中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-17

论文引用格式： Yang Jiaojiao, Sun Qiang, Zhu Xiatian. Interpretable invertible disentanglement and adaptive fusion for multimodal emotion recognition[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260004. (杨皎皎, 孙强, 朱霞天. 多模态情感识别的可解释可逆解耦与自适应融合[J/OL]. 中国图象图形学报,XXXX:1-17. DOI: 10.11834/jig.260004.)[DOI:10.11834/jig.260004]

多模态情感识别的可解释可逆解耦与自适应融合

杨皎皎[1],孙强[1*],朱霞天[2]

1. 西安理工大学自动化与信息工程学院,西安 710048; 2. 英国萨里大学以人为本人工智能研究所&视觉、语音和信号处理中心,吉尔福德 GU2 7XH,英国

摘 要： **目的** 多模态情感识别旨在融合文本、视觉与语音等模态信息来识别对象情感状态。然而,模态间固有的异质性问题使得情感语义与模态特有噪声纠缠,限制模型的可解释性。而且,现有融合策略难以充分捕捉模态间情感语义特征的细粒度关联,导致融合表示判别性不足。为此,提出一种多模态情感识别的可解释可逆解耦与自适应融合方法。**方法** 设计IAMD(invertible attention mask-based disentanglement)模块,构建各模态特征表示与情感语义因子之间的可逆映射,并结合注意力掩码将隐式特征解耦为跨模态一致性的共享特征与保留各模态独有属性的特有特征。构建MIC(mutual information constraint)机制,使用互信息约束共享特征、特有特征和情感标签间的依赖关系,增强语义一致性建模同时减少模态噪声冗余。提出SGAFF(semantic-guided adaptive feature fusion)模块,利用共享特征的上下文信息对特有特征进行语义引导,实现共享与特有引导双分支的自适应融合。**结果** 在CMU-MOSI数据集上,相较于DLF(disentangled language focused)模型,本文模型在平均绝对误差(mean absolute error, MAE)和七分类准确率(7-class accuracy, Acc-7)指标上分别提升了2.4%和2.9%;在CMU-MOSEI数据集上,相较于TMBL(Transformer-based multimodal binding learning)模型,在MAE和皮尔逊相关系数(Pearson correlation coefficient, Corr)上分别提升了2.6%和1.7%;在UR-FUNNY数据集上,相较于MISA(modality-invariant and specific analysis)模型,在F_1值(F_1-score, F_1)上提升了5.6%。**结论** 所提方法实现了情感语义信息与模态特有噪声的可解释解耦,并促进跨模态情感语义特征的细粒度交互。该方法适用于模态异质性较复杂的多模态场景及对情感识别指标有较高要求的任务。本文代码已开源至https://doi.org/10.57760/sciencedb.j00240.00138。

关键词： 多模态情感识别;可解释;可逆解耦;注意力掩码;自适应融合

Interpretable invertible disentanglement and adaptive fusion for multimodal emotion recognition

Yang Jiaojiao[1], Sun Qiang[1*], Zhu Xiatian[2]

1. School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China; 2. The Surrey Institute for People-Centred Artificial Intelligence and the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey, Guildford GU2 7XH, UK

Abstract: Objective Multimodal emotion recognition aims to understand the emotional states of specific subjects by fusing data from multiple modalities such as text, vision, and audio. However, the inherent heterogeneity existing in the represen-

收稿日期: 2026-01-04; **修回日期:** 2026-04-02

***通信作者:** 孙强 qsun@xaut.edu.cn

基金项目: 陕西省重点研发计划项目(2025CY-JJQ-186);西安理工大学国际科技合作促进项目(2024GHCJ014)

Supported by: Key Research and Development Program of Shaanxi Province(2025CY-JJQ-186); International Science and

tation forms and distribution laws of different modality data leads to emotional semantics in the latent feature space often being entangled and coupled with modality-specific non-emotional noise. This phenomenon of feature entanglement not only hinders the model′s effective learning of key emotional features but also limits the interpretability ability of the model′s decision-making process. Furthermore, feature fusion strategies adopt simple concatenation operations or coarse-grained attention mechanisms, making it difficult to effectively capture fine-grained emotional semantic interaction cues between modalities in complex cross-modal contexts, ultimately resulting in the fused emotional representation lacking sufficient discriminability. To this end, an interpretable invertible disentanglement and adaptive fusion method for multimodal emotion recognition is proposed. **Method** First, in order to reduce the loss of semantic information during the feature learning phase and achieve structured feature disentanglement, an invertible attention mask-based disentanglement (IAMD) module is designed. Based on invertible neural networks (INN), a bidirectional invertible mapping structure is constructed between the latent representations of each modality′s features and emotional semantic factors, and an attention mask mechanism is combined to disentangle latent features in the channel dimension into two parts: one part capturing shared features with semantic consistency across modalities, and the other retaining specific features containing unique attributes of each modality. Secondly, to further enhance the disentanglement effect from an information-theoretic level, a mutual information constraint (MIC) mechanism is constructed. The semantic consistency of emotional features in the shared subspace is enhanced by calculating and maximizing the mutual information between shared features as well as between shared features and emotion labels. Meanwhile, by minimizing the mutual information between specific features and emotion labels conditioned on shared features, the model is constrained to strip modality-specific attributes not directly related to the emotion task into the specific feature subspace, thereby reducing the interference of modality-redundant noise on emotional semantics. Finally, addressing the issue of insufficient interaction during the feature fusion phase, a semantic-guided adaptive feature fusion (SGAFF) module is designed. The cross-modal consistent emotional semantic information captured in the shared subspace is utilized by this module as contextual cues to perform fine-grained semantic correction and guidance on modality-specific features through residual connections, and a dual-branch prediction structure is constructed, in which a gating mechanism is utilized to adaptively assign weights to the shared branch and the specific-guided branch, thereby enhancing the discriminability of the fused representation. **Result** Extensive comparative experiments and ablation studies were conducted on the CMU-MOSI, CMU-MOSEI, and UR-FUNNY datasets. Specifically, on CMU-MOSI, the model improved mean absolute error (MAE) and 7-class accuracy (Acc-7) by 2.4% and 2.9%, respectively, compared to the disentangled language focused (DLF) model. On CMU-MOSEI, it yielded improvements of 2.6% and 1.7% in MAE and Pearson correlation coefficient (Corr), respectively, compared to the Transformer-based multimodal binding learning (TMBL) model. Furthermore, on UR-FUNNY, the model improved the F_1-score (F_1) by 5.6% compared to the modality-invariant and specific analysis (MISA) model. In addition, detailed ablation experiments verified the necessity of the IAMD, MIC, and SGAFF modules for improving model performance. Feature visualization analysis based on t-distributed stochastic neighbor embedding confirmed that the model realized the effective separation of emotional semantics and modality noise in the latent space. Furthermore, fusion weight visualization confirmed that the model adaptively assigned higher contributions to the specific-guided branch, confirming the role of fine-grained complementary cues in final emotion judgment. **Conclusion** In summary, the proposed method achieves interpretable disentanglement of emotional semantic information from modality-specific noise based on INN and mutual information constraints, at the same time, through the semantic-guided adaptive fusion strategy, it realizes deep and fine-grained interactions between cross-modal emotional semantic features, thereby improving the accuracy and robustness of multimodal emotion recognition tasks in complex scenarios. Although the proposed method achieves significant progress in model interpretability, the introduction of invertible transformations and multiple mutual information constraints increases the model′s computational complexity. This method is applicable to multimodal scenarios with complex modal heterogeneity, as well as tasks that have strict requirements for quantitative emotion recognition metrics. Future work will focus on lightweight disentanglement for emotion recognition tasks, to further improve the inference efficiency and generalization ability of the model in scenarios with limited computational resources. The source code has been archived athttps://doi.org/10.57760/sciencedb.j00240.00138.

Key words: multimodal emotion recognition; interpretable; invertible disentanglement; attention mask; adaptive fusion

0 引 言

在大规模情感数据库日益普及的背景下，结合视觉、语音和文本等多模态数据以全面理解人类情感，已成为人工智能领域的关键课题之一（Poria等，2017）。作为实现这一目标的关键任务，多模态情感识别近年来受到广泛关注，该方向正持续向更深入的情感理解、更有效的跨模态建模场景发展（陶建华等，2024）。与此同时，随着数字人和智能机器人技术的不断发展，多模态情感理解在医疗陪护、自然人机交互等场景中的应用价值进一步凸显（赵思成等，2025）。

尽管多模态情感识别在近年来取得了显著进展，然而，模态间固有的异质性问题使得现有深度学习方法难以实现有效的跨模态特征学习与融合，成为制约性能提升的核心瓶颈（Al-Tameemi等，2023）。一方面，在特征学习阶段，由于各模态数据在表示形式、分布规律上存在显著差异，导致特征表示中的情感语义信息往往与模态特有的非情感噪声（如背景杂音、无关视觉纹理等）发生纠缠，使得模型难以学习到更为有效的情感特征，从而限制模型的可解释性；另一方面，在特征融合阶段，现有方法通常对多模态特征以简单拼接形式或基于注意力机制进行融合，往往难以在复杂的上下文中捕捉到细粒度的语义信息，从而导致融合后的情感表示判别性不足（王善敏等，2025）。因此，如何设计有效的特征学习机制以从异质数据中实现情感语义信息与模态特有噪声的分离，同时构建更具判别性的融合策略以充分实现模态间的细粒度情感语义交互，成为当前多模态情感识别亟需突破的关键。

在特征学习层面，为学习到更为有效的情感特征，早期研究（Poria等，2016）直接从原始的或浅层的特征中学习高级情感表示，对各模态进行独立编码，但这种方式忽略了模态间潜在的语义关联与互补性。后续工作则通过引入注意力机制（衡红军等，2022）与Transformer结构（孙强等，2024）来捕捉模态内时序语义的长距离依赖关系。尽管这些方法在一定程度上能够隐式地学习情感特征，缓解模态间的异质性问题，但模型的黑箱特性往往导致特征潜在空间语义模糊，限制了模型的可信度。为此，部分学者引入了解耦表示学习（Wang等，2024），其核心思想是在表示层识别并分离原始数据中的潜在语义，将变化因子解耦为独立的潜在变量，从而获得更具解释性的特征。然而，现有解耦方法（Hazarika等，2020；Li等，2023）则多聚焦于特征子空间投影，导致情感语义的边界不够清晰。近期有些工作（Zeng等，2024a；Wang等，2025）在解耦框架的设计与优化层面进行了系列改进。此外，在情感生成任务中，有研究通过在向量量化变分自编码器框架下引入分层解耦以及内外部条件协同约束，从而提升生成的3D人脸情感表达的稳定性（陈胜等，2026）。然而，这些方法在解耦过程缺乏显式的结构化约束，导致关键情感语义信息丢失，使得决策过程难以追溯。

在特征融合层面，有学者通过尝试使用基于注意力的分组交互机制（Huang等，2024）来增强模态间的信息融合，但这种方法关注于模态级别的全局交互，在建模模态内部语义特征之间的动态关联方面存在一定困难。另一方面，部分工作从模型结构本身出发，将多模态特征对齐至统一的预训练嵌入空间（Song等，2025）或者设计层次化策略（王健等，2025）来融合来自不同模态的互补信息。同时，也有（罗渊贻等，2024）尝试采用自适应权重机制的方式来调整各模态对最终情感融合决策的影响。总体来看，上述方法仍倾向于将各模态特征作为一个整体进行粗粒度的融合，未能充分考虑模态内与模态间细粒度情感特征间的语义关联，从而限制了模型对细微情感变化的表达能力。

针对上述两个方面的挑战，本文提出一种面向多模态情感识别的可解释可逆解耦与自适应融合方法。具体而言，首先，设计可逆注意力掩码解耦模块（invertible attention mask-based disentanglement，IAMD），将各模态通过Transformer编码得到的特征作为输入，通过可逆神经网络（invertible neural networks，INN）的双向映射特性，结合注意力掩码机制将隐式特征显式解耦为具有跨模态一致性的共享特征与保留各模态独有属性的特有特征，减少了情感语义信息的丢失。其次，构建了互信息约束机制（mutual information constraint，MIC），通过最大化共享特征间以及共享特征与情感标签之间的互信息，增强共享特征子空间中的情感语义一致性，同时在给定共享特征的条件下，最小化特有特征与情感标签之间的互信息，减少特有特征子空间中的噪声冗余。而且，还设计了一种语义引导的自适应特征融

合模块(semantic-guided adaptive feature fusion, SGAFF),使用共享子空间所捕获的细粒度情感语义信息作为引导,通过残差连接对模态特有特征进行语义修正,并利用双分支预测结构,通过门控机制自适应地为共享分支与特有引导分支赋予权重来提高融合表示的判别性。

本文的主要贡献如下:1)设计了一个可逆注意力掩码解耦模块,将隐式特征显式解耦为共享特征与特有特征,减少了情感语义信息的丢失,进而增强决策过程的透明度;2)构建了一种互信息约束机制,通过互信息约束共享特征、特有特征和情感标签之间的依赖关系,增强情感语义一致性建模同时减少模态噪声冗余,提升模型可解释性;3)设计了一个语义引导的自适应特征融合模块,实现了跨模态情感语义特征间的细粒度交互,提升了多模态特征融合表示的判别性;4)CMU-MOSI、CMU-MOSEI和UR-FUNNY三个数据集上的实验结果表明,本文模型与多种经典方法相比具有显著优势,进一步证实了所提方法的可行性和有效性。

1 相关工作

1.1 基于深度学习的多模态情感识别

1.1.1 解耦表示学习

解耦表示学习旨在将多模态数据潜在空间的隐式表示分解为相互独立且具有明确意义的语义因子,以获得更具泛化性和解释性的表示。早期代表性方法主要聚焦于特征子空间投影,利用前馈神经网络将特征投影至模态不变子空间和模态特有子空间(Hazarika等,2020;Yang等,2022a)。但此类方法利用简单投影难以在复杂的模态异质性干扰下实现有效的特征解耦。在此基础上,有学者(Yang等,2022b)采用自注意力模块以增强模态特有特征,并通过分层跨模态注意力模块来捕获跨模态共性。然而,这些改进在异步序列处理上表现出色,但缺乏对特有特征中非情感噪声的约束。近期,一些研究进一步优化解耦框架的设计,如Zeng(2024a)等人提出基于松弛重构的解耦翻译网络以统一特征分布,Wang等人(2025)构建以语言为中心的解耦框架并引入几何度量优化模态共享与特有信息的分离,贾熹滨(2025)等人则通过元优化策略解耦领域不变与领域特定特征以增强跨域情感分类。尽管上述方法在性能上有所提升,但在将高维数据映射到特征子空间的过程中,会导致原始情感信息的丢失。为此,本文所提的IAMD模块,基于可逆神经网络构建隐式表示与情感语义因子之间的可逆映射,减少了解耦过程中情感语义信息的丢失,使得决策依据在特征空间中可追溯。

1.1.2 特征融合

多模态特征融合经历了从浅层融合到深度交互的过程。早期研究如张量融合网络(Zadeh等,2017)依赖外积操作捕捉模态间交互,虽模型直观,但难以建模模态间复杂的动态依赖。随着注意力机制的发展,研究者开始利用其捕捉模态间的深层交互关系,例如,通过引入定向成对的跨模态注意力机制,使一个模态的特征表示能够在时间维度上对另一模态的特征进行选择性关注(Tsai等,2019),实现无需显式对齐的特征融合。然而,该融合过程主要依赖于局部的注意力对齐,导致模态间特征交互不足。近年来,部分研究尝试从模型结构本身推进融合效果,例如,Lin等人(2023)基于多层感知机(multilayer perceptron,MLP)设计了极坐标与强度向量混合的交互策略实现跨模态交互。Song等人(2025)利用预训练的对比语言-图像预训练(contrastive language-image pre-training,CLIP)语义空间,实现对文本、语音和视觉模态的统一嵌入与融合。然而,此类方法倾向于将各模态特征视为整体进行粗粒度的融合,未能充分考虑模态内部与模态间细粒度情感特征的上下文依赖,从而限制了模型在复杂场景下的表达力。为此,所提的SGAFF模块,利用共享特征的情感语义作为上下文对模态特有特征进行引导,并结合门控机制实现双分支的自适应融合,从而提升融合表示的判别性。

1.2 可解释的深度学习方法

根据解释机制的结构性质,现有可解释深度学习方法可分为事后解释与事前结构化解释。事后解释方法通过梯度、扰动或可视化等手段对决策进行归因分析。例如,在单模态场景中的注意力归因(Li等,2024)、特征权重排序(Ma等,2022)以及多模态场景中应用梯度沙普利加性解释(Shapley additive explanations,SHAP)值量化特征贡献度的方法(Khalane等,2025)。然而,此类方法仅提供局部、表层的归因,无法触及模型内部的决策逻辑。相比之下,事前结构化解释方法则通过显式的结构设计使模型在

推理过程中实现透明决策。例如，Zadeh 等人(2018)通过图融合结构显式建模模态间的边权重关系，Yang等人(2025)则基于图信息瓶颈设计邻接可解释图神经网络，利用互信息约束图结构实现先验可解释。尽管如此，此类方法对跨模态情感语义一致性建模与潜在变量交互机制方面仍显不足。为此，本文所提解耦方法能够在特征空间中通过正向变换获得解耦后的共享和特有语义因子，也能够通过逆向变换重构回正向输入表示，并

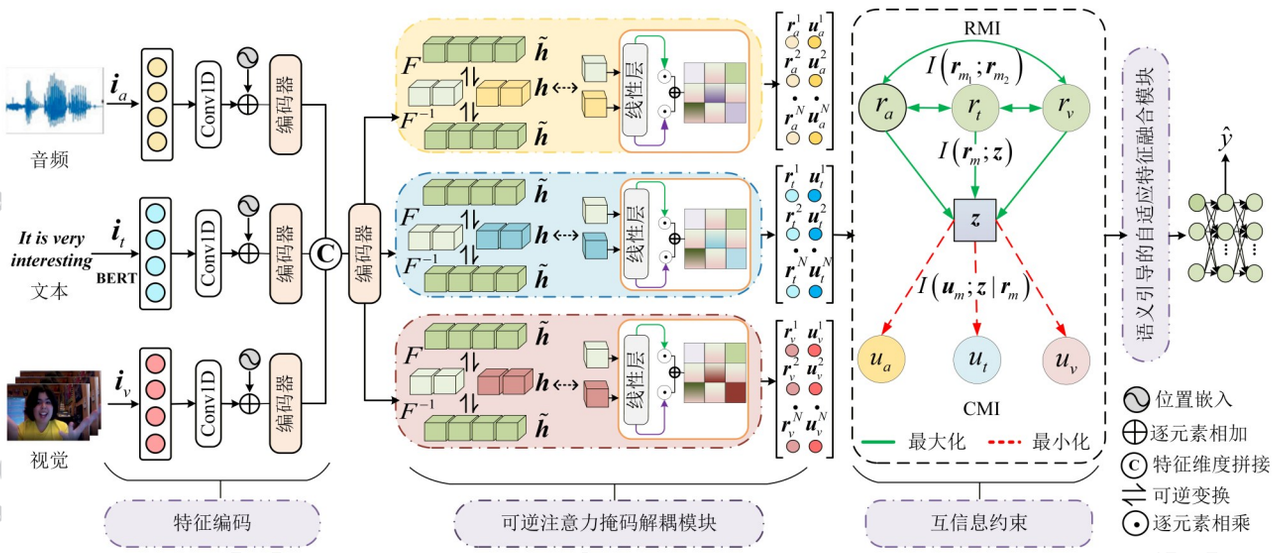


图1　模型整体结构

Fig. 1　Overall architecture of the model

使用互信息约束机制增强情感语义一致性建模，减少模态噪声冗余，进而增强决策过程的透明度。

2　本文方法

本节详细的介绍了提出的方法，图1为模型整体结构示意图，包括三种模态：音频(a)、视觉(v)和文本(t)。首先，使用Transformer编码器对文本、视觉与音频模态的上下文特征统一编码。其次，设计IAMD模块，针对每一个模态，建立隐式表示与情感语义因子之间的双向可逆映射关系，结合注意力掩码机制显式解耦出跨模态共享特征与模态特有特征。接着，构建MIC机制，利用相关互信息(relevant mutual information，RMI)与条件互信息(conditional mutual information，CMI)约束，从信息论层面增强共享特征语义一致性，减少模态冗余噪声。最后，设计SGAFF模块，利用共享特征的语义信息作为上下文对特有特征进行细粒度引导，并基于双分支结构实现自适应门控融合，最终输出情感预测结果。

2.1　多模态特征编码

本文为音频、文本与视觉模态设计了统一的特征编码流程，各模态输入序列通过一维卷积投影至统一维度，并使用位置编码输入至多层Transformer编码器(Tsai等，2019)，得到时序上下文特征$\boldsymbol{x}_m$。其中，文本模态额外使用BERT获取词向量表示，视觉与音频模态则直接输入帧级特征，表示为

$$\boldsymbol{x}_m = \mathrm{Transformer}\left(\mathrm{PE}\left(\mathrm{Conv1D}\left(\boldsymbol{i}_m\right)\right)\right) \tag{1}$$

式中，Conv1D(·)为卷积投影，PE(·)为位置编码，$\boldsymbol{i}_{m\in\{a,t,v\}}$代表各模态初始输入的特征。随后，将各模态的表示$\boldsymbol{x}_m$再次经过共享的Transformer编码器进行联合建模，得到各模态特征$\boldsymbol{X}_m$。

2.2　可逆注意力掩码解耦

受Esser等人(2020)提出的将隐式表示与人类可理解的语义概念之间建立一种可逆的翻译来实现可解释性这一思想启发，本文将可逆建模思想引入多模态情感识别任务，在此基础上设计了IAMD模块，以任意单模态为例，其结构如图2所示，该模块核心在于利用可逆神经网络构建原始隐式空间与解耦语义空间之间的双向映射，来确保解耦后的语义

特征保留了输入数据的原始信息，从而减少在特征解耦过程中丢失关键的情感语义线索。具体而言，正向变换函数通过可逆神经网络并结合注意力掩码机制，显式的将隐式表示 $\tilde{\boldsymbol{h}}$ 解耦为两部分：包含跨模态一致性情感信息的共享特征 $\boldsymbol{r}$ 与保留各模态独有属性的特有特征 $\boldsymbol{u}$。逆向变换函数则将解耦后的可解释语义特征映射回解耦前输入的特征空间。正向解耦与逆向重构过程定义为

$$\tilde{\boldsymbol{h}}=[\tilde{\boldsymbol{u}}_m;\tilde{\boldsymbol{r}}_m]\underset{F^{-1}}{\overset{F}{\rightleftharpoons}}\boldsymbol{h}=[\boldsymbol{u}_m;\boldsymbol{r}_m] \tag{2}$$

式中，F 为正向变换函数和注意力掩码交互，其作用是将隐式特征 $\tilde{\boldsymbol{h}}$ 映射为可解释的潜在语义特征 $\boldsymbol{h}$，[·] 表示拼接操作，F^{-1} 为逆向变换函数。

为了增强特征间的信息交互能力，本文采用基于 Householder 反射向量（Tomczak 等，2016）的正交混合与可逆变换来构建可逆块。首先，对任意单模态特征 $\boldsymbol{X}$ 分别由两层 MLP 网络得到与模态相关

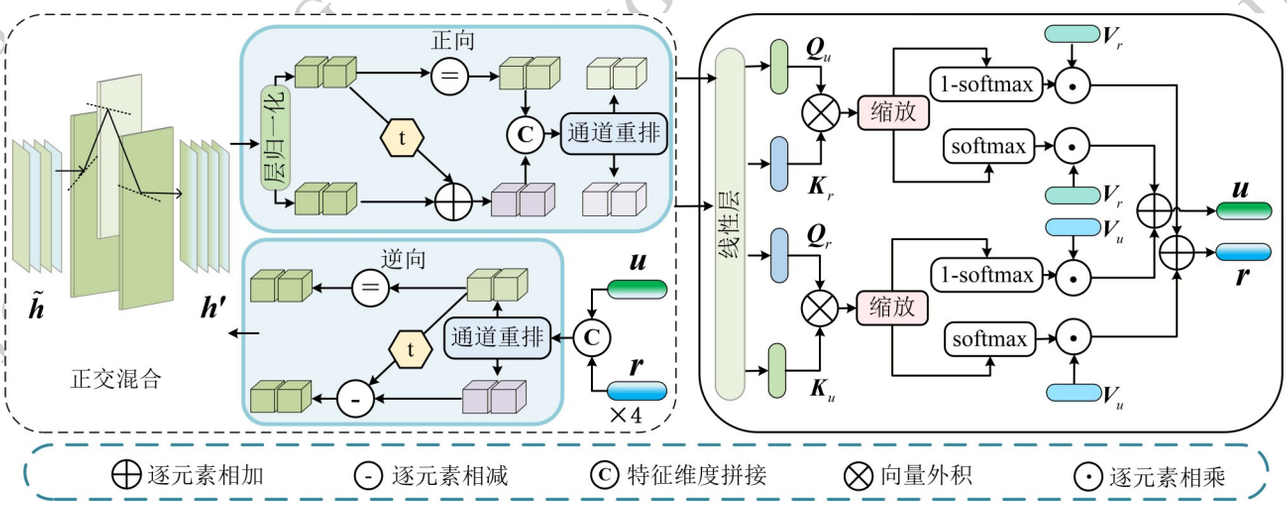


图2　可逆注意力掩码解耦模块

Fig. 2　Invertible attention mask-based disentanglement module

的特征 $\tilde{\boldsymbol{u}}$ 以及与任务相关的特征 $\tilde{\boldsymbol{r}}$，将两者拼接为隐式特征 $\tilde{\boldsymbol{h}}$。对于各模态输入特征 $\tilde{\boldsymbol{h}}$，引入一组可学习的反射向量 $\{\boldsymbol{v}_j\}_{j=1}^{J}$ 构造正交矩阵 $\boldsymbol{Q}$，实现通道维度上的正交混合，得到特征 $\boldsymbol{h}'$，该过程表示为

$$\boldsymbol{h}'=\boldsymbol{Q}\tilde{\boldsymbol{h}},\boldsymbol{Q}=\prod_{j=1}^{J}\left(\boldsymbol{I}-2\frac{\boldsymbol{v}_j\boldsymbol{v}_j^{\mathrm{T}}}{\|\boldsymbol{v}_j\|^2}\right) \tag{3}$$

式中，$\boldsymbol{v}_j\in\mathbf{R}$ 为第 j 个反射向量，J 表示反射次数，

本文中取 J 为 3，$\boldsymbol{I}$ 为单位矩阵，$\boldsymbol{v}_j^T$ 为 $\boldsymbol{v}_j$ 的转置向量。

其次，正交混合对输出的 $\boldsymbol{h}'$ 经层归一化后沿通道维度分割为两部分 $\boldsymbol{u}'$ 和 $\boldsymbol{r}'$，正向变换保持一部分特征 $\boldsymbol{u}'$ 不变，并将经过变换函数 $t(\cdot)$ 映射后的 $\boldsymbol{u}'$ 与另一部分特征 $\boldsymbol{r}'$ 相加，逆向变换则执行相反的减法操作，以恢复原始特征，构建的变换表示为

$$\bar{\boldsymbol{u}}=\boldsymbol{u}',\bar{\boldsymbol{r}}=\boldsymbol{r}'+t(\boldsymbol{u}') \tag{4}$$

$$\boldsymbol{u}'=\bar{\boldsymbol{u}},\boldsymbol{r}'=\bar{\boldsymbol{r}}-t(\bar{\boldsymbol{u}}) \tag{5}$$

式中，$\bar{\boldsymbol{u}}$ 和 $\bar{\boldsymbol{r}}$ 分别为正向变换得到的两部分特征，式（4）为正向变换，式（5）为逆向变换，变换函数 $t(\cdot)$ 由轻量级前馈网络构建，其网络结构由两个线性映射层与中间的高斯误差线性单元（gaussian error linear unit，GELU）激活函数构成。

为了增强跨通道的信息交互，引入通道重排操作得到 $\hat{\boldsymbol{h}}=[\hat{\boldsymbol{u}};\hat{\boldsymbol{r}}]\in\mathbf{R}^{B\times D}$，其过程表示为

$$\hat{\boldsymbol{h}}=\boldsymbol{P}_i\boldsymbol{c},\boldsymbol{P}_i\in\{0,1\}^{B\times D},\boldsymbol{P}_i\boldsymbol{P}_i^{\mathrm{T}}=\boldsymbol{I} \tag{6}$$

式中，$\boldsymbol{c}$ 为正向变换输出特征的拼接向量，表示为 $\boldsymbol{c}=[\bar{\boldsymbol{u}};\bar{\boldsymbol{r}}]$，$B$ 和 D 分别表示批量大小和特征维度，$\boldsymbol{P}_i$ 为固定但随机初始化的通道置换矩阵，在逆向过程中，模型使用其逆置换矩阵 $\boldsymbol{P}_i^{\mathrm{T}}$。

同时，在可逆变换的潜在空间中，结合注意力掩码机制以调整共享表示与特有表示之间的信息交互方式。该机制对情感同一模态得到 $[\hat{\boldsymbol{u}};\hat{\boldsymbol{r}}]$ 特征分别构造线性投影得到查询、键、值，接着计算 $\hat{\boldsymbol{u}}$ 对 $\hat{\boldsymbol{r}}$ 的解释权重 $\boldsymbol{\alpha}$ 以及 $\hat{\boldsymbol{r}}$ 对 $\hat{\boldsymbol{u}}$ 的解释权重 $\boldsymbol{\beta}$，并在每个通道维度上进行信息融合以实现特征间的互补更新，从而得到跨模态一致性情感信息的共享特征 $\boldsymbol{r}$ 和各模态

独有属性的特有特征 $\boldsymbol{u}$，该过程表示为

$$\begin{aligned}\boldsymbol{Q}_u &= \hat{\boldsymbol{u}}\boldsymbol{W}_{u,q}, \boldsymbol{K}_u = \hat{\boldsymbol{u}}\boldsymbol{W}_{u,k}, \boldsymbol{V}_u = \hat{\boldsymbol{u}}\boldsymbol{W}_{u,v}\\ \boldsymbol{Q}_r &= \hat{\boldsymbol{r}}\boldsymbol{W}_{r,q}, \boldsymbol{K}_r = \hat{\boldsymbol{r}}\boldsymbol{W}_{r,k}, \boldsymbol{V}_r = \hat{\boldsymbol{r}}\boldsymbol{W}_{r,v}\end{aligned} \tag{7}$$

$$\boldsymbol{\alpha} = \mathrm{diag}\left(\mathrm{softmax}\left(\frac{\boldsymbol{Q}_u\boldsymbol{K}_r^{\mathrm{T}}}{\sqrt{d_K}}\right)\right)$$

$$\boldsymbol{\beta} = \mathrm{diag}\left(\mathrm{softmax}\left(\frac{\boldsymbol{Q}_r\boldsymbol{K}_u^{\mathrm{T}}}{\sqrt{d_K}}\right)\right) \tag{8}$$

$$\boldsymbol{r} = (1-\boldsymbol{\alpha})\odot\boldsymbol{V}_r + \boldsymbol{\beta}\odot\boldsymbol{V}_u, \boldsymbol{u} = (1-\boldsymbol{\beta})\odot\boldsymbol{V}_u + \boldsymbol{\alpha}\odot\boldsymbol{V}_r \tag{9}$$

式中，$\boldsymbol{Q}_u$, $\boldsymbol{K}_u$, $\boldsymbol{V}_u$ 分别表示由 $\hat{\boldsymbol{u}}$ 映射得到的查询、键和值向量，$\boldsymbol{Q}_r$, $\boldsymbol{K}_r$, $\boldsymbol{V}_r$ 分别表示由 $\hat{\boldsymbol{r}}$ 映射得到的查询、键和值向量，$\boldsymbol{W}_{u,q}$, $\boldsymbol{W}_{u,k}$, $\boldsymbol{W}_{u,v}$ 为 $\hat{\boldsymbol{u}}$ 的查询、键、值权重矩阵，$\boldsymbol{W}_{r,q}$, $\boldsymbol{W}_{r,k}$, $\boldsymbol{W}_{r,v}$ 为 $\hat{\boldsymbol{r}}$ 的查询、键、值权重矩阵，$\mathrm{diag}(\cdot)$ 表示取矩阵对角元素的操作，$\odot$ 表示逐元素相乘操作，d_K 为键向量维度。

此外，为实现多模态特征解耦的有效性，在解耦过程中设计四个损失函数约束。首先，为避免在语义解耦过程中丢失关键模态特征，本文在输入空间引入重构约束 L_{RE}，表示为

$$L_{\mathrm{RE}} = \sum_{m\in\{a,t,v\}}\left\|\tilde{\boldsymbol{h}}_m - \boldsymbol{x}_m\right\|_2^2 \tag{10}$$

式中，$\|\cdot\|_2^2$ 表示为 L_2 范数的平方。

其次，设计可逆约束 L_{IN} 以确保潜在空间特征的可逆性，该约束通过最小化解耦过程中逆向重构特征与正向输入特征之间的分布差异，强化隐式表示与情感语义因子间的双向映射关系，并减少解耦过程中的情感语义信息的丢失，公式表示为

$$L_{\mathrm{IN}} = \left\|F^{-1}\left(\left[\boldsymbol{u}_m;\boldsymbol{r}_m\right]\right) - \left[\tilde{\boldsymbol{u}}_m;\tilde{\boldsymbol{r}}_m\right]\right\|_2^2 \tag{11}$$

式中，$F^{-1}(\cdot)$ 为逆向变换函数。

接着，为了提高共享特征间的语义对齐，同时增强特有特征的独立性，防止特有特征与共享特征发生语义混淆。引入基于中心矩差异（central moment discrepancy，CMD）（Zellinger等，2017）和基于希尔伯特施密特独立性准则（Hilbert Schmidt independence criterion，HSIC）（Greenfeld等，2020）的约束机制，即相似性约束 L_{S} 和独立性约束 L_{H}。其中，L_{S} 通过最小化共享特征分布之间的中心矩距离实现语义对齐，该过程表示为

$$\mathrm{CMD}_K(M,N) = \frac{1}{|b-a|}\left\|\mathrm{E}(M) - \mathrm{E}(N)\right\|_2 + \frac{1}{|b-a|^K}\left\|C_K(M) - C_K(N)\right\|_2 \tag{12}$$

$$L_{\mathrm{S}} = \frac{1}{3}\sum_{(m_1,m_2)}\mathrm{CMD}\left(\boldsymbol{r}_{m_1},\boldsymbol{r}_{m_2}\right) \tag{13}$$

式中，M、N 为在区间 $[a,b]$ 上的随机样本，$\|\cdot\|_2$ 为 L_2 范数，$\mathrm{E}(\cdot)$ 表示期望，$C_K(\cdot)$ 表示第 K 阶中心矩，在本文，K 的值为5，m_1 和 m_2 代表不同模态。

同时，独立性损失 L_H 旨在最小化模态内共享特征与特有特征之间以及模态特有特征之间的互相关性，表示为

$$\mathrm{HSIC}(M,N) = (B-1)^{-2}\mathrm{Tr}\left(\boldsymbol{H}\boldsymbol{G}_M\boldsymbol{H}\boldsymbol{G}_N\right) \tag{14}$$

$$L_{\mathrm{H}} = \frac{1}{3}\sum_{(m_1,m_2)}\mathrm{HSIC}\left(\boldsymbol{u}_{m_1},\boldsymbol{u}_{m_2}\right) + \frac{1}{3}\sum_{m}\mathrm{HSIC}\left(\boldsymbol{r}_m,\boldsymbol{u}_m\right) \tag{15}$$

式中，$\boldsymbol{G}_M$, $\boldsymbol{G}_N$ 为样本间的格拉姆矩阵，$\mathrm{Tr}(\cdot)$ 表示矩阵的迹运算，$\boldsymbol{H}$ 为中心化矩阵。

2.3 互信息约束机制

为了进一步提升解耦的效果和可解释性，受（Han等，2024）提出的基于互信息的可解释多模态解耦框架启发，本文进一步探索了特征间的依赖关系，构建了MIC约束机制，如图1所示。针对共享特征设计RMI，旨在最大化共享特征间及其与情感标签间的依赖程度，使得各模态的共享特征聚合到同一情感语义中心，针对特有特征设计CMI，旨在给定共享特征的条件下最小化其与标签的冗余关联。通过RMI与CMI约束，从信息论层面增强情感特征的语义一致性并减少模态噪声冗余。

为了确保解耦后的共享特征 $\boldsymbol{r}_m$ 真正捕获了跨模态的一致性情感语义，本文采用信息噪声对比估计（information noise-contrastive estimation，InfoNCE）下界（Oord等，2018）最大化不同模态共享特征间以及共享特征与情感标签 y 之间的互信息。具体而言，对各模态共享特征及各模态共享特征和标签均施加约束，定义的相关互信息损失 L_{R} 为

$$L_{\mathrm{R}}(X;Y) = \sum_{(X,Y)} -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s\left(x_i,y_i\right)}}{\sum_{j=1}^{B}e^{s\left(x_i,y_j\right)}} \tag{16}$$

式中，x_i 和 y_i 表示第 i 个正样本对，y_j 表示批中的负样本，$s(x,y) = f(x)^T h(y)$ 为相似度打分函数，$f(\cdot)$ 和 $h(\cdot)$ 为MLP打分网络，$\boldsymbol{z} = \mathrm{Linear}(y) \in \mathbf{R}^D$ 为情感

标签嵌入，$(X;\ Y)=\left\{(\boldsymbol{r}_{m_1};\ \boldsymbol{r}_{m_2}),(\boldsymbol{r}_m;\ z)\right\}$。

为了保证特有特征$\boldsymbol{u}_m$只保留与情感任务非直接相关的模态专属属性。本文设计了条件互信息约束L_{C}，旨在最小化已知共享特征$\boldsymbol{r}_m$条件下，特有特征$\boldsymbol{u}_m$与情感标签y之间的互信息，条件互信息为

$$I(\boldsymbol{u}_m;\ z|\boldsymbol{r}_m)=\mathrm{E}p(\boldsymbol{u}_m,z,\boldsymbol{r}_m)\left(\log\frac{p(\boldsymbol{u}_m,z|\boldsymbol{r}_m)}{p(\boldsymbol{u}_m|\boldsymbol{r}_m)\,p(z|\boldsymbol{r}_m)}\right)\tag{17}$$

式中，$\boldsymbol{u}_m$代表各模态特有特征，z是情感标签的嵌入表示，$\boldsymbol{r}_m$是各模态共享特征，$p(\cdot)$表示概率分布，若$I(\boldsymbol{u}_m;\ z|\boldsymbol{r}_m)$越小，代表在已知$\boldsymbol{r}_m$后，$\boldsymbol{u}_m$对$z$条件独立，即特有特征未残留与任务标签直接相关的噪声信息，解耦更有效。

由于真实分布$p(\boldsymbol{u}_m|\boldsymbol{r}_m)$，$p(z|\boldsymbol{r}_m)$难以直接获得，本文采用判别式变分近似，构造二元判别器$\varphi(\boldsymbol{u}_m,z,\boldsymbol{r}_m)\in(0,1)$来区分联合样本与条件乘积分布样本，因此，条件互信息可以表示为

$$I(\boldsymbol{u}_m;\ z|\boldsymbol{r}_m)=\mathrm{E}_1\left(\log\frac{\varphi(\boldsymbol{u}_m,z,\boldsymbol{r}_m)}{1-\varphi(\boldsymbol{u}_m,z,\boldsymbol{r}_m)}\right)-\mathrm{E}_2\left(\log\frac{\varphi(\boldsymbol{u}_m,z',\boldsymbol{r}_m)}{1-\varphi(\boldsymbol{u}_m,z',\boldsymbol{r}_m)}\right)\tag{18}$$

式中，E_1和E_2表示分别表示对联合样本与条件乘积分布样本的期望，判别器$\varphi(\cdot)$采用MLP架构实现，其输入为特有特征$\boldsymbol{u}_m$、共享特征$\boldsymbol{r}_m$与标签嵌入z的拼接向量，其网络架构包含三个线性层和一个输出层，线性层之间均采用修正线性单元（rectified linear unit，ReLU）激活，输出层将高维特征映射为二元判别分数，用于估计样本来自联合分布$p(\boldsymbol{u}_m,z|\boldsymbol{r}_m)$和条件乘积分布$p(\boldsymbol{u}_m|\boldsymbol{r}_m)\,p(z|\boldsymbol{r}_m)$的概率，$z'$表示在同一条件$\boldsymbol{r}$的邻域内通过K近邻（k-nearest neighbors，KNN）采样选取的难负样本得到的标签嵌入。最小化该项损失L_{C}，表示为

$$L_{\mathrm{C}}=\sum_{m\in\{a,t,v\}}I(\boldsymbol{u}_m;\ z|\boldsymbol{r}_m)\tag{19}$$

最后，CMI损失L_{MI}为L_{R}与L_{C}损失之和。

$$L_{\mathrm{MI}}=L_{\mathrm{R}}+L_{\mathrm{C}}\tag{20}$$

式中，L_{MI}表示互信息总损失，L_{R}和L_{C}分别表示相关互信息损失与条件互信息损失。

2.4 语义引导的自适应特征融合

在获得解耦特征后，针对现有融合方法难以捕捉细粒度情感语义信息的问题，本文设计了SGAFF模块，如图3所示。该模块利用共享子空间捕获的情感共享语义特征$\boldsymbol{r}_m$作为下文引导，通过残差结构对各模态的特有特征$\boldsymbol{u}_m$进行细粒度语义修正，减少特有子空间中的特征偏差。具体而言，首先，在各模态内部利用包含线性层和前馈网络的残差结构对共享特征$\boldsymbol{r}_{m\in\{a,t,v\}}$进行增强得到$\boldsymbol{r}^e_{m\in\{a,t,v\}}$，再通过线性残差结构进行交互，得到共享语义引导后的特有特征$\boldsymbol{u}^g_{m\in\{a,t,v\}}$，该过程表示为

$$\boldsymbol{u}^g_m=\boldsymbol{u}_m+\boldsymbol{W}_m\left[\boldsymbol{u}_m;\ \boldsymbol{r}^e_m\right]\tag{21}$$

式中，$\boldsymbol{u}^g_m$表示经共享语义引导后的特有特征，$\boldsymbol{W}_m$表示线性映射权重矩阵。

然后，为了提高最终融合表示的判别性，本文设计了双分支预测结构，共享分支$g(\boldsymbol{s})$基于增强后的共享特征预测情感倾向，经共享语义引导的特有分支$g(\boldsymbol{p})$基于修正后的特有特征捕捉互补细节。最后，通过门控机制自适应加权集成两路预测结果，实现动态平衡关系。其预测过程表示为

$$\begin{aligned}\hat{y}&=\gamma\odot g(\boldsymbol{s})+(1-\gamma)\odot g(\boldsymbol{p})\\\gamma&=\sigma\left(\mathrm{MLP}\left(\left[\boldsymbol{u}^{g'}_m;\ \boldsymbol{r}^{e'}_m\right]\right)\right)\end{aligned}\tag{22}$$

式中，$\sigma(\cdot)$为Sigmoid激活函数，$\gamma\in(0,1)$为权重系数，$g(\cdot)$为MLP预测头，$\hat{y}$为最终预测结果。

2.5 联合优化损失函数

情感识别任务的核心目标是捕获与任务标签紧

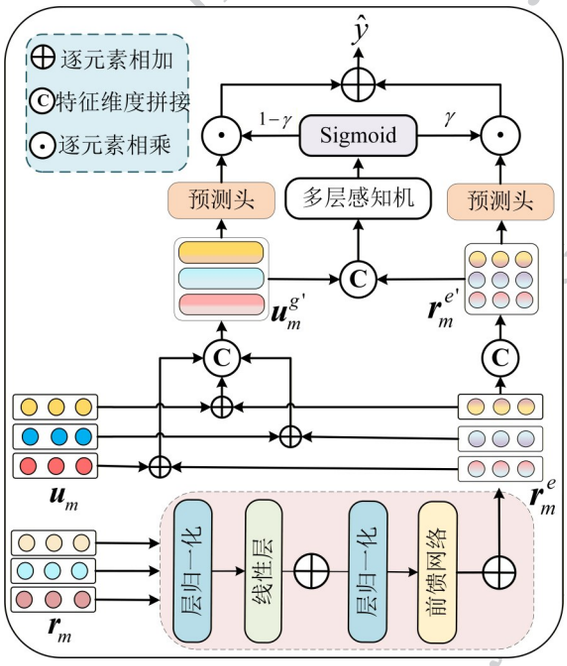


图3 语义引导的自适应特征融合模块

Fig. 3 Semantic-guided adaptive feature fusion module

密相关的判别特征，为直接优化模型在情感识别任务上的性能，本文引入任务损失L_T作为主要监督，在回归任务中采用均方误差作为任务损失，在分类任务中则采用交叉熵损失。此外，本文采用联合约束优化框架，以此有效整合不同模态间的互补信息，从而实现更精确的情感建模。总优化目标是将前述多个损失项组合为联合优化目标函数，表示为

$$L = L_T + \lambda_1 L_{IN} + \lambda_2 L_S + \lambda_3 L_H + \lambda_4 L_{RE} + \lambda_5 L_{MI} \quad (23)$$

式中，L为总损失，$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$为平衡前述各项损失贡献的权重超参数，通过验证集调优获得，并在训练过程中保持固定。

3 实验结果与分析

3.1 数据集与评价指标

在本文中，选择CMU-MOSI、CMU-MOSEI和UR-FUNNY这三个数据集对模型进行验证，本文采用了官方数据集划分，以保证在基线之间进行公平对比，数据集的划分方式如表1所示。其中，CMU-MOSI是多模态情感识别中被广泛使用的数据集，其内容来源于YouTube，由89个独白视频组成，每个视频片段涵盖一种情感状态，这些状态经人工标注获得的标签取值范围由-3到+3，共分为7个等级。CMU-MOSEI数据集是针对CMU-MOSI

数据集的扩充，包含1000个视频片段。UR-FUNNY提供了16514个来自TED演讲的多模态话语样本，每个话语都被标记为幽默/非幽默的二元标签。为保证与现有方法的可比性，本文采用多模态情感识别主流中一致使用的评价指标进行性能评估。在CMU-MOSI和CMU-MOSEI上，使用了二分类准确率（2-class accuracy，Acc-2）、F_1分数（F_1-score，F_1）和七分类准确率（7-class accuracy，Acc-7）、平均绝对误差（mean absolute error，MAE）和皮尔逊相关系数（Pearson correlation coefficient，Corr），需要注意的是，Acc-2和F_1分数的计算方式有两种：负/非负（包含零）和负/正（排除零）。在UR-FUNNY数据集上，使用了Acc-2和F_1。

表1 数据集的统计信息

Table 1 Statistics of the datasets

数据集	训练集	验证集	测试集	总和
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
UR-FUNNY	10598	2626	3290	16514

3.2 实验设置

在NVIDIA 4090GPU上使用PyTorch框架进行所有实验。训练过程中，所有参数均通过AdamW进行优化。对于文本特征，CMU-MOSI、CMU-MOSEI以及UR-FUNNY数据集均采用BERT（Devlin等，2019）预训练模型。对于上述三个数据集，模型的超参数设置如表2所示。在对比实验中，为了验证模型性能改进的稳健性，本模型及可复现的基线模型均使用5种不同的随机种子（42，1，11，111，1111）进行独立训练与测试，最终结果以均值±标准差的形式表示。本文的对比结果来源分为两类：对于已公开代码且可在相同数据设置下复

表2 不同数据集模型超参数设置

Table 2 Hyperparameter configuration for different datasets

参数类别	CMU-MOSI	CMU-MOSEI	UR-FUNNY
批次大小	32	32	32
学习率	6e-5	6e-5	4.5e-5
BERT学习率	2e-5	2e-5	5e-5
权重衰减	5e-5	5e-5	5e-5
Transformer层数 v/a/t	2/2/3	2/2/3	2/2/3
Transformer头数 v/a/t	2/2/4	2/2/4	2/2/4
Dropout	0.1	0.1	0.1
$\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5$	0.1，0.05，0.3，0.2，0.1	0.05，0.05，0.2，0.1，0.1	0.05，0.05，0.2，0.1，0.1

现的基线模型，在与本文一致的数据划分、输入特征与评测脚本下重新训练并测试，所得结果在表中以“•”标记；对于未公开代码或难以在相同配置下复现的模型，本文直接引用其原文报告结果，确保比较的公平性。

3.3 对比试验及分析

表3展示了在CMU-MOSI和CMU-MOSEI数据集上本模型与多种经典模型的对比结果。在CMU-MOSI中，与未采用解耦机制的模型相比，本文模型具有显著优势。相较于张量融合网络（Zadeh等，2017）和低秩融合模型（Liu等，2018），本文的MAE分别降低了19.9%和20.8%，相比于基于Transformer强交互的模型（Tsai等，2023），本文的MAE降低了17.1%，同时Corr提升了14.3%。这是因为此类模型往往将包含噪声的原始特征直接投影交互，导致特征空间中情感语义与模态噪声纠缠，而本文的IAMD模块与MIC机制，通过双向可逆映射在潜在空间显式解耦了跨模态共享特征与模态特有特征，有效减少了模态异质性干扰。与经典解耦模型（MISA（Hazarika等，2020），MFSA（Yang等，2022b），DMD（Li等，2023），DLF（Wang等，2025））相比，本模型依然保持性能优势。相比于基于对抗解耦的MFSA，本文MAE降低了15.7%，与DLF模型相比，在Corr上提升了1.7%。不同于MFSA模型依赖不稳定的对抗训练，本文的MIC机制从信息论角度最小化特有特征与标签的互信息，减少模态冗余噪声。本模型在Acc-7上为最优，这是因为DMD模型的自回归重构存在信息损耗，而本文基于可逆神经网络减少了信息丢失。相较于依赖强化文本模态作用的DLF模型，本文所设计的SGAFF模块并不预设主模态，利用共享语义动态修正特有特征，并通过双分支门控机制自适应分配权重，这种策略使得模型在处理非语言模态信息时更加鲁棒。

在CMU-MOSEI中，本模型在衡量情感强度的回归指标上取得了最优性能，并在分类指标上保持了竞争力。与基于Transformer的模型（Tsai等，2023）相比，本模型的MAE降低了8.4%，Corr提升了10.8%，这是因为本文的IAMD模块与MIC机制避免了特有模态冗余噪声对情感语义的干扰，确保模型在大规模数据下能学习到更为关键的情感特征。与解耦类模型（DLF（Wang等，2025），TMBL（Huang等，2024），DMD（Li等，2023））相比，本模型在回归精度上有一定优势，这得益于对解耦后特征的更优约束与利用。相较于DLF模型，本模型在MAE

上表现相当，在Corr上提升了2.0%。相较于TMBL模型，本模型在MAE上进一步降低了2.6%，在Corr上提升了1.7%，这表明本模型预测的情感强度值与真实标签更为接近。值得注意的是，虽然TMBL在Acc-2上略高，但其MAE显著高于本文，说明其虽能判断情感极性，但在情感强度的预测上偏差较大。观察到DMD模型在Acc-7指标上略高于本模型，是因为DMD采用了图知识蒸馏策略，倾向于学习平滑的类间概率分布，这在处理模糊边界的分类任务时具有一定优势，而本文通过MIC机制与IAMD模块，在连续情感空间的建模上更加精准。

在UR-FUNNY数据集上，表4结果表明，本模型在Acc-2和F_1上均取得了最优结果，展现了在捕捉细粒度情感信息语义的独特优势。相较于经典的模型TFN（Zadeh等，2017）和解耦模型MISA（Hazarika等，2020），本文的Acc-2分别提升了7.2%和1.3%。对比FDMER模型（Yang等，2022b），其利用对抗学习分离公共与私有特征，然而，对抗训练往往难以收敛，本文引入的MIC机制提供了比对抗学习更稳定的独立性约束，从而学习到更有效的情感特征。FRDIN模型（Zeng等，2024b）通过动态路由网络寻找最优交互路径，但缺乏对特征语义的显式引导，本文的SGAFF模块利用共享情感语义

作为上下文去动态修正特有特征，使得模型能够精准捕捉到语境与表达不符的幽默瞬间，从而实现了比FRDIN模型更精准的判别。

为进一步验证模型性能提升的稳健性，本文在各数据集上采用5个随机种子（42，1，11，111，1111）重复训练测试，并在可复现的最佳基线模型上进行了配对t检验，对于仅提供结果难以复现的模型，进行数值层面的对比分析。在CMU-MOSI数据集上，相较于DLF模型，本模型在MAE、Corr、Acc-2以及Acc-7指标上均达到统计显著水平（$p<0.05$）。在F_1（负/非负）指标上，未达到显著水平（$p=0.06$），而F_1（负/正）指标达到显著水平，这说明在去除模糊样本的标准二分类设置下，本模型的性能优势更为稳健。在CMU-MOSEI上，本模型在Corr、Acc-2及Acc-7指标上相较于DLF模型达到统计显著水平，而在MAE（$p=0.08$）及F_1（负/非负）（$p=0.53$）指标上未达到显著差异，表明在部分粗粒度分类指标上两者性能接

表3　CMU-MOSI和CMU-MOSEI数据集上不同模型的对比结果

Table 3　Performance comparison of different models on the CMU-MOSI and CMU-MOSEI datasets

模型	CMU-MOSI					CMU-MOSEI				
	MAE	Corr	Acc-2	F_1	Acc-7	MAE	Corr	Acc-2	F_1	Acc-7
TFN-BERT	0.901	0.698	-/80.8	-/80.7	34.9	0.593	0.700	-/82.5	-/82.1	50.2
LMF	0.912	0.668	76.4/-	75.7/-	32.8	0.623	0.677	-/82.5	-/82.1	48.0
MulT	0.871	0.698	-/83.0	-/82.8	40.0	0.580	0.703	-/82.5	-/82.3	51.8
Liu et al.	0.769	0.783	-/83.7	-/84.2	-	0.573	0.741	-/85.0	-/85.0	-
MFSA	0.856	0.722	-/83.3	-/83.7	41.4	0.574	0.734	-/83.8	-/83.6	53.2
DMD	-	-	-/83.5	-/83.5	41.9	-	-	-/84.8	-/84.7	**54.6**
PS-Mixer	0.794	0.748	80.3/82.1	80.3/82.1	44.3	0.537	0.765	83.1/**86.1**	83.1/**86.1**	53.0
TMBL	0.867	0.762	81.7/83.8	82.4/84.2	36.3	0.545	0.766	**84.2**/85.8	**84.8**/85.9	52.4
MER-CLIP	-	-	-/84.0	-/84.0	-	-	-	-/85.3	-/85.1	-
MISA •	0.828 ±0.004	0.728 ±0.011	79.9±0.33 / 81.4±0.31	79.9±0.18 / 81.4±0.20	40.8 ±0.25	0.555 ±0.004	0.750 ±0.004	84.0±0.16 / 85.2±0.28	83.8±0.13 / 85.0±0.19	50.5 ±0.36
DLF•	0.740 ±0.012	0.785 ±0.010	**83.8**±0.34 / 84.1±0.38	**83.7**±0.22 / 84.1±0.33	44.5 ±0.21	0.540 ±0.006	0.764 ±0.003	83.6±0.25 / 84.4±0.15	84.2±0.15 / 84.5±0.22	52.3 ±0.34
本模型	**0.722** ±0.004†	**0.798** ±0.003†	83.2±0.25† / **85.3**±0.26†	83.1±0.22 / **85.3**±0.21†	**45.8** ±0.14†	**0.531** ±0.003	**0.779** ±0.002†	**84.2**±0.16† / 85.2±0.13†	84.2±0.17 / 85.2±0.10†	53.2 ±0.28†

注：加粗、下划线字体表示各列最优、次优结果，标有•表示在相同设置下利用公开源代码和原始超参数复现的结果，-表示原文中无结果，†表示与可复现最佳基线模型相比具有统计显著性且$p<0.05$，其余数据直接引自相关文献。

近。在UR-FUNNY上，本模型在Acc-2与F_1指标上相较于MISA模型均达到统计显著水平（$p<0.05$）。

3.4　消融实验及分析

3.4.1　各模态对模型性能的影响

为研究各模态对模型性能的影响，本文将不同单模态进行组合，在CMU-MOSI数据集上进行模态消融实验，表5结果表明，文本在单模态中性能最优，高于视觉与音频模态，说明文本携带最直接的语义线索是情感识别的主要依据。而单独使用视觉或音频模态时性能下降，表明非语言模态的独立判别能力有限。在双模态组合中，文本加视觉组合表现最佳，说明视觉模态在文本语义基础上提供了有益

表4 UR-FUNNY数据集上不同模型的对比结果

Table 4 Performance comparison of different models on the UR-FUNNY dataset

模型	Acc-2	F_1
TFN-BERT	68.57	-
TFN	64.71	-
MISA	70.61	-
FDMER	<u>71.87</u>	-
FRDIN	71.82	-
MISA•	67.80±0.20	<u>67.91</u>±0.18
本模型	**71.91**±0.16†	**71.71**±0.13†

注:加粗、下划线字体表示各列最优、次优结果,标有•是按照原文公开代码复现获得的实验结果,-表示原文中无结果,†表示与可复现最佳基线模型相比具有统计显著性且$p<0.05$,其余数据直接引自相关文献。

的补充。相比之下,音频加视觉组合未能带来显著性能提升,表明缺乏语义支撑的模态间协同较弱。最终,三模态融合最优指标最多,这表明三模态间存在互补关系,能够提升整体情感建模的可靠性。

表5 CMU-MOSI数据集上不同模态配置的消融实验结果

Table 5 Ablation results of different modality configurations on the CMU-MOSI dataset

模态设置	MAE	Corr	Acc-2	F_1	Acc-7
v	1.411	0.133	55.7/58.1	39.8/42.7	16.1
t	0.726	0.786	80.6/82.7	80.5/82.5	44.7
a	1.450	0.041	41.9/44.3	24.8/27.2	15.2
a+v	**0.720**	0.124	55.7/58.2	39.7/42.6	16.0
t+v	0.724	0.797	**83.3/85.5**	82.1/84.0	45.6
a+t	0.723	0.797	81.3/83.1	81.2/83.1	44.9
a+t+v	0.722	**0.798**	83.2/85.3	**83.1/85.3**	**45.8**

注:加粗字体表示各列最优结果。

3.4.2 各模块对模型性能的影响

为验证方法内部核心模块的有效性,分别移除IAMD、MIC和SGAFF模块,结果如表6所示。当移除IAMD模块后,模型的整体性能下降,这说明该模块通过双向可逆映射显式解耦了共享与特有特征,有效保留了模态情感相关特征,若缺失该结构,模型的回归拟合能力大幅减弱。去除MIC机制后,模型性能退化,尤其在Corr与分类指标上下降更为明显,说明在缺乏信息论层面的约束时,解耦后的共享特征难以充分对齐跨模态一致的情感语义,而特有特征中也更易残留与任务相关的冗余噪声,表明MIC机制在强化共享子空间语义一致性、抑制无关噪声干扰方面发挥了重要作用。当去除SGAFF后,Acc-2与F_1均下降明显,这说明简单的特征拼接无法有效整合解耦后的异构信息,而共享语义对特有特征的动态引导在融合阶段能增强特征判别性。综合来看,完整模型在各项指标上均达到最优,三者协同构成模型性能提升的核心机制。

表6 CMU-MOSI数据集上不同模块的消融实验结果

Table 6 Ablation results of different components on the CMU-MOSI dataset

模块设置	MAE	Corr	Acc-2	F_1	Acc-7
w/o IAMD	0.755	0.767	81.5/83.5	81.4/83.5	45.4
w/o MIC	0.759	0.762	81.3/83.7	81.1/83.1	44.5
w/o SGAFF	0.756	0.762	81.3/83.5	81.1/83.5	44.5
本模型	**0.722**	**0.798**	**83.2/85.3**	**83.1/85.3**	**45.8**

注:加粗字体表示各列最优结果。

3.4.3 损失函数对模型性能影响

为评估不同损失函数对模型性能的影响,本文在CMU-MOSI数据集上进行逐项消融实验,分别移除相似性损失L_S、独立性损失L_H、重构损失L_{RE}、可逆约束损失L_{IN}以及互信息损失L_{MI},由表7可见,各类损失均对模型性能产生正向贡献,完整模型在多数指标上取得最优,表明多项约束在联合优化时能协同提升回归与分类性能。

表7 CMU-MOSI数据集上训练所用损失函数消融实验结果

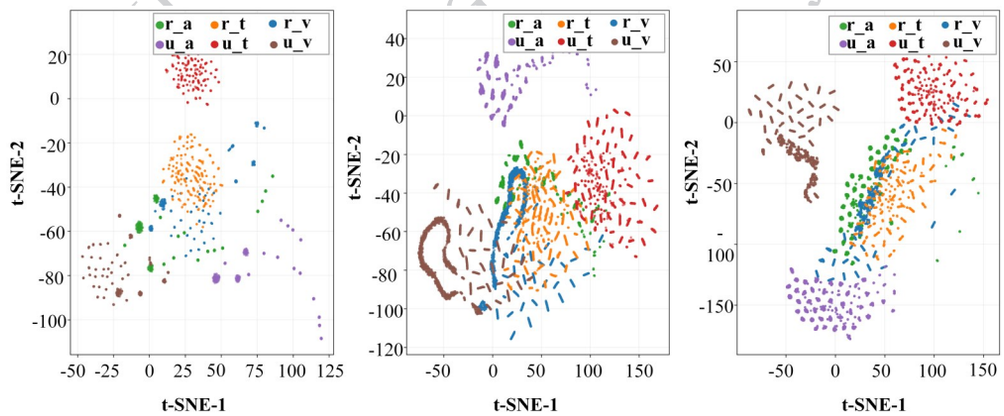
Table 7 Ablation results of different loss functions on the CMU-MOSI dataset

设置	MAE	Corr	Acc-2	F_1	Acc-7
w/o L_S	0.745	0.786	83.3/83.9	83.2/84.3	43.2
w/o L_H	0.737	0.784	82.1/83.1	**83.9**/84.3	44.1
w/o L_{RE}	0.738	0.796	**83.6**/84.8	83.7/84.2	44.9
w/o L_{IN}	0.741	0.792	83.4/84.1	82.2/82.5	**46.3**
w/o L_{MI}	0.746	0.785	82.5/83.2	83.5/83.6	43.5
本模型	**0.722**	**0.798**	83.2/**85.3**	83.1/**85.3**	45.8

注:加粗字体表示各列最优结果。

3.4.4　解耦表示重要性对比分析

为进一步对模型中共享表示与特有表示的作用进行量化分析，本文参照MISA和MFSA解耦模型在多模态情感识别任务中的表示重要性分析方法，



(a) CMU-MOSI (b) CMU-MOSEI (c) UR-FUNNY

图4　不同数据集上t-SNE特征解耦可视化

Fig. 4　t-SNE visualization of feature disentanglement on different datasets((a)CMU-MOSI;(b)CMU-MOSEI;(c)UR-FUNNY)

并与其对比，测试了仅使用共享特征和仅使用特有特征实现预测时的模型性能，并计算其相对于完整模型的性能相对下降幅度，由表8可见，三类模型在移除共享或特有特征后均出现性能下降，说明共享与特有特征对预测均有贡献。与MISA与MFSA相比，本模型在移除任一特征时均表现出更明显的性能退化。具体而言，当移除模态特有特征仅保留共享特征参与融合时，MFSA的Acc-2下降了2.5%，相比之下，本模型的Acc-2相对下降了3.6%；而且在Corr指标上，相较于MISA与MFSA分别下降了3.4%和1.9%，本模型下降幅度最大达到4.6%，说明本模型对特有分支的补充信息依赖程度更高。当移除模态共享特征，仅保留特有特征参与融合时，性能下降程度更为显著，在Acc-2指标上，MFSA仅下降了1.80%，而本模型则下降了4.8%，在Corr上相较于MISA与MFSA分别下降了5.9%和1.3%。这一结果客观验证本模型的IAMD模块与MIC机制的协同有效性，从而使模型获得了更高的判别效果。

表8　在CMU-MOSI数据集上不同解耦模型的解耦表示重要性对比结果

Table 8　Comparative results on the importance of disentangled representations for different models on the CMU-MOSI dataset

设置	MAE	Corr	Acc-2	F_1	Acc-7
MFSA(w/o共享)	0.871	0.712	81.8	82.4	39.2
MFSA(w/o特有)	0.898	0.708	81.2	81.5	38.6
MFSA	0.856	0.722	83.3	83.7	41.4
MISA(w/o共享)	0.858	0.716	–	–	–
MISA(w/o特有)	0.850	0.735	–	–	–
MISA	0.783	0.761	83.4	83.6	42.3
本模型(w/o共享)	0.788	0.746	81.2	81.1	42.1
本模型(w/o特有)	0.779	0.761	82.2	82.4	43.1
本模型	**0.722**	**0.798**	**85.3**	**85.3**	**45.8**

注：加粗字体表示各列最优结果，–表示原论文中无结果，其余数据直接引自相关文献。

3.5 可视化分析

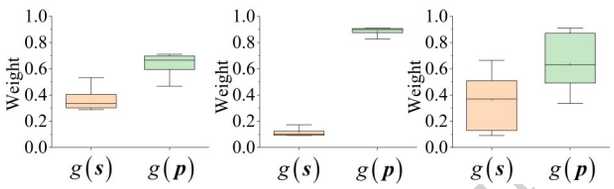
3.5.1 特征解耦可视化分析

为进一步验证本模型在特征层面的解耦效果，采用t分布随机邻域嵌入(t-distributed stochastic neighbor embedding，t-SNE)对模型学习到的共享特征与模态特有特征进行可视化分析。图4展示了在三个数据集上的可视化结果，其中绿、橙、蓝三类点分别对应音频、文本与视觉模态的共享特征，紫、红、褐三类点分别对应音频、文本与视觉模态的特有特征。从结果可以看出，不同数据集上均呈现出一致的解耦分布模式：共享特征在嵌入空间中聚集，而各模态的特有特征在不同方向上分散分布且各自独立，这说明所提出的模型能够在潜在空间中有效地将跨模态一致的情感语义与模态特有的属性信息进行分离，实现了更具可解释性的多模态情感表示学习。

3.5.2 融合权重分布可视化分析

图5展示了在不同数据集上，融合阶段门控输出的共享分支$g(\boldsymbol{s})$与共享引导的特有分支$g(\boldsymbol{p})$权重分布情况的箱线图。可以观察到，CMU-MOSI、CMU-MOSEI和UR-FUNNY三个数据集在融合阶段的门控权重分布趋势基本一致，模型均倾向于赋予经共享引导后特有分支$g(\boldsymbol{p})$更高的分布权重，这一现象验证了本文SGAFF模块的有效性，这表明共享特征提供了语义一致性基础，但最终的情感依赖于经过共享语义修正后的特有特征。证实了特有特征中包含的细粒度线索(如语调的微弱抖动、面部

的瞬间微表情)是区分相似情感类别的关键，而



(a) CMU-MOSI (b) CMU-MOSEI (c) UR-FUNNY

图5 不同数据集上双分支权重分布可视化结果

Fig. 5 Visualization results of dual-branch weight distributions on different datasets((a)CMU-MOSI;(b)CMU-MOSEI;(c)UR-FUNNY)

单独的共享语义可能过于平滑，难以捕捉这些高频细节。总体来看，共享分支提供跨模态一致的情感语义约束，而共享引导的特有分支负责引入具有判别价值的细粒度互补信息，两者通过可学习的门控机制进行自适应加权，使模型能够根据不同数据集与任务特性动态调整融合策略，从而提升整体预测的判别性。

3.6 模型复杂性分析

为了进一步客观评估所提模型的计算效率与复杂度，本文在CMU-MOSI数据集上对比了代表性基线模型的参数量、浮点运算量以及推理延迟，并同步报告七分类准确率。为保证对比公平，所有模型的实验测试结果均不包含预训练的BERT文本特征提取器，结果如表9所示。本模型的参数量达到8.63M，相较于基准模型MISA有明显增加。结合消融实验可知，这一增加主要归因于IAMD、MIC以及SGAFF关键模块。然而，尽管参数规模显著增加，本模型的浮点运算量仍与基准模型相近，表明本模型的网络结构在前向推理的主干路径上并未带来乘加计算负担。在实际运行效率方面，本模型的推理

表9 CMU-MOSI数据集上计算效率与复杂度对比

Table 9 Comparison of computational efficiency and complexity on the CMU-MOSI dataset

模型	参数量(M)	浮点运算量(G)	推理延迟(ms)	Acc-7
MulT	**2.57**	1.9	21.0	40.0
MISA	3.10	**1.7**	**20.5**	40.8
DLF	4.53	1.8	21.6	44.5
本模型(w/o IAMD)	7.28	1.9	23.5	45.4
本模型(w/o MIC)	5.57	1.9	22.5	44.5
本模型(w/o SGAFF)	7.39	1.9	23.7	44.5
本模型	8.63	1.9	24.5	**45.8**

注：加粗字体表示各列最优结果。

延迟为24.5ms，较MISA和DLF分别增加了4.0ms和2.9ms，这主要归因于引入复杂解耦机制带来的额外算子调度开销。综合来看，尽管本文方法引入了更复杂的结构导致参数规模上升，但在不显著增加浮点运算量与推理延迟的前提下，取得了更高的识别准确率，这种计算代价投入对于实现更具判别性的情感表示具有正面意义。

4 结 论

本文针对多模态情感识别任务中特征表示的情感语义信息与模态特有噪声纠缠以及跨模态融合机制判别性不足的问题，提出了一种面向多模态情感识别的可解释可逆解耦与自适应融合方法，基于可逆神经网络以及互信息约束实现了情感语义信息与模态特有噪声的可解释解耦，并促进跨模态情感语义特征的细粒度交互，提高了模型性能。核心贡献在于，基于可逆神经网络设计了IAMD模块，实现了跨模态一致性的共享特征与保留各模态独有属性的特有特征的分离，避免情感语义信息的丢失。构建了MIC机制，从信息论角度建模特征间依赖关系，增强情感语义一致性同时减少模态特有的噪声冗余。此外，在解耦基础上设计了SGAFF模块，利用共享语义信息作为上下文先验，对特有特征进行细粒度引导与自适应融合，提升特征融合表示的判别性。在CMU-MOSI、CMU-MOSEI和UR-FUNNY数据集上的实验表明，该方法在情感分类准确率与连续情感回归上均取得优越的性能，消融实验与可视化分析证实了所提方法的可行性和有效性。

尽管本文方法在性能上取得了显著提升，由于引入可逆变换与互信息约束，虽然增强了模型的可解释性与解耦效果，但同时增加了模型的计算复杂度。未来工作将进一步研究面向情感识别任务的轻量化解耦与融合机制，进一步提升模型在计算资源受限场景下的泛化能力。

参考文献(References)

Al-Tameemi I K S, Feizi-Derakhshi M R, Pashazadeh S and Asadpour M. 2023. Interpretable multimodal sentiment cla-ssification using deep multi-view attentive network of im-age and text data. IEEE Access, 11: 91060-91081 [DOI: 10.1109/ACCESS. 2023. 3307716]

Chen S, Sun Q and Zhu X T. 2026. Emotion-controllable 3D talking face generation with hierarchical disentanglement-guided VQ-VAE[J/OL]. Journal of Image and Graphics, 1 -15 (陈胜, 孙强, 朱霞天. 2026. 分层解耦引导的情感可控VQ-VAE 3D说话人脸生成方法[J/OL]. 中国图象图形学报, 1-15 [DOI: 10.11834/jig.250451]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Esser P, Rombach R and Ommer B. 2020. A Disentangling in-vertible interpretation network for explaining latent repres-entations//Proceedings of the IEEE/CVF Conference on C-omputer Vision and Pattern Recognition. Seattle, USA: IEEE: 9223-9232 [DOI: 10.1109/CVPR42600.2020.00924]

Greenfeld D and Shalit U. 2020. Robust learning with the Hilbert-Schmidt Independence Criterion//Proceedings of the 37th International Conference on Machine Learning. PMLR: 3759-3768 [DOI: 10.17760/d20382827]

Han Z, Luo T, Fu H, Zhou J and Zhang C. 2024. A principl-ed framework for explainable multimodal disentanglement. Information Sciences, 675: 120768 [DOI: 10.1016/j.ins.2024.120768]

Hazarika D, Zimmermann R and Poria S. 2020. MISA: Modality-invariant and specific representations for multimodal sentiment analysis//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1122-1131 [DOI: 10.1145/3394171.3413678]

Heng H J and Xu T B. 2022. Attention sentiment analysis m-odel based on multi-scale convolution and gating mechanism. Journal of Computer Applications, 42(9): 2674-2679 (衡红军, 徐天宝. 2022. 基于多尺度卷积和门控机制的注意力情感分析模型. 计算机应用, 42(9): 2674-2679) [DOI: 10.11772/j. issn. 1001-9081. 2021081448]

Huang J, Zhou J, Tang Z, Lin J and Chen C Y C. 2024. TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. Knowledge-Based Systems, 285: 111346 [DOI: 10.1016/j.knosys.2023.111346]

Jia X B, Li C, Wang L, Zhang M, Liu X J, Zhang Y Y and Wen J K. 2025. A multimodal cross-domain sentiment a-nalysis algorithm based on feature disentanglement meta-o-ptimization. Journal of Computer Research and Developm-ent, 62(11): 2697-2709 (贾熹滨, 李宸, 王珞, 张沐晨, 刘潇健, 张旸旸, 温家凯. 2025. 基于元优化特征解耦的多模态跨域情感分析算法. 计算机研究与发展, 62(11): 697-2709) [DOI: 10.7544/issn1000-1239. 202440624]

Khalane A, Makwana R, Shaikh T and Ullah A. 2025. Evaluating significant features in context-aware multimodal emotion recognition

with XAI methods. Expert Systems, 42(1): e13403 [DOI: 10.1111/exsy.13403]

Li Y, Chan J, Peko G and Sundaram D. 2024. An explanation framework and method for ai-based text emotion analysis and visualisation. Decision Support Systems, 178: 114121 [DOI: 10.1016/j.dss.2023.114121]

Li Y, Wang Y and Cui Z. 2023. Decoupled multimodal distill-ing for emotion recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 6631-6640 [DOI: 10.1109/CVPR52729.2023.00641]

Lin H, Zhang P, Ling J, Yang Z, Lee L K and Liu W. 2023. PS-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. Information Processing & Management, 60(2): 103229 [DOI: 10.1016/j.ipm.2022.103229]

Liu Z, Shen Y, Lakshminarasimhan V B, Liang P P, Zadeh A and Morency L P. 2018. Efficient low-rank multimodal fusion with modality-specific factors//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Comput-ational Linguistics: 2247-2256 [DOI:10.18653/v1/P18-1209]

Luo Y Y, Wu R, Liu J F and Tang X L. 2024. Multimodal sentiment analysis method based on adaptive weight fusion. Journal of Software, 35(10): 4781-4793 (罗渊贻, 吴锐, 刘家锋, 唐降龙. 2024. 基于自适应权值融合的多模态情感分析方法. 软件学报, 35(10): 4781-4793) [DOI: 10.13328/j.cnki.jos.006998]

Ma Y and Wang W. 2022. MSFL: Explainable multitask-based shared feature learning for multilingual speech emotion r-ecognition. Applied Sciences, 12 (24): 12805 [DOI: 10.3390/app122412805]

Oord A, Li Y Z and Vinyals O. 2018. Representation learning with contrastive predictive coding [EB/OL]. [2018-07-10]. https://arxiv.org/pdf/1807.03748.pdf

Poria S, Cambria E, Howard N, Huang G B and Hussain A. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing, 174: 50-59 [DOI: 10.1016/j.neucom.2015.01.095]

Poria S, Cambria E, Bajpai R and Hussain A. 2017. A review of affective computing: from unimodal analysis to multi-modal fusion. Information Fusion, 37: 98-125 [DOI: 10.1016/j. inffus. 2017.02.003]

Song Y and Cho S. 2025. Leveraging CLIP encoder for multi-modal emotion recognition//Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vis-ion. Tucson, USA: IEEE: 6115-6124 [DOI: 10.1109/WACV61041.2025.00596]

Sun Q and Wang S Y. 2024. Self-supervised multimodal emotion recognition combining temporal attention mechanism and unimodal label automatic generation strategy. Journal of Electronics & Information Technology, 46(2): 588-601 (孙强, 王姝玉. 2024. 结合时间注意力机制和单模态标签自动生成策略的自监督多模态情感识别. 电子与信息学报, 46(02): 588-601) [DOI: 10.11999/JEIT231107]

Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics, 29(6): 1607-1627 (陶建华, 范存航, 连政, 吕钊, 沈莹, 梁山. 2024. 多模态情感识别与理解发展现状及趋势. 中国图象图形学报, 29(06): 1607-1627) [DOI: 10.11834/jig.240017]

Tomczak J M and Welling M. 2016. Improving variational auto-encoders using householder flow [EB/OL]. [2016-11-29]. https://arxiv.org/pdf/1611.09630.pdf

Tsai Y H H, Bai S, Liang P P, Kolter J Z, Morency L P and Salakhutdinov R. 2019. Multimodal Transformer for unal-igned multimodal language sequences//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 6558-6569 [DOI: 10.18653/v1/P19-1656]

Wang J, Zhao X M, Wang C L, Zhang S Q and Zhao S C. 2025. Deep feature interaction and hierarchical multimodal fusion for emotion recognition. Application Research of Computers, 42(7): 1978-1985 (王健, 赵小明, 王成龙, 张石清, 赵舒畅. 2025. 基于深度特征交互与层次化多模态融合的情感识别模型. 计算机应用研究, 42(07): 1978-1985) [DOI: 10.19734/j.issn.1001-3695.2024.11.0487]

Wang P, Zhou Q, Wu Y, Chen T and Hu J. 2025. DLF: Dise-ntangled-language-focused multimodal sentiment analysis// Proceedings of the AAAI Conference on Artificial Intelli-gence. Philadelphia, USA: AAAI Press: 21180-21188 [DOI: 10.1609/aaai. v39i20.35416]

Wang S M, Liu C G, Chen S Y and Liu Q S. 2025. A survey of multimodal emotion recognition from facial expressions, audios, and language. Journal of Image and Graphics, 30(6): 2120-2138 (王善敏, 刘成广, 陈胜宇, 刘青山. 2025. 面向表情、语音和语言的多模态情感识别综述. 中国图象图形学报, 30(6): 2120-2138) [DOI: 10.11834/jig.250168]

Wang X, Chen H, Tang S, Wu Z and Zhu W. 2024. Disentan-gled representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 9677-9696 [DOI: 10.1109/TPAMI.2024.3420937]

Yang D, Huang S, Kuang H, Du Y and Zhang L. 2022a. Disentangled representation learning for multimodal emotion recognition//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM: 1642-1651 [DOI: 10.1145/3503161.3547754]

Yang D, Kuang H, Huang S and Zhang L. 2022b. Learning modality-specific and -agnostic representations for asynchr-onous multimodal language sequences//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM: 1708-1717 [DOI: 10.1145/3503161.3547755]

Yang H，Chen C L P，Chen B and Zhang T. 2025. Improving the interpretability through maximizing mutual informatio-n for EEG emotion recognition. IEEE Transactions on Af-fective Computing，16（2）：744-757［DOI：10.1109/TAFFC.2024.3463469］

Zadeh A，Chen M H，Poria S，Cambria E and Morency L P. 2017. Tensor fusion network for multimodal sentiment an-alysis//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen，D-enmark：Association for Computational Linguistics：1103-1114［DOI：10.18653/v1/D17-1115］

Zadeh A A B，Liang P P，Poria S，Cambria E and Morency L P. 2018. Multimodal language analysis in the wild：CMU-MOSEI dataset and interpretable dynamic fusion graph //Proceedings of the 56th Annual Meeting of the Associati-on for Computational Linguistics. Melbourne，Australia：A-ssociation for Computational Linguistics：2236-2246［DOI：10.18653/v1/P18-1208］

Zellinger W，Grubinger T，Lughofer E，Natschläger T and Saminger-Platz S. 2017. Central Moment Discrepancy（CMD）for domain-invariant representation learning［EB/OL］.［2017-02-27］. https://arxiv.org/pdf/1702.08811.pdf

Zeng Y，Yan W，Mai S and Hu H. 2024a. Disentanglement tr-anslation network for multimodal sentiment analysis. Infor-mation Fusion，102：102031［DOI：10.1016/j.inffus.2023.102031］

Zeng Y，Li Z，Chen Z and Ma H. 2024b. A feature-based restoration dynamic interaction network for multimodal sentiment analysis. Engineering Applications of Artificial Intelligence，127：107335［DOI：10.1016/j.engappai.2023.107335］

Zhao S C，Feng Y F，Zhang Z C，Sun B，Zhang S P，Gao Y，Yang J F，Liu M，Yao H X and Wang Y N. 2025. Res-earch advancements on emotionally and intellectually inter-grated digital humans and robotics. Journal of Image and Graphics，30（6）：2139-2160（赵思成，丰一帆，张知诚，孙斌，张盛平，高跃，杨巨峰，刘敏，姚鸿勋，王耀南. 2025. 情智兼备数字人与机器人研究进展. 中国图象图形学报，30（6）：2139-2160）［DOI：10.11834/jig.240780］

作者简介

杨皎皎，女，硕士研究生，主要研究方向为多模态情感计算。E-mail：2230321189@stu.xaut.edu.cn

孙强，通信作者，男，副教授，主要研究方向为情感计算与人智交互。E-mail：qsun@xaut.edu.cn

朱霞天，男，高级讲师，主要研究方向是可扩展机器学习，计算机视觉，多模态GenAI。E-mail：xiatian.zhu@surrey.ac.uk