



第五届中国情感计算大会
The Fifth Chinese Conference on Affective Computing

迈向共情智能：多模态情感分析 挑战与机遇初探

李勇-东南大学

2025-07-18

汇报提纲

- 多模态情感分析-问题定义与研究内容
- 多模态情感识别-研究背景及核心挑战
- 课题组相关进展-单模态、多模态情感识别研究进展
- 未来研究方向-大模型时代的多模态情感识别等

问题定义

□ 人类通过多种通道感知世界

- 视、听、触、嗅、味

□ 模态 (Modality)

- 事物发生或被感知的途径

□ 多模态 (Multimodal)

- 涉及多种模态的研究问题 (异质、互通)



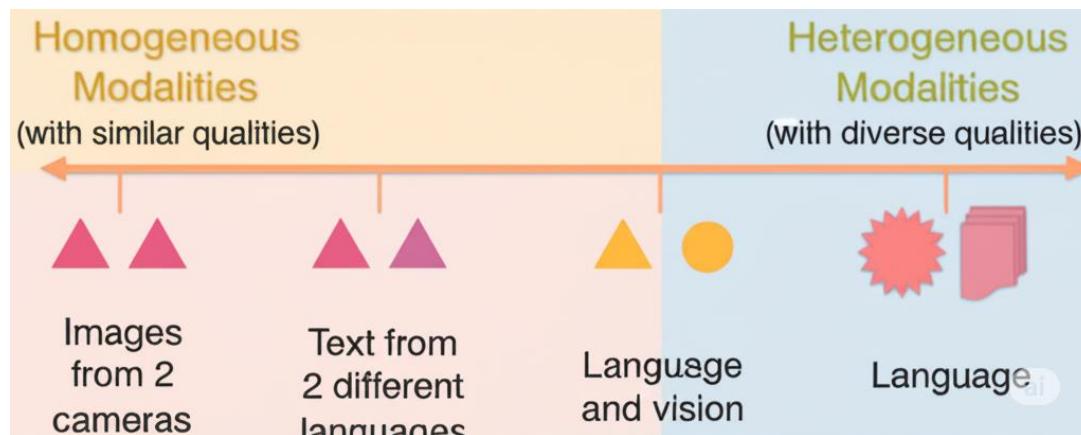
视觉

听觉

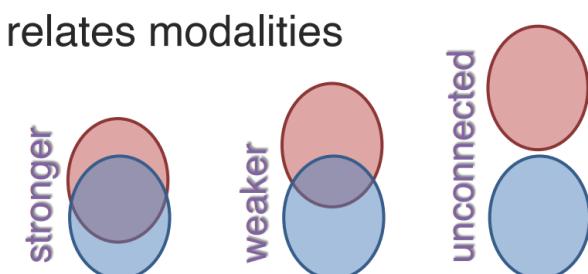
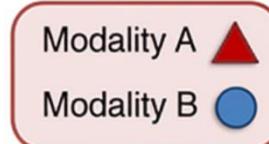
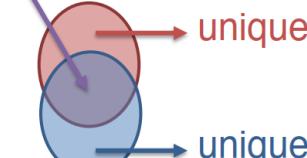
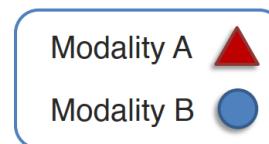
触觉

嗅觉

味觉



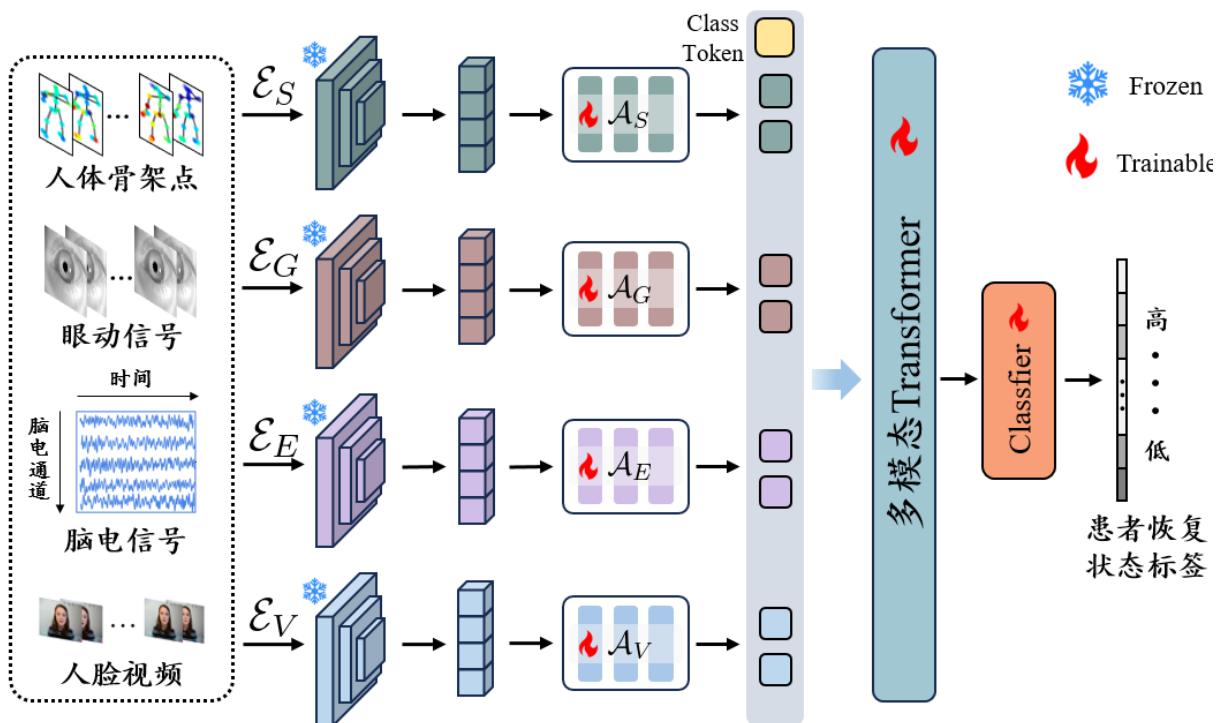
Connected: Shared information that relates modalities



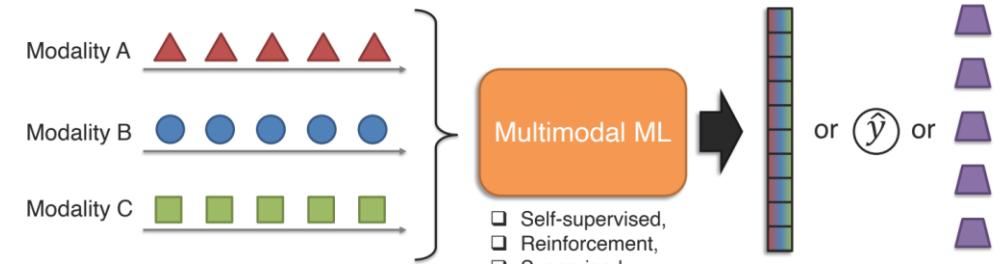
Interacting: process affecting each modality, creating new response

问题定义

- 多模态（机器）学习 (Multimodal (machine) learning)
- 构建模型使其可以处理多种模态的信息以及信息间的联系



基于多模态输入的患者恢复状态预测



典型多模态机器学习范式

*文本、语音、视觉等研究边界正在模糊和弱化

代表性应用

应用场景

教育、医疗、刑侦、军事等领域



数字人生成



智慧养老



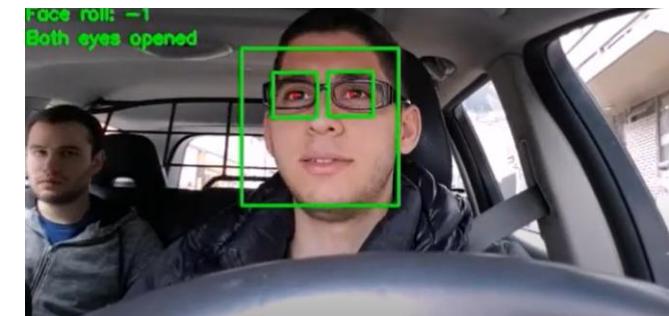
士兵作战意图同步



智慧医疗：患者恢复程度智能评估



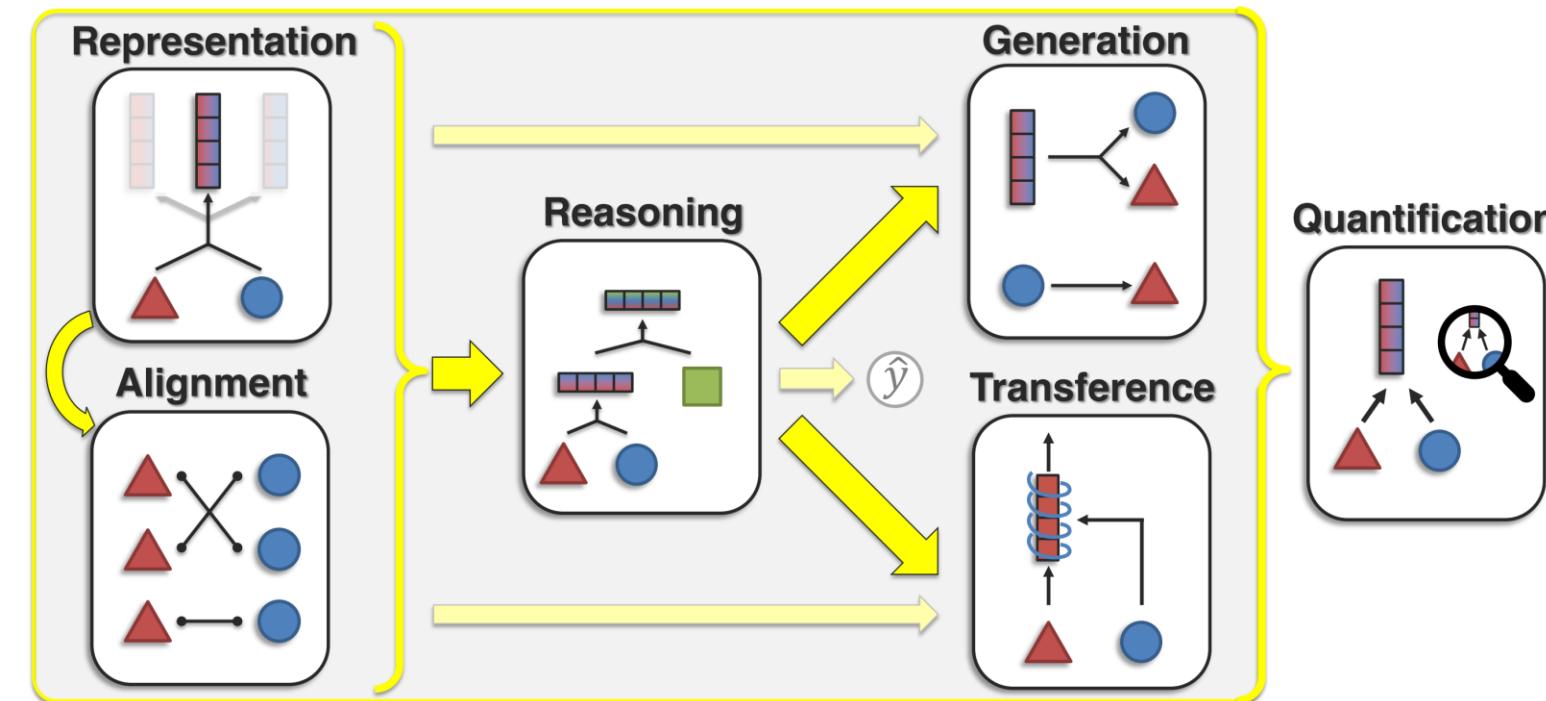
刑侦测谎



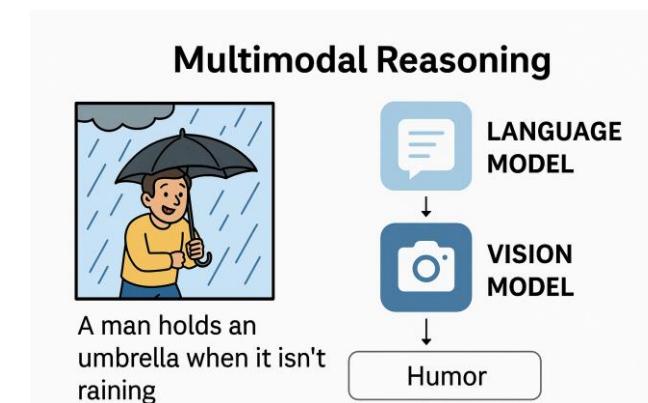
疲劳驾驶检测

挑战问题和研究内容

□ 多模态学习：表征、对齐、推理、生成、迁移和量化



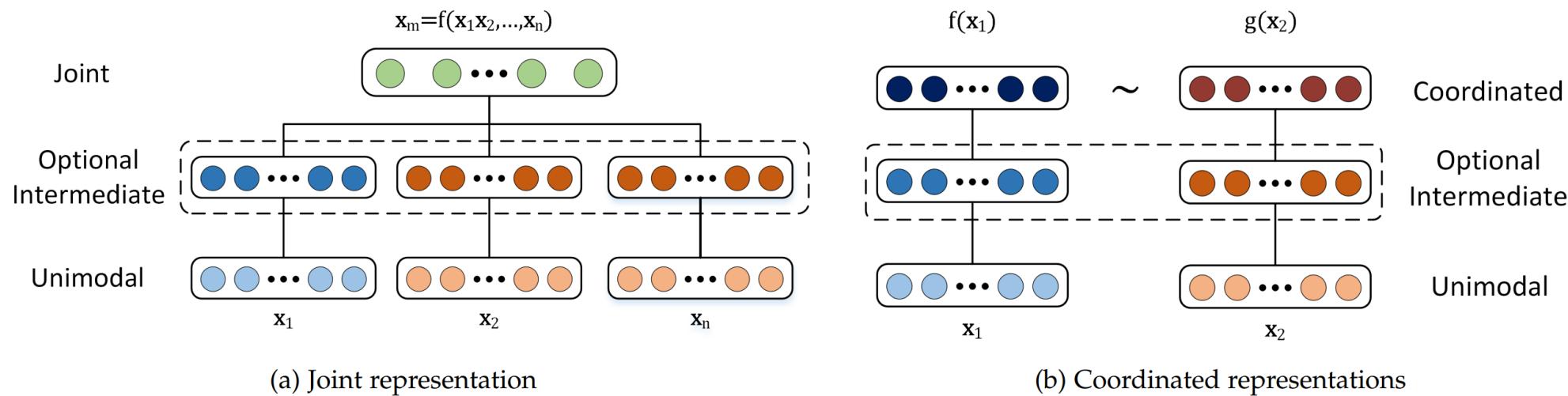
- Louis-Philippe Morency
- 卡内基梅隆大学计算机学院语言技术研究所，副教授



挑战问题和研究内容

□ 多模态学习：表征 (Representation)

- 研究多模态数据的表示方式，使其可以：（1）充分利用模态的互补性；
（2）尽可能消除冗余。
- 表示空间的相似关系应如实反映概念空间的相似性；
- 即使部分模态的信息缺失仍然容易获取表示；
- 根据已知模态的信息可填充或推算缺失模态的表示。



联合表示深度融合模态信息，但对缺失模态敏感。协调表示处理缺失模态灵活，但难以捕获深层交互，且需额外对齐。

挑战问题和研究内容

□ 多模态学习：对齐 (Alignment)

- 研究如何在多个模态中寻找并确定不同模态内子元素的直接对应关系。

显式对齐方法

以对齐为优化目标，核心问题是定义和计算相似性

- 无监督多模态对齐：以预设的序列关系作为约束条件

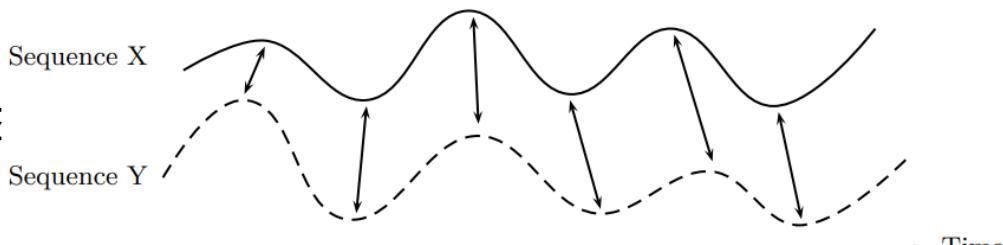
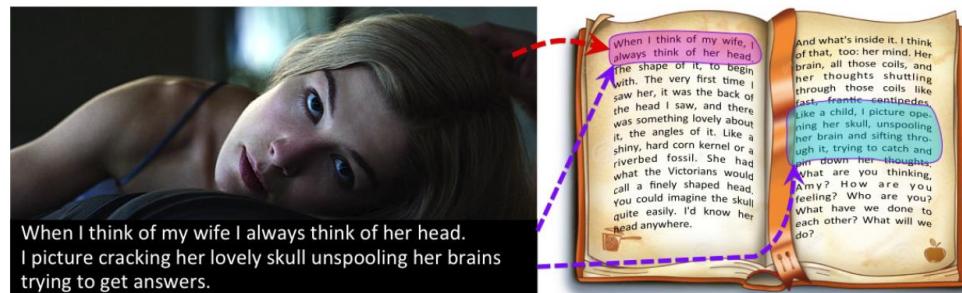


Fig. 4.1. Time alignment of two time-dependent sequences. Aligned points are indicated by the arrows

- 监督/弱监督多模态对齐：以全部/部分子元素对作为监督信号

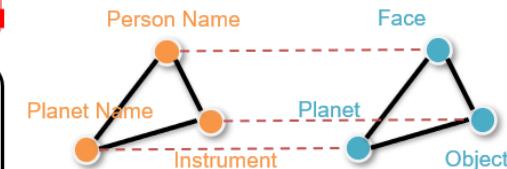
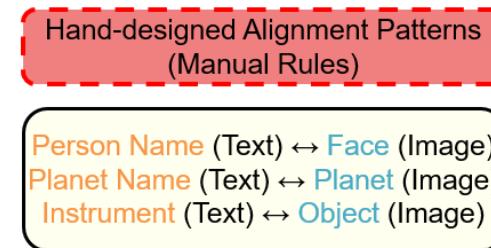


视觉和文本语义对齐

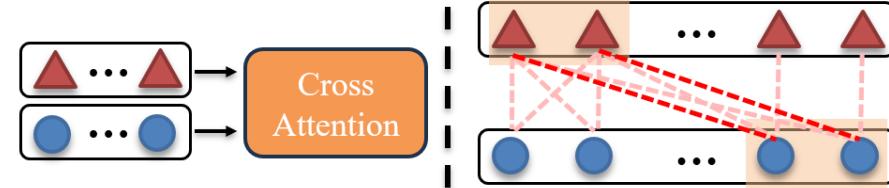
隐式对齐方法

对齐作为下游任务的中间步骤出现

- 基于图模型的多模态对齐：需手工设计子元素对齐模式



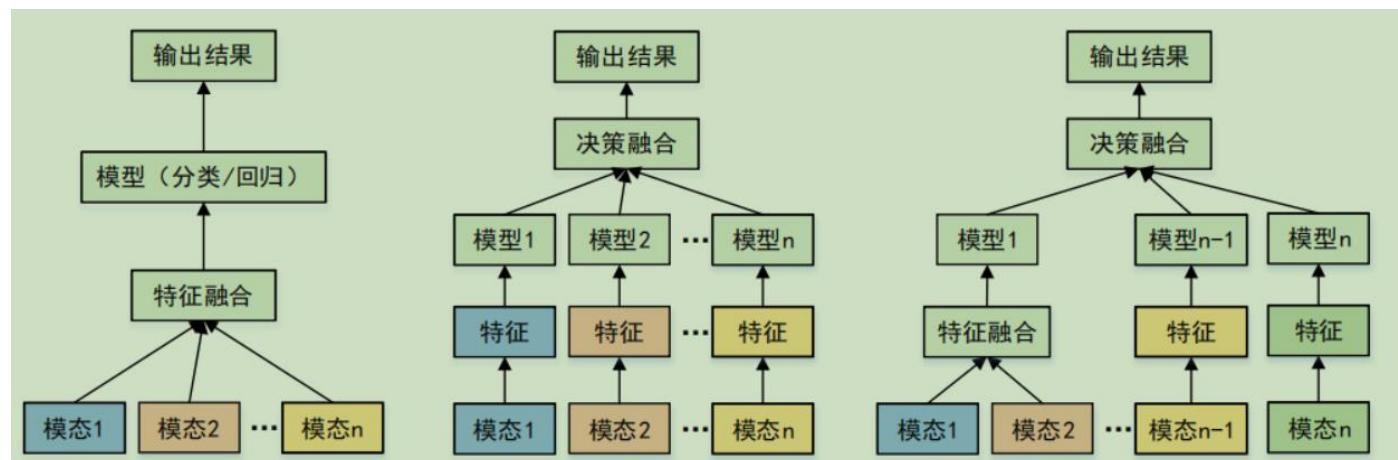
- 基于神经网络的多模态对齐：一般基于注意力机制实现对齐



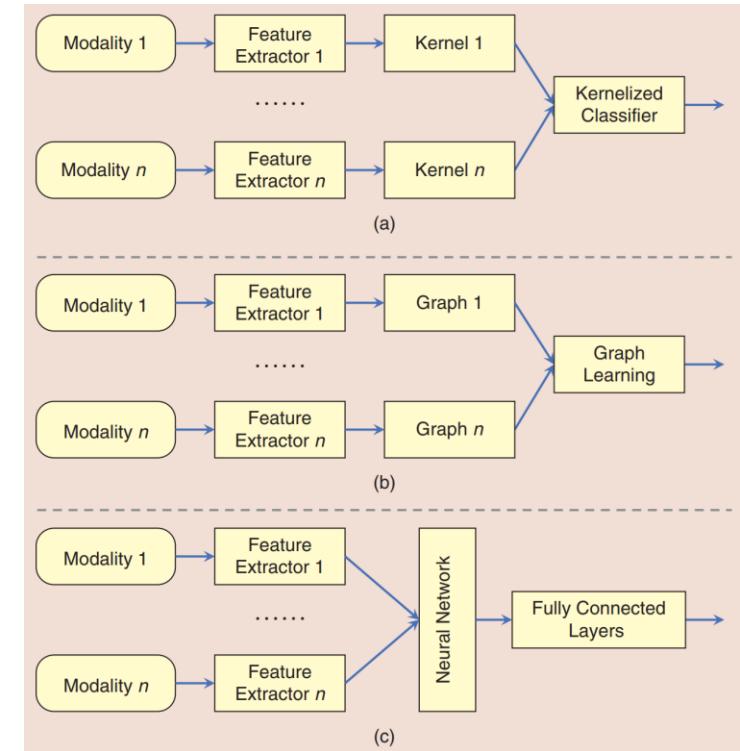
挑战问题和研究内容

□ 多模态学习：融合 (Fusion)

- 研究如何将不同模态的信息融合在一起以获得更准确的标签或连续值预测。
- 模型无关的融合
 - 前期融合（特征级融合）
 - 后期融合（决策级融合）
 - 混合式融合
- 模型相关的融合
 - 基于多核学习的融合
 - 基于图模型的融合
 - 基于神经网络的融合



模型无关融合：优点是通用性强，但可能丢失深层交互。

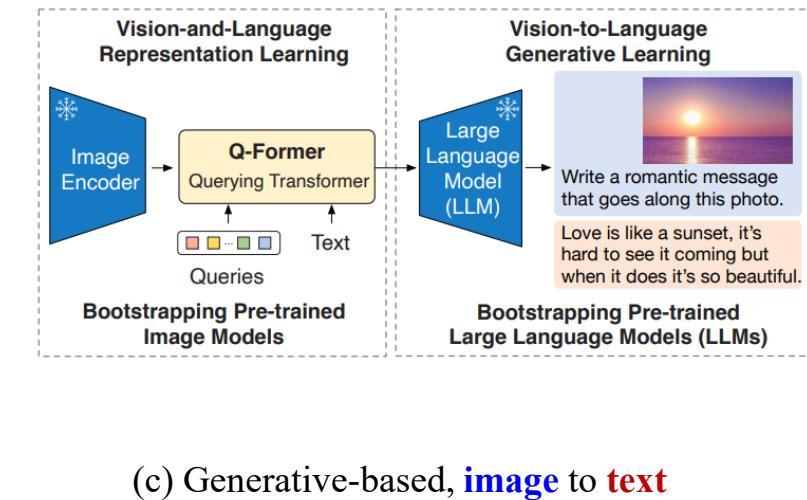
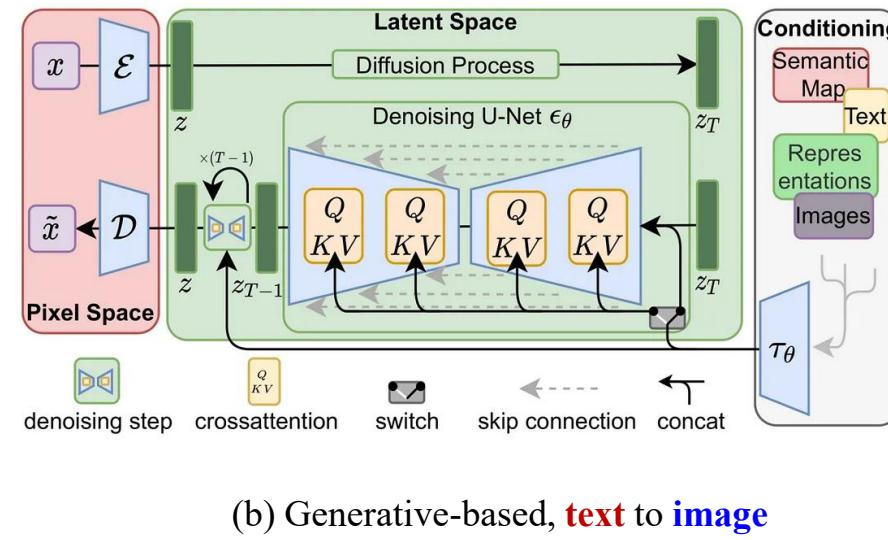
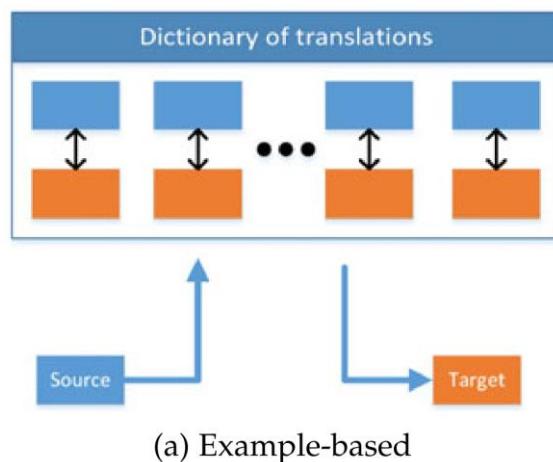


模型相关融合：优点是能深度捕获交互，
但通用性差，依赖特定模型。

挑战问题和研究内容

□ 多模态学习：翻译（Translation）

- 研究如何将数据从一种模态翻译（映射）到另一种模态。
- 基于实例的方法：（1）基于检索的方法；（2）基于检索结果混合的方法。
- 基于生成的方法：（1）基于语法的生成；（2）基于编解码生成；（3）连续生成。



挑战问题和研究内容

□ 多模态学习：协同学习 (Co-learning)

- 研究如何从一个资源丰富的模态及其表示和预测模型向另一个资源匮乏的模态迁移知识。
- 并行式协同学习：源域和目标域可直接映射
 - 协同训练：通过少量共存标签学习生成更多
 - 迁移学习
- 非并行协同学习：不可直接映射
 - 迁移学习
 - 概念限定
 - 零样本学习
- 混合式协同学习：存在中间模态

数据噪声 标签缺失或不可靠

面部区域 肌肉运动 副词

The inner corners of the eyebrows are lifted slightly, the skin of the glabella and forehead above it is lifted slightly and wrinkles deepen slightly and a trace of new ones form in the center of the forehead;



激活AU1的人脸图像

AU1的文本描述

汇报提纲

- 多模态情感分析-问题定义与研究内容
- 多模态情感识别-研究背景及核心挑战
- 课题组相关进展-单模态、多模态情感识别研究进展
- 未来研究方向-大模型时代的多模态情感识别等

研究背景

口 人工智能的发展

计算智能

能存、会算



表示、计算、存储与
人机输入/出等

感知智能

能听、会说
能看、会认



文本内容识别、图像
识别、语音识别等

认知智能

能理解、会思考
有情感

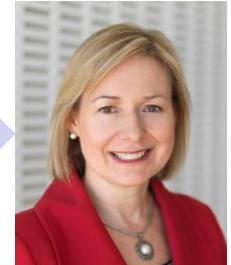


知识数据双驱动、认
知推理、决策智能等

研究背景

□ 国际前沿研究

□ 美国工程院院士、IEEE Fellow、麻省理工学院Rosalind Picard教授
如果机器不具有感知和表达情绪的能力，那么它就无法通过图灵测试
，也就意味着不具有真正意义上的智能。



□ 美国工程院院士、ACM/AAAS Fellow、斯坦福大学李飞飞教授



下一步人工智能的发展，需要加强对情感、情绪的了解。
情绪、情感，是人工智能未来的方向。

心智世界模型赋予AI代理理解人类内心状态的能力，是实现自然、
情感、智能人机交互的核心机制。构建“心智世界模型”(*mental world model*)”，包括对用户情绪和情感状态的建模，是未来人
机协作的关键能力之一。



研究背景

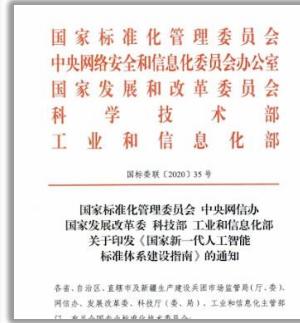
□ 情智兼备是国内外科技产业的重大需求

国家《新一代人工智能发展规划》



- 开发面向老年人的**情感陪护助手**
- 开发具有**情感交互功能**的智能助理产品

国家新一代人工智能标准体系建设指南

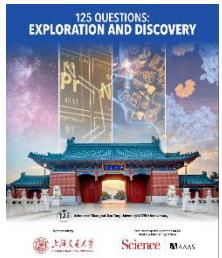


- **情感分析**是自然语言领域的**重点建设标准**
- 表情识别、情感交互人机交互领域的**重点建设标准**

2021年

Science: 125个科学问题

人工智能领域**第二项问题**: 探索与发现



是否有可能创建**有感知力的和有情感的机器人**

2024年

英国国家创新署《2040将对英国经济社会产生重大影响的50项新兴技术》第一项技术



人工智能识别情绪和表情，
开发能够识别和解释人类情感的人工智能技术

2024重大科学问题解读

情智兼备
数字人与机器人的研究

数字人

指通过计算机软件模拟出的具有人类特征和行为的虚拟实体，具有自己的思维、情感和个性。

机器人

包含机械结构、传感器、执行器和控制系统等多个部分的智能化系统。

2024重大科学问题之首
情智兼备数字人与机器人的研究

研究背景

口 行为模态情感分析相关数据集（部分）

数据库名称	模态	数据形式	数据规模	标注类型	采集机构	发布时间	发展趋势：
IEMOCAP [1]	图像、文本、音频	视频片段	10039	情感类别	南加州大学	2008	➤ (规模) 从小到大
ICT-MMMO [2]	图像、文本、音频	视频片段	370	情感类别	卡内基梅隆大学	2013	➤ (标注) 由粗到精
MSP-IMPROV [3]	图像、文本、音频	视频片段	8438	情感类别	德克萨斯大学达拉斯分校	2016	
CMU-MOSI [4]	图像、文本、音频	视频片段	2199	情感强度	卡内基梅隆大学	2016	
CHEAVD [5]	图像、音频	视频片段	7030	情感类别	中科院自动化所	2017	
CMU-MOSEI [6]	图像、文本、音频	视频片段	23453	情感强度 情感类别	卡内基梅隆大学	2018	
MELD [7]	图像、文本、音频	视频片段	13708	情感类别	新加坡国立大学	2018	
CH-SIMS [8]	图像、文本、音频	视频片段	2281	情感类别	清华大学	2020	
M ³ ED [9]	图像、文本、音频	视频片段	24449	情感类别	中国人民大学	2022	
MER2023 [10]	图像、文本、音频	视频片段	78178	情感强度 情感类别	中科院自动化所	2023	
MERR [11]	图像、文本、音频	视频片段	33105	情感类别 情感描述	深圳技术大学	2024	
EMER-Coarse [12]	图像、文本、音频	视频片段	115595	情感类别 情感描述	中科院自动化所	2024	

研究背景

口 生理信号情感分析相关数据集 (部分)

Database	Modality	Data	Emotion Annotation	Institute	Year	Publication Name
MAHNOB-HCI[1]	EEG, ECG, GSR, SKT, EOG, respiration, face body video, Audio	27 subjects;	9 classes, VA,	University of Geneva, Switzerland	2011	TAC
DEAP[2]	EEG, EOG, EMG, GSR, BVP, SKT, Respiration, Face Video	32 subjects;	VA liking, dominance, familiarity	Queen Mary University of London	2012	TAC
RECOLA[3]	ECG, GSR, Audio, Face Video	46 subjects;	VA	Université de Fribourg, Switzerland	2013	FG conference
DECAF[4]	MEG, EOG, ECG, EMG, Face Video	46 subjects	VA, dominance	University of Trento, Italy	2015	TAC
BP4D+[5]	ECG, GSR, BVP, Respiration, Face Video, Thermal	140 subjects	10 emotions, AU	Binghamton University, USA	2016	CVPR
DREAMER[6]	EEG, ECG	23 subjects	VA, dominance	University of the West of Scotland, UK	2018	IEEE JBHI
AMIGOS[6.1]	EEG, ECG, GSR, Audio, Video, Depth	40 subjects	VA, personality traits	Queen Mary University of London	2018	TAC
SEED-V[7]	EOG; ECG	20 subjects;	5 classes	上海交通大学	2019	Conference on Neural Engineering
CASE[8]	ECG, BVP, EMG, GSR, Respiration, Skin Temperature	30 subjects	VA	Institute of Robotics and Mechatronics, DLR, Germany	2019	Scientific data
BU-EEG[9]	EEG; Face Video	29 subjects	7 classes, AU, pain	Binghamton University, USA	2020	FG conference
MGEED[10]	Images, Depth; OMG, EEG, ECG	17 subjects; 150K facial images;	6 emotions, VA	University of Portsmouth, United Kingdom	2023	TAC
Mixed-ER [11]	EEG, Face Video, GSR, PPG	73 subjects	3 emotions	清华大学	2024	Scientific Data

发展趋势:

- EEG, ECG, GSR
- 行为与生理信号耦合

生理信号类别	英文名称	英文缩写
脑电图	Electroencephalogram	EEG
肌电图	Electromyogram	EMG
心电图	Electrocardiogram	ECG
眼电图	Electrooculogram	EOG
心率变异性	Heart rate variability	HRV
皮肤电反应	Galvanic skin response	GSR
皮肤电应答	Electrodermal response	EDR
皮肤电活动	Electrodermal activity	EDA
血压信号	Blood pressure	BP
皮肤温度	Skin temperature	ST
呼吸模式	Respiration pattern	RSP
光电容积脉搏波	Photoplethysmogram	PPG
眼动信号	Eye movement	EM
脉搏信号	Pulse rate	PR
血氧饱和度	Oxygen saturation	SpO2

摘自《基于生理信号的情感计算研究综述》,
自动化学报, 2021

多模态情感计算-情感模态

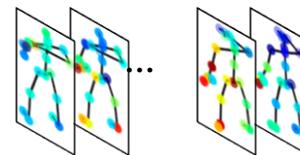
□ 情感计算多模态数据-显性情感线索

- 人脸表情：一个或多个脸区域/单元的孤立运动或运动组合
- 眼球运动：眼睛是心灵的窗户
- 语言语音：说话者通过使用不同文字、语调、声音大小和节奏来表达他们的意图
- 行为：将紧握的拳头推到空中，通常被视作表达胜利或欣喜的姿势
- 步态：与悲伤和满足等低激活度情感相比，愤怒和兴奋等高激活度情感与快速运动更相关

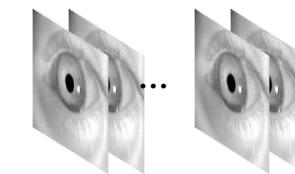
- 脑电
- 心电
- 体温
- 脉搏
-



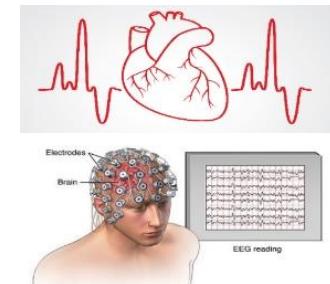
人脸表情



人体动作骨架点



眼动信号



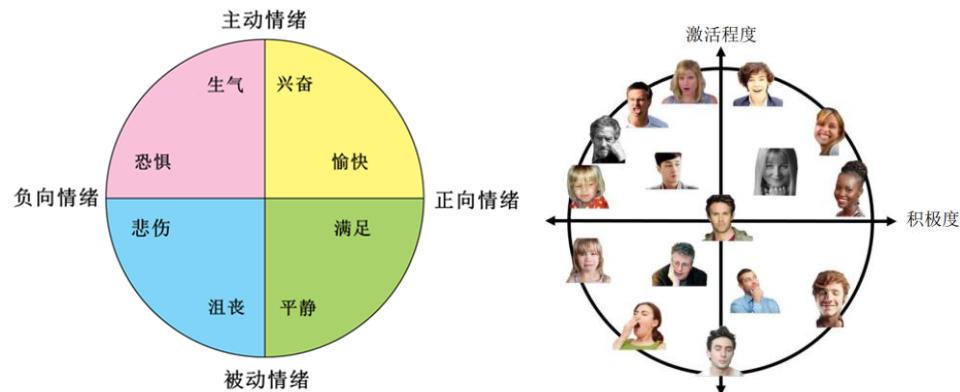
脑电信号

多模态情感识别-情感定义

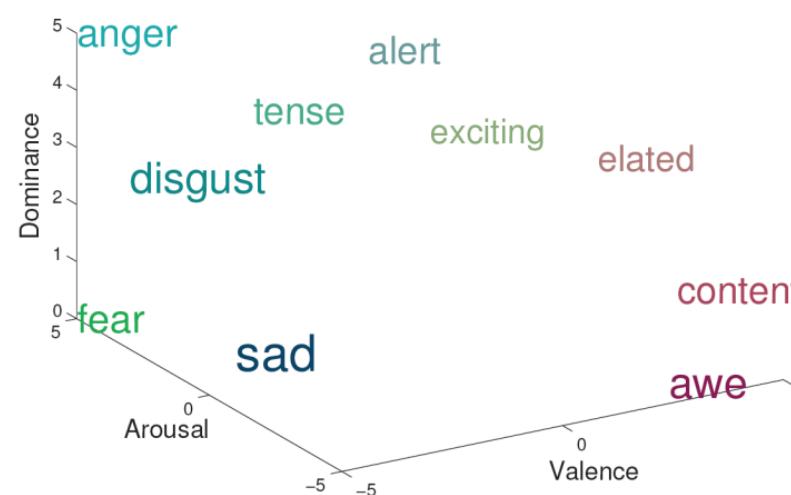
□ 心理学对情感没有统一、严格的定义，多采用定性的分析方法。情感类别越来越多样化和细粒度。

□ 心理学情感模型

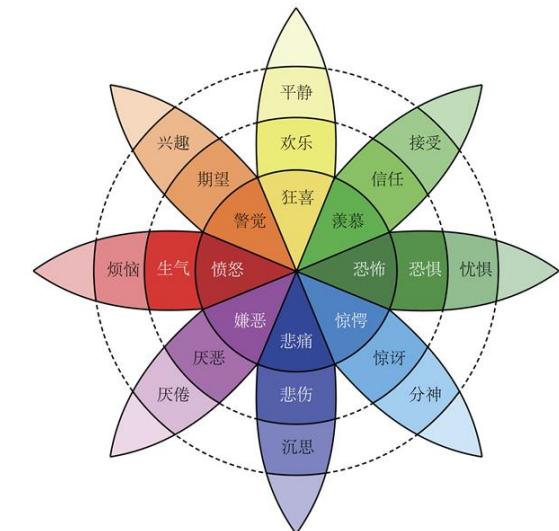
- 离散情感状态：Ekman六类（高兴、悲伤、恐惧、厌恶、愤怒、惊讶）
- 连续情感空间：二维情感模型（愉悦度（Valence）和激活度（Arousal））



二维情感模型



三维情感模型



情绪轮模型

离散情感直观易懂但缺乏细微度，难表强度；连续情感捕捉细节和强度但标注复杂，理解不直观，维度选择有争议。

多模态情感识别-情感定义

□ 心理学情感模型

Model	Type	Emotion states/dimensions
Ekman	CES	happiness, sadness, anger, disgust, fear, surprise
Mikels	CES	amusement, anger, awe, contentment, disgust, excitement, fear, sadness
Plutchik	CES	(× 3 scales) anger, anticipation, disgust, joy, sadness, surprise, fear, trust
Parrott	CES	a tree hierarchical grouping with primary, secondary and tertiary emotion categories
Sentiment	CES	positive, negative, (and neutral)
VA(D)	DES	valence-arousal(-dominance)
ATW	DES	activity-temperature-weight

- 离散情感类别 (CES): 便于用户理解和标注, 但描述能力有限
- 连续情感空间 (DES): 描述能力强, 但不易于理解

	CES	DES
understandability	easy	difficult
describability	limited	unlimited
perspective	qualitative	quantitative
examples	Mikels, Plutchik	VAD
granularity	coarse-grained	fine-grained
AICA tasks	classification, retrieval	regression, retrieval

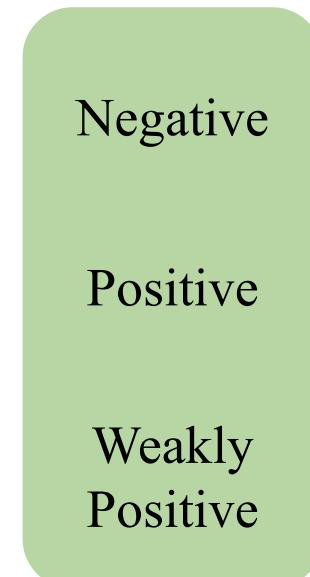
多模态情感识别-优势

□ 数据互补性

- 来自不同模态的线索可以相互增强或补充。例如，如果我们看到一个好朋友的帖子：“今天天气真好！”，那么这个朋友很有可能是在表达一种积极情感；但是如果还配有一张暴风雨的图片，我们就能推断出这段文字实际上是一种反讽，在表达一种消极情感。



I was moved to tears.

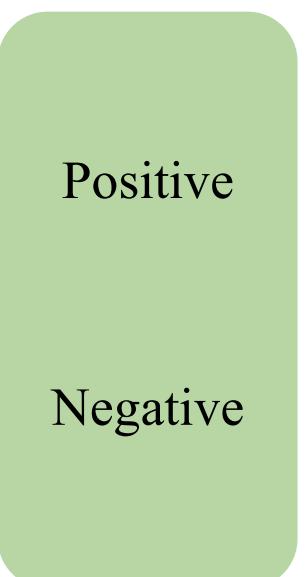


单模态情感 多模态情感

今天天气真好！



Justin
Sullivan/Getty
Images

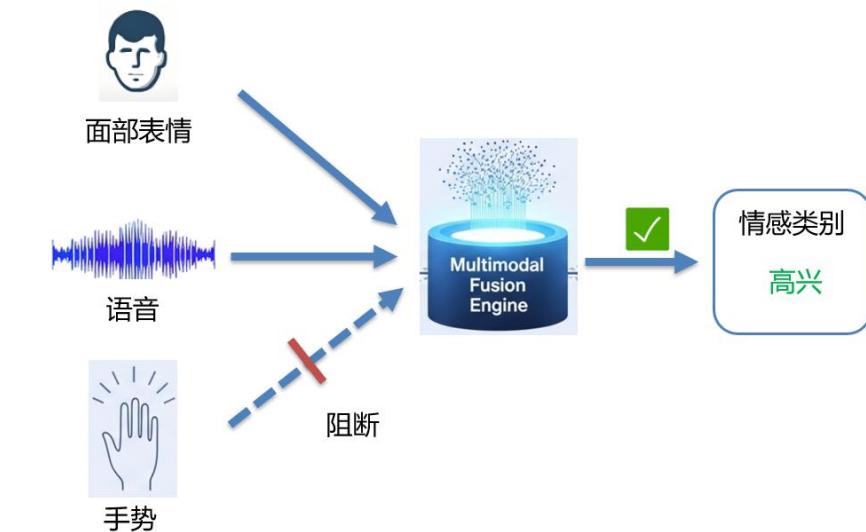


单模态情感 多模态情感

多模态情感识别-优势

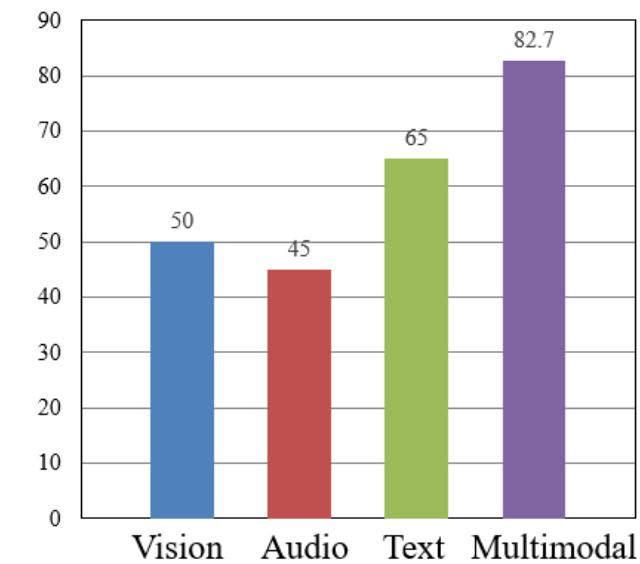
□ 模型鲁棒性

- 数据采集过程可能受突发因素的影响，如传感器设备故障，造成一些数据模态无法使用，这在非实验室场景尤其普遍。



□ 性能优越性

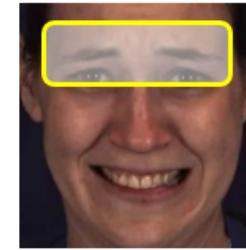
- 联合考虑不同模态的互补信息能带来更好的识别性能。与最优先单模态相比，多模态情感识别获得更好的性能提升。



核心挑战

□ 模态失衡

- 不同的模态可能对诱发情感有不同程度的贡献。如一篇在线新闻可能文字长度很长，包含很多详细信息，但只有一两张插图。



激活AU1的人脸图像

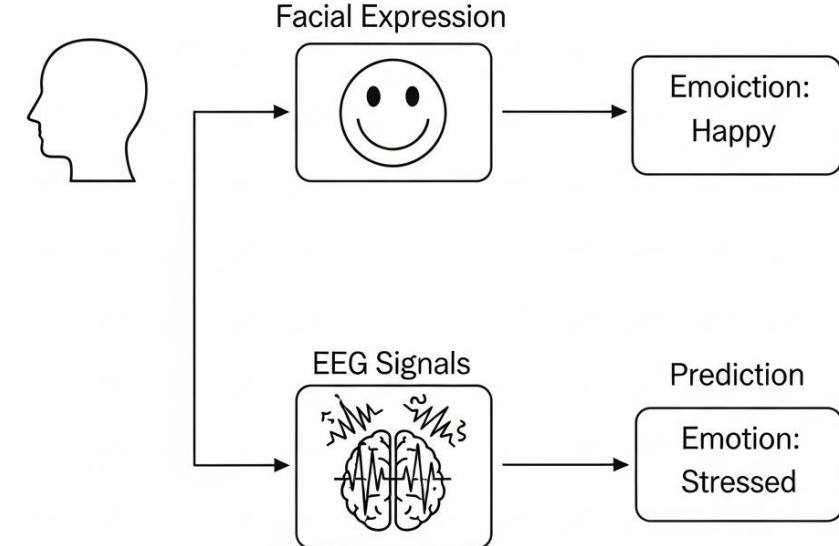
面部区域 肌肉运动 副词

The inner corners of the eyebrows are lifted slightly, the skin of the glabella and forehead above it is lifted slightly and wrinkles deepen slightly and a trace of new ones form in the center of the forehead;

AU1的文本描述

□ 模态冲突/不一致

- 例如人脸表情和语言很容易被抑制或隐藏以逃避检测，但由中枢神经系统控制的 EEG 信号可以反映人类无意识的身体变化。



核心挑战

□ 数据缺失

- 脑电图传感器可能会记录到有噪声的信号，甚至无法记录到任何信号；摄像机在夜间无法捕捉到清晰的人脸表情；用户可能会发布一条只包含图片（没有文字）的推文。

□ 标签缺失与噪声

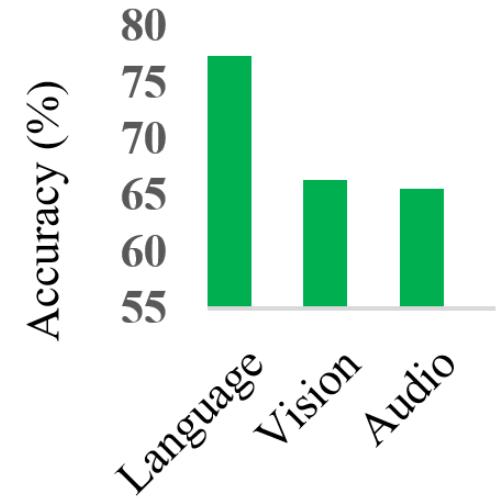
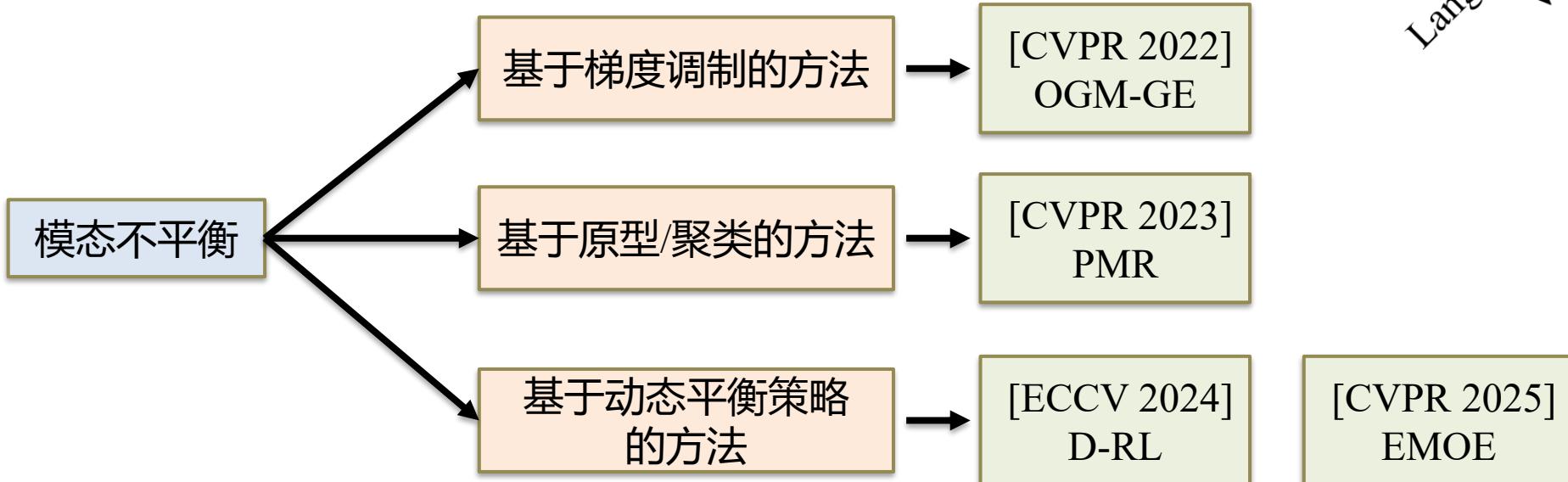
- 拥有大量的数据，却只有很少甚至没有情感标签的问题。随着情感需求的日益多样化和细粒度，可能某些情感类别有足够的训练数据，其他情感类别却没有（在情感类别多样化细粒度的情况下更是如此）。
- 一种替代人工标注的解决方案是利用社交推文的标签或关键词作为情感标签，但这种标签是不完整的、有噪声的。

多模态不平衡

□ 多模态不平衡问题：模型偏向优势模态，抑制弱势模态

- 模型过度依赖信息丰富的优势模态（如文本）
- 忽视或抑制音频、视觉等弱势模态
- 难以充分发挥模态互补性，限制模型性能上限

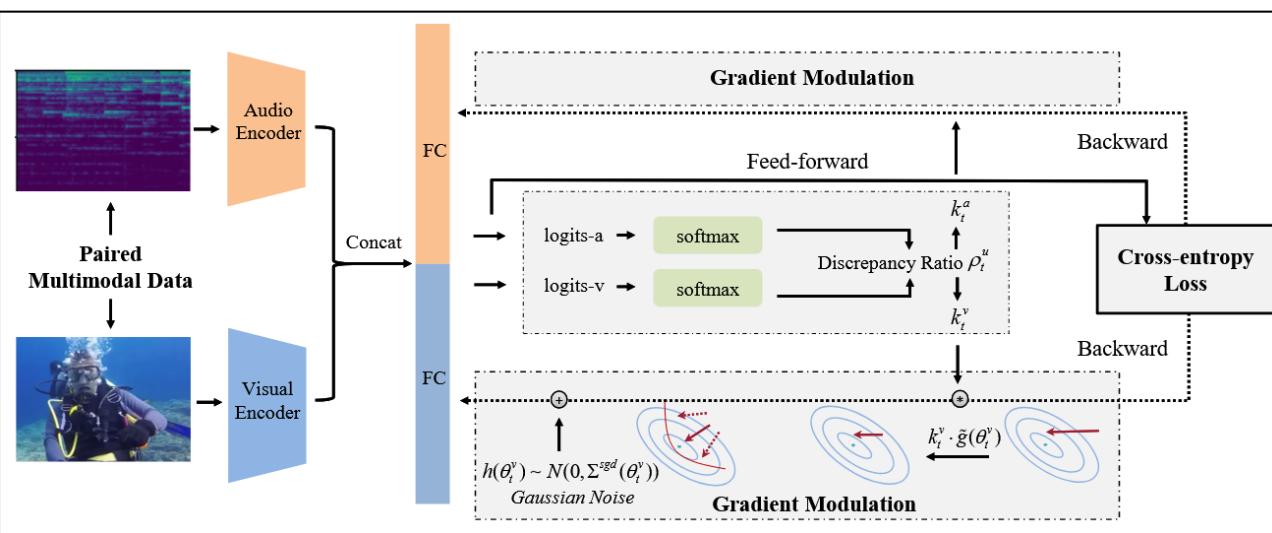
□ 现有方法



多模态不平衡

□ 基于梯度调制的方法

- 通过动态监测不同模态对当前学习目标的贡献差异**自适应调整不同模态的梯度**，从而实现多模态平衡学习。

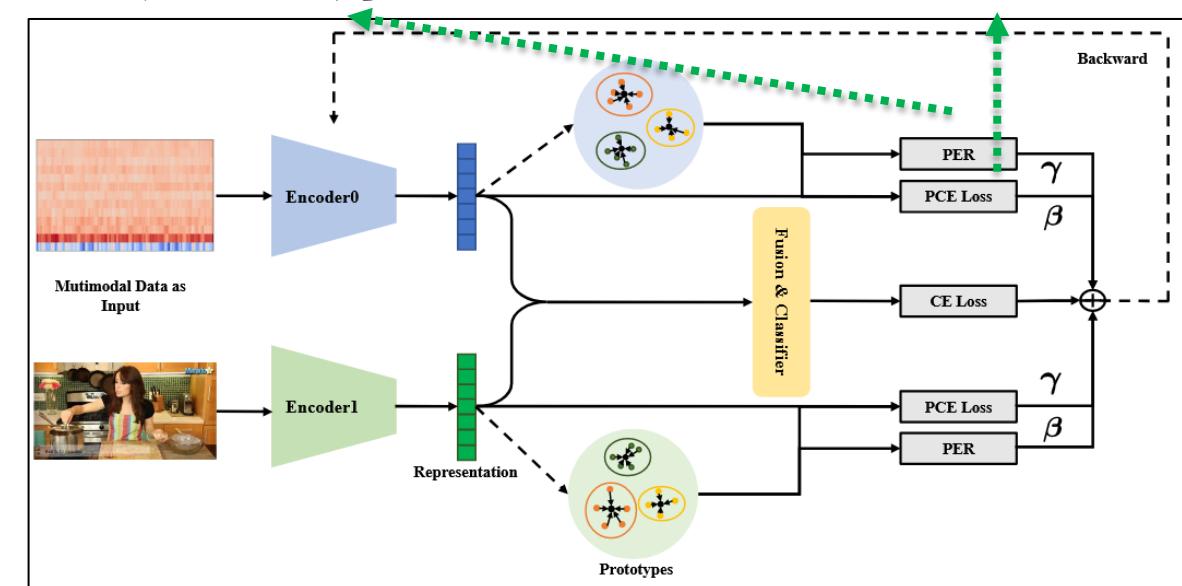


[CVPR 2022] OGM-GE

□ 基于原型的方法

- 通过引入**原型分类损失**，鼓励模态特征向对应的类别原型聚拢，加速弱势模态的学习；同时采用**原型熵正则化**抑制主导模态过快收敛。

原型熵正则化



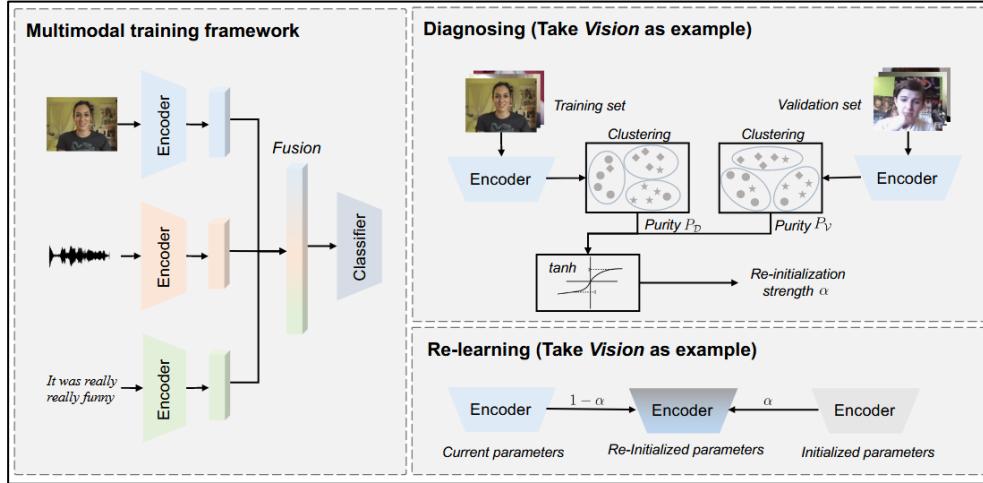
[CVPR 2023] PMR

[1] Xiaokang Peng et al. Balanced multimodal learning via on-the-fly gradient modulation, **CVPR 2022**.

[2] Yunfeng Fan et al. PMR: Prototypical Modal Rebalance for Multimodal Learning, **CVPR 2023**.

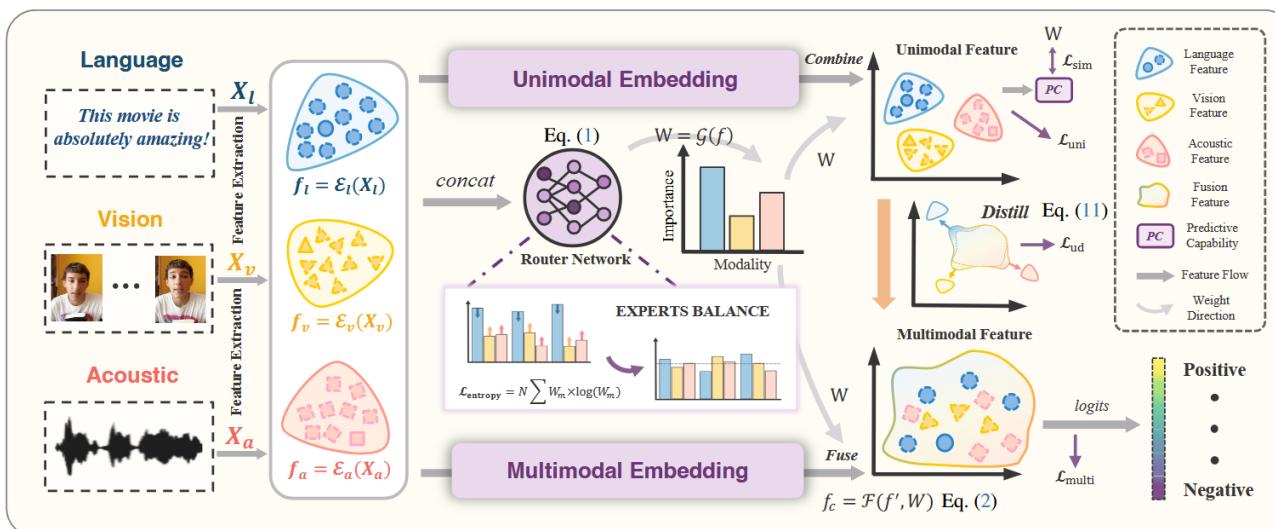
多模态不平衡

□ 基于动态平衡策略的方法



□ 通过周期性地评估每个模态的真实学习状态，并基于评估结果自适应调整编码器参数

D-RL [ECCV 2024]



□ 通过构建路由网络为每个样本的不同模态动态分配权重，实现自适应融合，并引导选择性知识蒸馏。

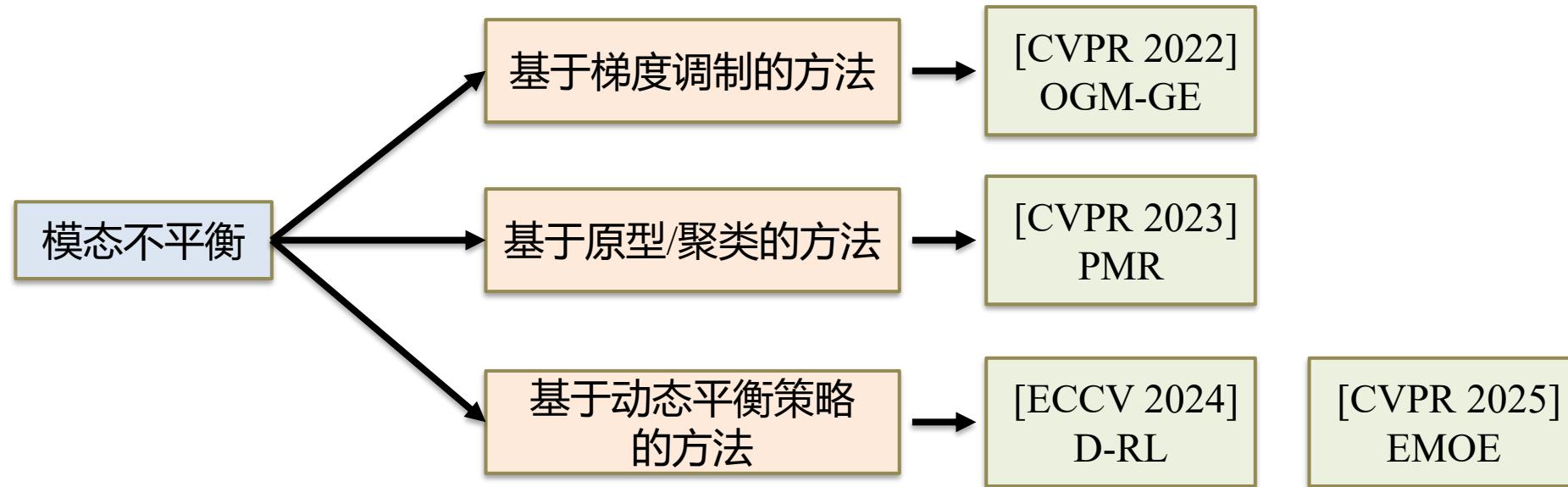
EMOE [CVPR 2025]

[1] Yake Wei et al. Diagnosing and Re-learning for Balanced Multimodal Learning, [ECCV 2024](#).

[2] Yiyang Fang et al. EMOE: Modality-Specific Enhanced Dynamic Emotion Experts, [CVPR 2025](#).

多模态不平衡

口 多模态不平衡问题：模型偏向优势模态，抑制弱势模态



过度追求模态间的“绝对平衡”可能导致削减多模态互补性，损害模型整体性能；
性能、平衡度和计算成本三者间难以取得的完美平衡。

多模态语义对齐

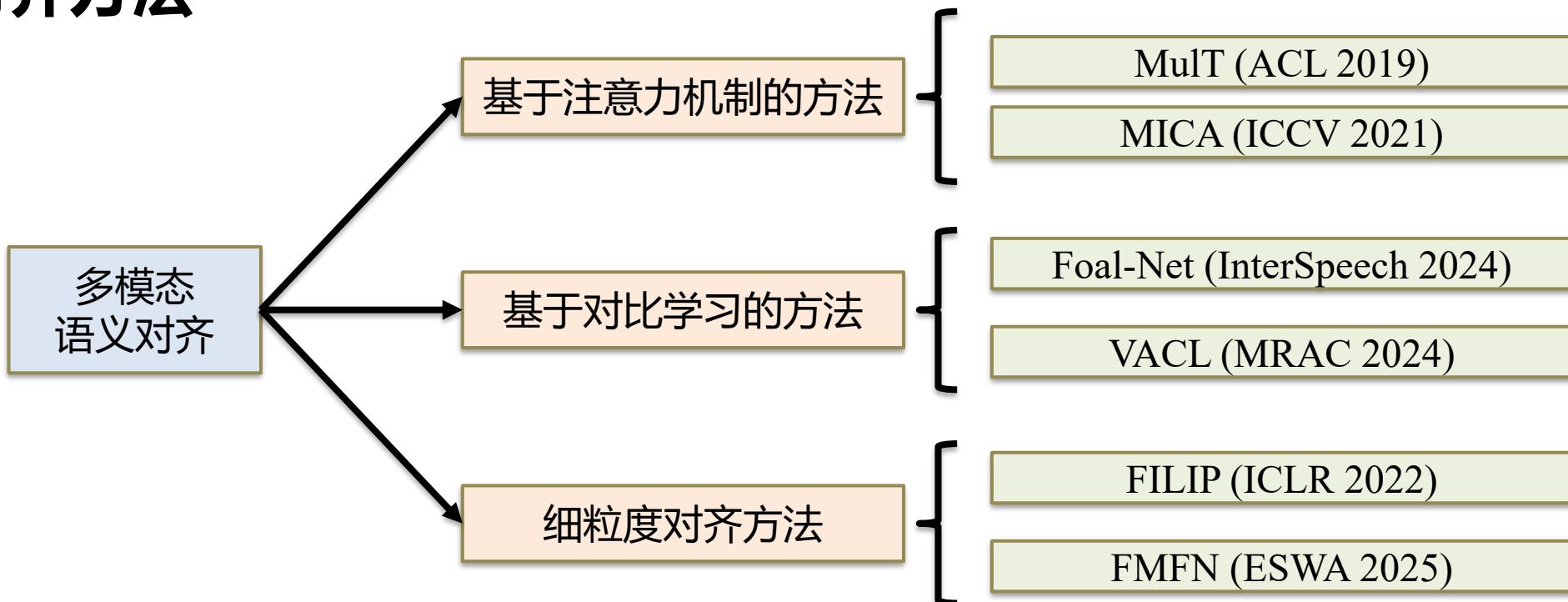
□ 多模态语义对齐

- 在不同模态的元素之间建立情感语义层面的对应关系

□ 挑战

- 如何有效挖掘异构模态中的情感线索，并实现语义层面的跨模态对齐

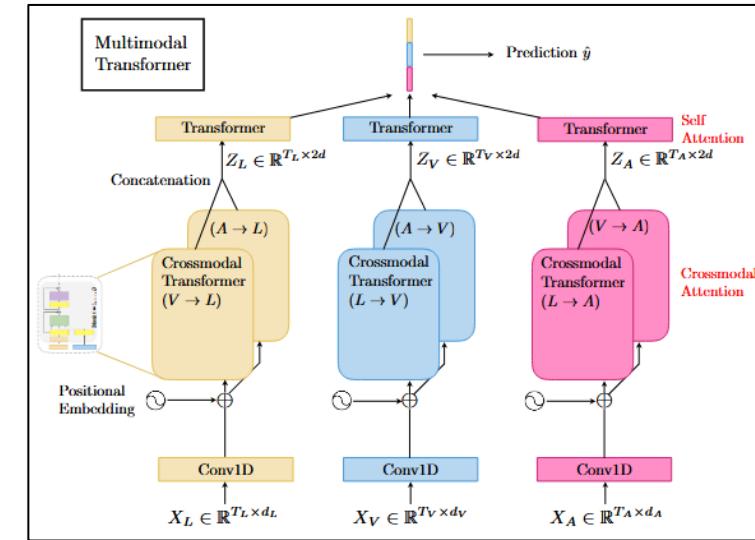
□ 语义对齐方法



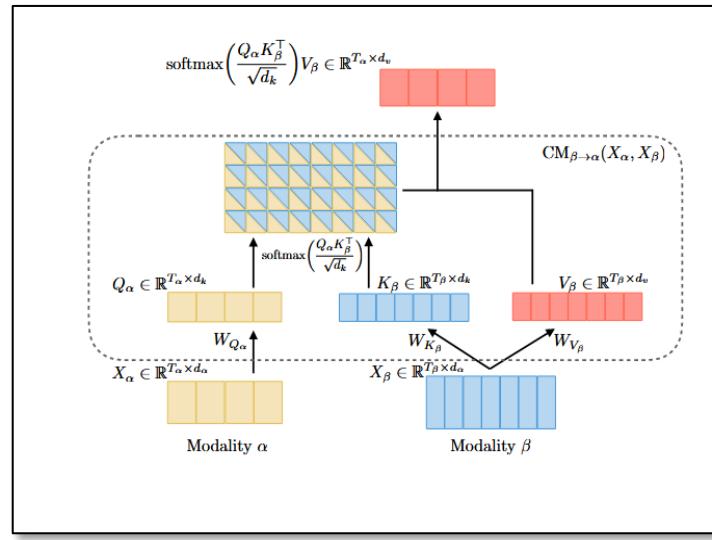
多模态语义对齐

□ 基于注意力的方法

□ 通过引入跨模态注意力机制，建模不同模态间异步的跨模态语义关联。

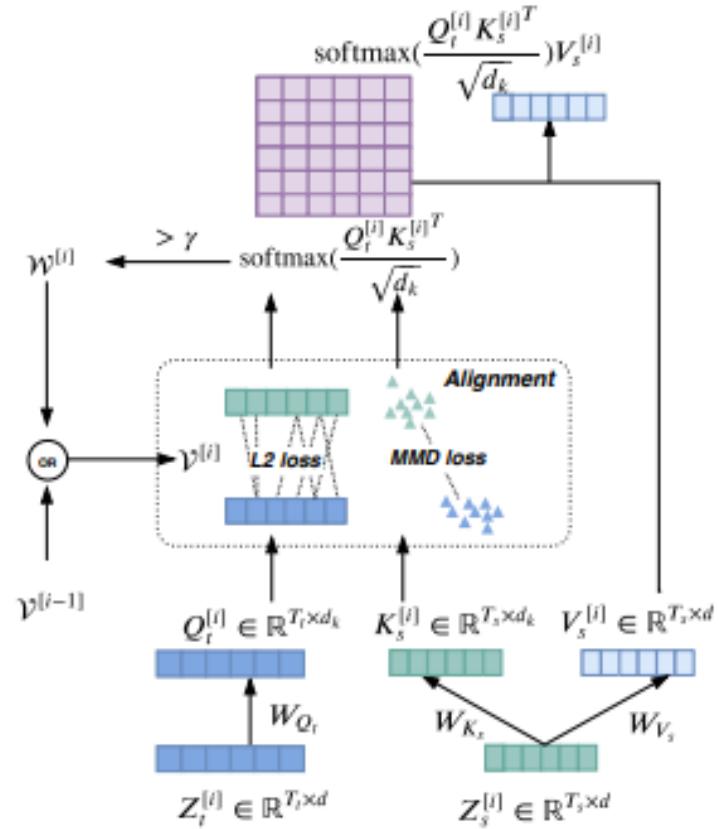


MuLT (ACL 2019)



跨模态注意力：从模态 β 到模态 α

□ 在分布和特征级别对齐来自不同模态的Query和Key特征，以学习到更可靠、准确的跨模态依赖/对齐关系



MICA (ICCV 2021)

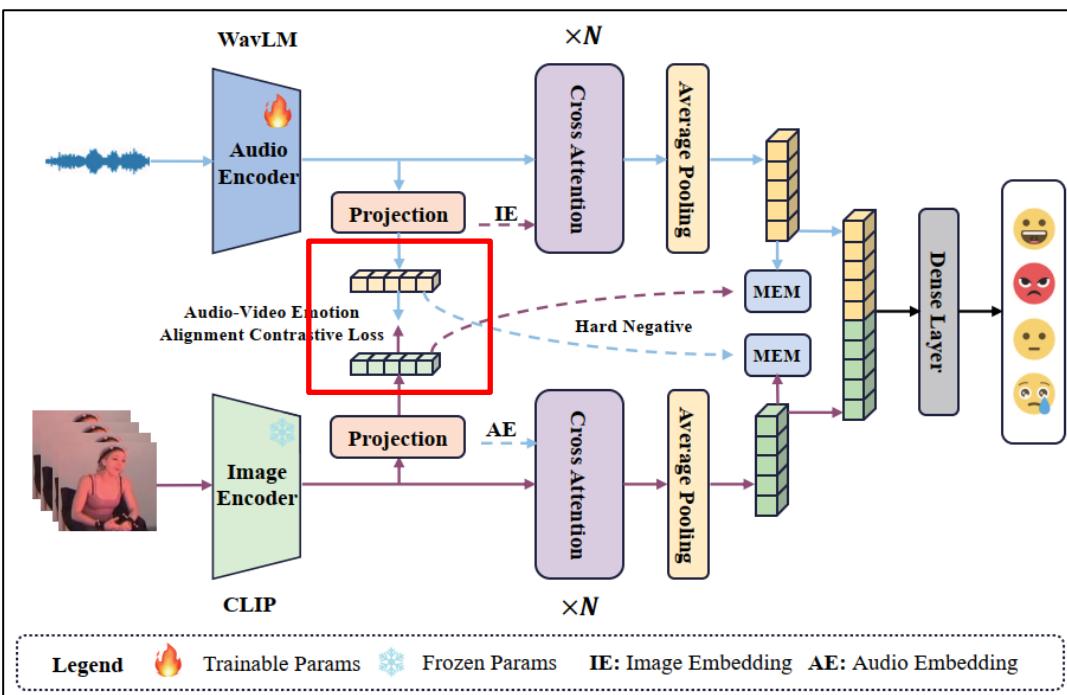
[1] Yao-Hung Hubert Tsai et al. Multimodal Transformer for Unaligned Multimodal Language Sequences, *ACL 2019*.

[2] Tao Liang, et al. Attention is not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion. *ICCV, 2021*.

多模态语义对齐

□ 基于对比学习的方法

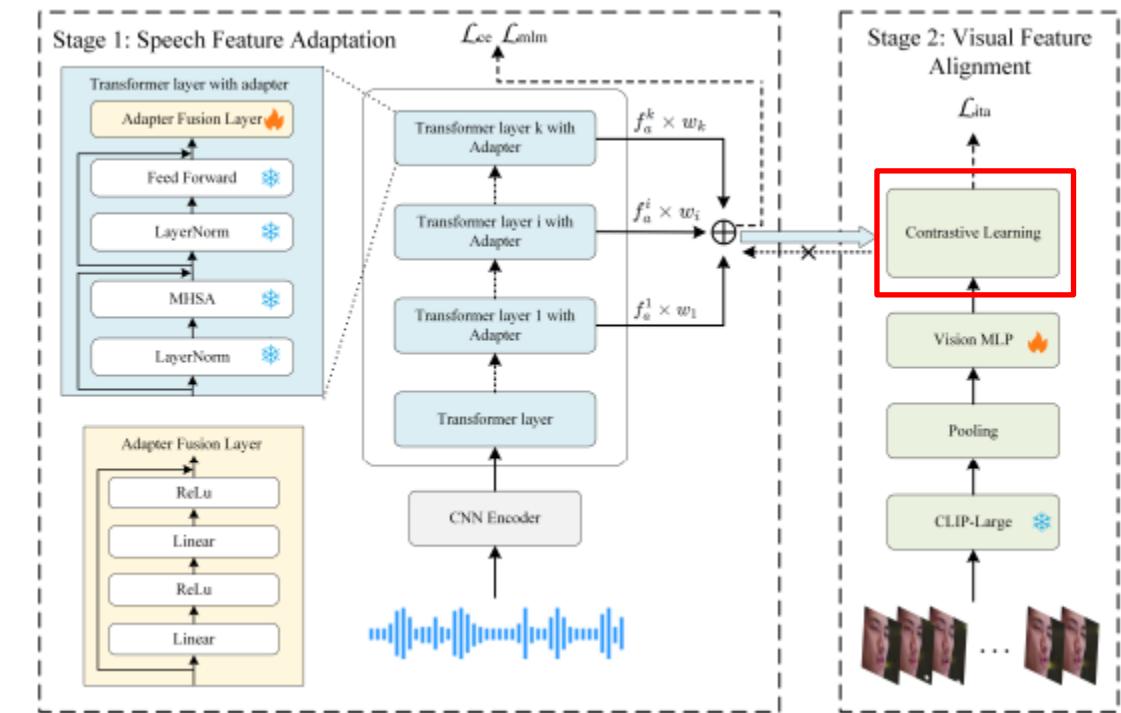
- 通过音频和视频间的对比学习，显示增强跨模态的情感语义一致性



Foal-Net (InterSpeech 2024)

- [1] Qifei Li et al. Enhancing Modal Fusion by Alignment and Label Matching for Multimodal Emotion Recognition, *InterSpeech 2024*.
- [2] Zhixian Zhao, et al. Improving multimodal emotion recognition by leveraging acoustic adaptation and visual alignment. *MRAC, 2024*.

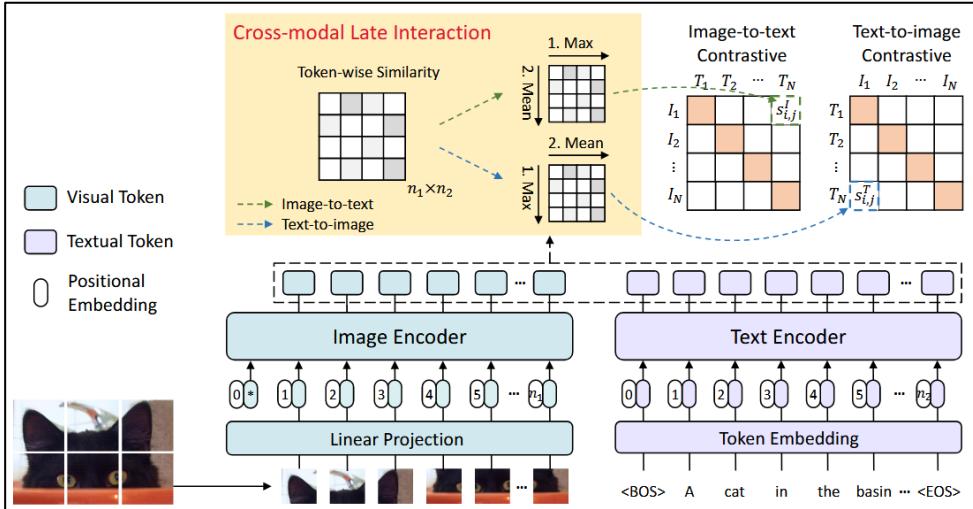
□ 采用对比学习，整将加权组合后的多层次音频特征与视觉特征进行对齐，以实现更精确的跨模态匹配。



VA-HCL (MRAC 2024)

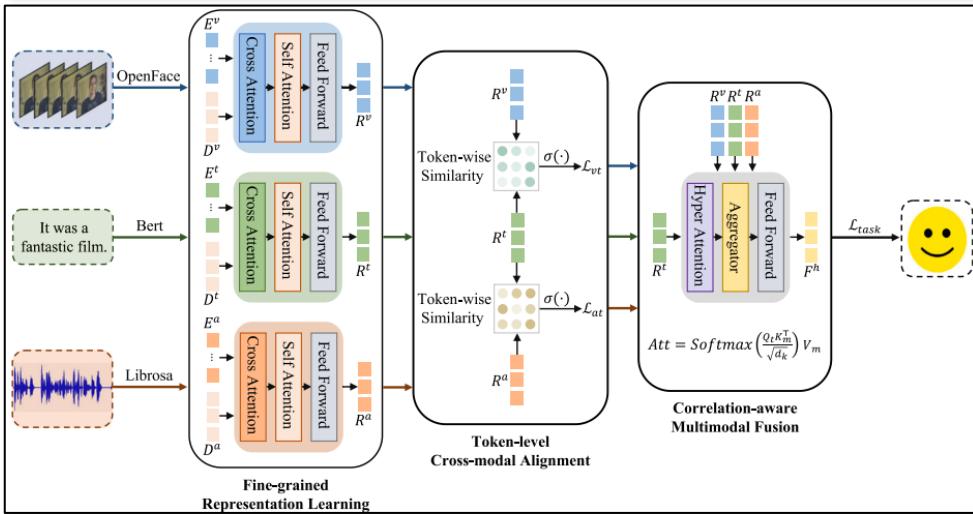
多模态语义对齐

□ 细粒度对齐方法



□ 通过跨模态交互机制，促进模型实现图像和文本间的**细粒度语义对齐**，增强其对物体属性与细节的感知能力。

FILIP (ICLR 2022)



□ 以文本模态为中心，通过**细粒度对比学习**将不同模态的特征都和文本特征对齐，获取更精炼、细粒度的情感表征。

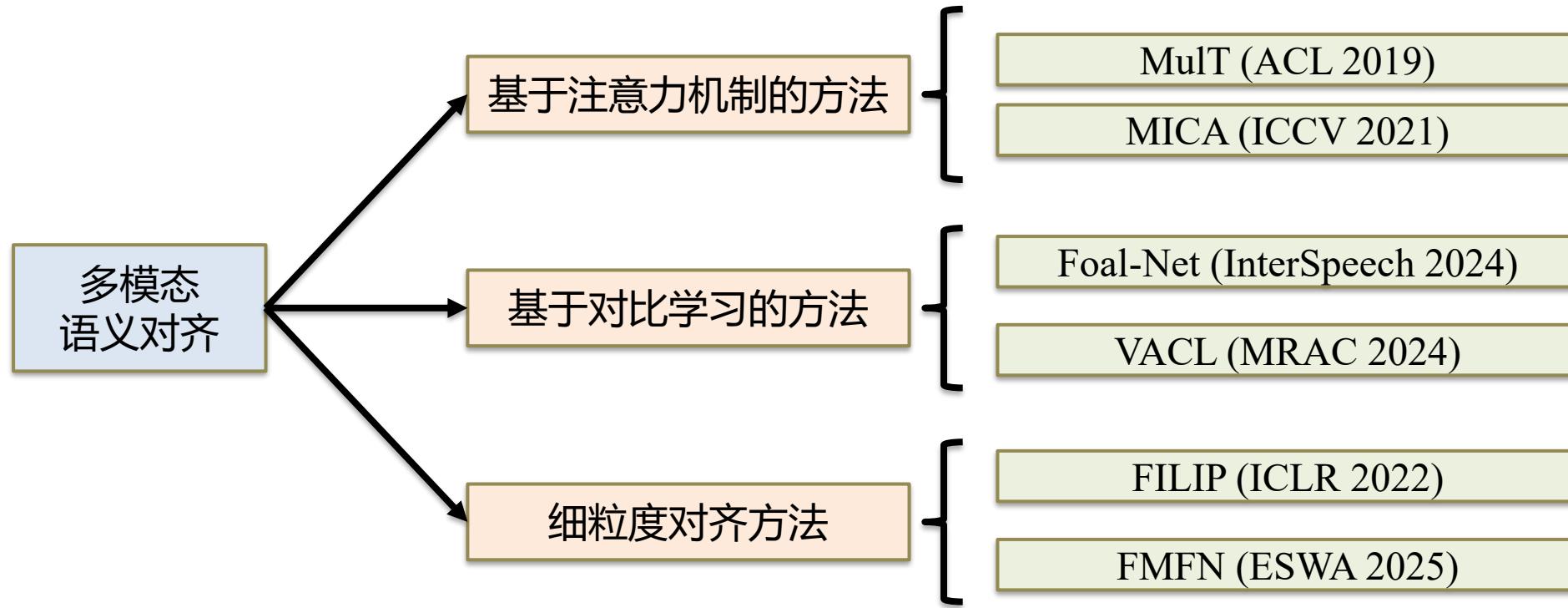
FMFN (ESA 2025)

[1] Lewei Yao et al. FILIP: Fine-grained interactive language-image pre-training, *ICLR 2022*.

[2] Xiang Li, et al. Learning fine-grained representation with token-level alignment for multimodal sentiment analysis, *ESA 2025*.

多模态语义对齐

□ 多模态语义对齐



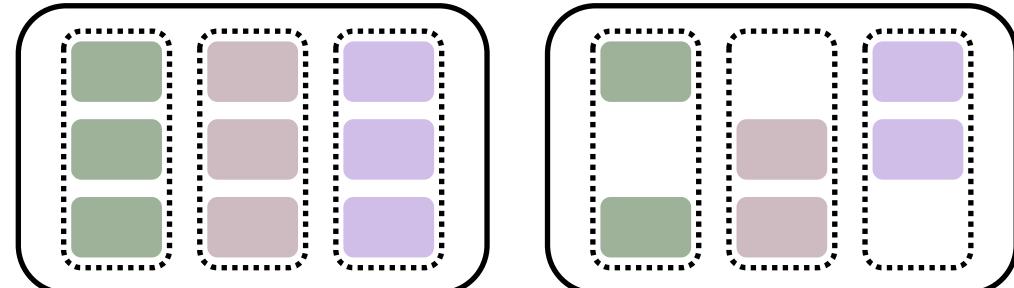
为实现对多模态数据的更细致与整体理解，未来研究应考虑：

1. 耦合多模态“平衡表征”与“语义对齐”；
2. 结合基础模型，设计多层次、多粒度、渐进式情感语义对齐方法

不完备多模态学习

□ 不完备多模态学习

- 研究如何构建对不完整模态数据具有鲁棒性的模型



□ 现有方法

基于补全的方法

思想: 根据已有的数据来填补或生成缺失的数据

- GCNet (Zheng Lian, et al.) TPAMI 2023
 - 图神经网络
- ShaSpec (Hu Wang, et al.) CVPR 2023
 - 特征解耦
- DiCMoR (Wang Yuanzhi et al.) ICCV 2023
 - 分布一致性

基于非补全的方法

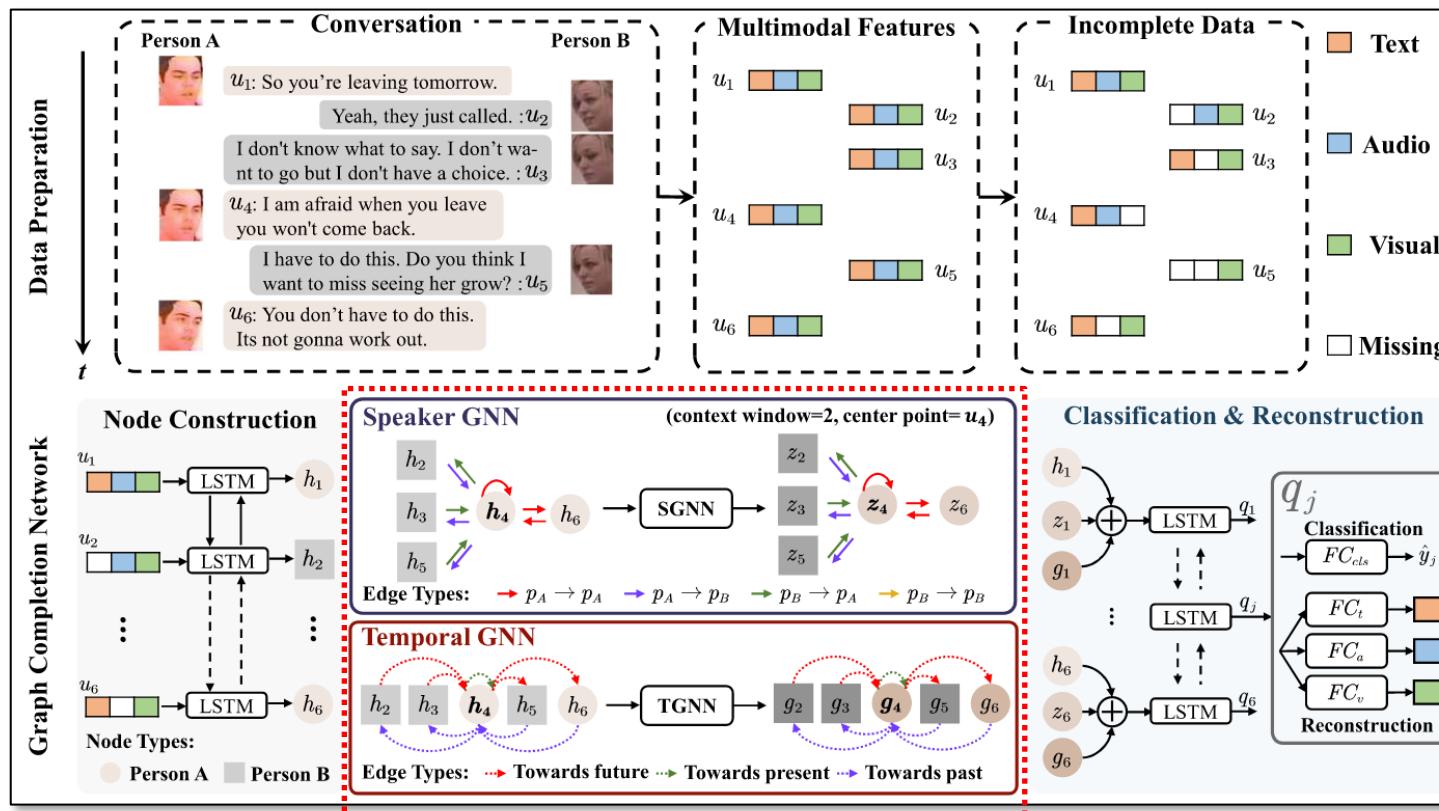
思想: 学习对数据缺失不敏感的、鲁棒性更强的模型

- PB-MSA (Yi-Lun Lee, et al.) CVPR 2023
 - 提示学习
- HRLF (Mingcheng Li, et al.) NeurIPS 2024
 - 知识蒸馏, 对抗学习
- CorrKD (Mingcheng Li, et al.) CVPR 2024
 - 知识蒸馏, 对比学习

不完备多模态学习

□ 基于补全的方法

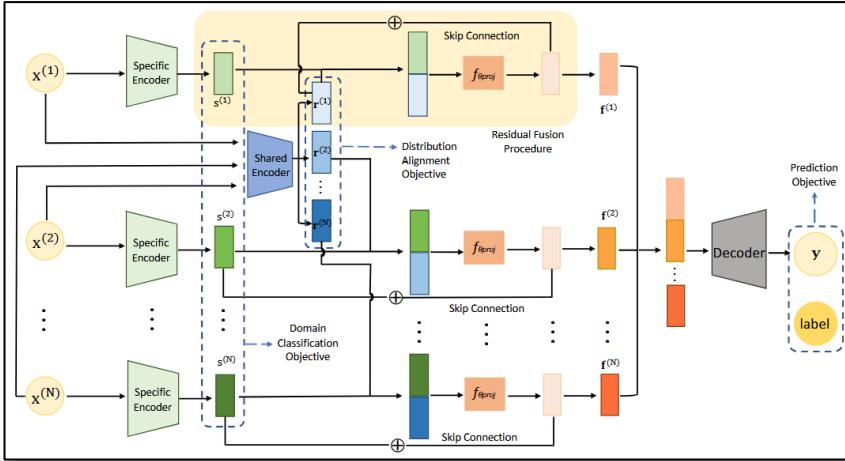
- 通过图神经网络显式建模时序关系和对话者信息
- 联合优化分类和重构任务，增强模型的鲁棒性



- **优点：**充分利用完整和不完整的多模态数据
- **缺点：**采用固定大小的上下文窗口构建图，可能无法捕捉长距离依赖关系

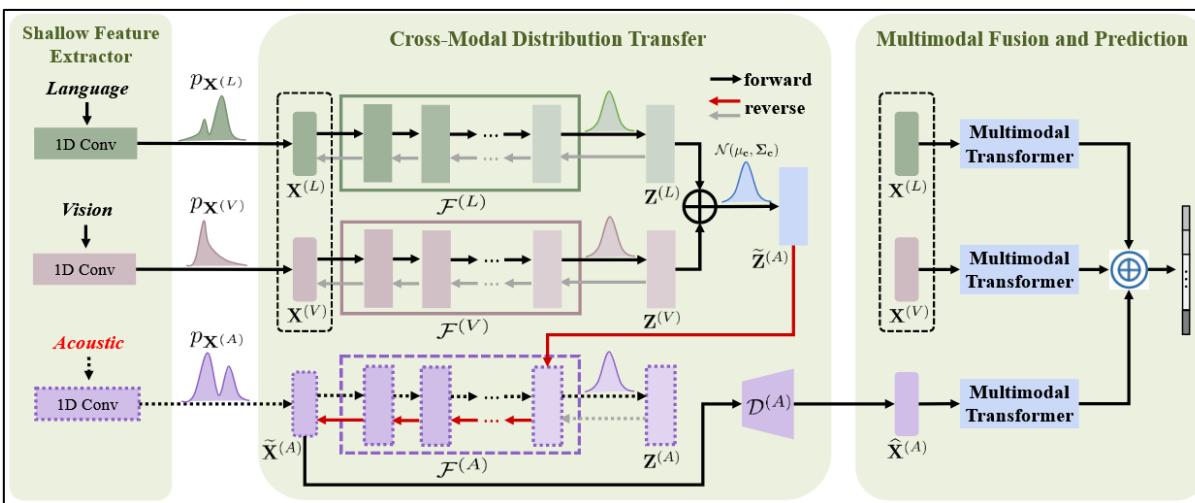
不完备多模态学习

□ 基于补全的方法



□ 通过**特征解耦**，将每个模态分解为共享特征和私有特征，利用可用模态的共享特征平均值估计缺失模态。

ShaSpec (CVPR 2023)



□ 从**分布一致性**的角度，先将分布从可用模态的迁移到缺失模态，再从中采样并恢复缺失模态

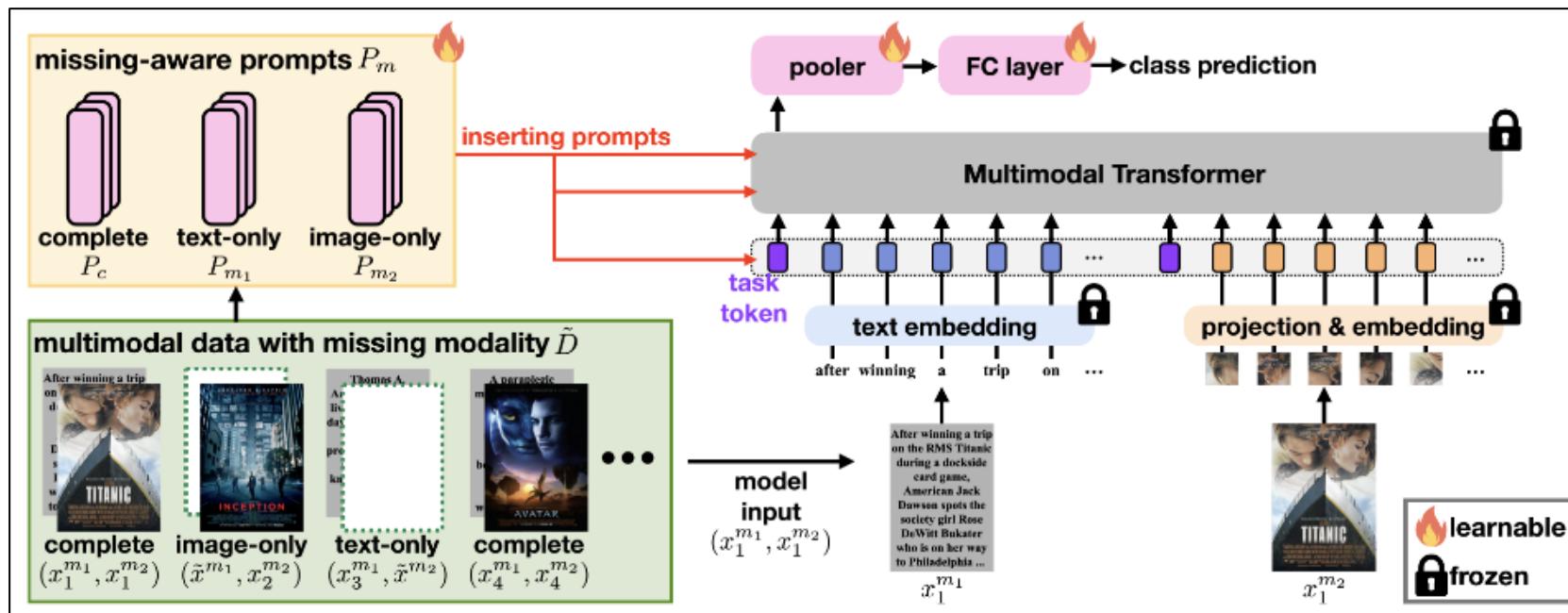
DiCMoR (ICCV 2023)

- [1] Hu Wang et al. Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling, **CVPR 2023**.
[2] Wang Yuanzhi et al. Distribution-consistent modal recovering for incomplete multimodal learning, **ICCV 2023**.

不完备多模态学习

□ 基于非补全的方法

- 通过提示学习为不同模态缺失情况学习专属感知提示
- 将提示注入预训练模型，弥补模态缺失的信息损失



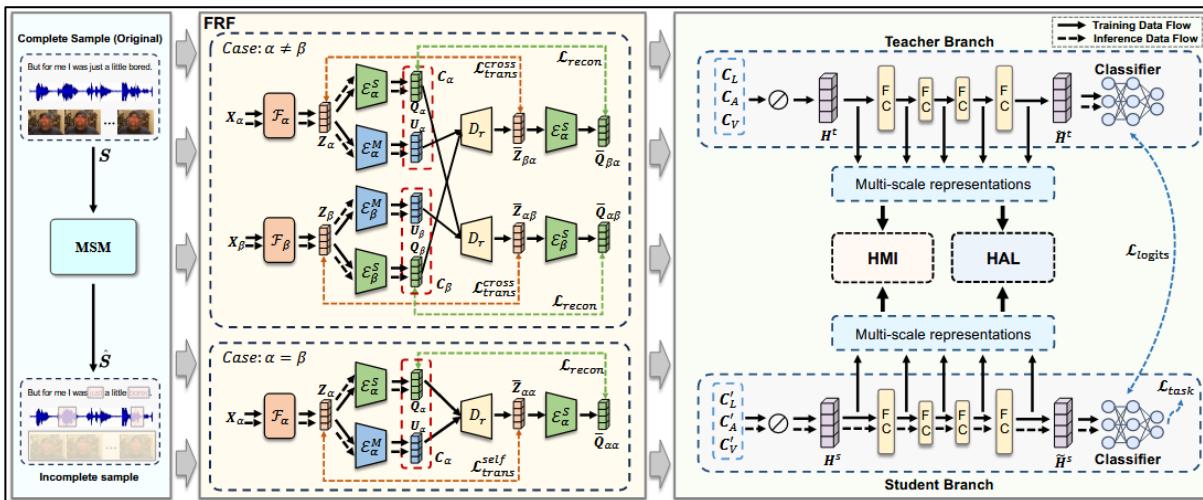
- **优点:** 能够同时处理训练和测试阶段模态缺失的情况
- **缺点:** 无法应对模态内部分缺失的场景

不完备多模态学习

□ 基于非补全的方法

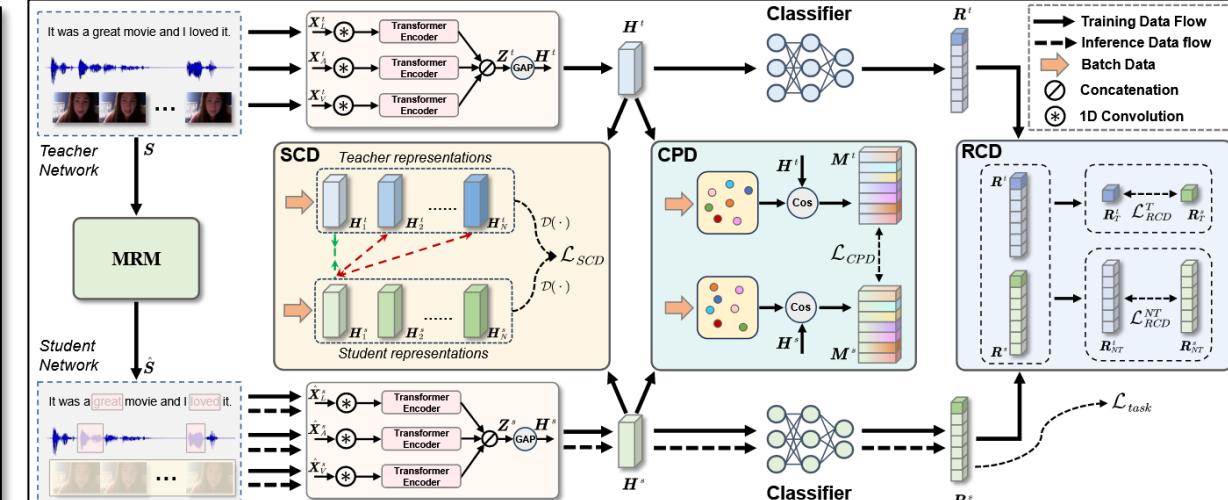
- 通过知识蒸馏将在完备模态上训练的“教师网络”的知识高效迁移到在不完备模态上训练的“学生网络”

➤ 层次化蒸馏：利用互信息最大化和对抗学习，在网络多个层级上对齐师生模型的语义信息和特征分布



HLRF (NeurIPS 2024)

➤ 关联性蒸馏：利用知识蒸馏促进学生模型充分学习样本间、类别间及最终决策的关联信息



CorrKD (CVPR 2024)

不完备多模态学习

□ 不完备多模态学习

- 研究如何构建对不完整情感模态数据具有鲁棒性的模型

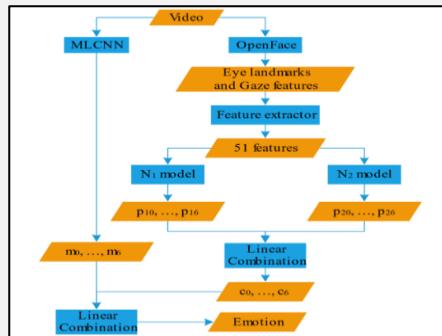
从增强复杂场景多模态情感分析模型稳健性角度，未来研究应考虑：

- 基于**检索增强**框架的情感模态缺失处理：检索增强框架（RAG）通过外部知识库填补模态缺失，**增强“补全”模态的保真度**；
- 考虑**组合式的缺失模态**：输入情感模态复杂多样，有必要设计**可组合的单模态表示**，以解决缺失模态组合数量爆炸问题，并支持处理各种未见模态缺失组合。

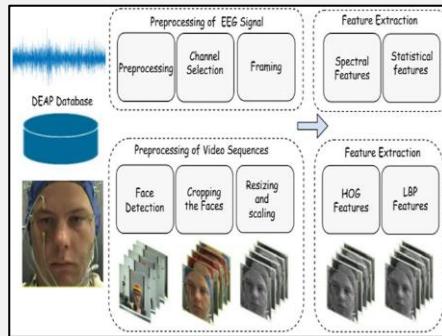
生理和行为模态

□ 传统机器学习方法

- 依赖手动特征工程
- 非端到端训练



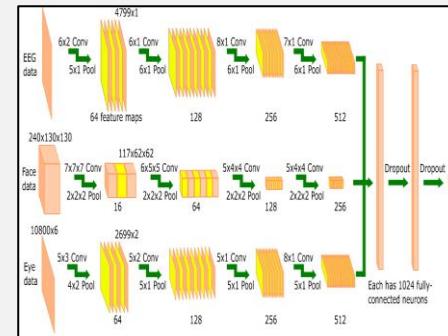
E-MLCNN [ICMLSC 2019]



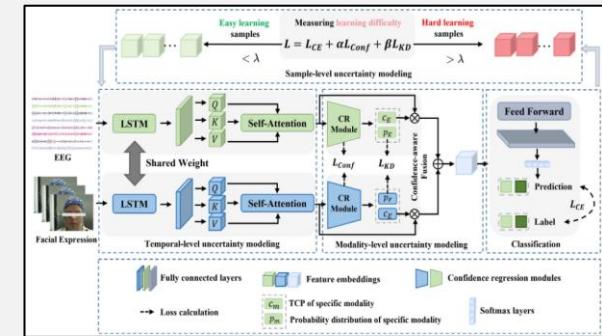
EVC [J Supercomput 2023]

□ 基于深度学习的方法

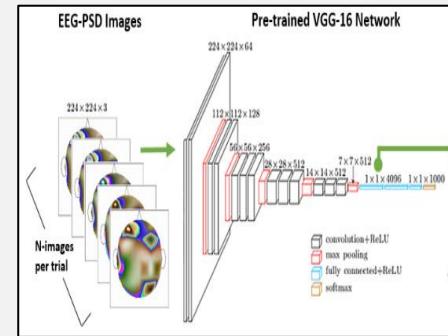
- 自动特征提取和端到端学习
- 利用注意力机制建模跨模态复杂依赖关系
- 通过不确定性动态调整模态权重，实现高效融合



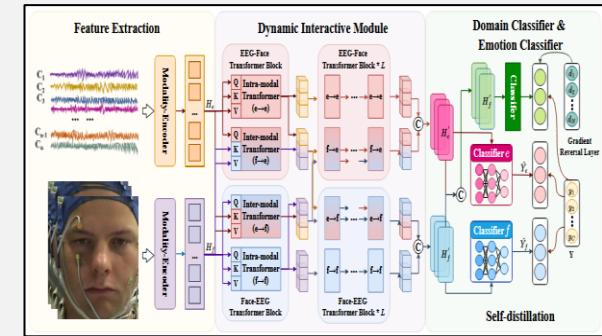
MBCNN [IF 2022]



DCMCF-Net [TAC 2024]



EEG-PSD [TAC 2022]

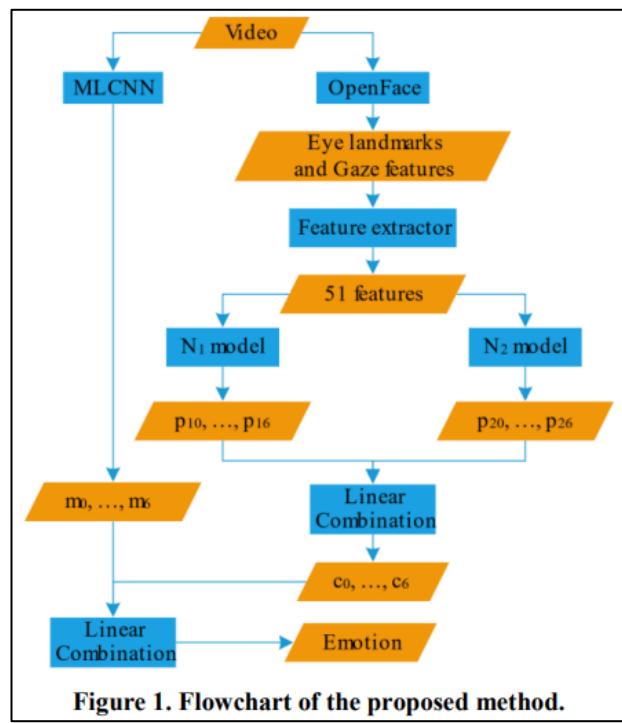


DISD-Net [TMM 2025]

生理和行为模态

□ 传统机器学习方法

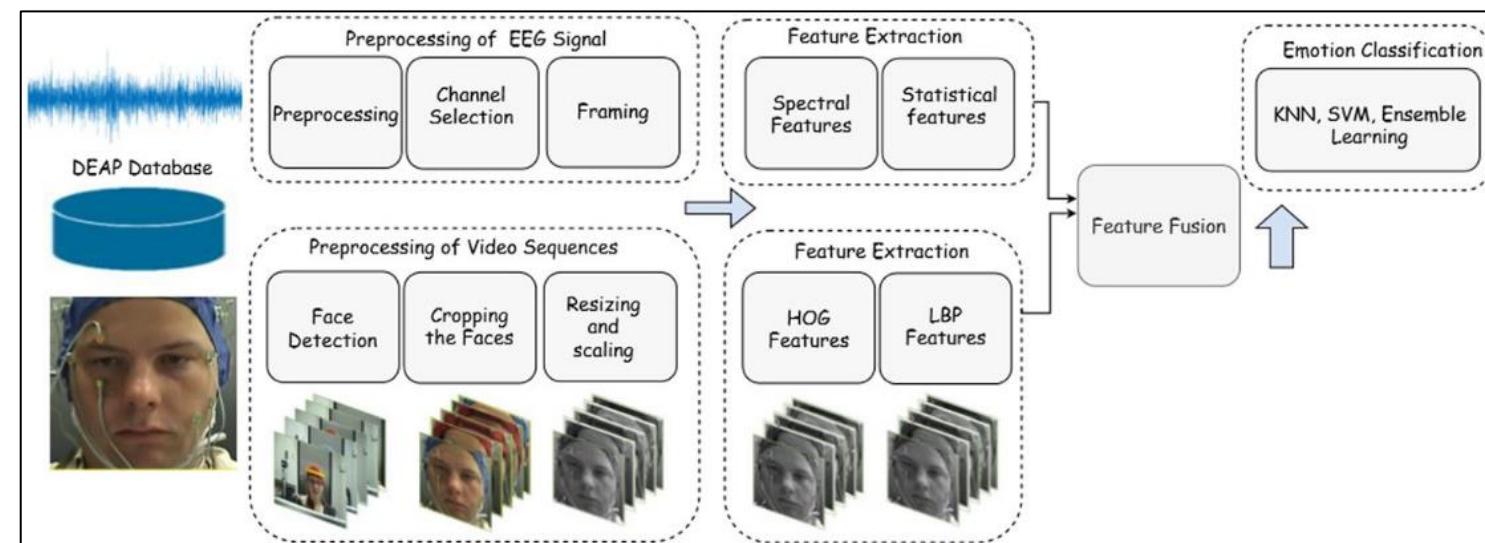
□ 采用传统的特征工程方法提取眼动特征



E-MLCNN [ICMLSC 2019]

□ 采用机器学习方法提取EEG和人脸特征，再进行分类

- 生理模态：对EEG信号进行预处理并提取频谱和统计特征
- 行为模态：通过HOG和LBP方法提取人脸形状与纹理特征



EVC [J Supercomput 2023]

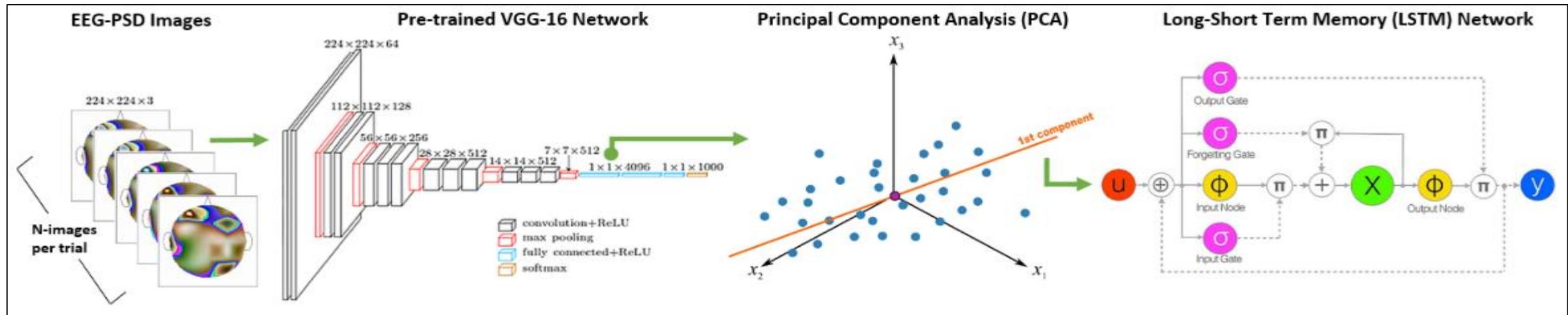
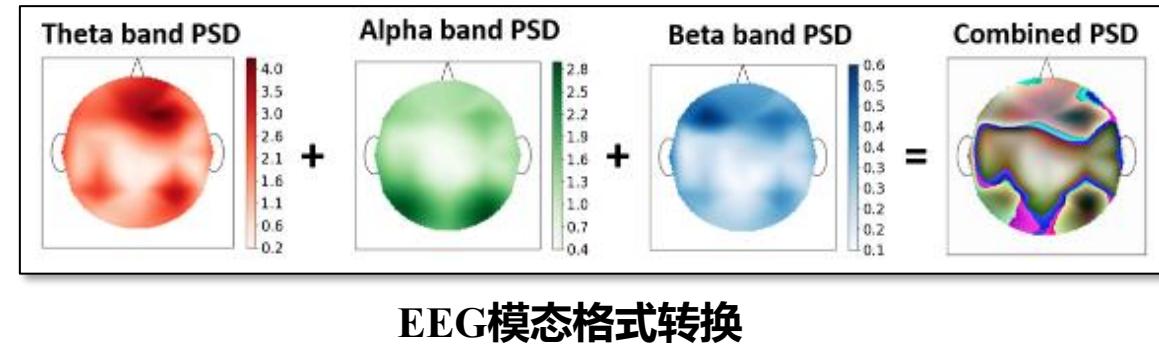
[1] Thong Van Huynh, et al. Emotion Recognition by Integrating Eye Movement Analysis and Facial Expression Model, *ICMLSC, 2019*.

[2] Anam Moin et al. Emotion recognition framework using multiple modalities for an effective human–computer interaction, *J Supercomput 2023*.

生理和行为模态

□ 基于深度学习的办法

- 通过将不同类型的生理信号统一转换为图像格式，并借助预训练CNN模型提取特征。
- 采用LSTM网络捕捉模型学习到EEG信号和人脸表情之间的跨模态依赖关系

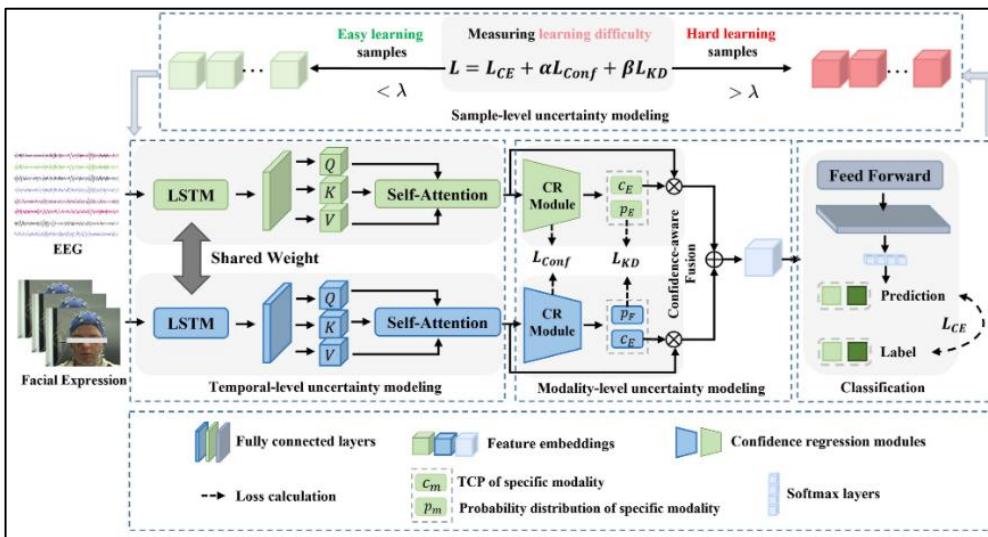


生理和行为模态

□ 基于深度学习的办法

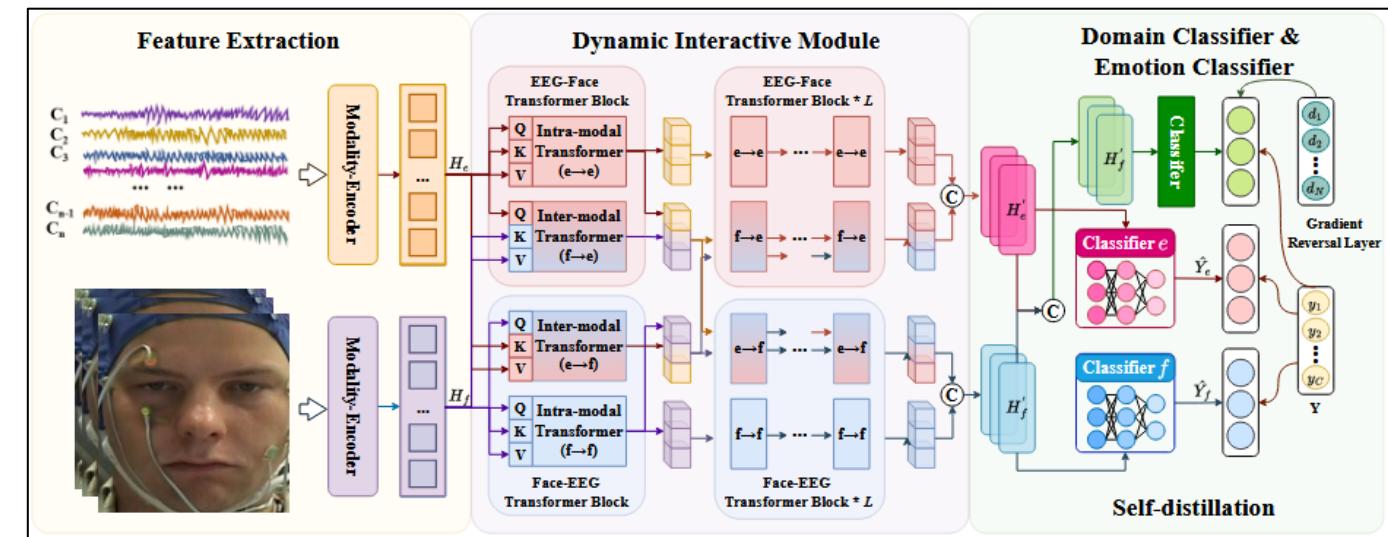
□ 从时间、模态、样本三个层次系统性建模不确定性

➤ 通过置信度感知融合机制，高效利用EEG模态和面部表情的互补信息



DCMCF-Net [TAC 2024]

□ 方法：通过动态交互、知识自蒸馏和域适应模块，有效捕捉跨模态依赖关系、强化特征表示并消除被试差异，提升模型泛化能力。



DISD-Net [TMM 2025]

[1] Qi Zhu, et al. Dynamic Confidence-Aware Multi-Modal Emotion Recognition, **TAC 2024**.

[2] Cheng Cheng et al. DISD-Net: A Dynamic Interactive Network with Self-distillation for Cross-subject Multi-modal Emotion Recognition, **TMM 2025**.

汇报提纲

- 多模态情感分析-问题定义与研究内容
- 多模态情感识别-研究背景及核心挑战
- 课题组相关进展-单模态、多模态情感识别研究进展
- 未来研究方向-大模型时代的多模态情感识别等

课题组相关进展

Yong Li, Jiabei Zeng, Shiguang Shan*. Learning Representations for **Facial Actions** from Unlabeled Videos, **IEEE TPAMI**, 2022. 视频情感分析

Yong Li, Menglin Liu et al. Decoupled Doubly Contrastive Learning for Cross Domain **Facial Action Unit Detection**, **IEEE TIP**, 2025.

Yi Ding et al. EmT: A novel transformer for generalized cross-subject **EEG emotion recognition**, **TNNLS**, 2025. 脑电情感分析

Yi Ding#, Yong Li#, et al. EEG-Deformer: A Dense Convolutional Transformer for **Brain-computer Interfaces**, **JBHI**, 2024. 脑电情感分析

Lifan Xia, Yong Li*, et al. Collaborative Contrastive Learning for Cross-Domain **Gaze Estimation**, **PR**, 2024. 人脸视线估计

Yong Li, Menglin Liu et al. Counterfactual discriminative **micro-expression recognition**, **Visual Intelligence**, 2024.

Yong Li, Shiguang Shan*. Contrastive Learning of Person-independent Representations for **Facial Action Unit Detection**, **IEEE TIP**, 2023.

Ximan Li, et al. Compound **expression recognition** in-the-wild with au-assisted meta multi-task learning, **CVPRW**, 2023

Yong Li, Antoni Chan, et al. Use of online therapy session data to develop behavioural markers for cognitive outcomes in non-pharmacological intervention, **Alzheimer's & Dementia**, 2023.

Yong Li, Shiguang Shan*. Meta Auxiliary Learning for **Facial Action Unit Detection**, **IEEE TAC**, 2023.

Yong Li, Yi Ren et al. Beyond Overfitting: Doubly Adaptive Dropout for Generalizable **AU Detection**, **IEEE TAC**, 2025. 图像情感分析

Yong Li, Jiabei Zeng, et al. Self-supervised representation learning from videos for **Facial Action Unit Detection**, **CVPR**, 2019.

Zili Wang, Lingjie Lao, Xiaoya Zhang, Yong Li*. Context-dependent **Emotion Recognition**, **ChinaMM**最佳海报奖, 2022. 场景情感分析

李勇, 曾加贝, 山世光*. 面部动作单元检测方法进展与挑战, **中国图象图形学学报**, 2020 (入选 2021 年中国图象图形学报优秀论文)

Yong Li, Jiabei zeng, Shiguang Shan, Xilin Chen. Occlusion aware **facial expression recognition** using cnn with attention mechanism, **IEEE TIP**, 2019. **ESI 高被引**

Yong Li, Yuanzhi Wang, et al. Decoupled **Multimodal** Distilling for **Emotion Recognition**, **CVPR(Highlight)**, 2023. 多模态情感分析

Yuanzhi Wang, Yong Li*, et al. Incomplete **multimodality-diffused emotion recognition**, **NeurIPS**, 2023.

Yuanzhi Wang, Zhen Cui*, Yong Li*. Distribution-Consistent Modal Recovering for Incomplete **Multimodal** Learning, **ICCV**, 2023. 多模态情感分析

Decoupled Hierarchical Distillation for **Multimodal Emotion Recognition**. **IEEE T-PAMI** (under review)

Hierarchical **Vision-Language** Interaction for **Facial Action Unit Detection**. **IEEE TAC** (under review)

单模态
情感识别

↑
↑

多模态
情感识别

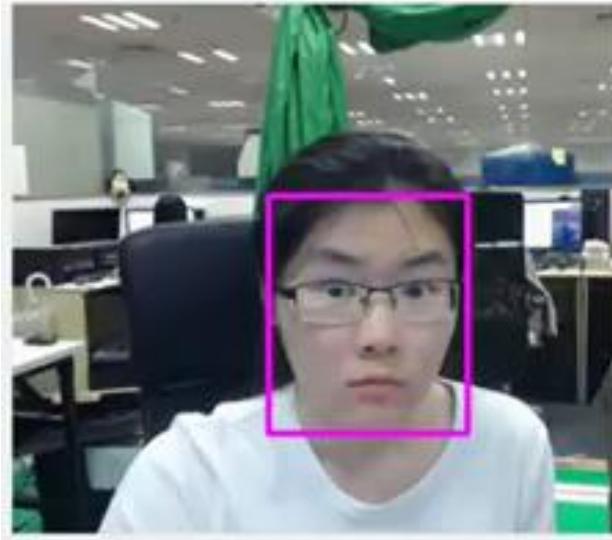
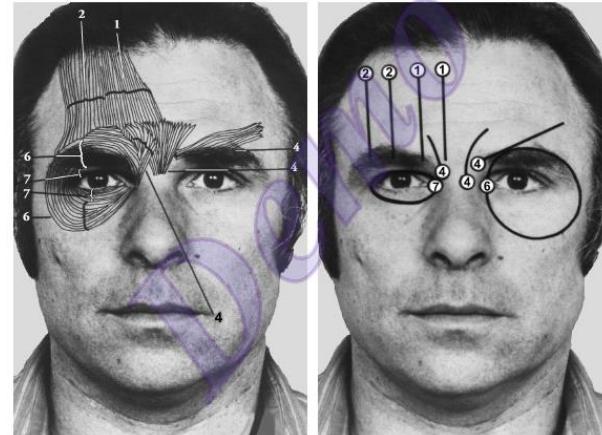
↑

单模态情感识别

□ 面部动作检测

□ AU

➤ Action Unit



AU1 内眉上扬 Inner Brow Raiser
AU2 外眉上扬 Outer Brow Raiser
AU3 眉毛下压 Brow Lowerer
AU5 上眼睑上扬 Upper Lid Raiser
AU6 直视抬起 Chaste Raiser
AU7 瞳孔紧缩 Lid Tightener
AU9 鼻子皱起 Nose Wrinkler
AU10 上唇上扬 Upper Lip Raiser
AU12 嘴角上扬 Lip Corner Puller
AU14 收紧嘴角 Dimpler
AU15 嘴角下拉 Lip Corner Depressor
AU16 下唇压紧 Lower Lip Pressor
AU17 下巴抬起 Chin Raiser
AU18 嘴巴皱起 Lip Pucker
AU20 嘴巴伸展 Lip Stretch
AU22 嘴巴收缩 Lip Tightener
AU23 嘴巴压紧 Lip Pressor
AU25 上下嘴唇分开 Lips Parting
AU28 下唇下拉 Lip Drop



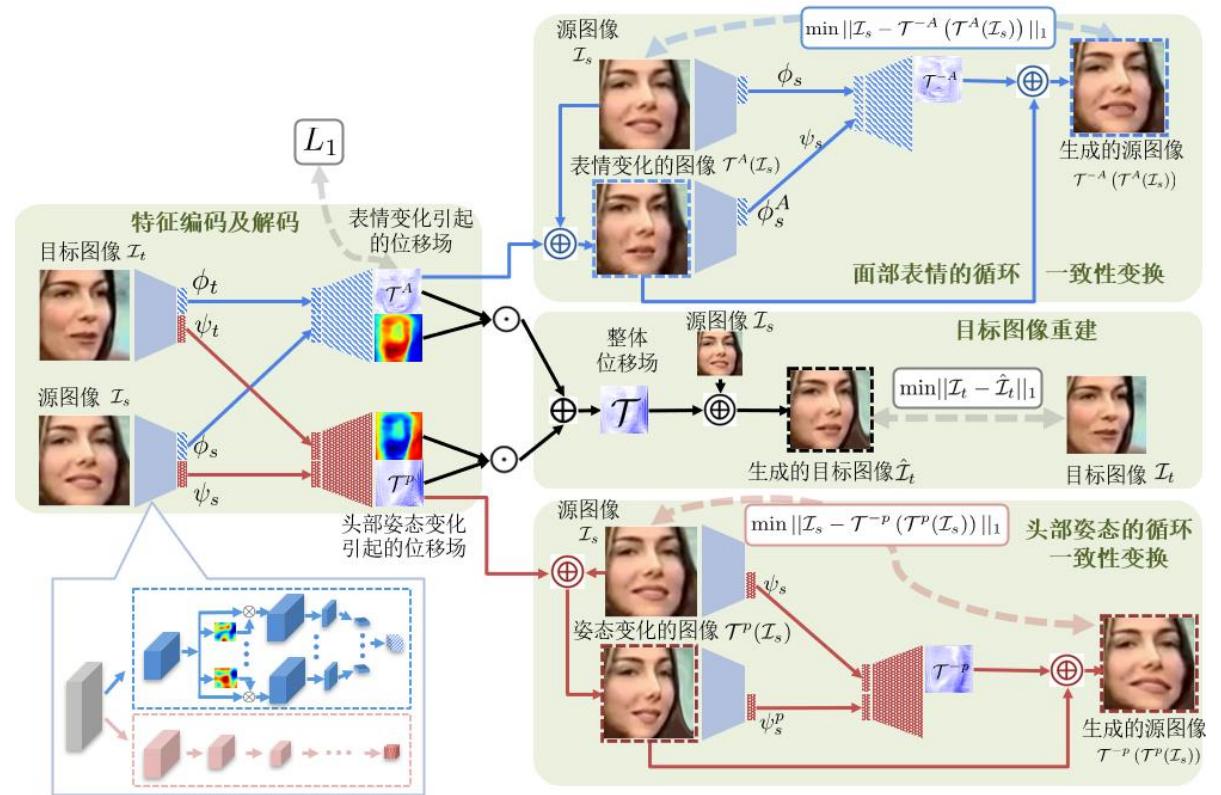
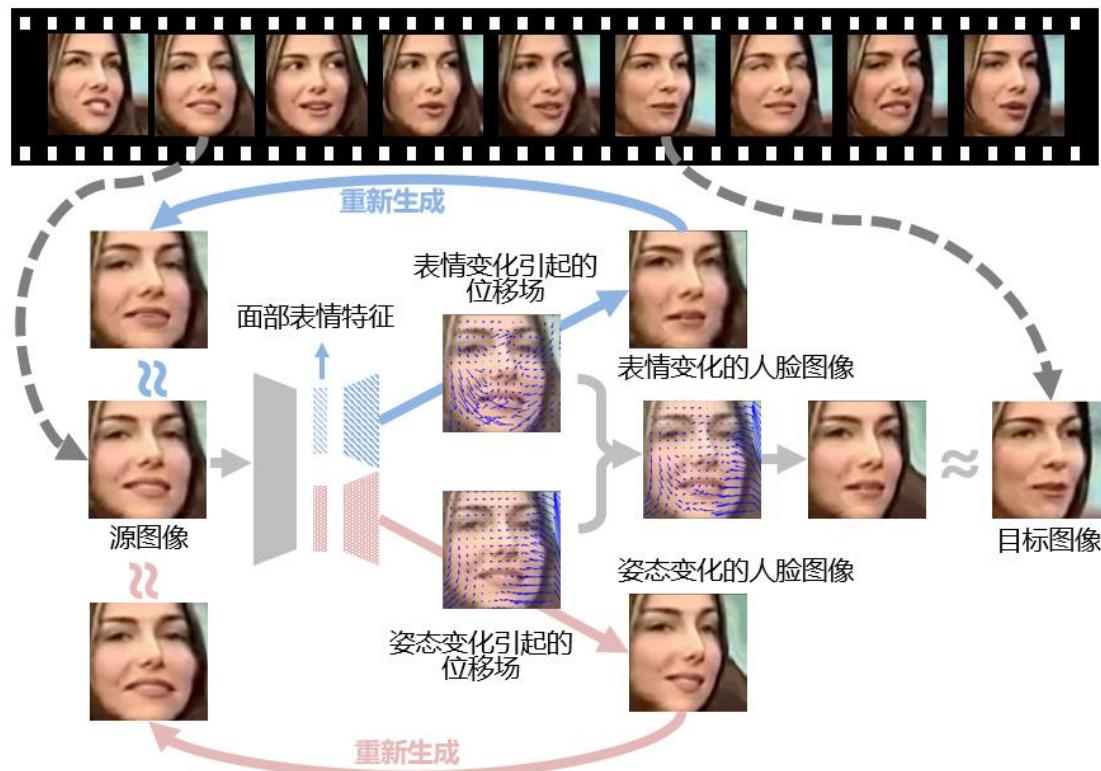
上半脸动作单元AU					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
内眉上扬	外眉上扬	眉毛下压	上眼睑上扬	脸颊抬起	眼睑收紧
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
眼睑下垂	眯眼	闭眼	半眯眼	眨眼	半眨眼

下半脸动作单元AU					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
鼻子蹙皱	上唇上扬	人中抬起	嘴角上扬	腮颊鼓起	收紧嘴角
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
嘴角下拉	下嘴唇下坠	下巴缩紧	噘嘴	嘴唇舒展	龇牙
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
嘴唇收缩	嘴唇压紧	上下嘴唇分开	下颌下拉	张嘴	抿嘴

单模态情感识别

□ 工作一：面向AU检测的生成式自监督学习

□ 方法框架：基于自监督思想的面部表情特征学习



单模态情感识别

□ 工作一：面向AU检测的生成式自监督学习

□ 实验：8层轻量级网络(无监督)+单FC层分类**媲美复杂方法**

	Methods/AU	1	2	4	6	7	10	12	14	15	17	23	24	ave
Descriptor	Handcrafted [21]*	43.4	40.7	43.3	59.2	61.3	62.1	68.5	52.5	36.7	54.3	39.5	37.8	50.0
	ResNet-80 face	39.3	40.6	38.5	64.2	67.5	71.0	65.3	57.2	37.8	51.3	35.1	32.6	49.9
	VGG emotion	46.4	36.3	49.6	76.0	77.6	80.2	87.8	60.8	40.4	59.1	43.7	48.2	58.8
Supervised	AlexNet [52]*	40.3	39.0	41.7	62.8	54.2	75.1	78.1	44.7	32.9	47.3	27.3	40.1	48.6
	DRML [2]*	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
	EAC-Net [3]*	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
	ROI [4]*	36.2	31.6	43.4	77.1	73.7	85.0	87.0	62.6	45.7	58.0	38.3	37.4	56.4
	JAA-Net [5]*	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
Self-supervised	SplitBrain [26]	39.0	32.0	39.7	72.9	70.6	78.2	83.7	57.8	37.3	53.6	32.3	45.1	53.5
	DeformAE [47]	39.5	34.5	40.8	70.5	68.4	76.3	82.9	60.7	23.1	54.1	34.3	43.1	52.3
	Fab-Net [24]	43.3	35.7	41.6	72.9	63.0	75.9	83.5	57.7	26.5	48.2	33.6	42.4	52.0
	TAE (w/o L_1 , w/o att.)	40.6	38.2	44.7	71.9	67.5	76.0	81.3	61.6	34.0	54.7	39.4	46.7	54.7
	TAE (w/ L_1 , w/o att.) [53]	43.1	32.2	44.4	75.1	70.5	80.8	85.5	61.8	34.7	58.5	37.2	48.7	56.1
	TAE (w/ L_1 & att., $K = 2$)	42.0	41.9	44.5	74.2	73.5	80.6	85.9	60.4	41.4	61.3	44.1	46.0	57.9
	TAE (w/ L_1 & att., $K = 4$)	40.7	42.1	52.7	71.0	73.7	79.7	85.0	62.1	44.0	62.9	43.7	48.1	58.8
	TAE (w/ L_1 & att., $K = 8$)	47.0	45.9	50.9	74.7	72.0	82.4	85.6	62.3	48.1	62.3	45.9	46.3	60.3
	TAE (w/ L_1 & att., $K = 16$)	44.2	43.5	50.1	73.8	73.5	81.5	85.1	62.8	44.4	60.5	45.3	45.9	59.2

BP4D数据集 F1 score

媲美监督
方法



单模态情感识别

突出性成果

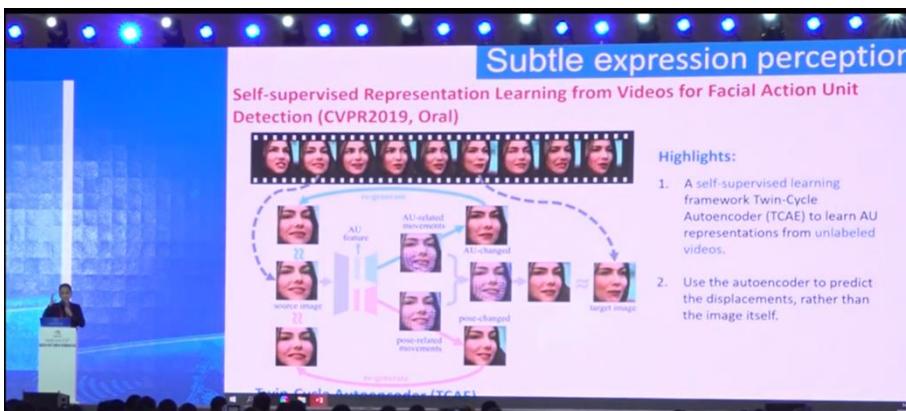
创新总结：构建基于大规模无标注视频的表情自监督学习框架，突破细微表情标注数据不足导致的泛化瓶颈

挑战

如何克服细微表情运动单元检测任务中标注数据匮乏难题

创新点

提出对近邻人脸视频帧之间的“刚体和柔性变换”进行解耦、重建及交叉验证，有效解决了监督表情数据匮乏问题(CVPR'19, T-PAMI'22)



在VALSE'19大会上，北京邮电大学教授在演讲中称：如何在这种非监督情况下，从海量视频里自动的挖掘出这种小的面部动作，这是个非常本质的问题。这个是CVPR 2019的一篇文章，它做出了特别突破性的工作。

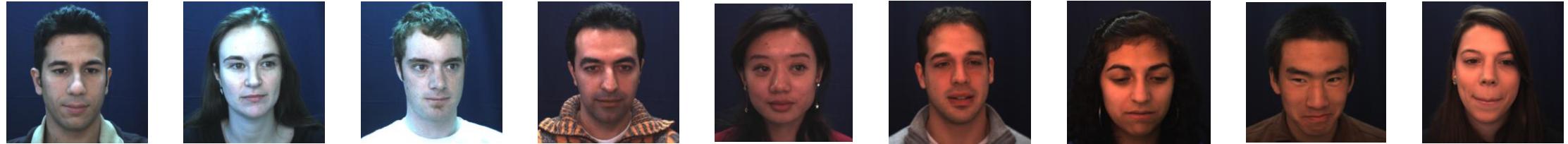
a few dozen. Additionally, to achieve current state-of-the-art performance, most of the published methods have made adaptations to their CNN architectures to utilize additional features for the representation learning [9], [42], [31]. These



MIT博士，微软研究员Daniel McDuff称我们的工作是“目前的先进算法”。

单模态情感识别

□ 工作二：面向AU检测的判别式自监督学习：削减个体差异性

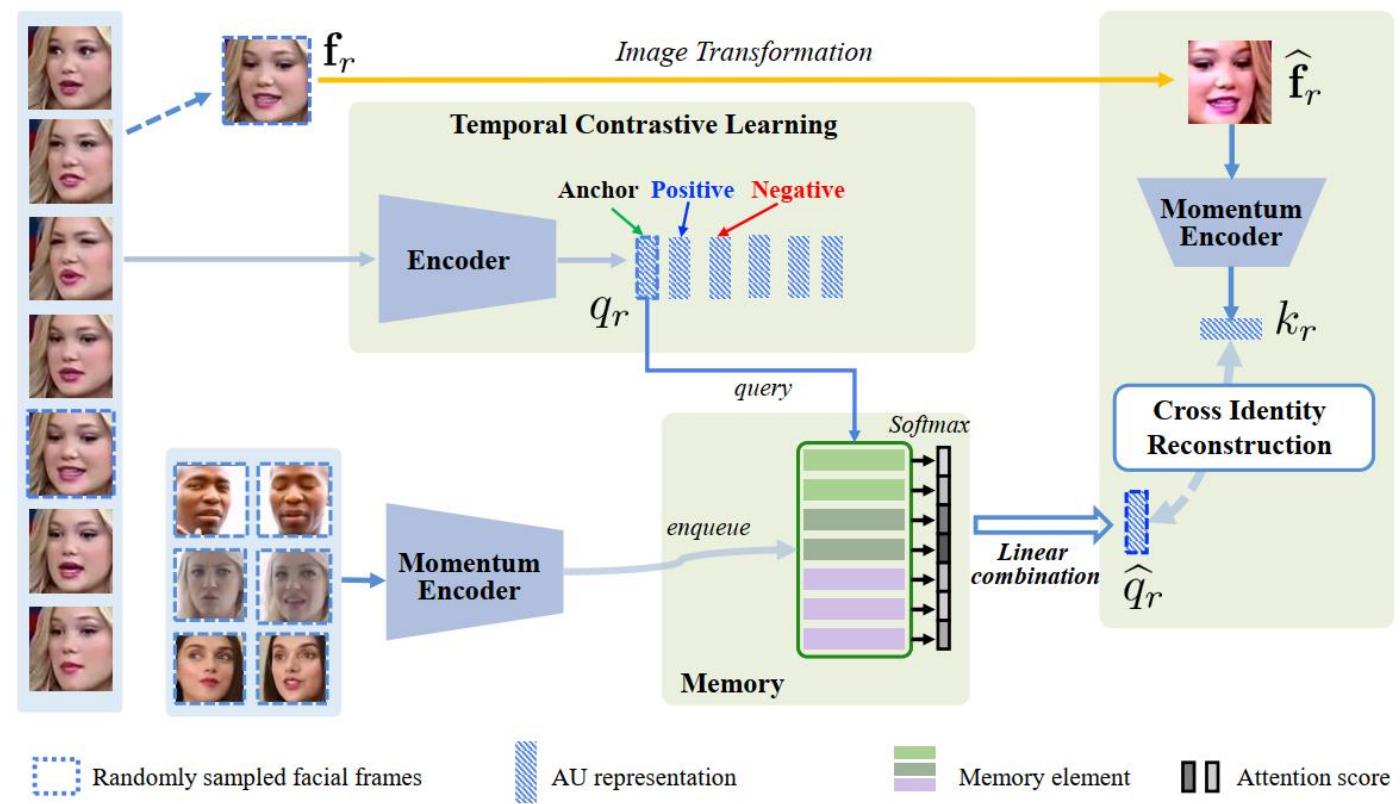
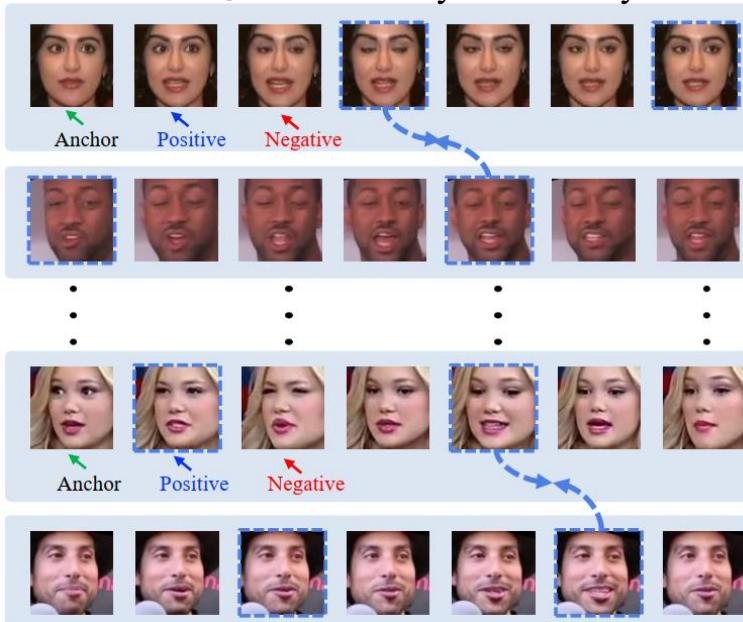


□ Intra-video对比学习

➢ 基于三元组对比学习

□ Inter-video对比学习

➢ 基于cross-identity consistency



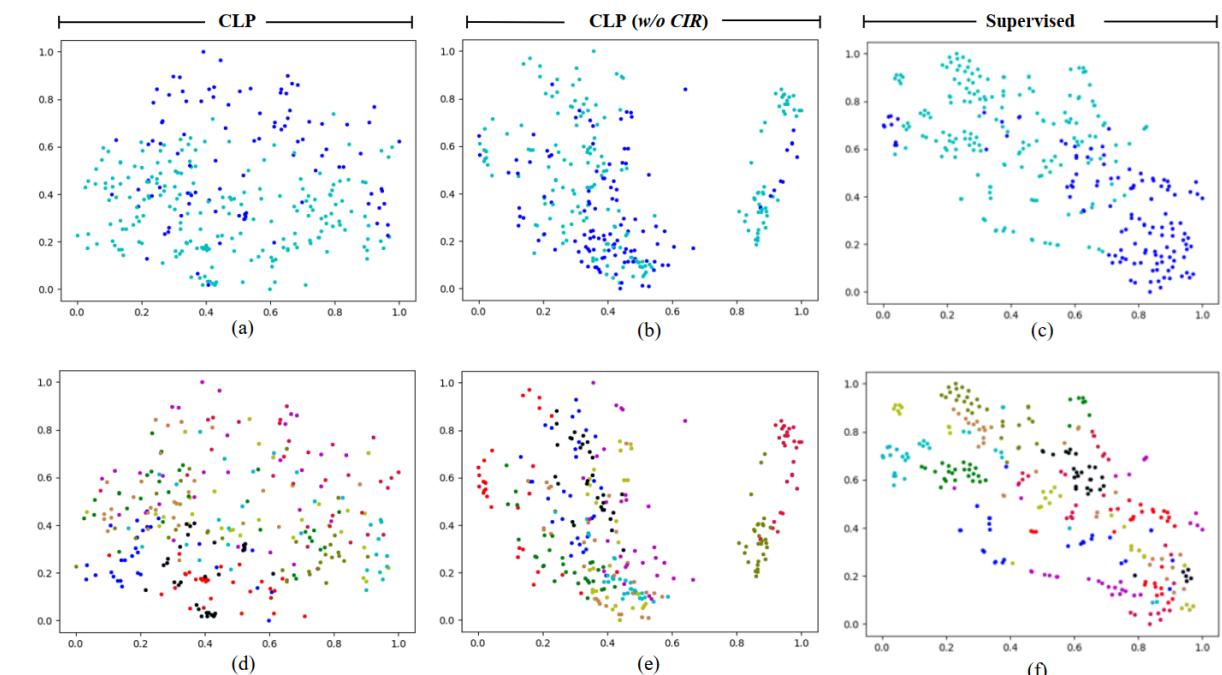
单模态情感识别

□ 工作二：面向AU检测的判别式自监督学习：削减个体差异性

Methods/AU	1	2	4	6	9	12	25	26	ave
ROI [7]*	41.5	26.4	66.4	50.7	80.5	89.3	88.9	15.6	48.5
EAC-Net [7]*	41.5	26.4	66.4	50.7	80.5	89.3	88.9	15.6	48.5
ATF [38]*	45.2	39.7	47.1	48.6	32.0	55.0	86.4	39.2	49.2
IdenNet [39]*	25.5	34.8	64.5	45.2	44.6	70.7	81.0	55.0	52.6
DSIN [8]*	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
JAA-Net [9]*	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
AU-RCNN [35] *	32.1	25.9	59.8	55.3	39.8	67.7	77.4	52.6	51.3
ARL [36] *	43.9	42.1	63.6	41.8	40.0	76.2	95.2	66.8	58.7
SRERL [10]*	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
UGN [57] *	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
SEV-Net [11]*	55.3	53.1	61.5	53.6	38.2	71.6	95.7	41.5	58.8
HMP-PS [12]*	38.0	45.9	65.2	50.9	50.8	76.0	93.3	67.6	61.0
Jacob <i>et al.</i> [37]*	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
SO-Net [40]*	33.8	44.5	70.3	57.6	39.7	78.2	86.7	57.3	58.5
PIAP [58]*	50.2	51.8	71.9	50.6	54.5	79.7	94.1	57.2	63.8
ResNet-34	38.0	33.1	51.8	46.2	34.2	65.4	85.4	56.9	51.3
Lu <i>et al.</i> [42]*	18.7	27.4	35.1	33.6	20.7	67.5	68.0	43.8	39.4
Fab-Net [41]*	15.5	16.2	43.2	50.4	23.2	69.6	72.4	42.4	41.6
TAE [44]*	21.4	19.6	64.5	46.8	44.0	73.2	85.1	55.3	51.5
EmoCo [17]*	34.3	31.9	63.9	52.5	44.0	77.0	78.3	44.2	53.3
CLP (Ours)	42.4	38.7	63.5	59.7	38.9	73.0	85.0	58.1	57.4

DISFA数据集, F1 score

低于/媲美
监督方法



- 上图：两种颜色分别表示AU12是否激活
- 下图：按个体可视化

单模态情感识别

□ 工作三：面向AU检测的跨域自监督学习

□ 方法动机

- 跨域细微表情运动单元检测性能衰退严重
- 学习领域无关面部动作单元特征表示



Commercial systems, including iMotions, Affectiva and Noldus, profess to recognize AU and facial expressions. Considering the relatively low cross-domain generalizability of the state-of-the-art, we urge caution in applying such systems to new domains. Use in new domains should first be validated on a subset of manually annotated video. If systems fail this validation step, re-training is recommended. This is not possible with current commercial systems but is an option with OpenFace and the CNN used here.

来源：*Crossing Domains for AU Coding: Perspectives, Approaches, and Measures, FG 2020, IEEE T-BIOM, 2020*

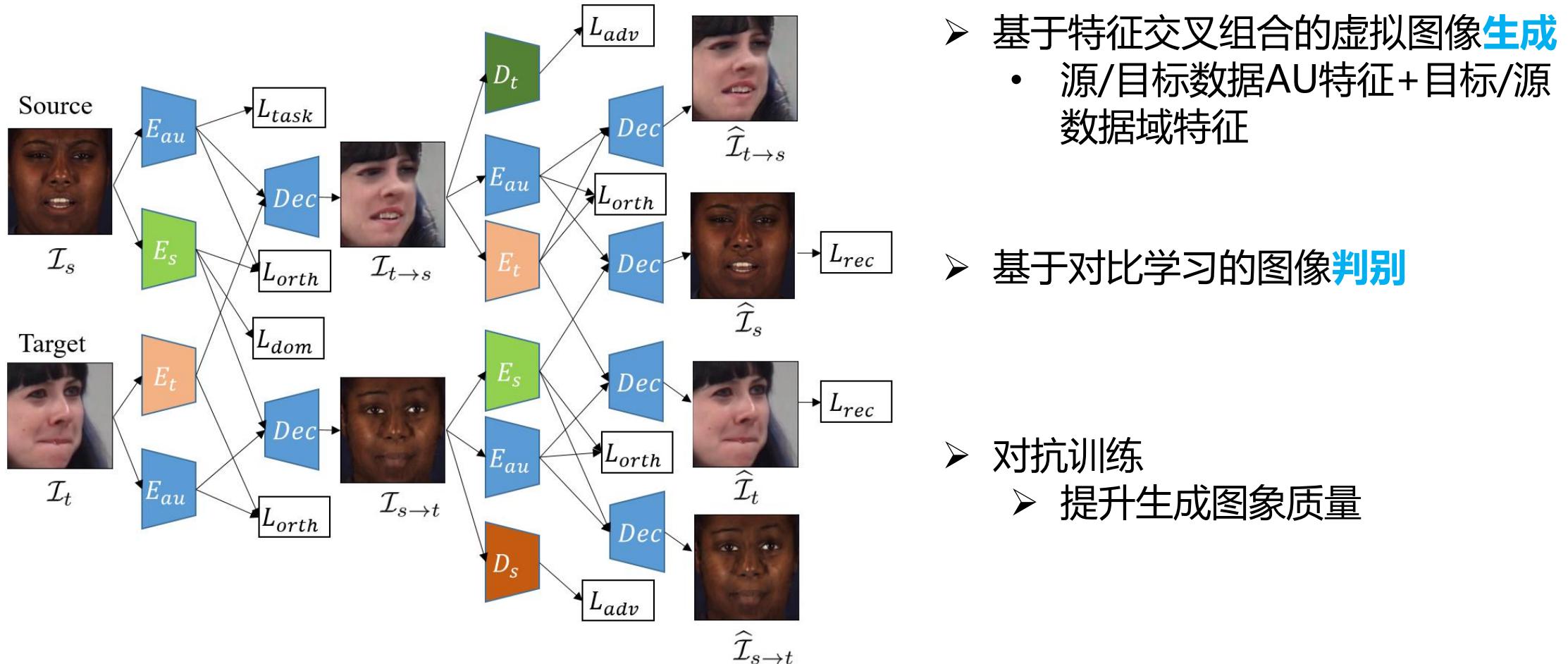
□ 难点

- 当前“风格-内容”解耦范式无法胜任：AU只是人脸内容的一部分，与身份等因素交叉纠缠！
- 训练数据不足：个体数量有限

单模态情感识别

□ 工作三：面向AU检测的跨域自监督学习

□ 框架设计：结合生成和判别式自监督学习



单模态情感识别

□ 工作三：面向AU检测的跨域自监督学习

□ 框架设计：结合生成和判别式自监督学习

授权发明专利：一种基于双重对比学习的跨域表情运动单元检测方法，ZL202311142081.4

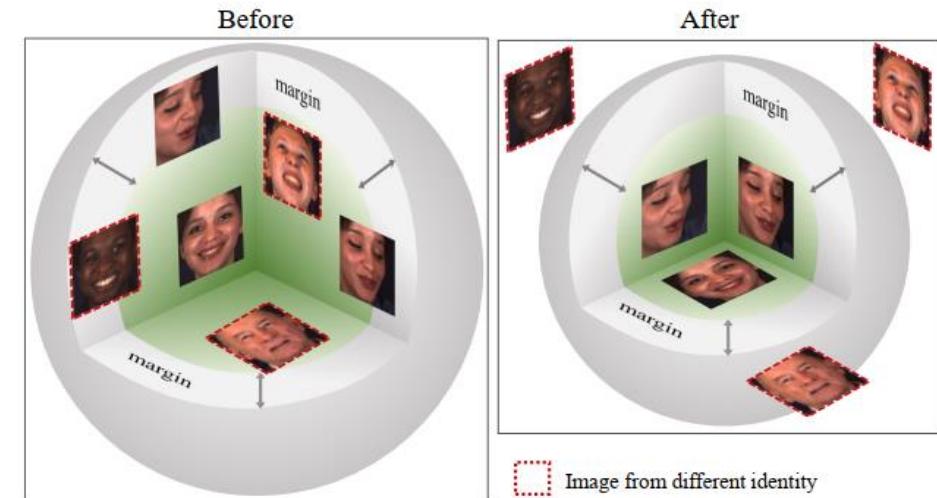
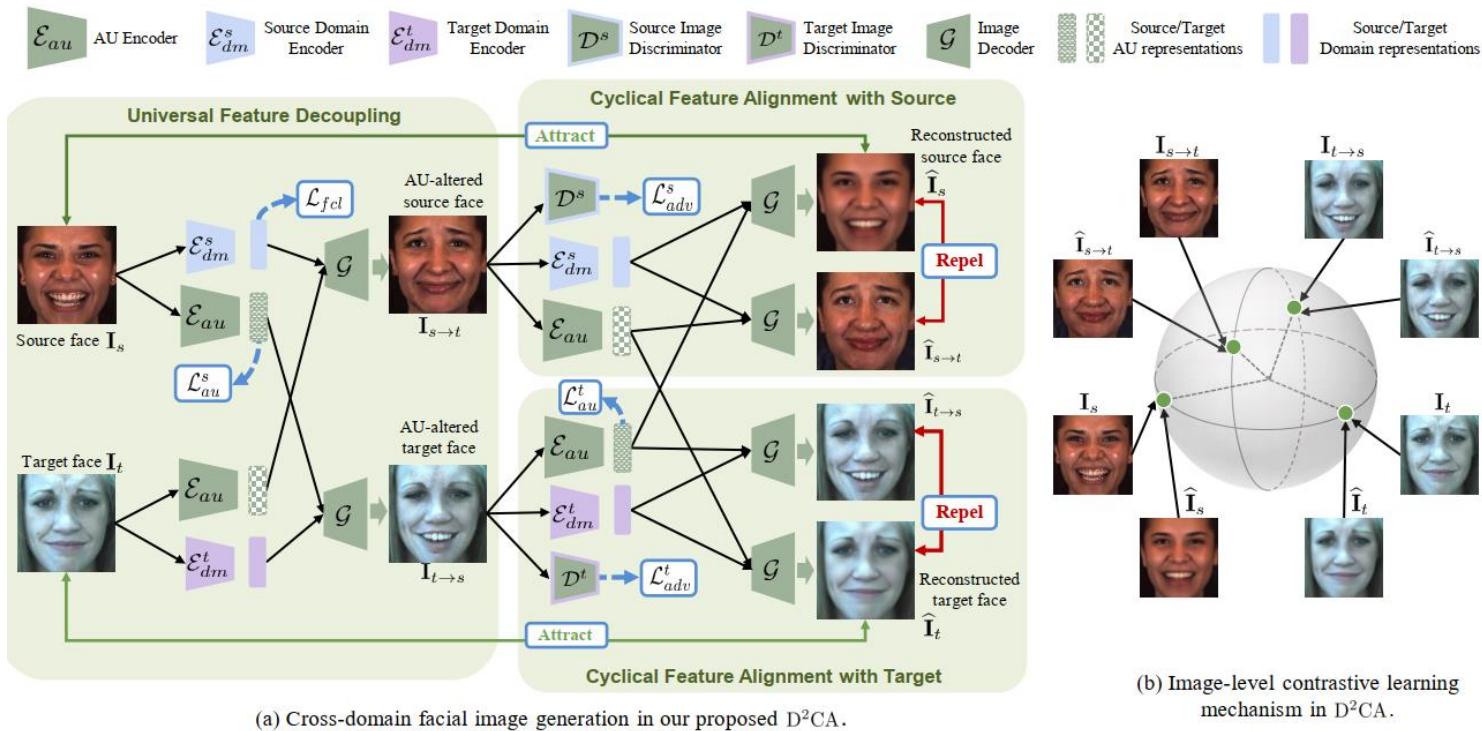
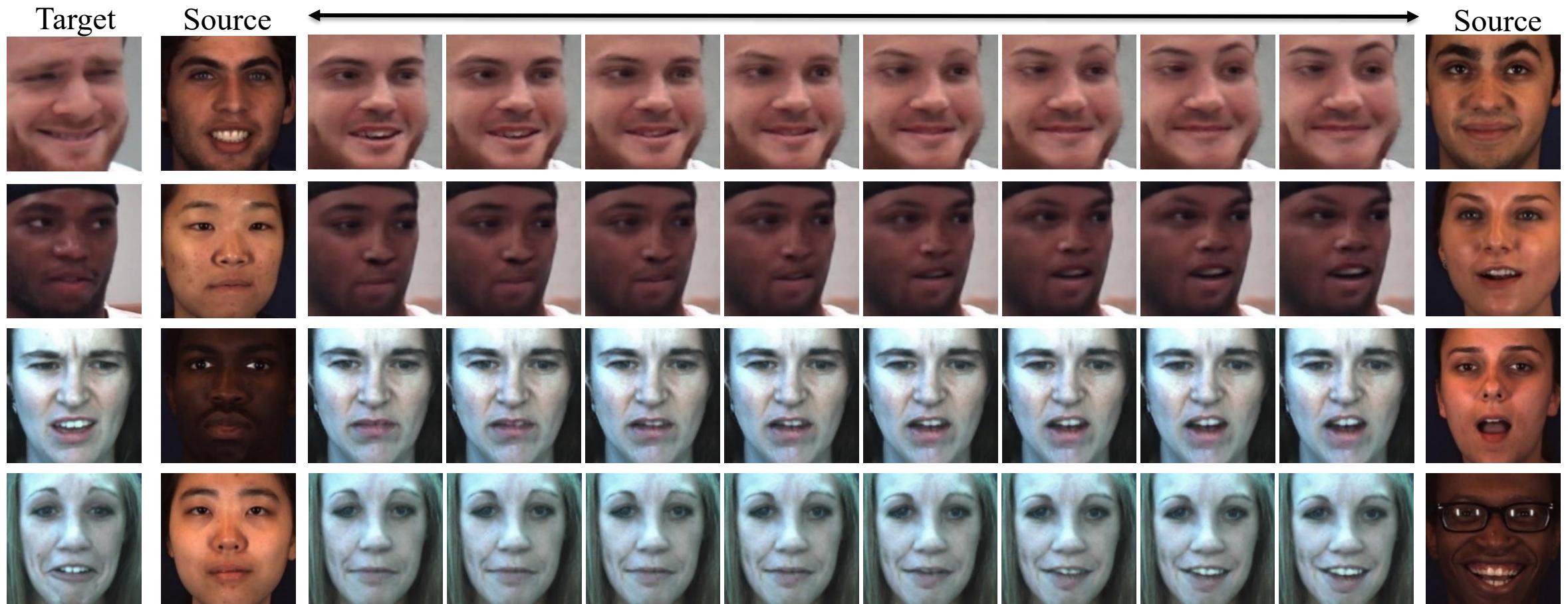


Fig. 4. Illustration of feature-level contrastive learning (FCL). Domain features from the same identity are pushed close.

单模态情感识别

□ 工作三：面向AU检测的跨域自监督学习

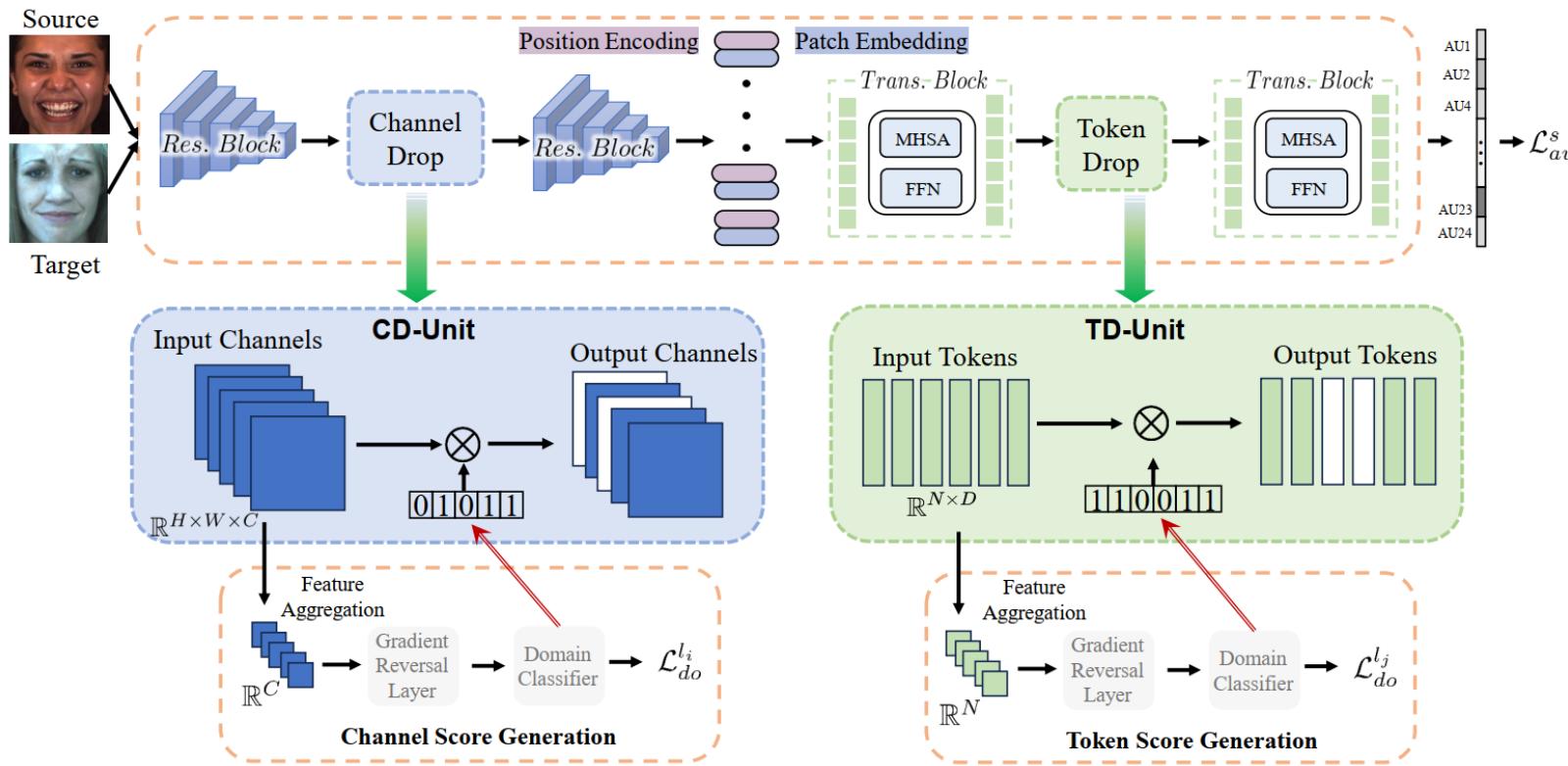
□ 实验：特征插值验证



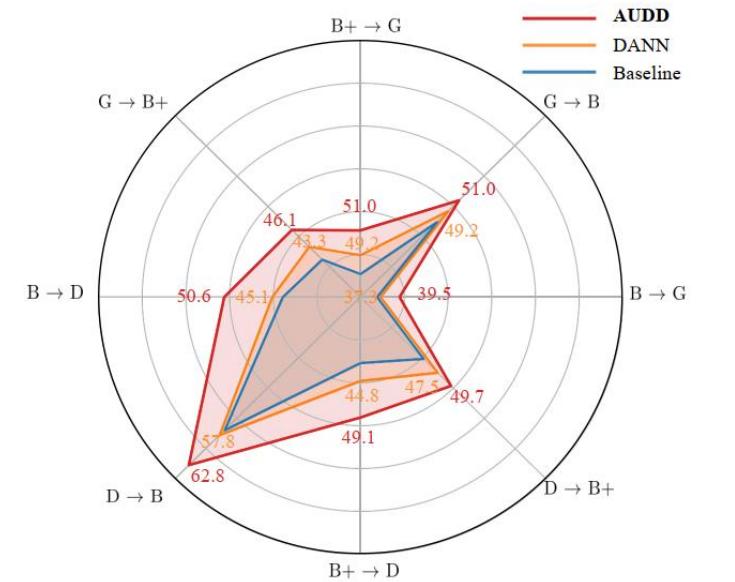
单模态情感识别

□ 工作四：面向AU检测的跨域自监督学习，从生成到判别

□ 方法：基于自适应跨域特征选择机制



- 混合CNN+Transformer架构
- 多尺度/粒度跨域特征自适应选择



多跨域场景测试性能对比

单模态情感识别

□ 工作四：面向AU检测的跨域自监督学习，从生成到判别

□ 方法：基于自适应跨域特征选择机制

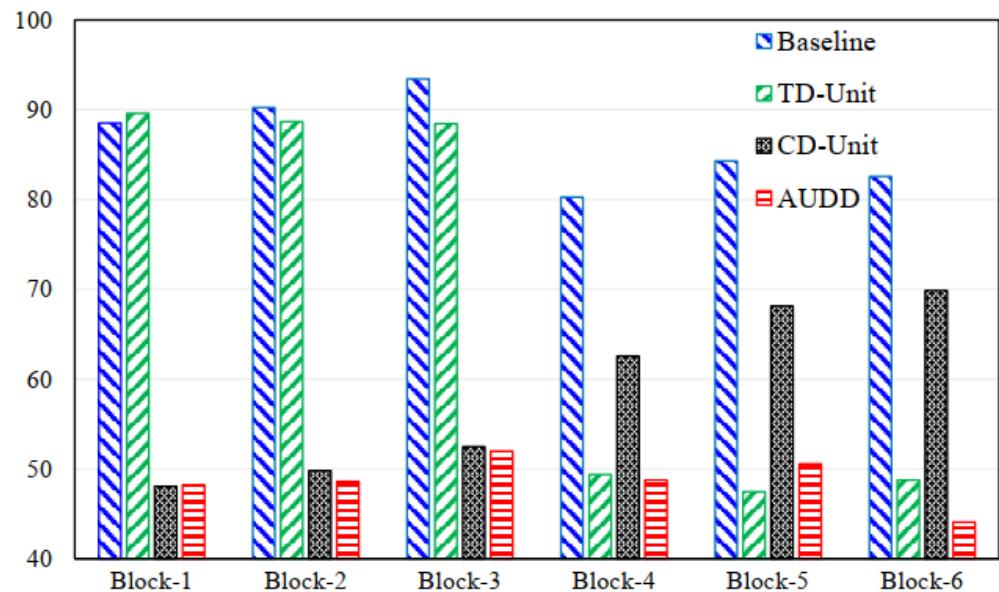
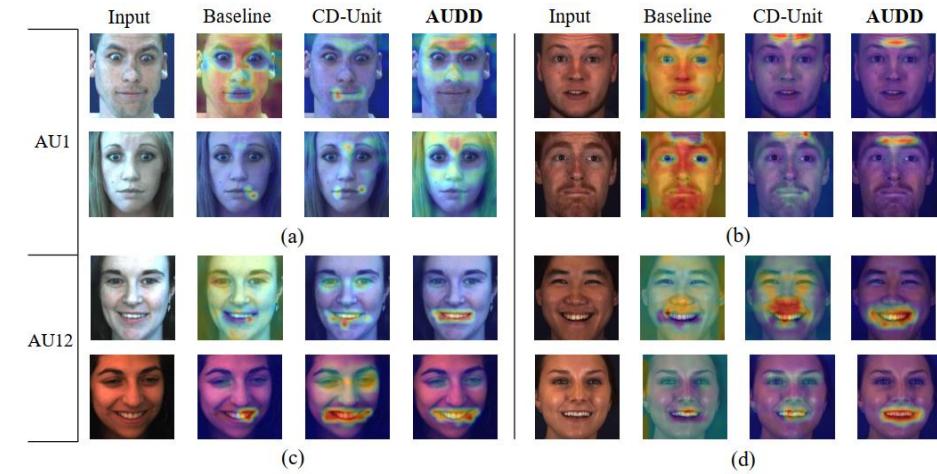
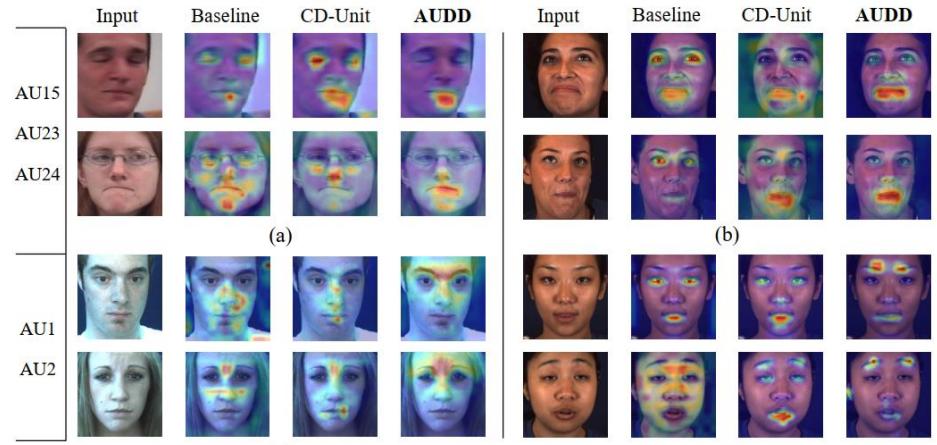


Fig. 4. Domain classification performance w.r.t. different blocks (Source: BP4D, Target:DISFA).

不同深度特征领域敏感度分析



跨域注意力图 (单个AU)



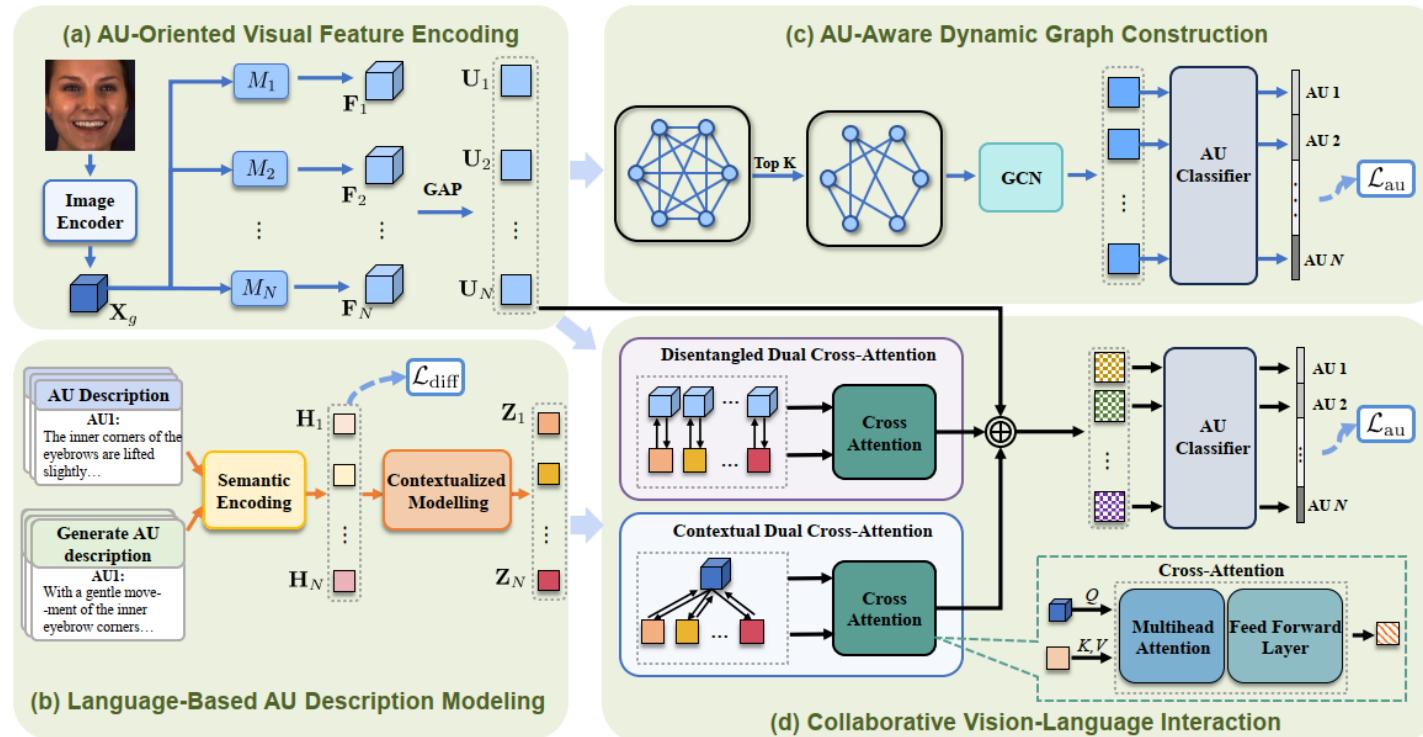
跨域注意力图 (组合AU)

多模态情感识别

□ 工作五：文本增强的多模态AU检测

□ 动机

- 挖掘AU的文本描述蕴含的丰富语义先验知识
- 学习更鲁棒的面部动作单元特征表示

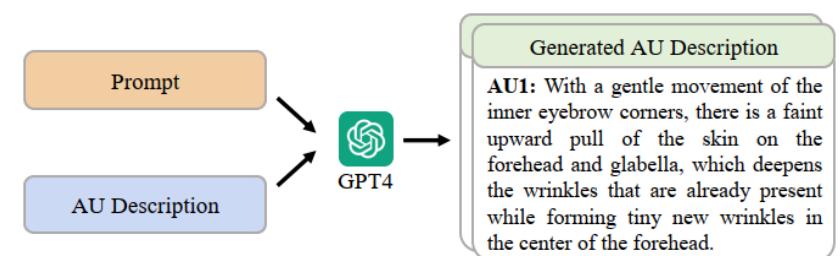


激活AU1的人脸图像

面部区域 肌肉运动

The inner corners of the eyebrows are lifted slightly, the skin of the glabella and forehead above it is lifted slightly and wrinkles deepen slightly and a trace of new ones form in the center of the forehead;

AU1的文本描述



基于大模型生成文本描述

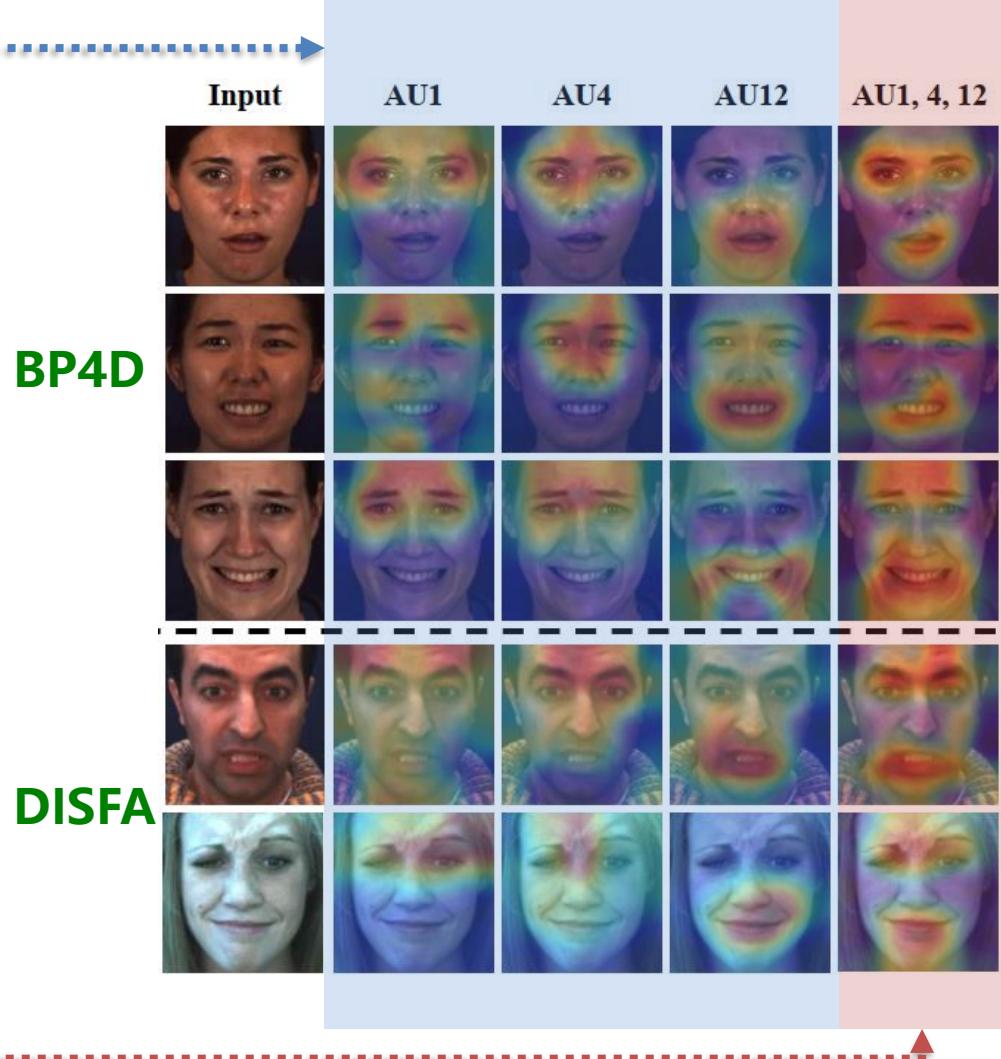
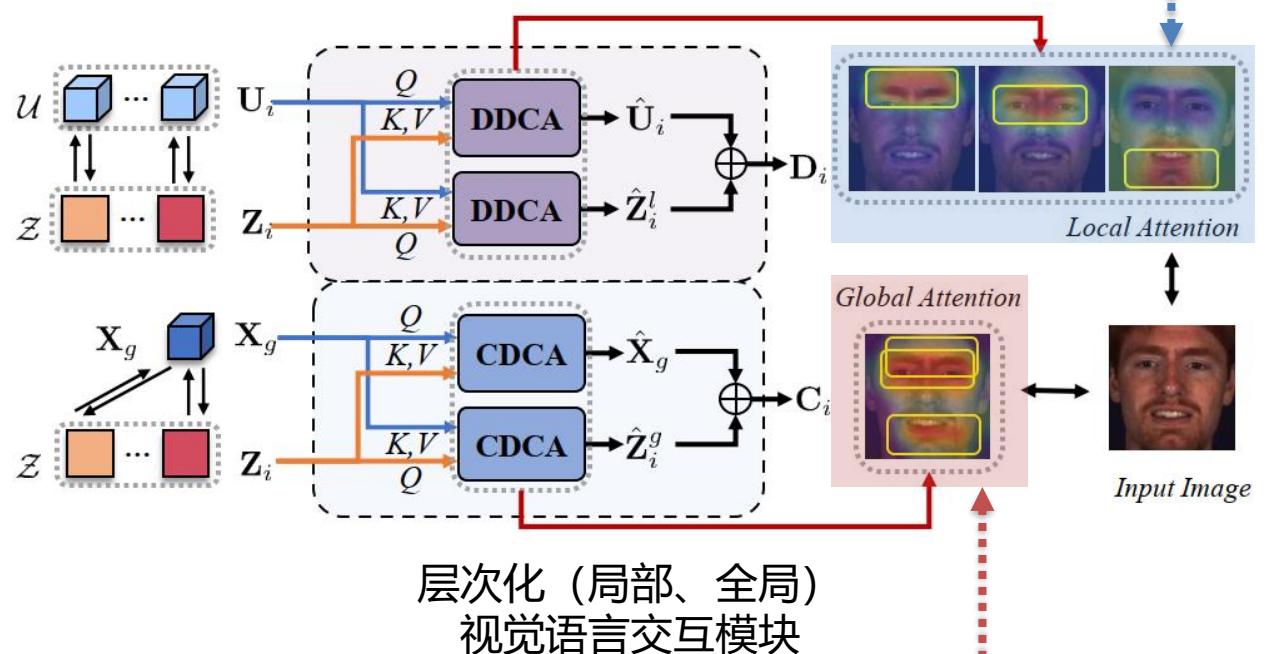
授权发明专利：**基于双重跨模态注意力的表情运动单元检测方法**, ZL202510634232.0

多模态情感识别

□ 工作五：文本增强的多模态AU检测

□ 实验结果

- 跨模态注意力图的可视化

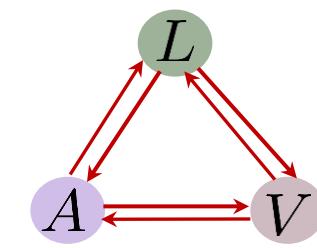
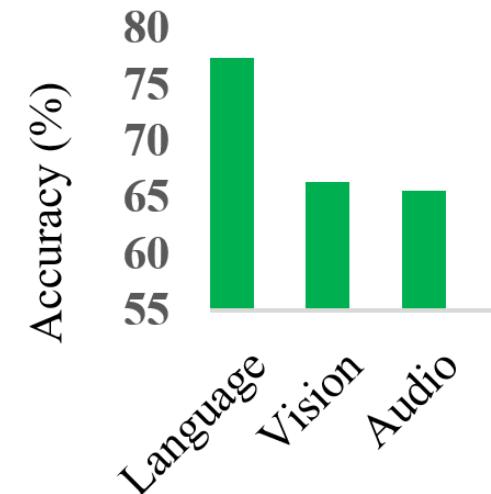
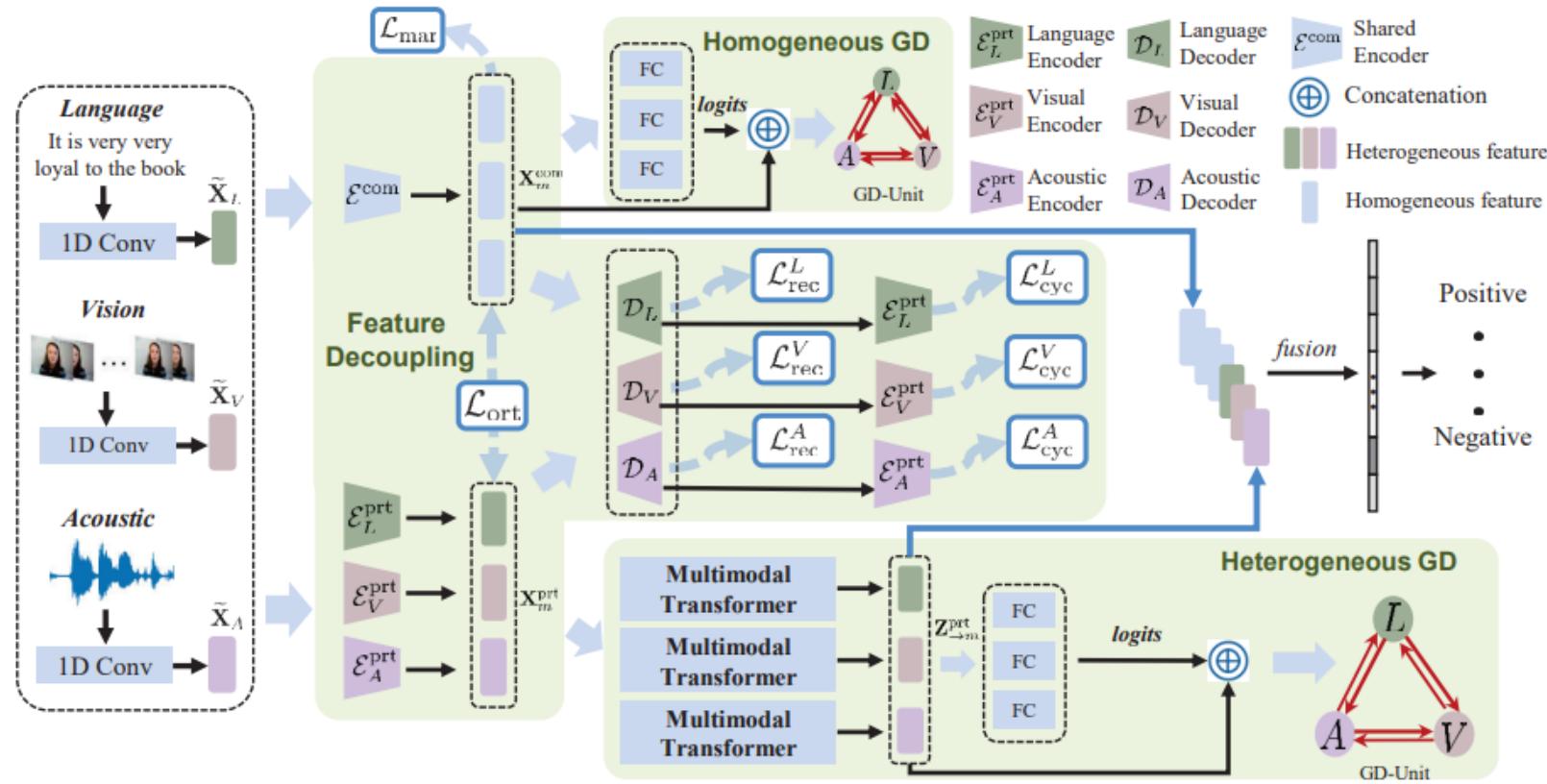




多模态情感识别

□ 工作六：面向跨模态互蒸馏的情感识别

□ 动机：语言模态占据绝对主导地位



Learnable Graph Edge:

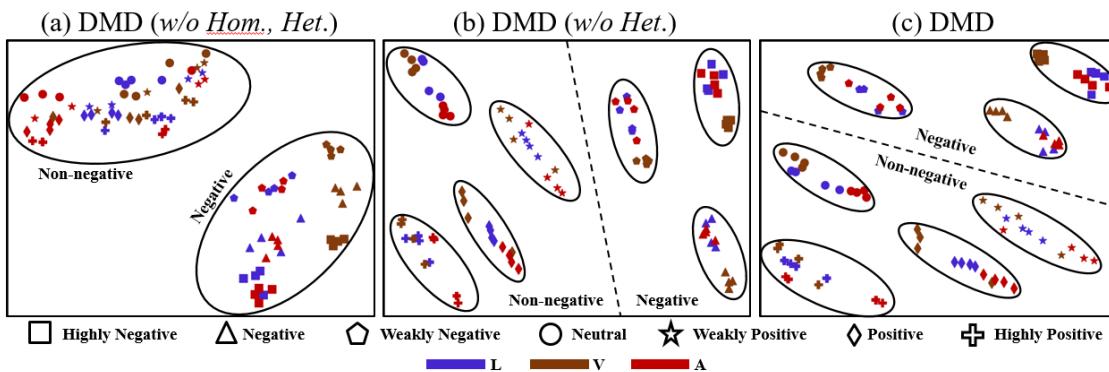
The graph edge means distillation strength. We encode the modality logits and the features into the graph edges:

多模态情感识别

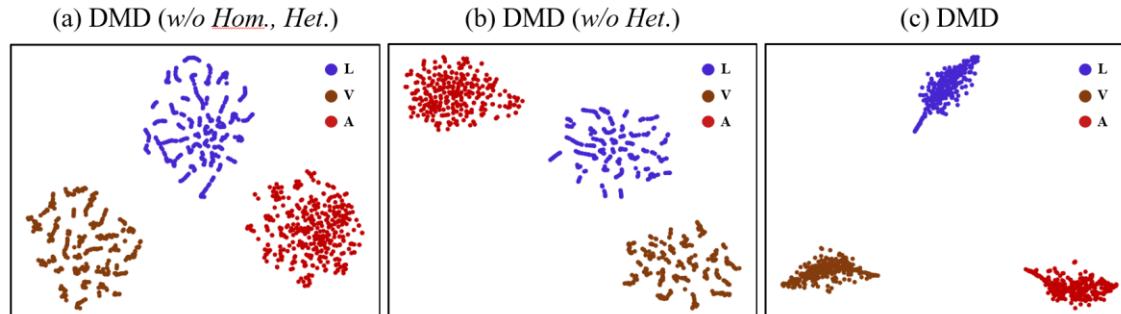
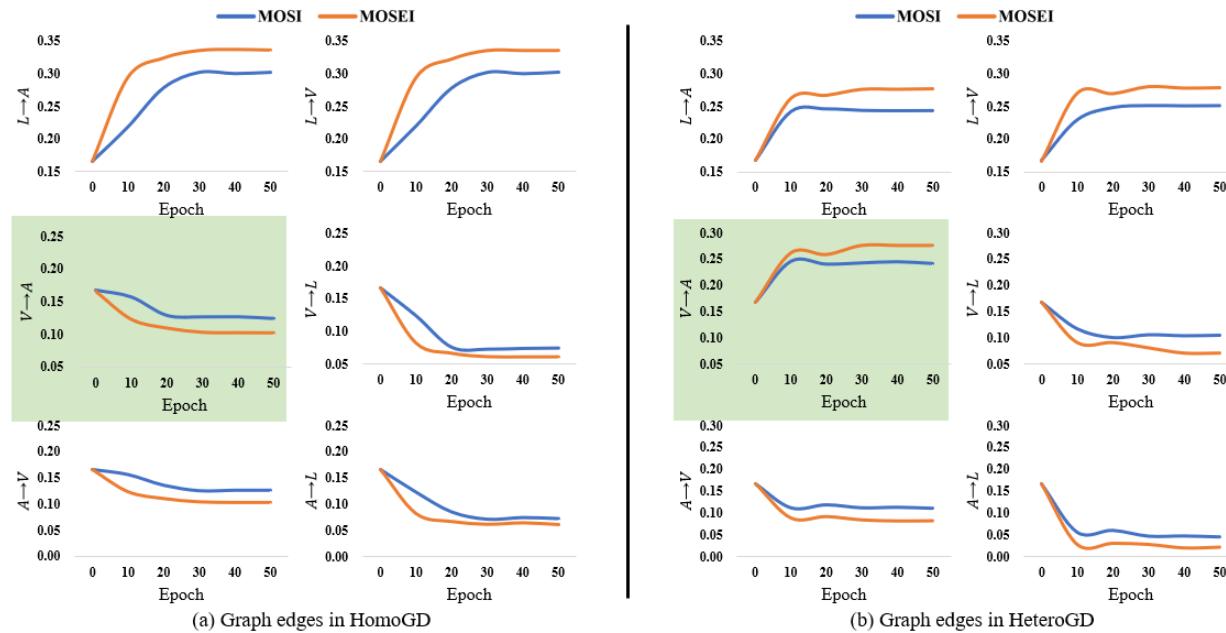
□ 工作七：面向跨模态互蒸馏的情感识别

□ 实验结果

授权发明专利：**一种基于特征解耦和图知识蒸馏的多模态情感识别方法**，ZL202310096857.7



同构空间特征：按**类别**聚类

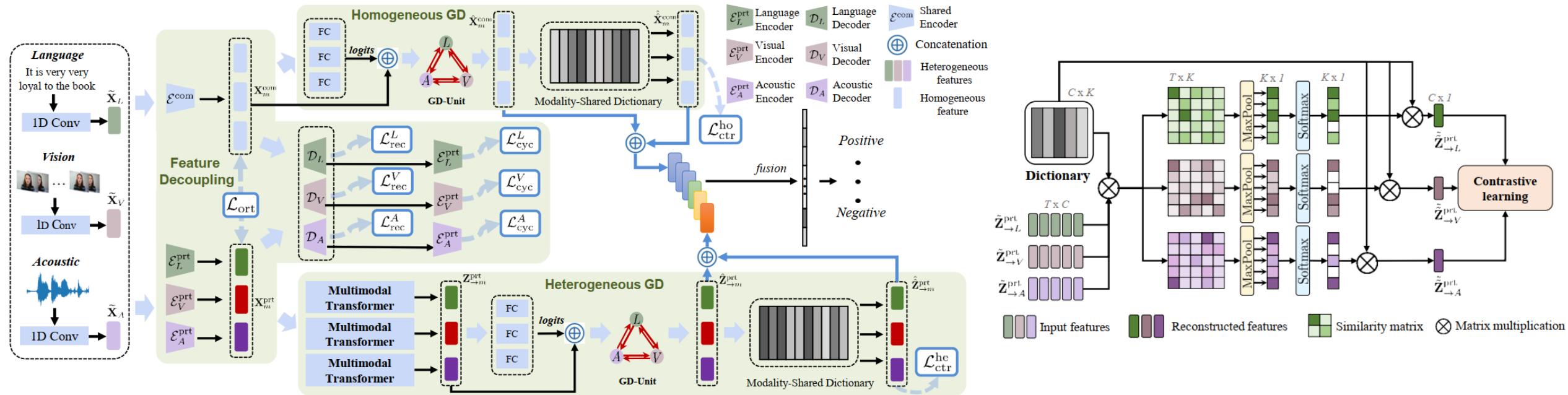


异构空间特征：按**模态**聚类

蒸馏权重可视化：
弱模态之间也可以互相促进！

多模态情感识别

- 工作七扩展：耦合多模态“均衡表示”与“语义对齐”
- 框架设计

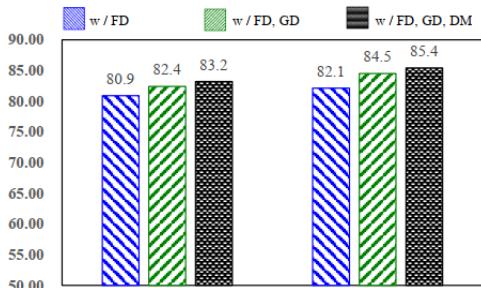


细粒度语义对齐：将不同模态特征尽可能映射到同一字典空间；

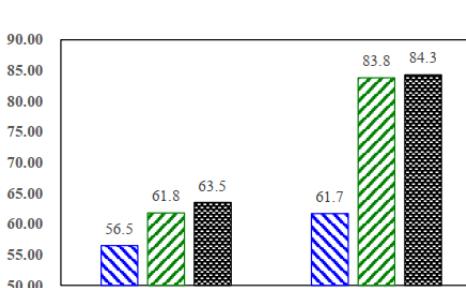
多模态情感识别

□ 工作七扩展：耦合多模态“均衡表示”与“语义对齐”

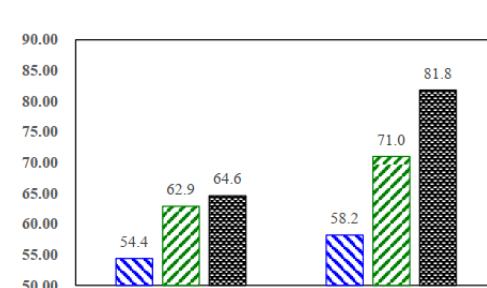
□ 实验分析



(a) Language modality

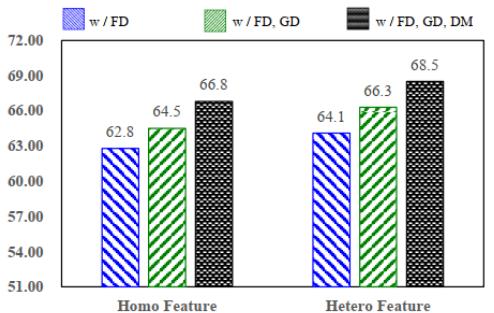


(b) Visual modality

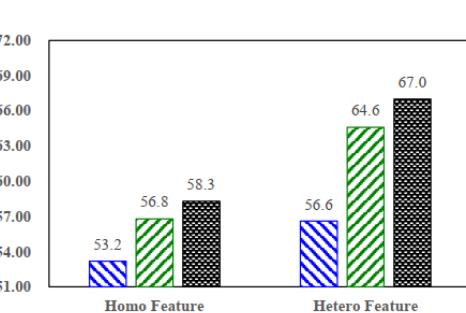


(c) Acoustic modality

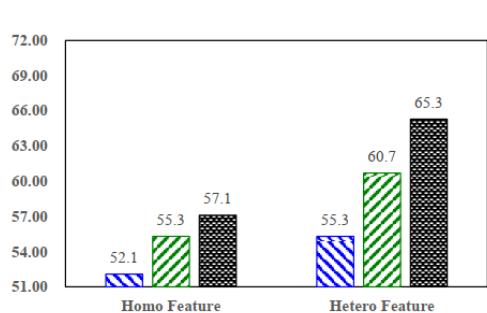
MOSEI数据集单模态识别精度



(a) Language modality

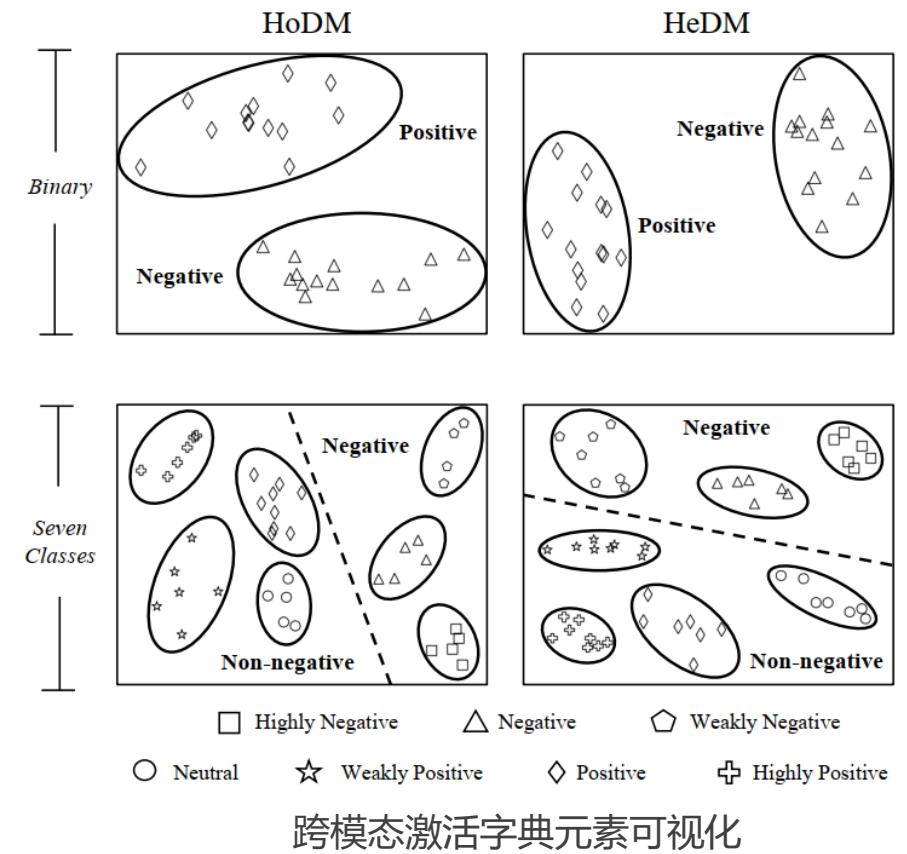


(b) Visual modality



(c) Acoustic modality

UR-FUNNY数据集单模态识别精度



跨模态激活字典元素可视化

多模态情感识别

突出性成果

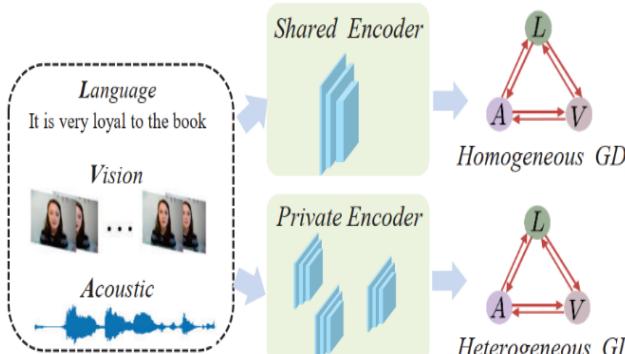
创新总结：构建自动化多模态互蒸馏框架，揭示异构多模态动态蒸馏机理，突破弱模态细微情感特征提取难题

挑战

如何自动、有效挖掘弱模态中蕴含的细微情感线索

创新点

提出“自动化多模态互蒸馏”算法，设计图蒸馏策略实现任意模态间互蒸馏，揭示异构多模态动态蒸馏机理，有效提升识别精度并具备解释性（CVPR’23）



中科大联合云知声在
Multimedia 2023的论文称：
受启发于我们的工作

3.2.4 *Graph Distillation (GD)*. Inspired by the success of using multimodal approaches in emotion recognition task [22], transferring knowledge between different modalities is beneficial to the task. Therefore, we propose to apply it to the task of humor detection. We define the different modality as node in graph, and the distillation strength from modality i to modality j is denoted as edge $\omega_{i \rightarrow j}$ connecting the corresponding nodes. We consider the

丹麦哥本哈根大学在
COLING 2024的论文称：
受启发于我们的工作

4.2.2. Subspace Constraint

Despite performing the aforementioned process, feature disentangling cannot be thoroughly guaranteed. There exists the potential for information to freely permeate between feature representations, whereby all modality information may be solely encoded in H_m^{hete} , which renders homogeneous (modality-agnostic) multi-modal features meaningless. Inspired by Li et al. (2023), we introduce a consistency constraint in the modality-agnostic subspace to strengthen the commonality across modalities, which is formulated as follows,

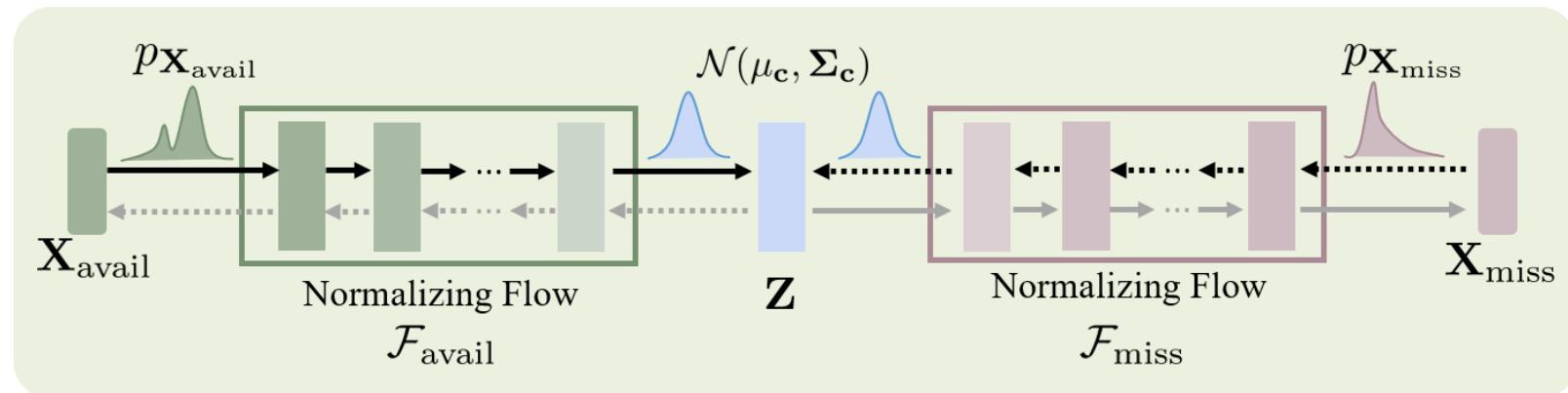
[1] Yong Li et al. Decoupled multimodal distilling for emotion recognition, CVPR 2023, Highlight, 被引163次

[2] Yong Li et al. Hierarchical Distillation of Cross-Modal Knowledge for Robust Emotion Recognition, T-PAMI, Under review

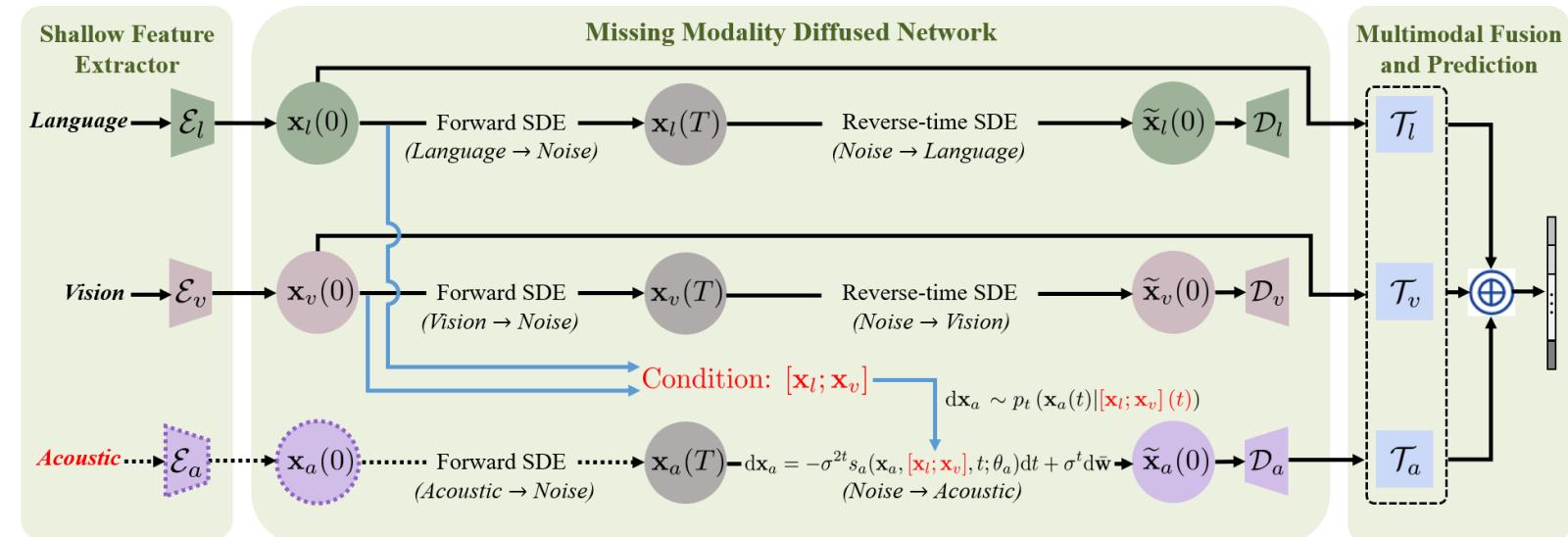
多模态情感识别

□ 工作八：基于分布一致模态补全的非完备多模态学习

□ 基于标准化流模型的
缺失模态生成



□ 基于多模态扩散生成的
不完备多模态情感分析



多模态情感识别

□ 工作八：基于分布一致模态补全的非完备多模态学习

□ 实验

- CMU提出的两个标准多模态数据集：CMU-MOSI[1] 和 CMU-MOSEI[2]
- 两种缺失模式：固定缺失模式和随机缺失模式



Examples of CMU-MOSEI

Table 1. Comparison on fixed missing protocol. The values reported in each cell denote ACC₂/F1/ACC₇. **Bold** is the best.

Datasets	Available	DCCA [1]	DCCAE [26]	MCTN [20]	MMIN [34]	GCNet [16]	DiCMoR (Ours)
CMU-MOSI	{L}	73.6 / 73.8 / 30.2	76.4 / 76.5 / 28.3	79.1 / 79.2 / 41.0	83.8 / 83.8 / 41.6	83.7 / 83.6 / 42.3	84.5 / 84.4 / 44.3
	{V}	47.7 / 41.5 / 16.6	52.6 / 51.1 / 17.1	55.0 / 54.4 / 16.3	57.0 / 54.0 / 15.5	56.1 / 55.7 / 16.9	62.2 / 60.2 / 20.9
	{A}	50.5 / 46.1 / 16.3	48.8 / 42.1 / 16.9	56.1 / 54.5 / 16.5	55.3 / 51.5 / 15.5	56.1 / 54.5 / 16.6	60.5 / 60.8 / 20.9
	{L, V}	74.9 / 75.0 / 30.3	76.7 / 76.8 / 30.0	81.1 / 81.2 / 42.1	83.8 / 83.9 / 42.0	84.3 / 84.2 / 43.4	85.5 / 85.4 / 45.2
	{L, A}	74.7 / 74.8 / 29.7	77.0 / 77.0 / 30.2	81.0 / 81.0 / 43.2	84.0 / 84.0 / 42.3	84.5 / 84.4 / 43.4	85.5 / 85.5 / 44.6
	{V, A}	50.8 / 46.4 / 16.6	54.0 / 52.5 / 17.4	57.5 / 57.4 / 16.8	60.4 / 58.5 / 19.5	62.0 / 61.9 / 17.2	64.0 / 63.5 / 21.9
	{L, V, A}	75.3 / 75.4 / 30.5	77.3 / 77.4 / 31.2	81.4 / 81.5 / 43.4	84.6 / 84.4 / 44.8	85.2 / 85.1 / 44.9	85.7 / 85.6 / 45.3
	Average	63.9 / 61.9 / 20.0	66.1 / 64.8 / 24.4	70.2 / 69.9 / 31.3	72.7 / 71.4 / 31.6	73.1 / 72.8 / 32.1	75.4 / 75.1 / 34.7
CMU-MOSEI	{L}	78.5 / 78.7 / 46.7	79.7 / 79.5 / 47.0	82.6 / 82.8 / 50.2	82.3 / 82.4 / 51.4	83.0 / 83.2 / 51.2	84.2 / 84.3 / 52.4
	{V}	61.9 / 55.7 / 41.3	61.1 / 57.2 / 40.1	62.6 / 57.1 / 41.6	59.3 / 60.0 / 40.7	61.9 / 61.6 / 41.7	63.6 / 63.6 / 42.0
	{A}	62.0 / 50.2 / 41.1	61.4 / 53.8 / 40.9	62.7 / 54.5 / 41.4	58.9 / 59.5 / 40.4	60.2 / 60.3 / 41.1	62.9 / 60.4 / 41.4
	{L, V}	80.3 / 79.7 / 46.6	80.4 / 80.4 / 47.1	83.2 / 83.2 / 50.4	83.8 / 83.4 / 51.2	84.3 / 84.4 / 51.1	84.9 / 84.9 / 53.0
	{L, A}	79.5 / 79.2 / 46.7	80.0 / 80.0 / 47.4	83.5 / 83.3 / 50.7	83.7 / 83.3 / 52.0	84.3 / 84.4 / 51.3	85.0 / 84.9 / 52.7
	{V, A}	63.4 / 56.9 / 41.5	62.7 / 59.2 / 41.6	63.7 / 62.7 / 42.1	63.5 / 61.9 / 41.8	64.1 / 57.2 / 42.0	65.2 / 64.4 / 42.4
	{L, V, A}	80.7 / 80.9 / 47.7	81.2 / 81.2 / 48.2	84.2 / 84.2 / 51.2	84.3 / 84.2 / 52.4	85.2 / 85.1 / 51.5	85.1 / 85.1 / 53.4
	Average	72.3 / 68.8 / 44.5	72.4 / 70.2 / 44.6	74.6 / 72.5 / 46.8	73.7 / 73.5 / 47.1	74.7 / 73.7 / 47.1	75.8 / 75.4 / 48.2

[1] Z. Amir, et al. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages.

[2] Z. Amir, et al. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph.

具体应用

□ 面向早期老年痴呆患者的康复训练精准评估



□ 研究背景

香港特别行政区研究资助局课题“科技增强的认知干预与照护支持效果评估”，面向早期阿尔兹海默症患者，探索并评估多种远程康复训练及互动策略的有效性。

□ 研究目标

探索多种康复训练护理策略的有效性，提升远程康复训练的智能化与精准化水平。

□ 技术路径

- 多模态行为数据建模
- 精细化视线轨迹估计

□ 成果贡献

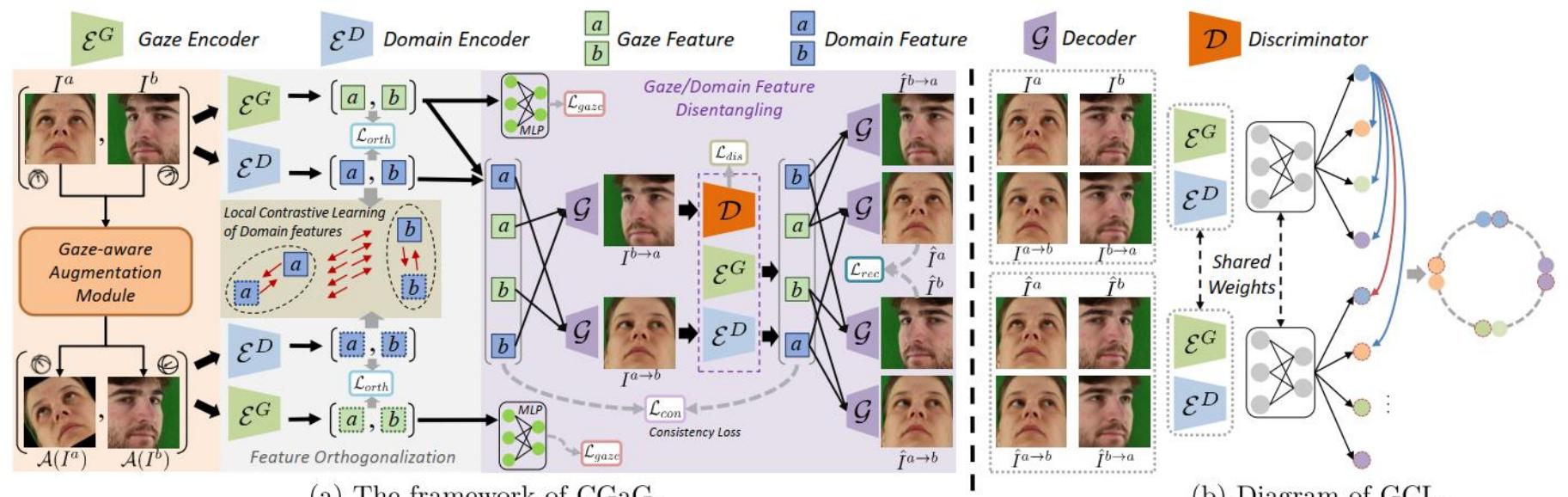
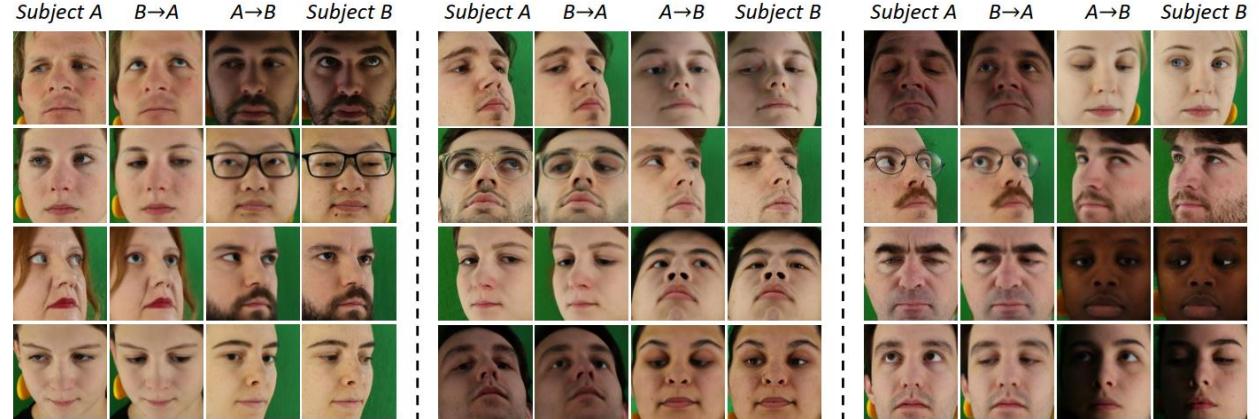
- 采集了174名患者、近2000段视频、约1630万张图像的康复训练数据
- 构建远程康复训练策略效果的量化评估体系
- 推动AI在老年认知障碍干预中的实际应用

具体应用

□ 聚焦复杂场景极端低标注条件下鲁棒视线估计

□ 难点梳理

- 图像质量非理想
尤其眼部区域视觉质量欠佳，
干扰因素多
- 标注稀缺



汇报提纲

- 多模态情感分析-问题定义与研究内容
- 多模态情感识别-研究背景及核心挑战
- 课题组相关进展-单模态、多模态情感识别研究进展
- 未来研究方向-大模型时代的多模态情感识别等

未来研究方向

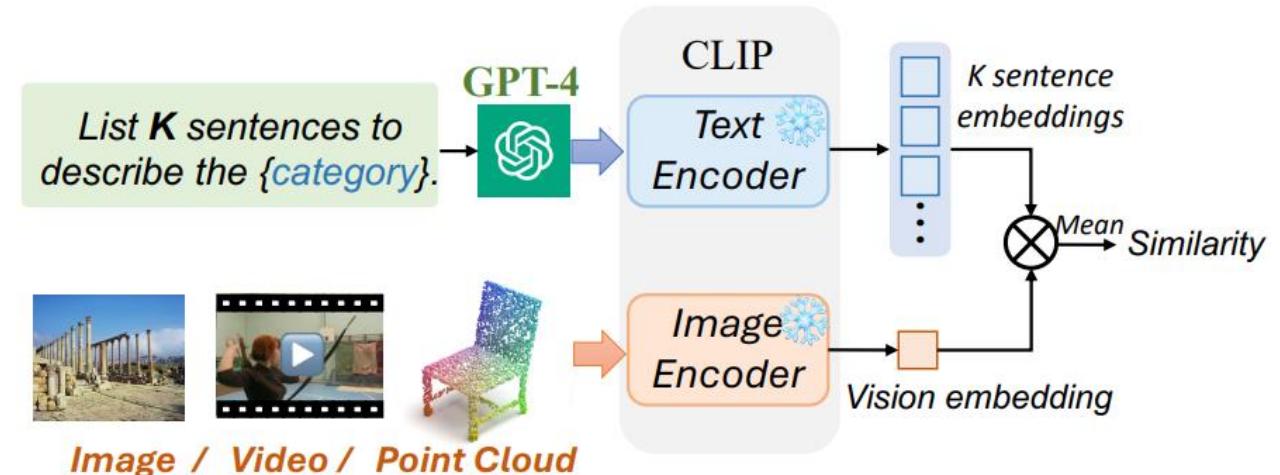
□ 大模型驱动的情感分析



摘自文献[1]

GPT-4V评测结论：

- AU识别精度高，评测来源单一
- 离散表情识别率低，采用Chain Of Thoughts有所提升
- 复合表情识别率一般



Dataset Backbone (#Param)	Real-world Affective Faces (RAF-DB)		
	Baseline	GPT Prompts	Top-1 Δ
CLIP ViT-B/32 (88M)	22.4 / 76.6	45.8 / 90.6	+23.4
CLIP ViT-B/16 (86M)	27.5 / 69.1	54.4 / 94.4	+26.9
CLIP ViT-L/14 (304M)	26.1 / 72.1	47.2 / 92.0	+21.1
EVA ViT-E/14 (4.4B)	31.0 / 90.9	54.9 / 93.7	+23.9
GPT-4V	68.7 / 93.8		

GPT-4V评测结论：

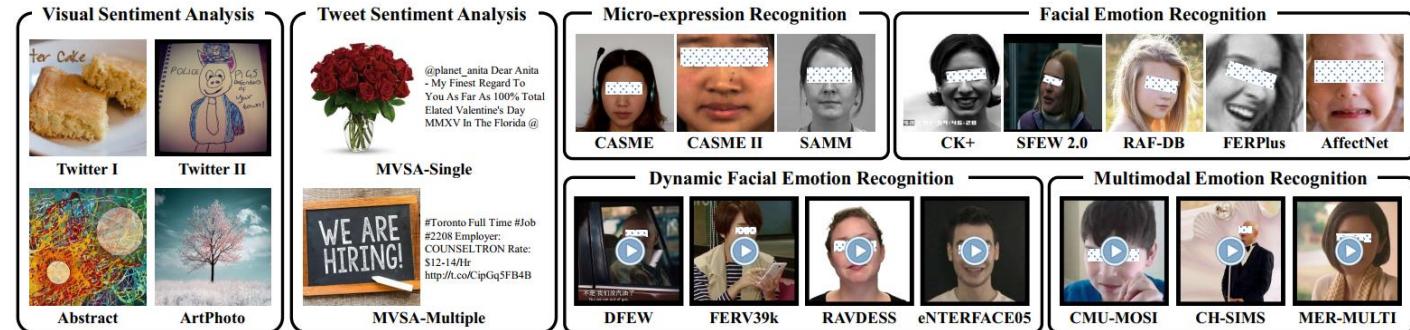
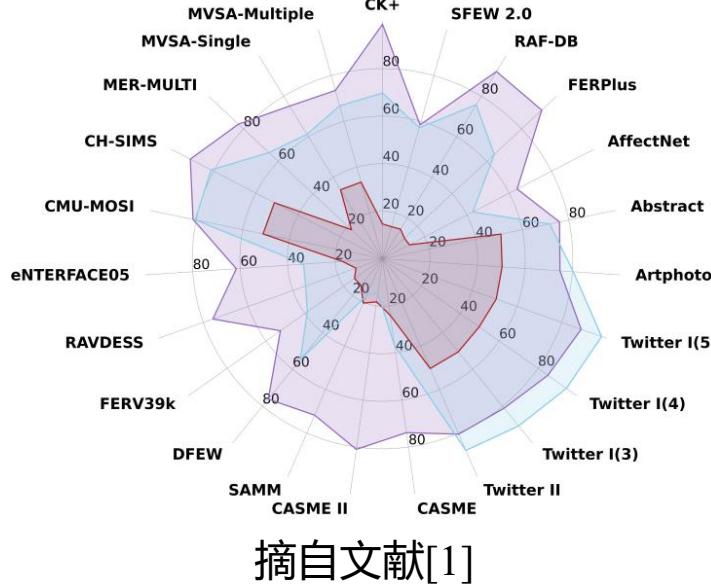
- 离散表情识别距离监督方法有很大差距

[1] GPT as Psychologist? Preliminary Evaluations for GPT-4V on Visual Affective Computing, *CVPRW 2024*

[2] GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition? *Arxiv, 2023*

未来研究方向

□ 大模型驱动的情感分析



Model	Twitter I (5/4/3)	Twitter II	Abstract	ArtPhoto
SentiBank [17]	71.32/68.28/66.63	65.93	64.95	67.74
PAEF [50]	72.90/69.61/67.92	77.51	70.05	67.85
DeepSentiBank [51]	76.35/70.15/71.25	70.23	71.19	68.73
PCNN [16]	82.54/76.52/76.36	77.68	70.84	70.96
VGGNet [52]	83.44/78.67/75.49	71.79	68.86	67.61
AR (Concat) [53]	88.65/85.10/81.06	80.48	76.03	74.80
Random	50.74/49.75/50.50	50.37	50.88	50.41
Majority	66.82/62.51/61.33	78.61	61.23	53.06
GPT-4V	97.81/94.63/90.71	87.95	71.81	80.40

Model	CMU-MOSI	CH-SIMS	MER-MULTI
MFM [62]	78.34	87.14	70.48
MISA [63]	79.08	89.82	82.42
MFN [64]	77.78	87.00	77.40
MMIM [65]	79.38	88.68	81.01
TFN [30]	79.63	90.56	82.52
MulT [66]	81.41	91.07	82.94
Random	51.33	51.23	17.87
Majority	42.37	50.47	10.40
GPT-4V	80.43	81.24	65.39

GPT-4V评测结论：

- 单模态视觉情感分析表现突出；
- 微表情以及细粒度视频情感识别表现不佳
- 预测结果方面呈现一定的不稳定性

单模态视觉情感分析

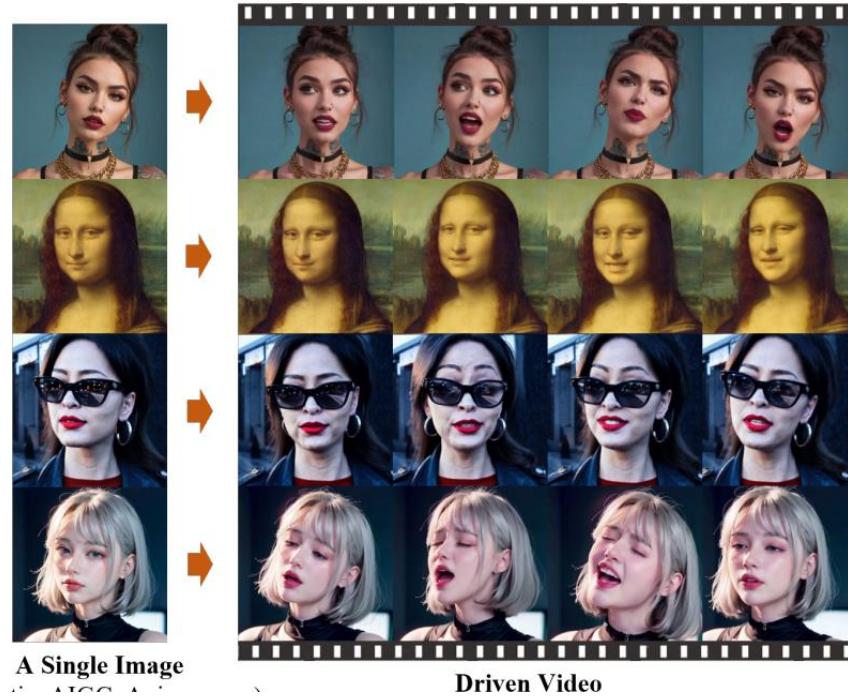
Model	RAF-DB	Model	FERPlus	Model	AffectNet
SCN [69]	87.03	RAN [70]	89.16	SCN [69]	60.23
EfficientFace [72]	88.36	SCN [69]	89.39	MA-Net [73]	60.29
MA-Net [73]	88.42	EAC [76]	89.64	ARM [77]	61.33
TransFER [80]	90.91	KTN [81]	90.49	DAN [79]	62.09
POSTER [83]	92.05	TransFER [80]	90.83	POSTER [83]	63.34
POSTER++ [85]	92.21	POSTER [83]	91.62	POSTER++ [85]	63.77
Random	14.57	Random	12.61	Random	12.73
Majority	38.64	Majority	35.75	Majority	12.50
GPT-4V	75.81	GPT-4V	64.25	GPT-4V	42.77

多模态情感分析

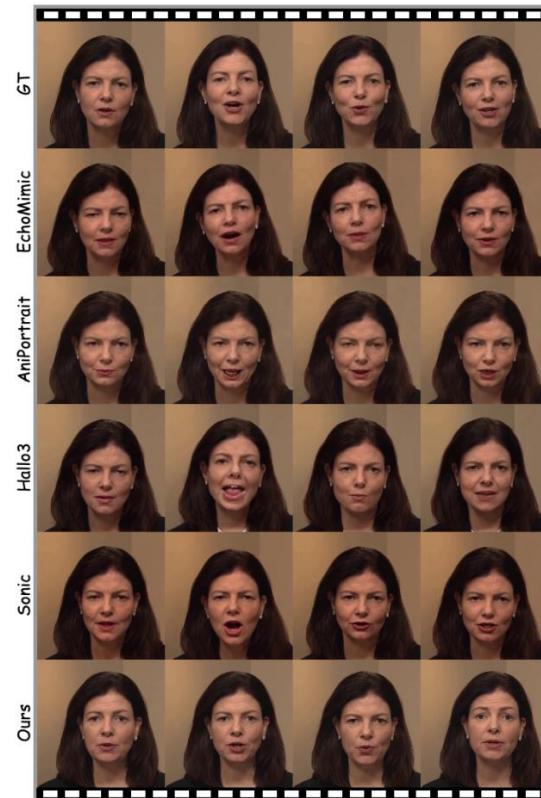
人脸表情识别

未来研究方向

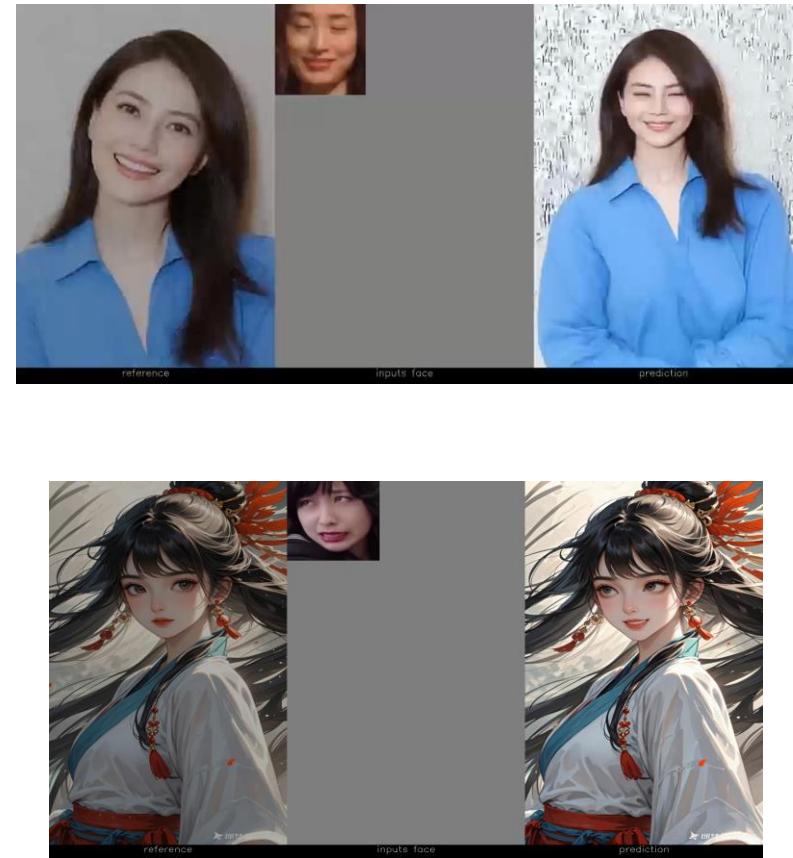
□ 数字人表情克隆/生成



Emo (ECCV 2024)[1] 系统评估了人脸表情的真实性 (新提出E-FID指标, 无法反映表情的时序自然性)



FantasyTalking[2] 分阶段学习“音频”与“视频”的对齐, 通过手动区域抠图的方式控制唇部运动与音频对齐, 基于序列人脸关键点方差控制表情



与字节跳动合作研发数字人

- [1] EMO: Emote Portrait Alive Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions, [ECCV 2024](#).
[2] FantasyTalking: Realistic Talking Portrait Generation via Coherent Motion Synthesis, [Arxiv 2025](#).

未来研究方向

□ 新的方法论

- 上下文和先验知识建模
 - 上下文信息，如会话和社会环境，会明显影响用户的情感体验。
 - 用户的先验知识，如个性和年龄，也与情感感知相关。
- 从未标记的、不可靠的、不匹配的情感信号中学习
 - 探索先进的机器学习技术，如自监督表示学习、动态数据选择和平衡、领域自适应、嵌入情感的特殊属性
- 行为和生理模态的紧密耦合
 - 抑郁症、焦虑症的检测及后续个性化音乐治疗
- 可信、可解释的情感克隆/生成
 - 表情/Gaze与其他信息（ID、姿态、年龄等）充分解耦
- 心智世界模型：面向心理感知的多模态大模型

未来研究方向

□ 上下文和先验知识建模

- 上下文信息，如会话和社会环境，会明显影响用户的情感体验。
- 用户的先验知识，如个性和年龄，也与情感感知相关。

Speaker A

Speaker B

阿杰，主持人马上就要公布冠军了！你觉得会是谁？

[期待]

快看，主持人拿到信封了！天啊！阿杰！冠军是你！

[兴奋]

不说，今天高手太多了。能进决赛我已经很满足了，冠军我想都不敢想。

[紧张]

.....我真没想到。

[惊喜]

年轻

年老



[担忧、紧张]

观看翼装飞行视频

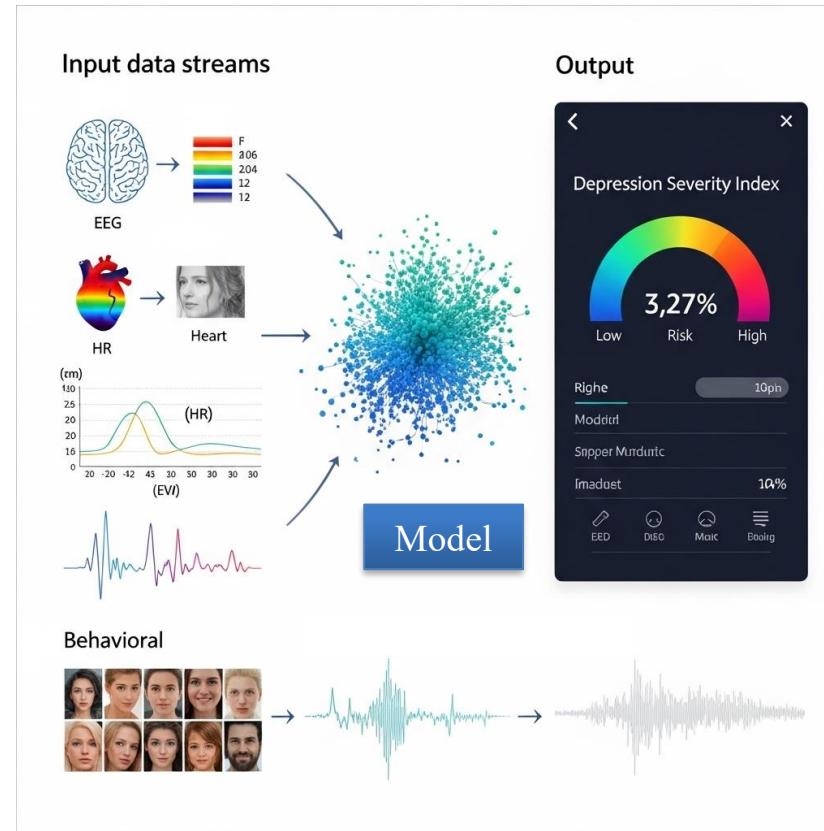
未来研究方向

□ 行为和生理模态的紧密耦合

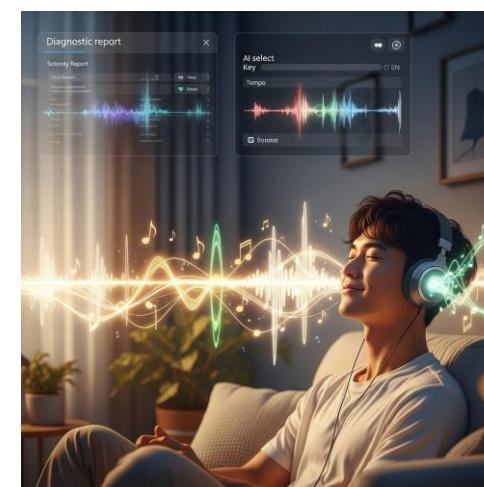
□ 抑郁症、焦虑症的检测及后续个性化音乐治疗

□ 心智世界模型：面向心理感知的多模态大模型

生理模态



精神类疾病无感诊断及免药物干预



个性化音乐治疗

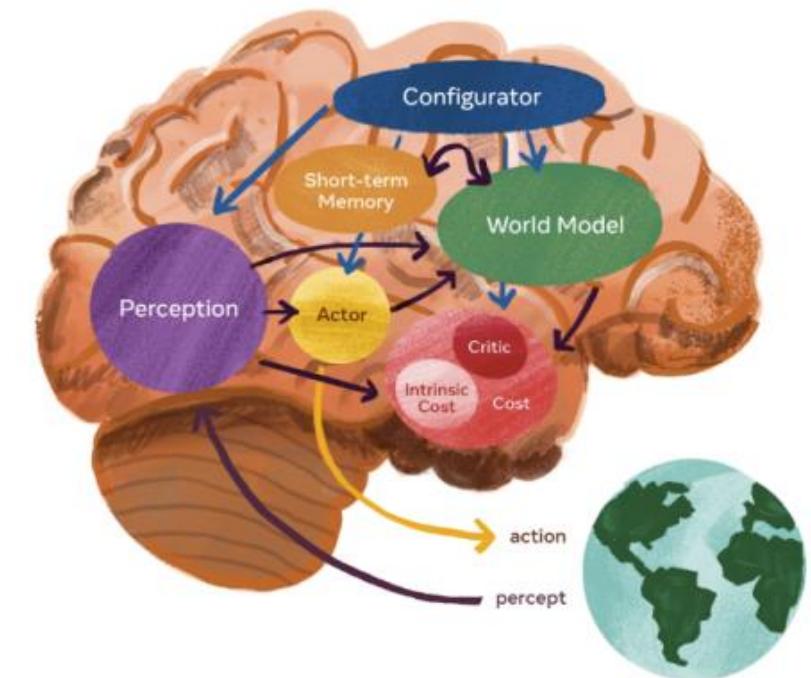


Figure 2 A modular system architecture for autonomous intelligence (LeCun, 2022)

多模态情感识别小结

方向

新方法论

实用设置

实际问题

应用



舆情



商业



健康



娱乐

方法

综述研究、单模态情感识别、多模态情感识别

情感模型、数据采集、情感标注、计算任务与框架、表征融合与学习

挑战

多模态固有挑战

数据缺失、标签缺失和
噪声、模态失衡、模态
冲突

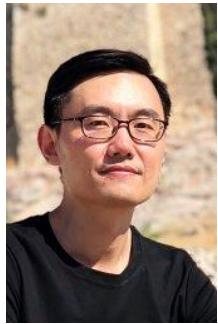
情感带来新挑战

情感鸿沟、情感主观性、
情感复杂性、情感模糊
性、情感微妙性

谢



山世光
计算所



耿新
东南大学



曾加贝
计算所



Antoni B. Chan
CityU



Cuntai Guan
NTU



王元植
南理工



任懿
南理工

谢

邮箱：yong.li@seu.edu.cn

主页：<https://mysee1989.github.io/>