Modeling and Simulation 建模与仿真, 2026, 15(5), 60-71
Published Online May 2026 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2026.155071

Hans汉斯

基于聚类优化与跨模态协同的多模态情感识别

徐金麟，魏　赟

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2026年4月15日；录用日期：2026年5月12日；发布日期：2026年5月21日

摘　要

多模态情感识别旨在通过融合文本、视频和音频等多种模态信息实现更准确的情感理解。现有方法在捕获跨模态互补信息时存在冗余问题，且难以建立有效的跨模态情感关联。研究提出了一种分层协同学习框架，通过特征聚类优化与跨模态协同学习机制解决上述问题。该方法首先采用聚类算法对多模态特征进行分组优化，结合注意力权重分配机制降低冗余并突出显著特征；随后设计跨模态协同学习模块，利用交叉注意力机制实现文本引导的初步学习以及音频与视频模态的相互引导学习，从而增强多模态表示能力。在MOSI和MOSEI两个公开数据集上的实验结果表明，所提方法在多个指标上取得具有竞争力或领先的性能，验证了该方法在提升多模态情感识别性能方面的有效性。

关键词

多模态情感识别，跨模态协同学习，交叉注意力

Clustering-Based Optimization and Cross-Modal Collaborative Learning for Multimodal Sentiment Analysis

Jinlin Xu, Yun Wei

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: April 15, 2026; accepted: May 12, 2026; published: May 21, 2026

Abstract

Multimodal emotion recognition aims to achieve more accurate emotion recognition by integrating information from multiple modalities, such as text, video, and audio. However, existing methods often suffer from feature redundancy when capturing cross-modal complementary information and

文章引用: 徐金麟, 魏赟. 基于聚类优化与跨模态协同的多模态情感识别[J]. 建模与仿真, 2026, 15(5): 60-71.
DOI: 10.12677/mos.2026.155071

struggle to establish effective cross-modal emotional correlations. To address these challenges, we propose a Hierarchical Collaborative Learning framework that combines clustering optimization with cross-modal collaborative learning. Specifically, a clustering algorithm is first applied to optimize the grouping of multimodal features, together with an attention-based weight allocation mechanism, to reduce redundant information and emphasize salient features. Subsequently, a cross-modal collaborative learning module is designed to perform text-guided initial learning and mutual guidance learning between audio and visual modalities through a cross-attention mechanism, thereby enhancing multimodal representation ability. Experimental results on the public MOSI and MOSEI datasets demonstrate that the proposed method achieves competitive performance across multiple evaluation metrics, validating the effectiveness of the proposed method in improving multimodal emotion recognition performance.

Keywords

Multimodal Emotion Recognition, Cross-Modal Collaborative Learning, Cross-Attention

1. 引言

多模态情感识别是自然语言处理和人工智能领域的重要研究方向，其核心任务是从文本、视频和音频等多种模态数据中识别和理解人类的情感状态。随着多媒体数据的爆炸式增长，单一模态的情感分析方法已难以满足实际应用需求。多模态情感识别通过整合不同模态的互补信息，能够更全面地理解复杂的情感表达，在智能医疗诊断[1]、个性化推荐系统[2]、人机交互系统[3]等场景中展现出巨大的应用潜力。

多模态数据的丰富性在带来信息优势的同时，也引入了显著的特征冗余。冗余特征不仅增加计算开销，还可能放大局部噪声权重并掩盖关键细节，从而削弱特征表示的判别性并降低模型泛化能力。针对该问题，已有研究从降维与特征过滤等角度开展探索：如通过低维映射[4]或多尺度滤波[5]抑制冗余，以及通过解耦重构[6]或层次解耦[7]分离共享与模态特定信息。然而，直接降维可能忽视模态间潜在交互并造成任务相关信息丢失，而解耦类方法在高维多模态场景下计算代价较高。

多模态情感识别面临的另一个主要挑战在于如何有效融合不同模态的信息。传统的特征融合方法通常采用简单的拼接或堆叠操作，将来自不同模态的特征向量组合成复合向量或矩阵[8] [9]。这类方法虽然实现简单，但往往忽略了各模态间的内在关联和差异性。近年来，一些研究尝试将文本模态作为主导模态，用于引导视频和音频模态的特征学习[10] [11]。然而，现有文本主导策略仍面临两方面不足：一是单向引导容易弱化音频/视频模态的模态特定线索，并可能诱发模态间信息冲突；二是缺乏音频与视频之间的显式交互与贡献建模机制，导致各模态对最终情感判断的贡献难以准确区分与自适应平衡，从而限制跨模态协同融合效果。

在跨模态交互建模方面，基于注意力的方法已成为主流。Ahmad 等人[12]采用基于注意力的图卷积网络建模模态内关系，Wang 等人[13]提出了自适应模态加权 Transformer，使用多个 Softmax 函数生成模态间的注意力权重。Zadeh 等人[14]利用多个注意力模块捕获跨模态情感上下文。这些研究虽然在不同程度上提升了跨模态交互能力，但在如何建立更有效的跨模态协同学习机制方面仍有改进空间。

针对上述问题，研究提出了一种分层协同学习框架(Hierarchical Collaborative Learning, HCL)，通过特征聚类优化和跨模态协同学习机制来提升多模态情感识别的性能。该框架首先利用 Transformer 将多模态

信息统一为单一格式，随后进行两个阶段的特征增强。第一阶段采用聚类优化与权重分配模块(Clustering Optimization and Weight Allocation, COWA)，通过 K-means 聚类算法对特征进行分组，结合注意力权重分配机制对特征进行重组，从而抑制跨模态冗余信息并放大显著特征。第二阶段设计跨模态协同学习模块(Cross-modal Collaborative Learning, CCL)，首先利用文本特征引导音频和视频特征的学习，随后通过音频和视频特征的相互引导实现深度交互学习，从而增强多模态表示能力。最后，采用跨模态融合 Transformer 进行最终的特征融合和情感预测。研究的主要贡献包括：

(1) 设计了一种分层协同学习框架 HCL，通过聚类优化和跨模态协同学习机制，实现了更鲁棒的多模态情感识别。

(2) 提出了聚类优化与权重分配模块 COWA，该模块通过聚类算法对特征进行智能分组，并结合注意力机制调整特征权重，有效降低冗余并突出不同模态间的互补信息。

(3) 提出了跨模态协同学习模块 CCL，该模块先利用文本模态引导音频和视频特征学习表示，并基于交叉注意力机制动态平衡各模态对情感判断的贡献，从而缓解模态信息冲突并提升多模态表示能力。

(4) 在 MOSI 和 MOSEI 两个广泛使用的公开数据集上进行了大量实验，结果表明 HCL 达到了当前最优性能，并通过消融实验和可视化分析验证了各模块的有效性和必要性。

2. 方法

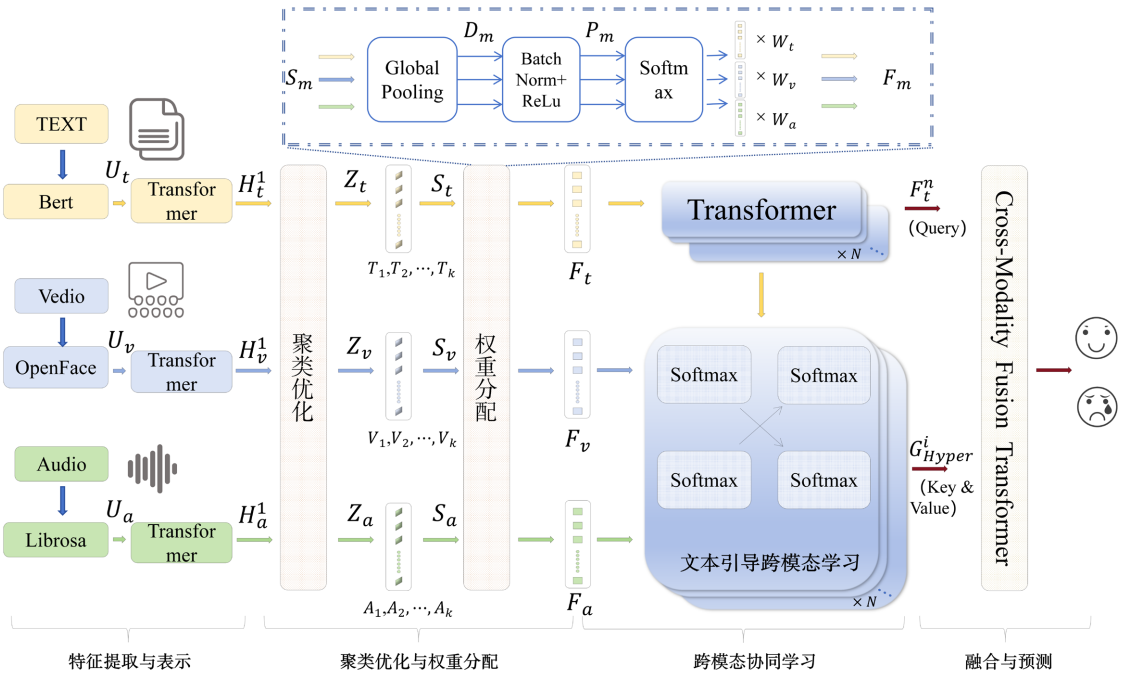

Figure 1. Overall framework
图 1. 总体框架图

2.1. 总体框架

图 1 展示了分层协同学习框架(HCL)的总体架构，该框架用于高性能多模态情感识别。HCL 由四个主要步骤组成：特征提取与嵌入、聚类优化与权重分配、跨模态协同学习、融合与预测。第一步通过 transformer 生成统一的多模态表示。第二步通过聚类优化降低冗余并增强显著的全局特征。第三步执行跨模态协同学习以增强特征表示。最后一步通过跨模态融合 transformer 进行特征融合和情感预测。

2.2. 特征提取与嵌入

利用 BERT、Librosa 和 OpenFace 提取原始特征序列。对于每个模态，提取的特征表示可表示为 $U_m\left(m\in\{Text,Audio,Video\}\right)$。引入三个独立的 Transformer，并为每个模态随机初始化一个低维 *token* H_m^0。然后，通过公式(1)获得统一的特征表示 H_m^1。

$$H_m^1=Transformer\left(concat\left(H_m^0\left|U_m\right.\right),\ \theta_{E_m^0}\right)\in\mathbb{R}^{T\times d} \tag{1}$$

Concat 表示拼接操作，初始化 token H_m^0 与相应的模态输入 U_m 拼接，θ 表示 Transformer 编码器的参数。在所提方法中，序列长度 T 和统一特征维度 d 分别设置为 8 和 128。

2.3. 聚类优化与权重分配

本研究采用聚类优化机制进行冗余消除和特征组织。首先，应用 K-means 算法对统一特征 H_m^1 进行聚类，得到 $\left(m_1,m_2,\cdots,m_K\right)$，聚类过程可表示为：

$$\begin{aligned}Z_m&=K-means\left(H_m^1,K\right)\\&=\left(m_1,m_2,\cdots,m_K\right),m\in\{T,A,V\}\end{aligned} \tag{2}$$

其中 Z_m 表示各模态的降维表示。实际应用中，将 K 设置为 8，与 H_m^1 的通道数相等。接下来，将各模态的聚类特征 Z_m 分别求和，得到新的表示 S_m。将 S_m 输入全局平均池化，生成整合各特征组信息的特征向量 D_m。

$$D_m=\frac{1}{\prod_{i=1}^{k}N_i}\sum_{(n_1,\cdots,\ n_k)}S_m\left(n_1,n_2,\cdots,n_k\right) \tag{3}$$

其中 N_i 表示特征图第 i 个空间维度的大小，k 是空间维度总数。$S_m\left(n_1,n_2,\cdots,n_k\right)$ 表示 S_m 在空间位置 $\left(n_1,n_2,\cdots,n_k\right)$ 处的值。然后，将 D_m 送入批归一化和 ReLU 激活函数，从而获得增强的联合表示 P_m。

$$P_m=ReLU\left(BatchNorm\left(\sum_{m\in\{T,\ A,\ V\}}D_m\right)\right) \tag{4}$$

随后，通过 softmax 激活函数，将 P_m 转换为注意力权重矩阵 W_m。该权重衡量特征组在整个多模态特征空间中的重要性。

$$W_m^K=\frac{\exp\left(P_m^K\right)}{\sum_{j=1}^{K}\exp\left(P_m^j\right)} \tag{5}$$

P_m^K 表示第 K 组的特征值，W_m^K 表示第 K 组的注意力权重。最后，将注意力权重矩阵 W_m 与特征组 Z_m 相乘，生成特征块 F_m。该重组过程在联合模态空间而非单一模态中将权重嵌入特征组。

2.4. 模态交互引导与融合

在多模态情感分析任务中，各模态间的信息密度呈现出明显的非对称特征。文本模态能够通过词汇语义及句法结构直接刻画情感极性，信息表达更为充分；而音频与视频模态分别依赖韵律语调和面部微表情，其情感线索较为隐式，且易受噪声干扰。为缓解该不对称性带来的表示偏差，本文设计了一种“先文本引导、后音视频协同”的跨模态协同学习策略，引入 Transformer 学习高级文本特征，采用交叉注意力机制实现跨模态引导学习，文本特征同时引导视频和音频模态的学习。以文本特征 F_t^i 作为 Query，分别以视频特征 F_v^i 和音频特征 F_a^i 作为 Key 和 Value，通过交叉注意力计算相似性矩阵 α 和 β：

$$\alpha = \text{softmax}\left(\frac{F_t^i W_Q \left(F_v^i W_K\right)^T}{\sqrt{d_k}}\right), \beta = \text{softmax}\left(\frac{F_t^i W_Q \left(F_a^i W_K\right)^T}{\sqrt{d_k}}\right) \tag{6}$$

其中 i 表示引导学习的层数，d_k 是注意力头的维度。随后，通过注意力加权获得引导特征：

$$G_v = \alpha V_v, G_a = \beta V_a \tag{7}$$

进一步，引导特征 G_a 和 G_v 作为新的 Query，分别与视频和音频特征进行交叉注意力计算，得到更新的相似性矩阵 α' 和 β'，实现音频与视频模态之间的交互学习。最后，通过残差连接融合各层引导特征，以缓解深度网络中的梯度问题：

$$G_{hyper}^n = G_{hyper}^{n-1} + \alpha' G_a + \beta' V_v, n \in \{1, 2, \cdots, N\} \tag{8}$$

最后，高级文本特征 F_t^n 作为 Query，高级引导特征 G_{hyper} 作为 Key 和 Value。多头注意力机制捕获互补信息。获得的综合特征输入分类器，生成最终输出。

3. 实验

3.1. 数据集

研究在两个广泛使用的多模态数据集上进行了实验，MOSI：该数据集提供来源于 YouTube 的情感强度数据，包含 2,199 个英语独白视频片段样本。每个样本标注有从−3 (非常消极)到+3 (非常积极)的连续尺度情感强度。训练集包含 1,284 个样本，验证集包含 229 个样本，测试集包含 686 个样本。MOSEI：该数据集包含更多英语独白视频样本，总计 22,856 个。标签范围为[−3, +3]。实验中，数据集划分为训练集、验证集和测试集，分别包含 16,326、1,871 和 4,659 个样本。

3.2. 评价指标

对于回归任务，报告平均绝对误差(MAE)和 Pearson 相关系数(Corr)作为性能指标。对于分类任务，采用 F1-Score (F1)、二分类准确率(Acc-2)和七级准确率(Acc-7)作为评估指标。

3.3. 实验设置

所有实验在配备单个 RTX 4080 GPU 的 Python 环境中使用 PyTorch 进行。批大小设置为 64，多头注意力机制使用 4 个注意力头。在每个数据集上使用 AdamW 优化器训练模型 100 个 epoch，模型学习率为 1e−4。

3.4. 对比方法

所提模型与多个多模态情感识别基线方法进行了对比，基线模型如下：

MFM [15]：基于生成－判别联合训练的多模态情感分析模型。

MISA [16]：利用独立子空间减少模态差距，增强情感和幽默检测的融合效果。

FDMER [17]：结合公共和私有编码器学习不变和特定特征，改善情感识别和幽默检测。

ALMT [4]：引入自适应超模态学习模块，通过多尺度文本引导提取有用特征并抑制无关信息。

SIMR [18]：使用先进的跨模态 transformer 学习与说话者无关的多模态表示，实现多样化交互。

MUTANet [19]：通过整合话语级特征和多模态情感损失增强表示学习，提高可区分性。

JTUM [20]：通过联合训练单模态和多模态特征，同时利用自注意力捕获时间信息，实现更好性能。

TMBL [21]：提出了具有双模态和三模态绑定机制以及细粒度卷积模块的框架。

PEST [22]：使用跨模态转换和动态传播模型将文本、视觉和声学特征映射到公共空间，实现更深层次交互。

DLF [23]：通过特征解耦与语言引导机制减少模态冗余与冲突并强化语言表征。

3.5. 性能对比

表 1 列出了所提方法与当前最优方法的对比结果。在 MOSI 数据集上，HCL 在 ACC-2、ACC-7、F1、MAE 和 Corr 方面取得最佳结果。在 MOSEI 数据集上，所提方法在 ACC-7、F1 和 MAE 方面取得最佳结果。ACC-2 和 Corr 值优于大多数基线模型，但略弱于 TMBL 和 JTUM。

将所提模型与仅基于文本引导学习的 ALMT 进行对比。在 MOSI 数据集上，所提模型在 ACC-2、F1、MAE 和 Corr 方面均优于 ALMT，也进一步证明了音频和视频模态之间协同学习的有效性。MFM、MISA 和 SIMR 方法旨在通过不同方法更好地学习多模态特征表示。与这些方法相比，所提的 HCL 表现最佳，表明该方法有效捕获并整合了跨模态的互补信息。这突显了 HCL 在学习跨模态特征表示方面的优越性。

Table 1. Comparative experiments on MOSI and MOSEI datasets
表 1. MOSI 与 MOSEI 数据集对比实验

模型	MOSI					MOSEI				
	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr
MFM	35.4	81.7	81.6	0.877	0.706	50.2	84.4	82.1	0.593	0.7
MISA	42.3	83.4	83.6	0.783	0.761	52.2	85.5	85.3	0.555	0.756
FDMER	44.1	84.6	84.7	0.724	0.788	54.1	86.1	85.5	0.536	0.773
ALMT	47.38	85.21	85.31	0.703	0.801	54.28	85.16	85.73	0.532	0.779
SIMR	-	86.10	86.10	0.706	0.795	-	82.9	82.9	0.58	0.696
MUTANet	-	84.9	84.9	0.708	0.798	-	85.9	85.2	0.537	0.764
JTUM	44.9	86.28	86.15	0.721	0.798	53.29	85.58	85.44	0.548	**0.796**
TMBL	-	83.84	84.29	0.867	0.762	-	**85.84**	85.29	0.545	0.766
PEST	-	86.1	86.1	0.723	0.796	-	85.3	85.1	0.542	0.761
DLF	47.08	85.06	85.04	0.731	0.781	53.90	85.42	85.27	0.536	0.764
所提方法	**47.45**	**86.8**	**86.68**	**0.701**	**0.802**	**54.43**	85.75	**85.86**	**0.53**	0.785

3.6. 模型复杂度分析

为评估所提方法的计算效率，表 2 报告了 HCL 与三个代表性基线模型的总可训练参数量和单样本推理浮点运算量(FLOPs)。所有模型均在相同硬件环境下测量，输入为单个样本的三模态特征序列。

Table 2. Model complexity comparison
表 2. 模型复杂度对比

模型	参数量(M)	FLOPs (G)
MFM	3.28	0.42
ALMT	6.42	1.08
TMBL	6.87	1.15
所提方法	5.51	0.98

由表 2 可知，HCL 的参数量为 5.51 M，FLOPs 为 0.98 G，均低于 ALMT 和 TMBL。与参数量最大的 TMBL (6.87 M/1.15 G)相比，HCL 的参数量降低 19.8%，FLOPs 降低 14.8%；与同样采用文本引导策略的 ALMT (6.42 M/1.08 G)相比，参数量和 FLOPs 分别降低 14.2%和 9.3%。结合表 1 的性能对比结果，HCL 在 MOSI 数据集上的 ACC-2 较 ALMT 高 1.59 个百分点，较 TMBL 高 2.96 个百分点，表明所提方法在更低的模型复杂度下取得了更优的识别性能。参数量最低的 MFM (3.28 M)虽然结构最为轻量，但其 ACC-2 仅为 81.7%，与所提方法 HCL 相差 5.1 个百分点。

3.7. COWA 模块的消融实验

为探讨聚类优化和权重分配模块在 HCL 中的作用，通过去除该模块进行了消融实验。如表 3 第一行所示，结果表明去除该模块导致 ACC-2 下降 4.2%，Corr 下降 1.3%。这些下降表明该模块通过聚类降低冗余和通过重新分配权重增强相关性方面发挥了重要作用，从而提高了多模态情感识别的鲁棒性。

同时，为了进一步评估 K-Means 聚类的效果，将 K-Means 聚类替换为按固定通道分组。如表 3 第二行所示，观察到该替换大幅降低了性能，证明 K-Means 聚类是有效的。这归因于 K-Means 聚类能够根据数据的分布特征自动识别和聚合相似特征，从而有效减少冗余信息并增强跨模态协同能力。

Table 3. Ablation experiments of COWA
表 3. COWA 的消融实验

方法	ACC-2	F1	Corr
w/o COWA	82.6	85.21	0.789
w/o K-Means	83.1	84.59	0.781
所提方法	**86.8**	**86.68**	**0.802**

3.8. CCL 模块的消融实验

为验证跨模态协同学习的有效性，在 MOSI 数据集上展示了去除各组件的消融结果。实验结果如表 4 所示。当去除视频和音频之间的协同学习模块时，观察到 ACC-2、F1 和 Corr 的下降。该结果表明协同学习模块在实现模态间互补信息交换方面起关键作用。此外，去除文本引导后，观察到性能再次下降，这支持文本引导能够有效提高音频和视频特征对情感信息的关注度。

Table 4. Ablation experiments of CCL
表 4. CCL 的消融实验

方法	ACC-2	F1	Corr
w/o 协同学习	82.5	83.35	0.785
w/o 文本引导	83.4	85.7	0.796
所提方法	**86.8**	**86.68**	**0.802**

为进一步研究不同引导模态的影响，在表 5 中对比了使用音频和视频作为引导模态的效果。当使用音频作为引导模态时，模型的 ACC-2/F1/Corr 分别下降 5.9%/5.4%/2.9%。当使用视频作为引导模态时，相应下降为 5.1%/3.6%/4.1%。对比结果清楚地表明，当使用文本模态作为引导模态时模型表现最佳。这可归因于文本模态相比音频和视频模态具有更高的信息密度和清晰度。

为探索最优层数，实验了 2 层、3 层和 4 层协同学习结构。如表 6 所示，3 层协同学习结构被证明是

最优选择，提供了最高的 ACC-2、F1 和 Corr 分数。这可能是因为 2 层结构无法捕获模态间的充分交互，而 4 层结构虽然能够捕获更复杂的特征关系，但导致了过拟合。

Table 5. Effect of different guidance modality
表 5. 不同引导模态的影响

方法	ACC-2	F1	Corr
文本引导(本文)	**86.8**	**86.68**	**0.802**
音频引导	81.6	81.96	0.778
视频引导	82.3	83.52	0.769

Table 6. Effect of different guidance layers
表 6. 不同引导层数的影响

方法	ACC-2	F1	Corr
2 层	82.3	84.7	0.756
3 层(本文)	**86.8**	**86.68**	**0.802**
4 层	84.7	84.66	0.792

3.9. 特征熵分析

图 2 显示了使用和不使用 COWA 模块的模型在 MOSI 数据集上的特征熵演化。使用 COWA 模块的模型在早期 epoch 中熵快速下降，稳定在显著较低的值，表明有效降低了冗余并增强了特征清晰度。相比之下，不使用 COWA 模块的模型稳定在较高的熵水平，反映了保留的冗余和较低的特征表示区分度。这些结果突显了 COWA 模块在精炼多模态特征、降低噪声和提高整体表示效率方面的作用。

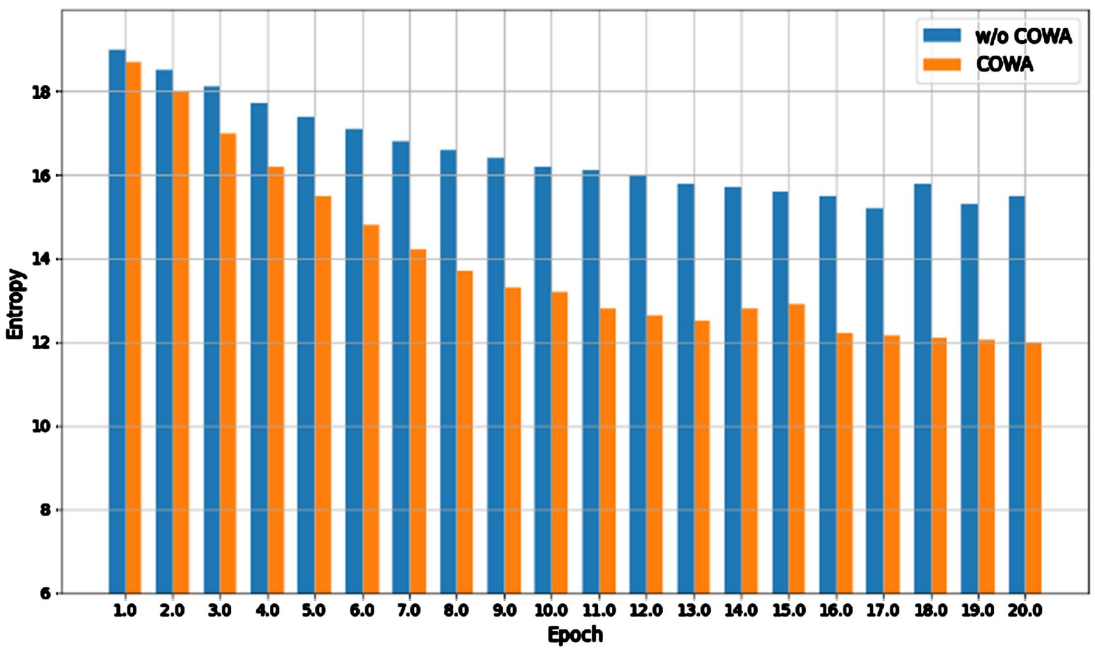

Figure 2. Feature entropy analysis
图 2. 特征熵分析

3.10. 注意力分布可视化

图 3 与图 4 分别给出了在文本引导机制作用下，音频模态与视频模态在 MOSI 数据集上最后一层交互引导学习层的注意力权重可视化。图 3 中出现较大面积的负权重区域(蓝色)，说明在文本语义的引导下，模型会在部分时间步主动抑制与文本情感线索一致性较弱的音频信息，从而避免噪声干扰并突出与文本语义相关的关键信号。相比之下，图 4 中权重分布更为集中，多个时间步呈现明显的高权重区域(红色)，表明文本提供的语义更容易激活与情感表达一致的视觉线索，使视频模态在相应时刻获得更强的关注度。

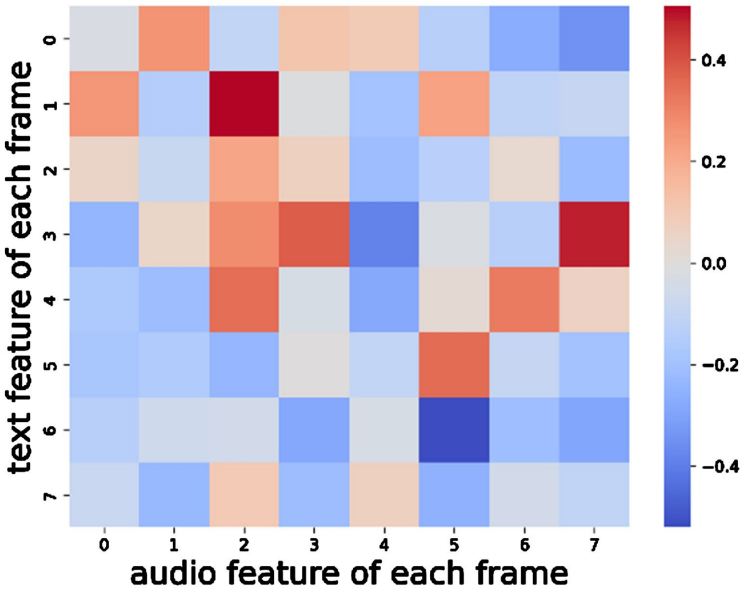

Figure 3. Visualization of attention weights from the audio modality under text guidance
图 3. 文本引导的音频注意力权重可视化

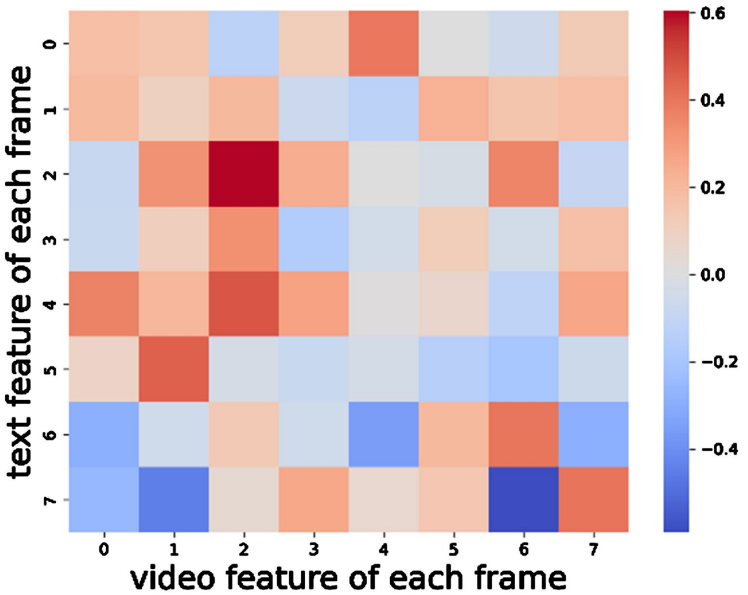

Figure 4. Visualization of attention weights from the video modality under text guidance
图 4. 文本引导的视频注意力权重可视化

3.11. 不同特征表示的可视化

图 5 给出了 CCL 模块内不同特征在三维空间的 t-SNE 可视化结果。可以看到，音频特征与视频特征在原始表示空间中存在明显间隔，表明两种模态的分布差异较大。同时，各模态内部样本点的离散分布也反映了模态内的表征多样性。相比之下，融合后的混合特征表示 G_{hyper} 呈现出更为紧凑的聚类结构，并在一定程度上连接了音频与视频两类特征的分布区域。这表明 CCL 模块能够有效缩小模态间的分布差异，为后续多模态融合提供更一致的特征空间。

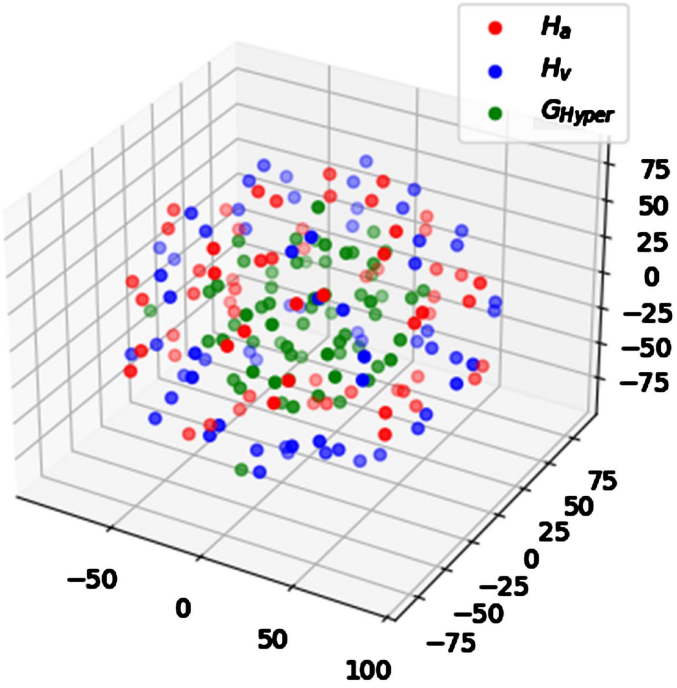

Figure 5. Visualization of different representations
图 5. 不同特征表示的可视化

4. 结束语

研究提出了一种分层协同学习框架(HCL)，通过聚类优化和跨模态协同学习机制来应对多模态情感识别中的挑战。通过整合聚类优化与权重分配(COWA)模块和跨模态协同学习(CCL)模块，该方法实现了鲁棒且可解释的情感识别。实验结果表明，HCL 在基准数据集上超过当前最优方法的性能，其有效捕获和利用跨模态互补信息的能力得到验证。此外，分层学习过程中对模态特定参数的依赖和计算开销可能阻碍模型在具有更多模态或更大特征空间的数据集上的可扩展性。为解决这些问题，未来工作将探索轻量级聚类技术以降低计算成本，并开发参数调优的自适应机制，以提高在各种模态和数据集上的泛化能力。

基金项目

国家重点研发计划资助(2021YFF0600605)。

参考文献

[1] Salloum, S., Alhumaid, K., Salloum, A. and Shaalan, K. (2024) Disease Discourse through Sentiment and Network Analysis. *Procedia Computer Science*, **244**, 23-29. https://doi.org/10.1016/j.procs.2024.10.174

[2] Cui, Y., Yu, H., Guo, X., Cao, H. and Wang, L. (2024) RAKCR: Reviews Sentiment-Aware Based Knowledge Graph Convolutional Networks for Personalized Recommendation. *Expert Systems with Applications*, **248**, Article 123403. https://doi.org/10.1016/j.eswa.2024.123403

[3] Xu, N., Mao, W. and Chen, G. (2019) Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 371-378. https://doi.org/10.1609/aaai.v33i01.3301371

[4] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y. and Yu, T. (2023) Learning Language-Guided Adaptive Hyper-Modality Representation for Multimodal Sentiment Analysis. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 756-767. https://doi.org/10.18653/v1/2023.emnlp-main.49

[5] Guo, Z., Ma, H. and Li, A. (2025) A Lightweight Finger Multimodal Recognition Model Based on Detail Optimization and Perceptual Compensation Embedding. *Computer Standards & Interfaces*, **92**, Article 103937. https://doi.org/10.1016/j.csi.2024.103937

[6] Fu, Y., Huang, B., Wen, Y. and Zhang, P. (2024) FDR-MSA: Enhancing Multimodal Sentiment Analysis through Feature Disentanglement and Reconstruction. *Knowledge-Based Systems*, **297**, Article 111965. https://doi.org/10.1016/j.knosys.2024.111965

[7] Li, Z., Huang, Z., Pan, Y., Yu, J., Liu, W., Chen, H., *et al*. (2024) Hierarchical Denoising Representation Disentanglement and Dual-Channel Cross-Modal-Context Interaction for Multimodal Sentiment Analysis. *Expert Systems with Applications*, **252**, Article 124236. https://doi.org/10.1016/j.eswa.2024.124236

[8] Park, S., Shim, H.S., Chatterjee, M., Sagae, K. and Morency, L. (2016) Multimodal Analysis and Prediction of Persuasiveness in Online Social Multimedia. *ACM Transactions on Interactive Intelligent Systems*, **6**, 1-25. https://doi.org/10.1145/2897739

[9] Xu, N. and Mao, W. (2017) MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis. *Proceedings of the* 2017 *ACM on Conference on Information and Knowledge Management*, Singapore, 6-10 November 2017, 2399-2402. https://doi.org/10.1145/3132847.3133142

[10] Liu, Z., Cai, L., Yang, W. and Liu, J. (2024) Sentiment Analysis Based on Text Information Enhancement and Multimodal Feature Fusion. *Pattern Recognition*, **156**, Article 110847. https://doi.org/10.1016/j.patcog.2024.110847

[11] Huang, C., Zhang, J., Wu, X., Wang, Y., Li, M. and Huang, X. (2023) TEFNA: Text-Centered Fusion Network with Crossmodal Attention for Multimodal Sentiment Analysis. *Knowledge-Based Systems*, **269**, Article 110502. https://doi.org/10.1016/j.knosys.2023.110502

[12] Ahmad, K.M., Liu, Q., Khalil, M.M.Y., Gan, Y., Khan, A.A., Liu, X., *et al*. (2024) Aspect-Specific Parsimonious Segmentation via Attention-Based Graph Convolutional Network for Aspect-Based Sentiment Analysis. *Knowledge-Based Systems*, **300**, Article 112169. https://doi.org/10.1016/j.knosys.2024.112169

[13] Wang, Y., He, J., Wang, D., Wang, Q., Wan, B. and Luo, X. (2024) Multimodal Transformer with Adaptive Modality Weighting for Multimodal Sentiment Analysis. *Neurocomputing*, **572**, Article 127181. https://doi.org/10.1016/j.neucom.2023.127181

[14] Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E. and Morency, L. (2018) Multi-Attention Recurrent Network for Human Communication Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 5642-5649. https://doi.org/10.1609/aaai.v32i1.12024

[15] Tsai, Y.H.H., Liang, P.P., Zadeh, A., *et al*. (2019) Learning Factorized Multimodal Representations. arXiv: 1806.06176.

[16] Hazarika, D., Zimmermann, R. and Poria, S. (2020) MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. *Proceedings of the* 28*th ACM International Conference on Multimedia*, Seattle, 12-16 October 2020, 1122-1131. https://doi.org/10.1145/3394171.3413678

[17] Yang, D., Huang, S., Kuang, H., Du, Y. and Zhang, L. (2022) Disentangled Representation Learning for Multimodal Emotion Recognition. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 1642-1651. https://doi.org/10.1145/3503161.3547754

[18] Wang, J., Wang, S., Lin, M., Xu, Z. and Guo, W. (2023) Learning Speaker-Independent Multimodal Representation for Sentiment Analysis. *Information Sciences*, **628**, 208-225. https://doi.org/10.1016/j.ins.2023.01.116

[19] Tang, Z., Xiao, Q., Zhou, X., Li, Y., Chen, C. and Li, K. (2023) Learning Discriminative Multi-Relation Representations for Multimodal Sentiment Analysis. *Information Sciences*, **641**, Article 119125. https://doi.org/10.1016/j.ins.2023.119125

[20] Li, M., Zhu, Z., Li, K., Zhou, L., Zhao, Z. and Pei, H. (2024) Joint Training Strategy of Unimodal and Multimodal for Multimodal Sentiment Analysis. *Image and Vision Computing*, **149**, Article 105172. https://doi.org/10.1016/j.imavis.2024.105172

[21] Huang, J., Zhou, J., Tang, Z., Lin, J. and Chen, C.Y. (2024) TMBL: Transformer-Based Multimodal Binding Learning Model for Multimodal Sentiment Analysis. *Knowledge-Based Systems*, **285**, Article 111346.

https://doi.org/10.1016/j.knosys.2023.111346

[22] Gan, C., Tang, Y., Fu, X., Zhu, Q., Jain, D.K. and García, S. (2024) Video Multimodal Sentiment Analysis Using Cross-Modal Feature Translation and Dynamical Propagation. *Knowledge-Based Systems*, **299**, Article 111982. https://doi.org/10.1016/j.knosys.2024.111982

[23] Wang, P., Zhou, Q., Wu, Y., Chen, T. and Hu, J. (2025) DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 21180-21188. https://doi.org/10.1609/aaai.v39i20.35416