints to extract task-relevant modality-specific information [10, 11, 33], while others leverage information-theoretic methods to quantify the task relevance of private information, further improving the interpretability and discriminative power of learned representations [14, 34, 35].

2.2 Modality Fusion

Morency et al. [36] identified five core challenges in multimodal learning: representation learning, modality conversion, modality alignment, co-learning, and modality fusion. Among these, modality fusion is essential for cross-modal knowledge integration. Existing fusion strategies can be broadly categorized into feature-level, decision-level, and interaction-based fusion. Feature-level fusion combines features from different modalities, such as through concatenation [37] or time-scale-aware integration [38], but increases the classifier’s burden in handling redundancy and modality misalignment. Decision-level fusion [39] integrates modality-specific predictions using methods like ensemble learning, weighted averaging, or voting, but often overlooks fine-grained interactions crucial for tasks like emotion recognition. Interaction-based fusion learns cross-modal relationships through attention mechanisms or latent space alignment. For instance, [40] proposed a multi-hop attention mechanism, allowing textual tokens to iteratively query audio features, thus enhancing fusion expressiveness.

Although attention-based fusion strategies are effective, they require task-specific queries to adapt to dataset variations, limiting unified multimodal modeling. To address this, [19, 20] proposed a bidirectional cross-attention mechanism to improve adaptability and generalization across datasets. However, while bidirectional cross-attention has proven effective, it may introduce noise when dealing with emotionally inconsistent modalities, leading to model confusion and performance degradation.

3 APPROACH

3.1 Problem Statement

Our proposed model takes raw speech and text as input, aiming to integrate acoustic information from speech and semantic information from text to comprehensively determine the conveyed emotion. The input speech and text are first processed by their respective encoders, resulting in speech sequence vectors $S_{seq} = \{s_1, s_2, \ldots, s_m\}$ and text sequence vectors $T_{seq} = \{t_1, t_2, \ldots, t_n\}$, where m and n denote the lengths of the encoded sequences. The outputs of the encoders are optimized through the **ORL Module**. The optimized representations are then fed into the **KF Module**, which outputs the final emotion probability vector $P \in \{p_1, p_2, \ldots, p_n\}$.

3.2 Model Architecture

Fig. 2 illustrates the proposed DRKF. It consists of two key components: the ORL Module, and the KF Module.

(1) The ORL Module comprises three components: Modality Encoding (ME), Progressive Augmentation (PA), and Decoupled Representations (DR). The ME integrates an acoustic encoder and a semantic encoder to extract modality-specific embeddings. The acoustic encoder, based on the pre-trained wav2vec2 model[1], transforms raw audio into acoustic embeddings, while the semantic encoder, leveraging the pre-trained RoBERTa model[2], encodes raw text into semantic embeddings. The PE employs two identically structured residual autoencoder networks, each consisting of five residual autoencoder blocks with six linear layers per block. The purpose of this component is to learn the optimal feature augmentation for the acoustic and semantic modalities. The DR includes contrastive training and emotion modality (Emodality) shuffling mechanisms. The contrastive training method eliminates task-irrelevant modality noise and facilitates the learning of decoupled representations. The Emodality shuffling mechanism restructures optimized modality pairs for downstream processing.

[1] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
[2] https://huggingface.co/FacebookAI/roberta-large

(2) The KF Module comprises three components: the Fusion Encoder (FE), Emotion Classification Submodule (EC) and the Emotion Discrimination Submodule (ED). The FE is a lightweight, self-attention-based encoder that identifies the dominant modality and integrates supplementary emotional information from other modalities. To address potential errors caused by incorrect dominant modality selection under emotionally inconsistent conditions, the ED enforces the fused representation to retain discriminative cues related to intermodal emotional discrepancies. This mechanism ensures that, even when the FE fails to select the optimal modality, the EC can still make accurate predictions by leveraging the preserved inconsistency information. Both ED and EC are implemented as two independent multilayer perceptrons (MLPs).

3.3 The ORL Module

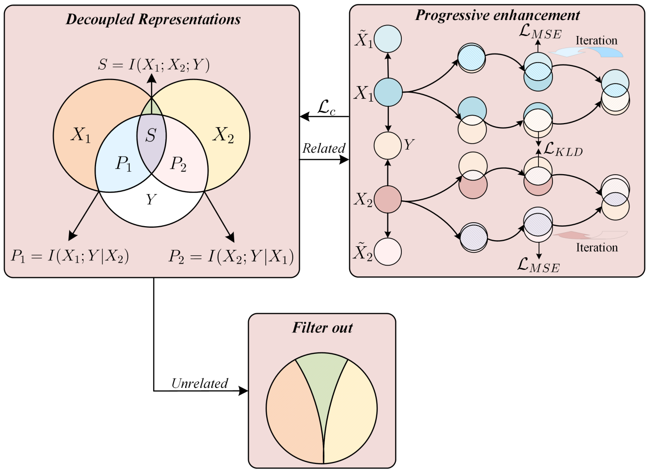


Figure 3: Decoupled Representations Learning Flowchart.

3.3.1 *The Decoupled Representations.*

We employed an information-theoretic-based decoupled representation approach to filter out task-irrelevant modality information and optimize task-relevant representations, including modality-shared information S and modality-specific information P. The task-relevant modality information can be expressed by the following formula:

$$I(X_1, X_2; Y) = S + P_1 + P_2 \tag{1}$$

Where, $I(X_1, X_2; Y)$ is the mutual information between the task variable Y and the modalities X_1 and X_2.

$$S = I(X_1; X_2) - I(X_1; X_2 \mid Y) \tag{2}$$

$$P_1 = I(X_1; Y \mid X_2), \quad P_2 = I(X_2; Y \mid X_1) \tag{3}$$

Where, S is the task-relevant modality-shared mutual information, P_1 and P_2 are the task-relevant modality-specific mutual information within each modality.

$$I(X_1; X_2) = \int p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)}\, dx_1\, dx_2 \tag{4}$$

$$I(X_1; X_2 \mid Y) = \int p(x_1, x_2, y) \log \frac{p(x_1, x_2 \mid y)}{p(x_1 \mid y)p(x_2 \mid y)}\, dx_1\, dx_2\, dy \tag{5}$$

Where, $I(X_1; X_2)$ represents the total mutual information between the two modalities X_1 and X_2, $I(X_1; X_2 \mid Y)$ represents the conditional mutual information between X_1 and X_2 given the task Y, reflecting task-irrelevant modality-shared mutual information.

Calculating mutual information, as shown in the formula above, requires a closed-form density function and a log-density ratio between the joint and marginal distributions in a manageable form. However, in real-world machine learning tasks, we only have access to samples from the joint distribution, making direct computation of mutual information difficult and forcing us to rely on approximation methods.

3.3.2 *The Progressive Augmentation.*

To address the challenge of directly computing mutual information, we introduce contrastive mutual information estimation, formulated as follows:

$$I\left(X_1; X_2\right) = \mathbb{E}x_1, x_2, x_2^- \left[\log \frac{\exp f\left(x_1, x_2\right)}{\sum_M \exp f\left(x_1, x_2^-\right)}\right] \tag{6}$$

$$I\left(X_1; X_2 \mid Y\right) = \mathbb{E}_{y, x_1, x_2, x_2^-}\left[\log \frac{\exp f(x_1, x_2, y)}{\sum_M \exp f(x_1, x_2^-, y)}\right] \tag{7}$$

In the above equations, M represents the batch size during training. The expectation operator $\mathbb{E}$ is taken over the joint distribution of positive and negative sample pairs, with or without conditioning on the label y. The function $f(\cdot, \cdot)$ measures correlation between inputs. Positive samples (x_1, x_2) or (x_1, x_2, y) share semantic alignment, while negative samples x_2^- or (x_2^-, y) are drawn from other instances in the batch.

Although contrastive mutual information estimation circumvents the challenges of direct computation, it still faces a critical limitation: the representation gap between modalities and task labels further increases the difficulty of mutual information estimation. As shown in Eq. (7), x_1 and x_2 follow continuous distributions in their respective feature spaces, whereas y is discrete. This modality-label distribution mismatch makes score estimation more challenging. To tackle these challenges, we propose a progressive modality augmentation strategy that guides the alignment between modalities and between modality and label distributions by iteratively learning the optimal augmented modality. The optimal augmented modality is defined as follows:

- Optimal augmented modality: When $I(X; Y) = I(X; \tilde{X}_1)$, $\tilde{X}_1$ is the optimal unimodal augmentation of X, which implies that the only information shared between X and $\tilde{X}_1$ is task-relevant, and that X and $\tilde{X}_1$ lie within the same subspace.

Our proposed progressive augmentation strategy, as illustrated in Fig. 3, is designed to learn the optimal augmented modality. Unlike traditional static feature augmentation methods, our proposed progressive augmentation strategy is a dynamic optimization approach based on original modality features, enabling a stepwise optimization process to achieve optimal modality augmentation. To guide the learning of the augmented features, we introduce two carefully designed constraints. First, by constraining the distribution discrepancy between the augmented modality and the original modality, we ensure that the augmented features reside in the same subspace as the original modality features, thus reducing modality heterogeneity. Second, by minimizing the discrepancy between the augmented modality and the task label distribution, we ensure that the augmented features are aligned with the task label distribution to the greatest extent. Finally, we control the model's overall performance by adjusting the weights of these two constraints.

The process of progressive augmentation can be described by the following equations:

We utilize a residual autoencoder f_{AE} to learn the optimal unimodal augmentation $\tilde{T}_{seq}, \tilde{T}_{cls}$ for the text feature sequence.

$$\tilde{T}_{seq} = f_{AE}(T_{seq}), \quad \tilde{T}_{cls} = \text{avg}(\tilde{T}_{seq}) \tag{8}$$

Where $T_{seq} \in \mathbb{R}^{d_n \times d_z}$ denotes the feature sequence output of the text encoder, with d_n as the sequence length and d_z as the dimensionality of each sequence element. $\tilde{T}_{cls} \in \mathbb{R}^{d_z}$ represents the global feature vector obtained by average pooling the sequence feature vector.

Similarly, the optimal unimodal augmentation for the speech feature sequence is calculated as follows:

$$\tilde{S}_{seq} = f_{AE}(S_{seq}), \quad \tilde{S}_{cls} = \text{avg}(\tilde{S}_{seq}) \tag{9}$$

Where $S_{seq} \in \mathbb{R}^{d_n \times d_z}$ denotes the feature sequence output of the speech encoder, with d_n as the sequence length and d_z as the dimensionality of each sequence element. $\tilde{S}_{cls} \in \mathbb{R}^{d_z}$ represents the global feature vector obtained by average pooling the sequence feature vector.

To ensure that the augmented modality remains in the same subspace as the original modality, we use the Mean Squared Error (MSE) loss to constrain the learning space of the augmented modality. Meanwhile, the Kullback-Leibler divergence (KLD) loss is employed to enforce the augmented modality to learn the distribution of the task labels. The calculation formula is as follows:

$$\mathcal{L}_{MSE} = \frac{1}{2M} \sum_{i=1}^{M} \left[\left\| S_{seq}^{i} - \tilde{S}_{seq}^{i} \right\|^2 + \left\| T_{seq}^{i} - \tilde{T}_{seq}^{i} \right\|^2 \right] \tag{10}$$

$$\mathcal{L}_{KLD} = \frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{C} y_i^c \log \frac{y_i^c}{\hat{y}_i^c} \tag{11}$$

$$\mathcal{L}_a = \alpha \cdot \mathcal{L}_{MSE} + \mathcal{L}_{KLD} \tag{12}$$

Here, S_{seq}^i and T_{seq}^i represent the original feature sequence vectors of the i-th sample in the speech and text modalities, respectively. $\tilde{S}_{seq}^i$ and $\tilde{T}_{seq}^i$ represent the augmented feature sequence vectors. M is the batch size, and C denotes the total number of emotion categories. The true label of the i-th sample in the fusion modality for category c is denoted as y_i^c, while $\hat{y}_i^c$ represents the predicted probability for the same sample and category. $\mathcal{L}_a$ refers to the final augmentation loss, and α is a hyperparameter.

3.3.3 Contrastive Mutual Information Estimation. The maximum task-relevant modality mutual information can be transformed into minimizing the negative of contrastive mutual information. Therefore, the objective function for our decoupled representation learning method is defined as follows:

$$z_S^i = g(S_{cls}^i), \tilde{z}_S^i = g(\tilde{S}_{cls}^i), z_T^i = g(T_{cls}^i), \tilde{z}_T^i = g(\tilde{T}_{cls}^i) \tag{13}$$

$$\mathcal{L}_{MI_{S_i}} = -\log \frac{\exp(\text{sim}(z_S^i, \tilde{z}_S^i)/\tau)}{\sum_{k=1}^{2M} [k \neq i] \exp(\text{sim}(z_S^i, \tilde{z}_S^k)/\tau)} \tag{14}$$

$$\mathcal{L}_{MI_{T_i}} = -\log \frac{\exp(\text{sim}(z_T^i, \tilde{z}_T^i)/\tau)}{\sum_{k=1}^{2M} [k \neq i] \exp(\text{sim}(z_T^i, \tilde{z}_T^k)/\tau)} \tag{15}$$

$$\mathcal{L}_{MI_{S_i \& T_i}} = -\log \frac{\exp(\text{sim}(z_S^i, z_T^i)/\tau)}{\sum_{k=1}^{M} \exp(\text{sim}(z_S^i, z_T^k)/\tau)} \tag{16}$$

In the above formulas, $g(\cdot)$ represents a projection function implemented via an MLP; $\text{sim}(\cdot,\cdot)$ represents the cosine similarity; $\mathcal{L}_{MI_{S_i}}$ and $\mathcal{L}_{MI_{T_i}}$ denote the intra-modality CMIE objective functions, while $\mathcal{L}_{MI_{S_i \& T_i}}$ corresponds to the intermodality CMIE objective function.

Finally, our CMIE optimization objective $\mathcal{L}_c$ is defined as follows:

$$\mathcal{L}_c = \frac{1}{M} \sum_{i=1}^{M} \left(\mathcal{L}_{MI_{S_i}} + \mathcal{L}_{MI_{T_i}} + \mathcal{L}_{MI_{S_i \& T_i}} \right) \tag{17}$$

3.4 The KF Module

3.4.1 The Fusion Encoder. To enable the proposed knowledge fusion module to effectively model the complex interactions between different modalities, we applied a specialized concatenation process to the multimodal input sequences, as shown in Equation 18.

$$X_{fusion} = f_{FE} \left(\text{Concat} \left(C_{cls}, S_{seq}, C_{sep}, T_{seq}, C_{sep} \right) \right) \tag{18}$$

Where, $Concat(\cdot)$ denotes the concatenation function, C_{cls} and C_{sep} represent the classification token vector and separator token vector, respectively. C_{cls} is designed to aggregate global information from the input features during the modality fusion process, while C_{sep} acts as a boundary to distinguish between the two modalities, facilitating the ED in learning modality-consistent information. Finally, $f_{FE}(\cdot)$ processes the concatenated sequence to generate the fused feature vector X_{fusion}.

3.4.2 The Emotion Discrimination Submodule. The ED is implemented as an MLP, trained on data generated through the Emodality shuffling process in the ORL module. Specifically, for each batch of data, samples from different modalities are randomly combined to create new speech-text pairs. If the original emotion labels of both modalities in a newly formed pair match, the emotional information is considered consistent (assigned a label of 1). Conversely, if the labels do not match, the emotional information is deemed inconsistent (assigned a label of 0). The optimization function for this module is as follows:

$$\mathcal{L}_b = -\frac{1}{M^2} \sum_{i=1}^{M^2} \left[Y_i \log \hat{Y}_i + (1 - Y_i) \log(1 - \hat{Y}_i) \right] \tag{19}$$

Where, M is the batch size, Y_i is the true label of the i-th sample in the binary emotion decoupling task, and $\hat{Y}_i$ represents the predicted probability of the i-th sample output.

3.4.3 The Emotion Classification Submodule. We employ an independent MLP to perform multimodal emotion classification. Specifically, the EC takes as input the correctly matched sample pairs generated by the Emodality shuffling process. The calculation processes for the emotion classification loss $\mathcal{L}_f$ is presented in Equation 20.

$$\mathcal{L}_f = -\frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{C} y_i^c \log \hat{y}_i^c \tag{20}$$

Where, M is the batch size, C is the number of emotion categories, y_i^c is the true label of the i-th sample in the fusion modality for

Table 1: Model Performance Comparison on the IEMOCAP Dataset

Models	Audio & Text Encoder	ACC (%)	WACC (%)	Avg (%)
GBAN [41]	CNN-LSTM	70.1	72.4	71.2
MSER-MVAM [42]	CNN-LSTM	74.2	75.4	74.8
MSER-CADF [43]	CNN-GRU	77.2	76.5	76.8
MCFN [44]	CNN-Roberta	77.8	76.0	76.9
SAMS [45]	BiGRU-Bert	78.1	76.6	77.3
LLMSER [46]	BiLSTM-Bert	78.3	78.1	78.2
KS-Transformer [47]	Wav2vec-Roberta	75.3	74.3	74.8
KBCAM [48]	Wav2vec2-Bert	77.0	75.5	76.2
DBT [49]	Wav2vec2-Roberta	78.9	77.8	78.3
Ours(ORKF)	Wav2vec2-Roberta	**80.7**	**79.9**	**80.3**
ΔSota	—	**↑ 2.28**	**↑2.30**	**↑2.55**

Notes: Bold values indicate the best performance. Underlined values denote the second-best performance. ΔSota represents the relative improvement of our proposed method compared to the second-best model. (↑) indicates an improvement over the second-best performance, where higher values are better. (↓) indicates a decrease relative to the best performance, where lower values are better.

category c, and $\hat{y}_i^c$ represents the predicted probability of the i-th sample in the fusion modality for category c.

Finally, our objective function, denoted as $\mathcal{L}$, is formally defined in Equation 21.

$$\mathcal{L} = \mathcal{L}_a + \beta \cdot \mathcal{L}_c + \gamma \cdot \mathcal{L}_f + \delta \cdot \mathcal{L}_b \tag{21}$$

Where β, γ, and δ are hyperparameters.

4 Experiment Settings

4.1 Datasets

We validated our proposed model on three publicly available multimodal datasets, IEMOCAP [50], MELD [51] and M3ED [52]. Specifically, IEMOCAP is a recorded dialogue dataset with emotion labels including anger, happiness, sadness, frustration, excitement, fear, surprise, disgust, and others. To ensure consistency with previous research, we focused on four emotion categories: happiness, sadness, anger, and neutral, where excitement was merged into the happiness category, resulting in a total of 5,531 samples. The experiments followed a five-fold leave-one-session-out strategy, Using Unweighted Accuracy (ACC), Weighted Accuracy (WACC), and their average (Avg) to evaluate model performance. MELD is a challenging multi-party conversation dataset, annotated with seven emotion labels. Unlike IEMOCAP, this dataset is divided into training, development, and test sets, providing a standardized training and evaluation strategy for models. WACC, Weighted F1 score (WF1) and Avg were used to assess the performance of the models on this dataset. M3ED is the first Chinese multi-label emotion dialogue dataset. The utterance-level emotion labels include seven categories: happiness, surprise, sadness, disgust, anger, fear, and neutral. Following previous studies, we used Precision, Recall, ACC, Micro-F1 (F1) score and the Avg as evaluation metrics to assess model performance.

4.2 Implementation Details

Our model is implemented using the PyTorch framework, with AdamW as the optimizer, a learning rate of 1e-5, and a batch size of 4. The output dimension of the projection head function in contrastive learning is set to 1024, and the multimodal fusion layer has 8 attention heads. The values of the loss function hyperparameters are set to 0.2, 0.2, 1.0, and 0.2, respectively. Our training is conducted on a Linux system with an A100 GPU, for a total of 100 epochs.

4.3 Baseline Models

To validate the effectiveness of the proposed method, we compared ORKF with the current advanced baseline methods. The baselines used to evaluate ORKF across different datasets are as follows. It is important to note that some baseline models are evaluated on multiple datasets, and we provide their descriptions only when they are mentioned for the first time.

Compared Methods for IEMOCAP Dataset: GBAN [41] with gated attention fusion; DIMMN [53] using dynamic memory interaction; MSER-CADF [43] with cross-attention fusion; MCFN [44] employing dual-stream temporal-spatial modeling; SAMS [45] aligning semantics across modalities; LLMSER [46] enhancing prompts in language models; KS-Transformer [47] using pre-trained feature extraction and early fusion; KBCAM [48] incorporating Bayesian attention with external knowledge; DBT [49] utilizing dual-branch Transformer with fine-tuning fusion. Compared Methods for MELD Dataset: MCSCAN [54] with parallel cross/self-attention; DIMMN [53] with dynamic memory integration; SACMA [55] integrating speaker-aware emotion recognition; SMCN [56] self-guided modality alignment; RMERCT [57] using Transformer-based cross-modal fusion; SMIN [58] semi-supervised multimodal learning; HiMul-LGG [59] hierarchical decision fusion strategy. Compared Methods for M3ED Dataset: MSCNN-SPU [60] integrating multi-scale CNN with statistical pooling; M-TLEAF [61] using bidirectional GRU and Transformer fusion; CARAT [62] employing contrastive feature reconstruction and aggregation.

4.4 Results

As shown in Table 1, the proposed ORKF method achieves the best overall performance on the IEMOCAP dataset, with an ACC of 80.7%, a WACC of 79.9%, and an average (Avg) of 80.3%. In terms of

Table 2: Model Performance Comparison on the MELD Dataset

Models	Audio & Text Encoder	WACC (%)	WF1 (%)	Avg (%)
MCSCAN [54]	CNN & LSTM-LSTM	N/A	59.2	59.2‡
DIMMN [53]	Attention-CNN	60.6	58.6	59.6
SACMA [55]	LSTM-TextCNN	62.3	59.3	60.8
MCFN [44]	CNN-Roberta	64.5	62.2	63.3
SMCN [56]	GRU-Bert	64.9	62.3	63.6
SAMS [45]	BiGRU-Bert	65.4	62.6	64.0
RMERCT [57]	WaveRNN-GPT	63.1	64.0	63.5
SMIN [58]	Wav2vec-Roberta	65.5	64.5	65.0
HiMul-LGG [59]	BiGRU-Roberta	66.2	65.1	65.6
Ours(ORKF)	Wav2vec2-Roberta	**66.7**	**65.4**	**66.0**
ΔSota	—	**↑ 0.75**	**↑0.46**	**↑0.60**

Notes: N/A indicates that the metric value was not provided in the original paper. ‡ denotes that the average value was calculated from known experimental results due to irreproducibility.

Table 3: Model Performance Comparison on the M3ED Dataset

Models	Audio & Text Encoder	Precision (%)	Recall (%)	ACC (%)	F1 (%)	Avg (%)
MSCNN-SPU † [60]	CNN-TextCNN	44.3	50.1	45.0	47.0	46.6
M-TLEAF † [61]	CNN-BERT	45.2	49.1	46.7	47.1	47.0
MCFN † [44]	CNN-Roberta	44.7	50.9	46.1	47.6	47.3
SAMS † [45]	BiGRU-Bert	47.1	51.5	**51.1**	49.2	49.7
CARAT† [62]	Transformer Encoder-Based	45.0	51.4	44.3	48.0	47.1
Ours(ORKF)	Wav2vec2-Roberta	**52.6**	**51.6**	50.6	**52.0**	**51.7**
ΔSota	—	**↑ 11.6**	**↑0.38**	**↓0.98**	**↑5.69**	**↑4.02**

Notes: † The results are obtained through our own reproduction experiments.

ACC and Avg, ORKF achieves relative improvements of approximately 2.28% and 2.55%, respectively, compared to the second-best method DBT (ACC of 78.9%, Avg of 78.3%). For the WACC metric, ORKF shows a relative improvement of about 2.30% over the second-best method LLMSER (WACC of 78.1%).

To further validate the performance of ORKF, we evaluated the model on the MELD dataset. The experimental results are presented in Table 2.

According to the experimental comparison results in Table 2, ORKF demonstrates superior performance even on the highly imbalanced MELD dataset. Specifically, ORKF achieves a WACC of 66.7, a WF1 of 65.4, and an average score of 66.0, all of which represent the best results among the compared methods. Although ORKF achieved SOTA performance on both the IEMOCAP and MELD datasets, it should be noted that these datasets are English datasets with single-label annotations, where task-relevant information may primarily rely on shared mutual information. To further validate the model's performance, we introduced the Chinese multi-label emotion recognition dataset M3ED for testing.

As shown in Table 3, even on the more complex multi-label Chinese emotion recognition dataset, ORKF demonstrates strong performance, achieving an F1 score of 52.0 and an average score of 51.7, reaching SOTA results.

Overall, the comparative experimental results demonstrate that ORKF effectively integrates information from multiple modalities, achieving robust emotion recognition.

4.5 Ablation Study

To evaluate the effectiveness of the proposed strategy, we conducted ablation experiments on the IEMOCAP, MELD, and M3ED datasets, with the results presented in Table 4.

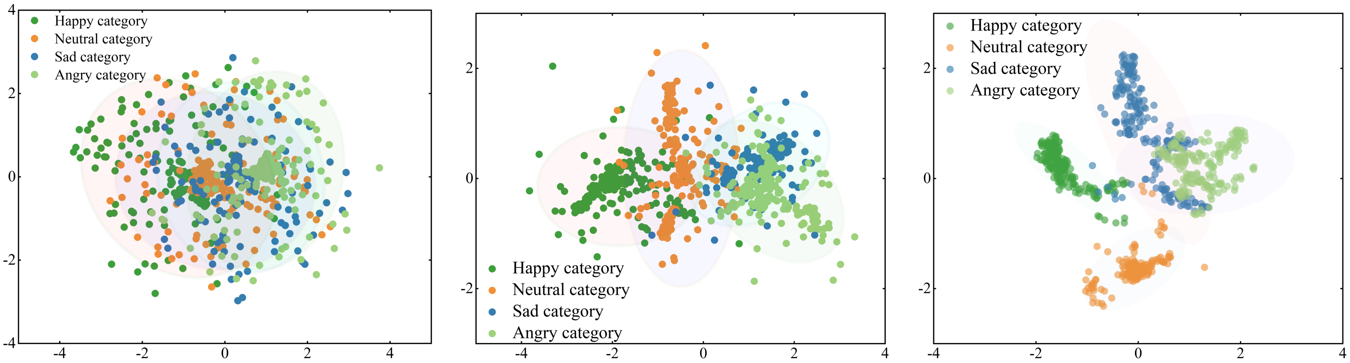
As shown in Table 4, the comparison between the first and fourth rows demonstrates that removing the ED leads to a noticeable performance drop across all three datasets, highlighting its critical role in helping the fusion module learn effective joint representations. The comparison between the second and third rows further validates the effectiveness of the proposed Progressive Contrastive Mutual Information Estimation (PCMI) approach in enhancing decoupled representation learning and improving overall model performance.

To further verify the effectiveness of the proposed decoupled representation learning strategy based on PCMI, we visualized the learned embeddings on the test set (session 2) of the IEMOCAP dataset using t-distributed Stochastic Neighbor Embedding (t-SNE), as shown in Figs. 4a–4c. From the figures, it can be observed that the

Table 4: Results for Strategy Analysis

Methods						IEMOCAP		MELD		M3ED	
BME	PCMI	CMIE	FE_c	FE_s	ED	ACC	WACC	WACC	WF1	ACC	F1
✓				✓		78.1	77.0	64.5	63.3	47.1	49.3
✓		✓		✓		78.4	77.5	64.7	63.4	47.3	49.9
✓	✓			✓		79.7	78.4	65.6	64.7	47.9	50.8
✓				✓	✓	79.0	78.1	65.1	63.5	48.1	51.1
✓	✓		✓		✓	78.2	76.3	64.3	63.2	46.7	48.5
✓	✓			✓	✓	**80.7**	**79.9**	**66.7**	**65.4**	**50.6**	**52.0**

Notes: The ✓symbol indicates that the corresponding method is applied. BME refers to the BiModal Encoder. PCMI represents the Progressive Contrastive Mutual Information Estimation. CMIE represents the Contrastive Mutual Information Estimation introduced by [14]. FE_c denotes the fusion encoder with a bidirectional cross-attention mechanism. FE_s denotes the fusion encoder with a self-attention mechanism. *ED* refers to the emotion discrimination submodule.



(a) Distribution of Different Emotion Categories before Using PCMI Method

(b) Distribution of Different Emotion Categories after Using CMIE

(c) Distribution of Different Emotion Categories after Using PCMI Method

Figure 4: Comparison of Emotion Category Distributions: PCMI vs. CMIE and Baseline.

decoupled representation learning strategy based on PCMI effectively facilitates the learning of well-structured and discriminative representations.

Finally, the comparison between the fifth and sixth rows shows that the proposed self-attention-based fusion encoder outperforms the traditional cross-attention fusion encoder, offering more substantial gains in emotion recognition accuracy.

5 Conclusion

To address the prevalent challenges of modality heterogeneity and emotional inconsistency across modalities in MER tasks, we propose a novel framework named Decoupled Representations with Knowledge Fusion (DRKF). The framework consists of two core modules: the Optimized Representation Learning (ORL) module and the Knowledge Fusion (KF) module. Specifically, the ORL module aims to decouple task-relevant modality-shared and modality-specific information while reducing inter-modality heterogeneity, thereby facilitating more effective multimodal fusion. The KF module is designed to learn a fusion representation that is sensitive to emotional discrepancies across modalities, which enhances the model's robustness in scenarios where emotional cues from different modalities are not aligned. Extensive experiments on three widely used benchmark datasets for multimodal emotion recognition demonstrate that DRKF outperforms several state-of-the-art models across multiple evaluation metrics, exhibiting strong performance and generalization capabilities.

Despite the promising results achieved by the proposed DRKF model on bimodal emotion recognition tasks, certain limitations remain. The current evaluation is limited to the audio-text bimodal setting, and has not yet been extended to trimodal or higher-order multimodal fusion scenarios. In future work, we plan to further explore the adaptability and scalability of DRKF in more complex multimodal input settings, such as those involving video, speech, and text, to better address the demands of real-world multimodal emotion recognition applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Award No. U22B2061) and the Natural Science Foundation of Sichuan, China (Award No. 2024NSFSC0496).

References

[1] Andreas Triantafyllopoulos, Björn W. Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, Ruibo Fu, and Jianhua Tao. 2023. An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era. *Proc. IEEE* 111, 10 (2023), 1355–1381. doi:10.1109/JPROC.2023.3250266

[2] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5666–5675. doi:10.18653/v1/2021.acl-long.440

[3] 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*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. doi:10.18653/v1/D17-1115

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Virtual Event, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[5] Simon Jenni, Alexander Black, and John Collomosse. 2023. Audio-Visual Contrastive Learning with Temporal Self-Supervision. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Washington, DC, USA, Article 898, 9 pages. doi:10.1609/aaai.v37i7.25967

[6] Hassan Akbari, Liangzhe 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. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., Red Hook, NY, USA, 24206–24221. https://proceedings.neurips.cc/paper_files/paper/2021/file/cb3213ada48302953cb0f166464ab356-Paper.pdf

[7] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. 2021. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (Proceedings of Machine Learning Research, Vol. 132)*, Vitaly Feldman, Katrina Ligett, and Sivan Sabato (Eds.). PMLR, Virtual Event, 1179–1206. https://proceedings.mlr.press/v132/tosh21a.html

[8] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Self-supervised Learning from a Multi-view Perspective. In *International Conference on Learning Representations (ICLR)*. OpenReview, Virtual Event, 10 pages. https://openreview.net/forum?id=-bdp_8Itjwp

[9] 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)* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 1642–1651. doi:10.1145/3503161.3547754

[10] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor Versatile Multi-Modal Learning for Multi-Label Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. AAAI Press, Vancouver, British Columbia, Canada, 9100–9108.

[11] Haoqin Sun, Shiwan Zhao, Xuechen Wang, Wenjia Zeng, Yong Chen, and Yong Qin. 2024. Fine-Grained Disentangled Representation Learning For Multimodal Emotion Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Seoul, South Korea, 11051–11055. doi:10.1109/ICASSP48485.2024.10447667

[12] Xulong Du, Xingnan Zhang, Dandan Wang, Yingying Xu, Zhiyuan Wu, Shiqing Zhang, Xiaoming Zhao, Jun Yu, and Liangliang Lou. 2024. Integrating Representation Subspace Mapping with Unimodal Auxiliary Loss for Attention-based Multimodal Emotion Recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. European Language Resources Association (ELRA), Torino, Italy, 9120–9130.

[13] Zhixiang Shen, Shuo Wang, and Zhao Kang. 2024. Beyond Redundancy: Information-aware Unsupervised Multiplex Graph Structure Learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Red Hook, NY, USA, 10 pages.

[14] Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., Red Hook, NY, USA, 32971–32998. https://proceedings.neurips.cc/paper_files/paper/2023/file/6818dcc65fdf3cbd4b05770fb957803e-Paper-Conference.pdf

[15] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, 9912–9924. https://proceedings.neurips.cc/paper/2020/file/f1748d6b0fd9c3c0b67a5f43a715af2b-Paper.pdf

[16] Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning Representations by Maximizing Mutual Information Across Views. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Red Hook, NY, USA, 15509–15520. https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

[17] 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*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. doi:10.18653/v1/D17-1115

[18] Joosung Lee and Wooin Lee. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5669–5679. doi:10.18653/v1/2022.naacl-main.416

[19] Yuntao Shou, Huan Liu, Xiangyong Cao, Deyu Meng, and Bo Dong. 2025. A Low-Rank Matching Attention Based Cross-Modal Feature Fusion Method for Conversational Emotion Recognition. *IEEE Transactions on Affective Computing* 16, 2 (2025), 1177–1189. doi:10.1109/TAFFC.2024.3498443

[20] Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14752–14766. doi:10.18653/v1/2023.acl-long.824

[21] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2023. Understanding and Constructing Latent Modality Structures in Multi-Modal Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 7661–7671.

[22] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[23] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1459–1462. doi:10.1145/1873951.1874246

[24] Flavio Carvalho, Gabriel Santos, and Gustavo Paiva Guedes. 2018. AffectPT-br: An Affective Lexicon Based on LIWC 2015. In *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*. IEEE, Puerto Varas, Chile, 1–5. doi:10.1109/SCCC.2018.8705251

[25] Jamilu Awwalu, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. 2019. Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter. *Neural Computing and Applications* 31 (2019), 9207–9220. doi:10.1007/s00521-019-04248-z

[26] Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. 2023. BERT-ERC: Fine-Tuning BERT Is Enough for Emotion Recognition in Conversation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*, Vol. 37. AAAI Press, Washington, DC, USA, 13492–13500. doi:10.1609/aaai.v37i11.26582

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[28] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates,

Inc., Red Hook, NY, USA, 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[29] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518. doi:10.1109/JSTSP.2022.3188113

[30] Ronghao Lin and Haifeng Hu. 2024. Adapt and explore: Multimodal mixup for representation learning. *Information Fusion* 105 (2024), 102216. doi:10.1016/j.inffus.2023.102216

[31] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. AudioCLIP: Extending CLIP to Image, Text and Audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, 976–980. doi:10.1109/ICASSP43922.2022.9747631

[32] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Salt Lake City, UT, USA, 3733–3742.

[33] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. 2018. Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc., Red Hook, NY, USA, 7354–7365. https://proceedings.neurips.cc/paper_files/paper/2018/file/f3ce96dfe0061d0e6e105b0b70e5aafb-Paper.pdf

[34] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 531–540. https://proceedings.mlr.press/v80/belghazi18a.html

[35] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. CLUB: A Contrastive Log-Ratio Upper Bound of Mutual Information. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Virtual Event, Article 166, 10 pages.

[36] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443. doi:10.1109/TPAMI.2018.2798607

[37] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2594–2604. doi:10.18653/v1/D18-1280

[38] Efthymios Georgiou, Charilaos Papaioannou, and Alexandros Potamianos. 2019. Deep Hierarchical Fusion with Application in Sentiment Analysis. In *Proceedings of Interspeech 2019*. ISCA, Graz, Austria, 3302–3306. https://api.semanticscholar.org/CorpusID:202736442

[39] Qiuju Zhang, Hongtao Zhang, Keming Zhou, and Le Zhang. 2023. Developing a Physiological Signal-Based, Mean Threshold and Decision-Level Fusion Algorithm (PMD) for Emotion Recognition. *Tsinghua Science and Technology* 28, 4 (2023), 673–685. doi:10.26599/TST.2022.9010038

[40] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech Emotion Recognition Using Multi-hop Attention Mechanism. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Brighton, United Kingdom, 2822–2826. doi:10.1109/ICASSP.2019.8683483

[41] Pengfei Liu, Kun Li, and Helen Meng. 2020. Group Gated Fusion on Attention-Based Bidirectional Alignment for Multimodal Emotion Recognition. In *Proceedings of Interspeech 2020*. ISCA, Shanghai, China, 379–383. doi:10.21437/Interspeech.2020-2067

[42] Lin Feng, Lu-Yao Liu, Sheng-Lan Liu, Jian Zhou, Han-Qing Yang, and Jie Yang. 2023. Multimodal speech emotion recognition based on multi-scale MFCCs and multi-view attention mechanism. *Multimedia Tools Appl.* 82, 19 (March 2023), 28917–28935. doi:10.1007/s11042-023-14600-0

[43] Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Saddik, and Soonil Kwon. 2024. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications* 245 (2024), 122946. doi:10.1016/j.eswa.2023.122946

[44] Xiaoheng Zhang and Yang Li. 2023. A Dual Attention-based Modality-Collaborative Fusion Network for Emotion Recognition. In *Proceedings of Interspeech 2023*. ISCA, Dublin, Ireland, 1468–1472. doi:10.21437/Interspeech.2023-523

[45] Mixiao Hou, Zheng Zhang, Chang Liu, and Guangming Lu. 2023. Semantic Alignment Network for Multi-Modal Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 9 (2023), 5318–5329. doi:10.1109/TCSVT.2023.3247822

[46] Jennifer Santoso, Kenkichi Ishizuka, and Taiichi Hashimoto. 2024. Large Language Model-Based Emotional Speech Annotation Using Context and Acoustic Feature for Speech Emotion Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Seoul, South Korea, 11026–11030. doi:10.1109/ICASSP48485.2024.10448316

[47] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, Jichen Yang, and Jianxin Pang. 2022. Key-Sparse Transformer for Multimodal Speech Emotion Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, 6897–6901. doi:10.1109/ICASSP43922.2022.9746598

[48] Zihan Zhao, Yu Wang, and Yanfeng Wang. 2023. Knowledge-aware Bayesian Co-attention for Multimodal Emotion Recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Rhodes Island, Greece, 1–5.

[49] Yufan Yi, Yan Tian, Cong He, Yajing Fan, Xinli Hu, and Yiping Xu. 2023. DBT: multimodal emotion recognition based on dual-branch transformer. *The Journal of Supercomputing* 79, 8 (2023), 8611–8633. doi:10.1007/s11227-022-05001-5

[50] 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 (2008), 335–359. doi:10.1007/s10579-008-9076-6

[51] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 527–536. doi:10.18653/v1/P19-1050

[52] Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 5699–5710. doi:10.18653/v1/2022.acl-long.391

[53] Jintao Wen, Dazhi Jiang, Geng Tu, Cheng Liu, and Erik Cambria. 2023. Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion* 91 (2023), 123–133. doi:10.1016/j.inffus.2022.10.009

[54] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal Cross-and Self-Attention Network for Speech Emotion Recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Toronto, ON, Canada, 4275–4279. doi:10.1109/ICASSP39728.2021.9413816

[55] Lili Guo, Yikang Song, and Shifei Ding. 2024. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowledge-Based Systems* 296 (2024), 111969. doi:10.1016/j.knosys.2024.111969

[56] Mixiao Hou, Zheng Zhang, and Guangming Lu. 2022. Multi-Modal Emotion Recognition with Self-Guided Modality Calibration. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, 4688–4692. doi:10.1109/ICASSP43922.2022.9747859

[57] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. 2021. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion. *Sensors* 21, 14, Article 4913 (2021), 16 pages. doi:10.3390/s21144913

[58] Zheng Lian, Bin Liu, and Jianhua Tao. 2023. SMIN: Semi-Supervised Multi-Modal Interaction Network for Conversational Emotion Recognition. *IEEE Transactions on Affective Computing* 14, 3 (2023), 2415–2429. doi:10.1109/TAFFC.2022.3141237

[59] Changzeng Fu, Fengkui Qian, Kaifeng Su, Yikai Su, Ze Wang, Jiaqi Shi, Zhigang Liu, Chaoran Liu, and Carlos Toshinori Ishi. 2025. HiMul-LGG: A hierarchical decision fusion-based local–global graph neural network for multimodal emotion recognition in conversation. *Neural Networks* 181 (2025), 106764. doi:10.1016/j.neunet.2024.106764

[60] Zixuan Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu. 2021. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Toronto, ON, Canada, 3020–3024. doi:10.1109/ICASSP39728.2021.9414286

[61] 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)*. IEEE, Singapore, 6917–6921. doi:10.1109/ICASSP43922.2022.9747723

[62] Cheng Peng, Ke Chen, Lidan Shou, and Gang Chen. 2024. CARAT: Contrastive Feature Reconstruction and Aggregation for Multi-Modal Multi-Label Emotion Recognition. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Vancouver, BC, Canada, Article 1626, 9 pages. doi:10.1609/aaai.v38i13.29374