自动化学报
ACTA AUTOMATICA SINICA

视听多模态学习综述

宣寒宇 陈强 吴之亮 韩向敏 董文祥 马楠

A Survey on Audio-visual Multi-modal Learning

XUAN Han-Yu, CHEN Qiang, WU Zhi-Liang, HAN Xiang-Min, DONG Wen-Xiang, MA Nan

---

视听多模态学习综述

宣寒宇[1, 2, 3] 陈 强[4] 吴之亮[5] 韩向敏[6] 董文祥[3] 马 楠[7]

摘 要 在人类信息获取过程中, 视听感知扮演着重要角色, 大脑通过整合视听信息, 形成统一、连贯且稳定的知觉体验. 视听多模态学习旨在模拟人类的视听多感官整合能力, 近年来受到研究者的广泛关注. 然而, 该领域在应用场景、任务目标和技术方法上呈现出显著的多样性, 目前尚且缺乏对视听多模态学习领域系统性回顾和分析的综合性中文综述. 基于人类的多感官整合机制在视听认知中的重要性以及不同视听多模态学习任务间的内在关联性, 提出一个统一框架, 将现有研究归纳为三类: 视听增强通过引入音频或视觉信息实现对初始单模态任务的增强效应; 跨模态交互旨在探索视听信息间的相互转换; 视听协作致力于探索视听信息的综合理解方法及其协同效应. 在此基础上, 对该领域中的最新研究进展进行系统性综述和总结. 此外, 深入剖析当前视听多模态学习研究所面临的五大核心共性问题和挑战——视听表征、对齐、转换、融合和共同学习; 探讨大模型背景下视听多模态学习的发展现状.

关键词 多感官整合; 视听多模态学习; 视听增强; 跨模态交互; 视听协作

引用格式 宣寒宇, 陈强, 吴之亮, 韩向敏, 董文祥, 马楠. 视听多模态学习综述. 自动化学报, xxxx, **xx**(x): x−xx

DOI 10.16383/j.aas.c250341 **CSTR** 32138.14.j.aas.c250341

A Survey on Audio-visual Multi-modal Learning

XUAN Han-Yu[1, 2, 3] CHEN Qiang[4] WU Zhi-Liang[5] HAN Xiang-Min[6] DONG Wen-Xiang[3] MA Nan[7]

Abstract In the process of human information acquisition, audio-visual perception plays crucial roles, with the brain integrating audio-visual information to form a unified, coherent, and stable perceptual experience. Audio-visual multi-modal learning (AVMML) aims to simulate human capacity for audio-visual multi-sensory integration, which has garnered significant attention from researchers in recent years. However, this field exhibits significant diversity in application scenarios, task objectives and technical methodologies. Currently, there is a lack of a comprehensive Chinese survey that systematically reviews and analyzes the field of AVMML. In this paper, we proposes a unified framework based on the importance of human multi-sensory integration mechanisms in audio-visual cognition and the intrinsic relationships among different AVMML tasks. Under our framework, existing AVMML researches can be categorized into three main types: Audio-visual enhancement, which improves the performance of initial uni-modal tasks by incorporating audio or visual information; Cross-modal interaction, which explores the mutual translation between audio and visual information; Audio-visual collaboration, which investigate comprehensive understanding methods and synergistic effects of audio-visual information. Building on this, this paper systematically reviews and summarizes the latest research progress in the AVMML field. Additionally, we provides an in-depth analysis of five core common issues and challenges faced by current AVMML research, covering audio-visual representation, alignment, translation, fusion, and co-learning; We also discusses the development status of AVMML in the context of large models.

Keywords multi-sensory integration; audio-visual multi-modal learning; audio-visual enhancement; cross-modal interaction; audio-visual collaboration

Citation Xuan Han-Yu, Chen Qiang, Wu Zhi-Liang, Han Xiang-Min, Dong Wen-Xiang, Ma Nan. A survey on audio-visual multi-modal learning. *Acta Automatica Sinica*, xxxx, **xx**(x): x−xx

收稿日期 2025-07-24 录用日期 2025-12-31

Manuscript received July 24, 2025; accepted December 31, 2025

国家自然科学基金 (62302006, 62502447, 62371013), 安徽省自然科学基金 (2308085QF221), 安徽省高校科研计划项目 (2024AH040012) 资助

Supported by National Natural Science Foundation of China (62302006, 62502447, 62371013), Anhui Natural Science Foundation (2308085QF221), and Anhui Provincial University Scientific Research Project (2024AH040012)

本文责任编委 林宙辰

Recommended by Associate Editor LIN Zhou-Chen

1. 安徽大学大数据与统计学院 合肥 230601 2. 中国科学技术大学人工智能与数据科学学院 合肥 230026 3. 合肥综合性国家科学中心数据空间研究院 合肥 230088 4. 安徽大学计算机科学与技术学院 合肥 230601 5. 浙江大学计算机科学与技术学院 杭州 310000 6. 清华大学软件学院 北京 100084 7. 北京工业大学信息科学技术学院 北京 100124

1. School of Big Data and Statistics, Anhui University, Hefei 230601 2. School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei 230026 3. Institute of Dataspace, Hefei Comprehensive National Science Center, Hefei 230088 4. School of Computer Science and Technology, Anhui University, Hefei 230601 5. School of Computer Science and Technology, Zhejiang University, Hangzhou 310000 6. School of Software, Tsinghua University, Beijing 100084 7. School of Information Science and Technology, Beijing University of Technology, Beijing 100124

1 背景与基础

1.1 视听多模态学习背景

1.1.1 发展背景

在人类感知过程中, 所获取的信息本质上是多种模态的. 心理学家 Treichler[1] 的研究表明, 人类从周围环境中获取的信息, 视觉和听觉分别占比 83% 和 11%. 尽管人类通过不同的感官通道接收信息, 但大脑并未将不同感官的信息孤立或碎片化. 相反地, 大脑左后侧的颞上沟能有效地整合来自多个感官通道内的多样化线索[2], 形成统一、连贯且稳定的知觉体验. 在神经科学领域, 这一过程被称作多种感官通道线索的知觉整合, 简称多感官整合[3]. 这种多感官刺激的“联合”作用, 将产生“整体大于局部之和”的效果.

如图 1 所示, 相机和麦克风这两种最常见的设备, 分别扮演着机器“眼睛”和“耳朵”的角色. 同时, 视频和音频也是互联网中广泛存在的数据形式. 全球最大的视频共享网站 YouTube 发布的最新数据表明, 人们每天观看的视频累计时长达 10 亿小时, 每天上传的视频累计时长达 72 万小时. 这些视频本质上是多模态的, 即既是可看的也是可听的.

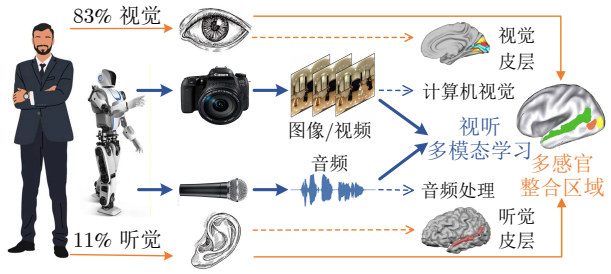


图 1 视听多模态学习示意图

Fig. 1 Diagram of audio-visual multi-modal learning

在我国, 随着互联网技术的崛起和移动设备的普及, 短视频平台 (如抖音和 Bilibili 等) 的蓬勃发展打破了传播的壁垒, 也激发着更多人的参与热情, 大量的短视频内容被创造和分享. 根据中国互联网络信息中心发布的最新数据, 我国网络视听用户规模达 10.26 亿人, 互联网音视频规模达 2 412 亿元.

在此背景下, 面对海量且天然耦合的视听数据, 视听多模态学习 (audio-visual multi-modal learning, AVMML) 作为一种新兴的技术范式应运而生. 作为人工智能 (artificial intelligence, AI) 领域的前沿研究方向, 该方向已成为学术界关注的热点, 其学术影响力与产业价值持续攀升. 如图 2 所示的年度文献统计分析结果表明, 近年来视听多模态学习相关研究成果的发表数量呈稳定上升态势. 如图 3(a) 所示, 其研究范畴涵盖了分类、分割、定位、生成、推理等多元化任务场景. 这一发展态势不仅彰显了视听多模态学习在突破单模态感知瓶颈方面的独特优势, 其强大的技术延展性, 也不断驱动着跨领域应用的创新与升级.

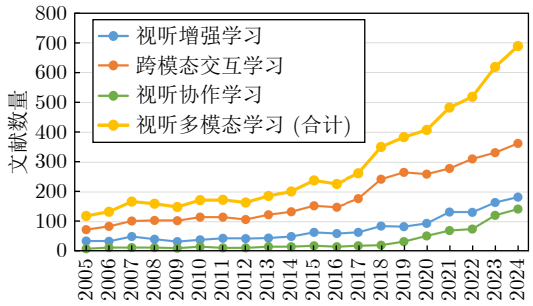


图 2 视听多模态学习文献年度统计分析图

Fig. 2 Annual statistical analysis chart of audio-visual multi-modal learning

1.1.2 应用背景

视听多模态学习作为 AI 技术体系的代表性范式, 通过视觉与听觉信息的深度耦合与协同处理, 实现对环境更全面、更精准的感知建模与语义理解. 这一技术范式正在重构传统行业的技术边界[4–5], 展现出显著的跨领域变革潜力.

在公共安全领域, 视听融合技术显著提升安防效能. 例如, 通过声学增强图像技术, 改善低光照环境下的取证质量; 行为识别系统结合声音特征与视频分析, 实现暴力事件智能预警; 声纹与人脸特征的双因子认证机制, 大幅提升身份识别的防伪能力.

在医疗健康领域, 多模态技术助力精准诊疗. 例如, 结合视听信号的情绪识别系统, 为自闭症患者提供实时情绪监测与干预支持; 结合骨传导音频与计算机视觉的场景重建技术, 提升视障人士的环境感知与自主导航能力; 结合内窥镜影像与手术器械声纹的多模态分析技术, 提升微创手术的操作精准度与安全性.

在教育信息化领域, 视听交互技术实现教学革新. 例如, 结合眼动追踪与语音情感分析, 实现对学习者状态的实时监测与量化评估; 结合唇部运动特征的声纹分离技术, 解决线上课堂中多人语音重叠难题; 结合视频语义理解与语音问答, 为深度知识推理提供有力支持.

在智能交通领域, 视听融合赋能感知跃升. 例如, 通过声学信号与视觉数据的协同分析, 增强复杂交通场景下障碍物检测精度; 基于音频事件触发的显著性跟踪技术, 大幅缩短交通事故等突发状况

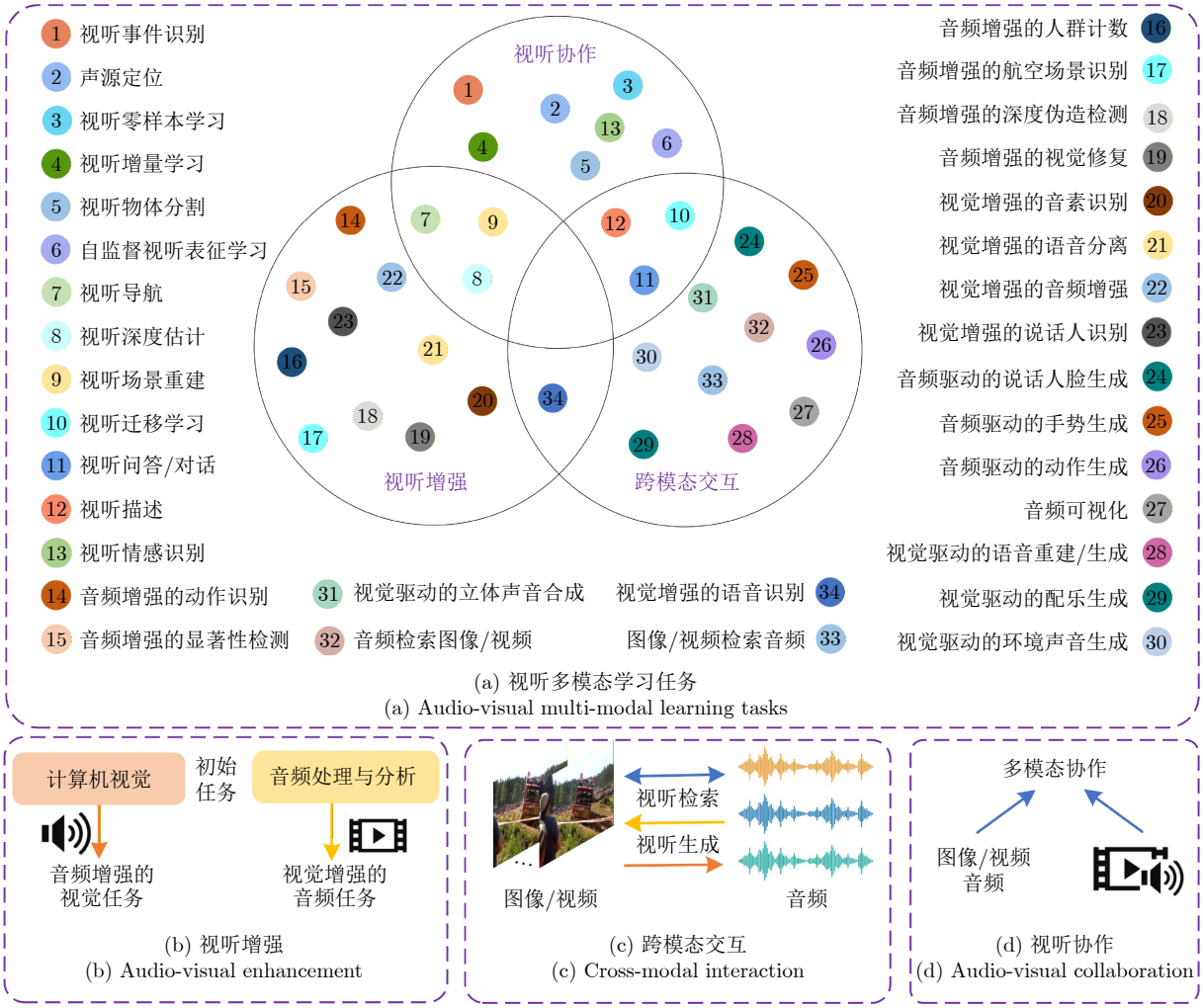


图 3 视听多模态学习任务及其分类示意图

Fig. 3 Diagram of audio-visual multi-modal learning tasks and their classification

的应急响应时间; 通过融合面部表情特征与语音特征, 提升疲劳驾驶检测准确率.

在数字娱乐领域, 视听生成技术创造沉浸体验. 例如, 音乐驱动的虚拟角色动作生成技术, 实现数字人与用户间的自然交互; 动态画面自适应的智能音频优化技术, 通过实时声场调节, 增强观众在虚拟环境下的临场真实感; 空间音频与视觉渲染的帧级精准同步技术, 缓解虚拟现实中的运动眩晕问题.

在工业制造领域, 视听协同技术推动质检与运维升级. 例如, 融合焊接声纹与高速视觉, 实现微米级焊缝缺陷的精准识别; 基于机械振动声学特征与红外视频的智能诊断技术, 实时预警轴承磨损故障; 结合语音指令与立体视觉, 实现基于自然语言的精密装配与校正.

在智慧农业领域, 多模态感知技术支持精准生产管理. 例如, 结合冠层图像与叶片振动声音, 实现病虫害早期识别与精准施药; 融合无人机遥感视频与田间声景, 评估水肥胁迫与灌溉决策优化; 结合畜禽行为视频与声音识别, 实时监测群体健康状态并预警异常状况, 提升养殖管理效率.

然而, 随着应用场景的不断拓展和深入, 视听多模态学习技术在实际落地过程中也面临诸多挑战. 具体表现在以下几个方面:

- 数据标注成本高: 以手术视频为例, 对手术过程中的音视频进行阶段、动作和器械标注, 需要资深医生参与, 人力成本极高, 且标注一致性难以保障;
- 数据异构性显著: 音频与视觉数据在信号形式、采样方式、空间结构和语义密度等方面存在显著差异;
- 模型复杂度高: 例如, 会议转录与发言人识别任务中, 需同时处理视频流与多路音频, 模型参数量大、推理延迟高, 难以在边缘设备中实时运行;
- 实际环境干扰多: 例如, 在安防监控中, 嘈杂

街道上的目标人声容易被环境噪声淹没, 视觉遮挡也会使人脸识别与唇读分析失效;

● 隐私保护难度大: 例如, 在线教育平台收集的学生课堂音视频, 既包含人脸、声纹等生物信息, 也反映学生的学习行为和状态, 其传输、存储与使用面临严格的合规要求.

1.2 视听多模态学习基础

1.2.1 认知基础

人类大脑通过多感官通道的协同处理, 实现复杂环境下的高效感知[2]. 其中, 视觉系统通过视网膜接收光学信息, 听觉系统则经由鼓膜获取声波信号. 这些感官输入经特异性神经通路, 传导至大脑皮层进行交互整合. 视觉感知主要依赖三类线索[6]: 生理线索通过视轴辐合运动, 编码物体动态特征; 双眼线索基于视网膜视差构建深度知觉; 单眼线索则包含透视几何、遮挡关系等空间信息. 而听觉感知则通过双耳时差定位声源方位, 并依据声强衰减模型估算距离[2, 6].

认知神经科学的相关研究表明[3], 尽管视听信号在物理特性上存在本质差异, 但大脑通过颞上沟后部等多感观整合区域的神经振荡同步机制, 将不同感官通道的知觉线索有效地合并为统一、连贯且强健的知觉体验. Calvert 等[7] 揭示了视听整合发生在神经处理的初级阶段. 其随后研究 [2] 进一步定位了视听整合的关键脑区–大脑皮层左侧颞上沟后部, 并观察到显著的视听整合效应: 当个体同时接收说话者的唇部运动与对应语音时, 左侧颞上沟区域的神经激活强度显著高于单一模态刺激条件下的激活水平. 这种视听增强效应不仅在严格控制的实验室环境中得以验证, 更在日常生活场景中表现出普遍性. 典型例证是“鸡尾酒会”场景, 个体可通过观察说话者的唇部运动, 显著提升对目标说话者语音识别的准确率.

此外, Bedny[6] 针对先天失明群体的研究发现, 尽管其从未接收过视觉感官的输入, 但其在执行阅读任务时, 视觉皮层仍可被激活, 这一现象表明视觉皮层具备功能重塑能力. 该发现进一步支持了感知皮层的跨模态感知特性: 即使缺乏某一模态的典型输入, 相关脑区仍可通过功能代偿参与其他模态的信息处理.

上述研究从不同角度阐释了多感官整合对视听认知的核心作用[8], 其重要性可归纳为:

● 多感官增强效应: 视听觉信息的协同作用显著提升环境感知的清晰度;

● 跨模态可塑性: 单一感官刺激可触发其他模态的神经资源代偿性分配;

● 多感官整合: 大脑皮层存在专门的视听线索整合功能区域.

这些发现不仅深化了人类对多感官信息整合神经机制的科学认知, 更从认知神经科学层面为视听多模态学习领域构建了基础.

1.2.2 理论基础

在信息论与多模态学习的交叉研究领域, Shannon[9] 的信息熵理论为信息的量化分析提供重要理论基础. 假设事件空间 F 的概率分布已知, 记事件 $x \in F$ 的发生概率为 $p(x)$, 其信息量可被表示成:

$$h(x) = -\log p(x) \tag{1}$$

该式表明, 事件 x 的信息量 $h(x)$ 与其发生概率 $p(x)$ 呈负对数关系.

信息以各种模态为载体, 在视听多模态信息处理框架下, 将事件空间 F 表示为听觉模态 a 和视觉模态 v 在空间 $\mathcal{S}$ 和时间 $\mathcal{T}$ 上的张量, 则个体在事件空间中的信息量 K 可被定义为:

$$K = \|T \odot (I_a + I_v)\| \tag{2}$$

其中, I_a 和 I_v 分别表示在听觉模态和视觉模态中所有事件信息量所构成的矩阵; $\odot$ 表示矩阵点乘; T 表示个体对视觉事件和听觉事件的注意力分配矩阵.

基于人类认知的有限容量假设, 注意力资源的归一化约束可表述为: 对于任意时空位置, 注意力分配矩阵 T 满足 $\|T\|_1 = 1$. 当感知模式从单一模态感知转变为视听觉的多模态联合感知时, 个体在感知过程中会动态调整注意力分配策略, 以实现信息量的最大化获取. 这一优化过程可被表示为:

$$K^* = \max_{\|T\|_1=1} \|T \odot (I_a + I_v)\| \tag{3}$$

由上述优化目标可知, 在多模态共现的时空事件场景中, 通过合理的注意力分配策略, 信息量可达到理论最大值. 因此, 机器系统可通过多模态数据整合, 实现对感知信息的更全面捕获与高效利用.

如表 1 所示, 基于上述信息优化理论, 视听多模态学习相较于传统单模态方法展现出以下核心优势:

● 多源异构数据处理: 能够并行处理图像、视频、音频等多种模态数据, 这些多源异构数据提供了更全面、更丰富的特征表示;

● 跨模态知识迁移: 知识可以从一种模态中获取, 也能被应用于另一种模态, 如从音频数据中获取的知识能被应用于视觉任务中;

● 噪声鲁棒: 不同模态的噪声特性形成天然互补, 如音频数据对光照变化不敏感, 视觉数据可修

表 1 视/听单模态学习与视听多模态学习
Table 1 Visual/audio uni-modal learning and audio-visual multi-modal learning

		数据处理能力	知识迁移能力	噪声鲁棒能力
计算机视觉/音频处理 (CV/AP)	单模态学习	仅能处理图像、视频、音频单一模态数据	知识只能从一种模态中学习并应用在该模态中	容易受到模态自身数据噪声的影响
视听多模态学习	多模态学习	能同时处理图像、视频、音频多模态数据	知识可以从一种模态获取，也能应用于不同模态	各模态数据噪声互不影响，且信息可相互补充

正声学混响效应.

1.3 本文思路及组织结构

1.3.1 本文思路

如图 2 所示, 作为 AI 领域的重要前沿方向, 视听多模态学习近年来在学术界获得了持续关注与快速发展. 然而, 尽管研究热度持续升温, 该领域仍面临两大核心挑战. 一方面, 如图 3(a) 所示, 当前研究多聚焦于特定任务场景, 缺乏统一的理论框架与方法论体系, 导致不同任务间的技术积累难以有效复用; 另一方面, 尽管国际顶会和期刊上已涌现大量高水平成果, 但面向中文语境的系统性综述仍存在空白. 这一现状严重制约了国内研究团队的知识体系构建和技术创新.

这种“任务驱动型研究”与“理论体系缺失”之间的结构性矛盾, 凸显了构建统一理论框架的必要性和紧迫性. 基于第 1.2 节对多感官整合机制在视听认知中的重要性分析, 以及不同视听多模态学习任务间的内在关联性研究, 本文系统梳理了现有视听多模态学习的相关研究成果, 并将其归纳为如图 3(b) ~ 图 3(d) 所示的三类:

1) 视听增强: 计算机视觉 (computer vision, CV) 和音频处理 (audio processing, AP) 作为 AI 领域的两大核心研究领域, 虽已取得显著进展, 但其单模态范式存在固有局限性 (如表 1 所示). 为此, 研究者开始探索在初始 CV/AP 任务中, 引入互补模态, 即音频增强的 (audio-enhanced) 视觉任务和视觉增强的 (vision-enhanced) 音频任务. 其本质是利用异构模态的互补性, 增强初始任务对复杂环境的适应能力.

2) 跨模态交互: 为赋予机器类似人类的跨模态可塑能力, 视听跨模态交互进一步探索模态间的代偿性交互机制[8]. 该方向致力于构建模态间的双向映射关系, 实现跨模态的信息转换, 主要涵盖跨模态语义检索和跨模态内容生成.

3) 视听协作: 如表 1 所示, 单一模态的信息往往难以满足实际应用需求, 视听协作旨在探索视觉与音频信息的综合理解方法及其协同效应, 以实现更高效的环境感知. 根据应用场景的不同, 视听协作可被细分为: 视听实例感知、视听场景理解、视听推理与交互和非传统的视听学习.

本文从上述三个核心维度出发, 对主流视听多模态学习任务及其文献进行系统性梳理、解耦与分类. 然而, 在实际分析中, 部分任务间仍存在多属性交叉现象. 这种分类边界的模糊性, 主要源于音视频数据在三个层面的本质耦合特性:

- 在物理层面体现为时空同步性, 如视频与音频的帧级对齐;
- 在语义层面表现为语义关联性, 如“狗吠”场景中视觉对象与声源的语义绑定;
- 在认知层面则反映为注意力分配的动态协同性, 如鸡尾酒会效应中的视听注意力分配.

这种多维耦合机制也体现出视听多模态学习任务间的内在复杂性. 从认知神经科学视角看[3], 人类视听整合过程本身存在功能区重叠, 如颞上沟既参与低级声学分析又承担高级语义整合. 因此, 本文在构建分类体系时, 一方面力求清晰展现各任务维度的技术特性, 同时也充分考量任务间的交叉与重叠关系. 这一设计旨在为后续研究提供一个更贴近任务实际属性、更具包容性的综述性框架.

如图 3(a) 所示, 具体来说, 视听深度估计、视听场景重建以及视听导航兼具视听增强与视听协作属性. 一方面, 这些任务通过音视频数据理解视听场景的空间信息; 另一方面, 这些任务旨在从二维时序数据中推断三维空间信息, 实现空间线索的维度跃迁. 视听问答/对话、视听描述涉及“视 & 听 $\rightarrow$ 文本”的语义转换, 视听迁移学习涉及跨模态知识传递. 因此, 这些任务兼具视听协作与跨模态交互属性. 此外, 视觉增强的语音识别利用视觉线索补偿音频信息, 其结果往往以文本的形式呈现. 因此, 这一任务同时具备视听增强与跨模态交互属性.

1.3.2 创新点及组织结构

与现有的视听多模态学习相关综述 [4–5] 相比, 本文的创新性与不同之处主要体现在以下四个方面:

- 涵盖 30 余个细分任务, 兼顾传统热点与新兴方向, 广度与深度并重;
- 作为首篇系统性中文综述, 本文致力于为国内研究者提供可直接参考的“任务–方法–核心问题–

前沿”闭环;

● 本文尝试以认知为切入点, 提出“视听增强–跨模态交互–视听协作”三分类体系框架, 并对任务交叉属性进行显式标注, 推动分类方式从机械割裂转向有机整合, 为视听多模态学习领域构建一个更贴合实际、包容性更强的综述框架;

● 在此分类体系框架基础上, 本文进一步凝练出视听多模态学习中所涉及的“表征–对齐–转换–融合–共同学习”五大共性科学问题, 为视听多模态学习领域提供可复制、可扩展的研究路线.

第 1 节已系统阐述了视听多模态学习的背景与基础, 本文其余部分的安排如图 4 所示: 第 2 ~ 4 节分别围绕视听增强、跨模态交互、视听协作三大核心维度, 系统回顾相关视听多模态学习任务和相关文献; 第 5 节总结了这些视听多模态学习任务所涉及的核心共性问题和挑战, 并阐述了大模型背景下的视听多模态学习研究进展; 第 6 节对本文进行总结和展望.

2 视听增强

如表 1 所示, 单模态信息往往视角单一、上下文有限, 难以全面捕捉复杂场景的语义特征. 通过向初始单模态任务注入另一种模态信息以实现“增强效应”, 本文将此类视听多模态学习归纳为视听增强, 其核心在于利用视听信息间的交叉验证与互补增强, 提升单模态任务的鲁棒性和性能上限. 依据在初始任务中主导模态类型的差异性, 视听增强可归纳为:

● 音频增强的视觉任务: 在视觉主导任务 (如动作识别) 中引入音频线索 (如环境声音), 利用时序同步性, 细化视觉语义理解 (如图 5(a) 所示);

● 视觉增强的音频任务: 在音频主导任务 (如语音分离) 中引入视觉信息 (如唇部运动), 通过多模态对齐, 优化声学特征表达 (如图 5(b) 所示).

2.1 音频增强的视觉任务

CV 作为研究机器“看”的科学, 其核心在于赋予机器对图像/视频的处理与理解能力. 然而, 在真实世界的开放场景中, 单一视觉模态的感知存在诸多固有局限性, 主要体现在:

● 环境干扰: 夜间、隧道等低光照条件导致图像细节丢失与信噪比骤降; 雨雾、雪尘等散射效应造成图像模糊与对比度降低.

● 目标遮挡: 行人、车辆等移动物体相互遮挡, 导致目标特征不完整.

● 硬件限制: 相机分辨率、帧率或动态范围不足时, 高速运动目标 (如车辆) 易产生运动模糊, 或高亮/暗部区域丢失纹理细节.

这些局限性表明, 视觉模态的信息密度存在边界. 为此, 在 CV 任务中引入音频信息, 通过视听信息的互补性, 增强初始视觉任务的理解能力, 成为一种创新策略. 如图 5(a) 所示, 音频增强的视觉任务涵盖动作识别、显著性检测、人群计数、航空场景识别、深度伪造检测和视觉修复. 此类视听多模态学习任务通过充分利用音频信息的独特优势, 如复杂环境中的方向性感知、遮挡情况下的持续性声学线索等, 有效弥补视觉信息的缺失或模糊, 从而提升视觉任务的表达深度与理解精度.

2.1.1 音频增强的动作识别

动作识别作为 CV 领域的关键任务之一, 旨在精准解析图像或视频中的动作. 传统动作识别方法主要依赖三种信息来提取动作特征[10]: 视觉信息捕

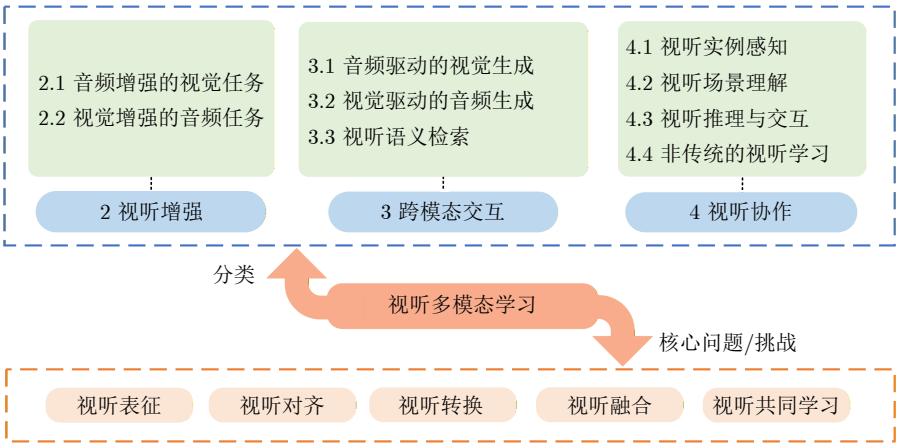


图 4 本文结构及其章节安排示意图

Fig. 4 Diagram of the paper structure and chapter organization

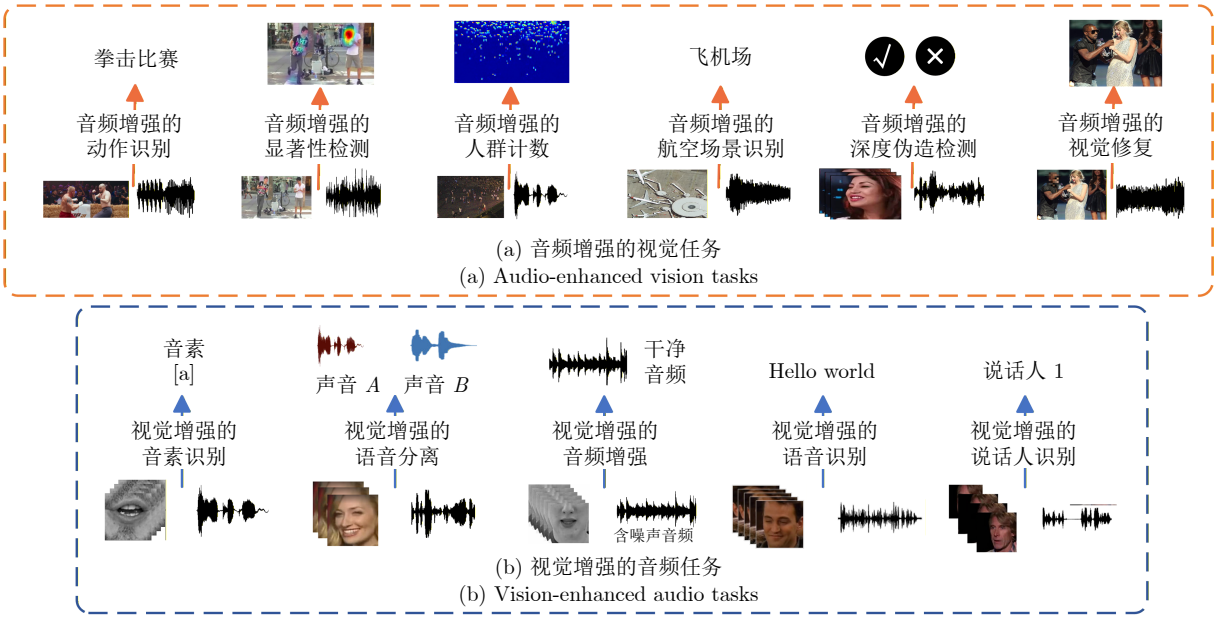


图 5 视听增强分类及其任务划分示意图

Fig. 5 Diagram of audio-visual enhancement categorization and its task division

捉到丰富的运动变化细节, 为动作识别提供直观的动态线索; 骨骼序列信息通过记录人体关节的运动轨迹和姿态变化, 以简洁有效的结构化形式表征行为动作; 深度信息反映物体在空间中的距离和位置关系, 为动作识别增添空间维度的感知.

在实际场景中, 特定动作往往与特定声音相伴, 如演唱时的歌声、舞蹈时的配乐等. 音频增强的动作识别任务应运而生, 其通过引入音频信息作为对视觉信息的有效补充, 使模型在复杂环境和遮挡情况下仍能保持较好的识别性能. 然而, 异质模态的引入也给动作识别任务带来以下两个核心挑战: 视觉和音频数据在特征空间、数据分布等方面存在差异, 需设计有效的融合机制, 以确保多模态信息能够互补而非相互干扰; 动作识别场景复杂多变, 不同模态和环境下的动作特征呈现多样性, 要求模型具备动态适应与泛化能力.

为应对这些挑战, 一些研究聚焦于跨模态特征交互与融合, 以提升模型的判别能力. Liu 等[11] 结合对比学习与注意力机制, 实现异构数据对齐和融合. Shaikh 等[12] 将音频中提取的表征转换至图像域, 与视觉表征深度融合, 形成统一的感知线索. Shahabinejad 等[13] 采用视频帧之间的运动线索作为指导, 实现视频片段级别的空间注意力机制, 有效提高模型训练和推理阶段的计算效率. 为进一步提升模型的通用性, Shaikh 等[14] 通过从音频信号的图像域中提取时间信息, 同时利用卷积神经网络 (convolutional neural network, CNN) 获取视频空间信息, 融合视听表征进行动作分类. Gao 等[15] 则运用音频信息作为代理, 借助师生蒸馏框架和基于注意力的长短期记忆 (long short-term memory, LSTM), 消除长期与短期的视觉冗余信息. Song 等[16] 通过输入的动作片段动态调整超平面位置, 有效缓解因不同模态特征多样性而导致的动作预测不一致性问题.

近期研究进一步聚焦于长时序建模、轻量化迁移和噪声鲁棒性等实际场景中的挑战. Chalk 等[17] 通过显式建模长视频中视听事件的时间范围和模态间关联, 实现对长视频中动作的高效识别. Wang 等[18] 通过在预训练模型中引入轻量级适配器, 实现向音频增强的动作识别任务的高效知识迁移. 针对“噪声标签”和“噪声对应”问题, Han 等[19] 通过噪声容忍对比学习与跨模态噪声估计机制, 提升模型在实际场景中的抗噪性.

2.1.2 音频增强的显著性检测

受人类注意力机制启发, 显著性检测旨在定位图像或视频序列中具有感知突出性的时空区域. 显著性检测[20] 主要包括视觉焦点预测和显著对象检测, 前者聚焦于建模人眼注视点分布规律, 后者则致力于精确分割视觉场景中最具辨识度的目标区域. 大量心理学研究表明[21], 听觉信息对视觉注意力具有显著影响. 例如, 在嘈杂环境中, 突发声响 (如高声说话者、火车的汽笛声等) 往往能迅速吸引人们的注意力.

近年来, 音频增强的显著性检测逐渐成为研究热点, 众多学者开始深入探讨听觉信息如何影响视

觉显著性检测. 音频增强的显著性检测结合视觉和听觉模态, 旨在更准确地模拟人类注意力机制. 其核心特点主要体现在以下两个方面: 需同时考虑空间 (物体位置、形状) 和时间 (运动、声音变化) 特征, 以全面捕捉显著性区域在时空维度上的变化规律; 部分实时应用场景对系统存在严格的因果性约束, 即模型只能利用当前及历史时刻的信息进行在线推理, 而不能依赖未来帧.

针对这些核心特点, 研究者从多模态融合与动态建模两个维度提出多种创新方法. 在特征融合策略方面, Tavakoli 等[22] 提出分别为音频和视觉数据训练独立的 3D ResNet, 并直接连接两者输出作为后期融合策略, 以解决传统视频显著性检测方法过度依赖单一视觉线索的问题. Min 等[23] 利用视听数据分别生成音频、空间和时间注意力图, 再融合这些注意力图构建最终的视听显著性图.

在时序建模与特征优化层面, Jiang 等[24] 引入跨网掩码和层次特征归一化策略, 结合 CNN 与 LSTM 的优势, 实现从单帧到多帧的显著性检测. 为了赋予模型因果特性, Jain 等[25] 在解码阶段运用线性插值和 3D 卷积技术, 实现从多层级特征中预测显著区域. Xiong 等[26] 考虑在特定时间片段内, 音频与视频信息可能出现不一致的情况, 采用视听特征对齐策略来调和模态间的异构性. Chang 等[27] 设计一种融合时序信息与多尺度视听特征的策略, 以提升联合表征的质量.

在跨模态动态建模与知识融合层面, Xie 等[28] 通过引入时空约束, 将跨模态对齐从静态特征匹配升级为动态关联建模, 利用自监督对比学习实现动态语义对齐. Chen 等[29] 构建了首个沉浸式视听眼动数据集, 并提出一种时空特征解耦方法, 实现高效的跨模态感知协同. Zhu 等[30] 引入弱监督学习范式, 利用跨模态 Transformer 生成伪注视图, 将类别标签转化为空间监督信息. 为克服纯数据驱动的局限性, Qiao 等[31] 提出一种知识引导方法, 通过超网络架构动态学习并融入运动先验知识. Xiong 等[32] 将显著性检测重构为条件生成任务, 利用扩散模型迭代去噪特性, 实现灵活的单/多模态自适应的显著性检测.

2.1.3 音频增强的人群计数

人群计数[33] 作为一项重要的 CV 任务, 其目标在于准确估计不同场景中的个体数量. 当前的人群计数方法主要可以分为两类: 基于密度图的方法和基于半监督学习的方法. 然而, 仅依赖视觉信息的人群计数方法常常受到物体或人物遮挡、弱光照等不利因素的限制. 在实际环境中, 声音与人数往往呈现正相关关系. 因此, 将声音作为第二模态的数据引入人群计数任务中, 可以有效缓解上述问题.

Hu 等[34] 首次提出将视觉特征和音频特征融合以进行人群计数, 并发布了基准视听数据集 DISCO. 在此基础上, Zou 等[35] 进一步考虑了各个方向的空间音频特征信息, 通过融合提取的空间音频特征来生成人群密度图. 为了更有效地处理人群场景中不同大小的物体, Hu 等[36] 利用多尺度视觉分支捕获视觉特征; 将音频信息转换为谱图信息, 并采用级联融合架构来估计人群密度图. 此外, Sajid等[37] 通过生成区域人群数量与局部重要性两种辅助信息, 构建出第三模态, 并利用跨模态注意力机制将其与视觉、音频特征融合.

音频增强的人群计数方法通过融合视觉和音频信息, 充分发挥了二者的互补优势. 视觉模态主要捕捉人体轮廓、头部特征等空间信息, 为任务提供精确的几何约束, 尤其在中低密度场景下可实现像素级精确定位. 音频模态则依据声学物理规律编码密度特征, 其毫秒级时序分辨率可有效补偿视觉运动模糊, 在动态场景中表现出独特优势. 此外, 特征表示的多样性进一步提升人群计数模型的适应能力.

2.1.4 音频增强的航空场景识别

航空场景识别作为一项重要的遥感图像分析任务, 其目标是识别航空拍摄场景图的类别. 航空场景通常具有广阔的视野、多样的地理特征以及动态变化的环境条件, 仅凭视觉信息易导致关键特征的遗漏或误判. 例如, 从高空俯瞰的机场、城市街区或乡村田野等场景, 视觉上差异细微, 单靠图像分析易混淆. 此外, 恶劣天气或光线条件可造成视觉特征退化, 会进一步削弱视觉方法的效果. 通常情况下, 声景信息与特定航空场景存在强关联性, 例如, 操场上可能听到孩子的嬉戏声. 因此, 将音频信息融入航空场景识别任务, 可以补充视觉信息的不足, 提供更丰富的上下文线索, 从而提升识别的准确性和鲁棒性.

Hu 等[38] 提出利用声音事件的知识来提高航拍场景识别的性能, 并构建基准数据集 ADVANCE. Heidler 等[39] 发布 SoundingEarth 基准数据集, 并设计一种自监督的方法, 使模型在无标注数据上学习到场景特异性表征. Sun 等[40] 基于层次化特征对齐机制, 实现时序信号到图像域的数据级转换, 并利用双流动态交互网络缓解视听模态的异构特征偏差. 这些方法在音频特征利用方面主要呈现三个特点: 通过注意力机制筛选与视觉强相关的声学事件, 抑制无关噪声; 基于信噪比感知设计融合机制, 在视觉退化时实现音频补偿; 借助预训练策略显著降

低对精细标注数据的依赖.

随着研究的深入, 研究视角逐渐从基础特征对齐扩展到细粒度语义理解. Han 等[41] 采用对象中心化建模范式, 提出一种槽注意力机制, 通过对物体进行语义解耦与门控融合, 缓解传统全局特征融合导致的语义干扰瓶颈. Khanal 等[42] 关注地理空间多尺度关联与不确定性量化, 将概率嵌入框架引入多分辨率卫星影像编码网络, 利用伪阳性样本抑制数据噪声, 实现声学场景分布的概率化表征. Corley 等[43] 通过建立影像尺寸调整与通道归一化基准体系, 为多源数据迁移学习提供可复现的预处理范式, 从数据工程层面解决跨域特征分布失配问题.

2.1.5 音频增强的深度伪造检测

深度伪造检测作为数字媒体取证领域的重要研究方向, 旨在识别通过 AI 技术生成的虚假视频内容, 特别是人脸替换、表情操纵等伪造形式. 现有方法可分为: 基于空间信息的方法[44] 聚焦于单帧图像的静态特征分析, 如人脸关键点偏移、光谱异常等; 基于时间信息的方法[45] 则通过捕捉帧间动态不一致性, 如头部运动轨迹异常、眨眼频率异常等. 音频模态的引入为该任务提供了关键补充信息, 可显著提升对伪造内容的鉴别能力.

音频增强的深度伪造检测任务的核心挑战主要体现在: 微观伪影的动态捕捉, 需识别像素级视觉异常和频谱级声学特征; 跨模态同步建模, 需精准分析唇动–语音时序关系; 复杂场景适应, 应对单模态输入、数据损坏等实际问题; 对抗性演化追踪, 需持续更新以应对迭代的伪造技术.

为应对深度伪造检测任务中的多模态融合挑战, 研究者提出了多种创新方法. Hashmi 等[46] 集成多种 Transformer 变体, 从多角度精准捕获音视频关键特征, 并通过投票策略对特征进行综合分析. Zhang 等[47] 利用多分支分类器保留模态特异性, 并通过跨模态注意力机制促进视听特征融合. Wang 等[48] 设计三阶段分层特征交互模块, 结合自适应模态选择策略实现特征语义的局部–全局语义传播. Wang 等[49] 利用可学习参数动态平衡视听模态的异质特征贡献度, 实现跨模态伪造线索的自适应聚合.

除了多模态特征融合, 时序维度的伪造痕迹捕捉也成为研究重点. Liu 等[50] 通过构建音频频谱与唇部运动的毫秒级对齐模型, 结合多尺度注意力机制揭示生成式伪造的时序缺陷. Astrid 等[51] 利用伪造样本库, 增强模型对局部时序偏移的敏感性. Koutlis 等[52] 采用特征金字塔结构, 实现帧级跨模态差异建模. Shahzad 等[53] 引入自监督学习和多尺度 CNN, 深入挖掘视频与音频间的时间相关性, 有效增强模型对视频时空伪影的捕获能力.

在表征学习与模态协同方面, Chugh 等[54] 提出音频与视频模态间的不和谐分数, 是识别伪造的关键线索; 构建基于模态不和谐分数的深度伪造检测框架, 独立学习真实与伪造数据的特征. Zou 等[55] 通过跨模态对齐损失与模态内对比学习, 构建多粒度约束体系缓解表征冲突. Bekheet 等[56] 通过揭示单模态方法的语义校验局限, 提出文本覆盖解析与动态行为识别的多模态视听联合范式. Nie 等[57] 采用冻结预训练 ViT 模型, 并结合潜在令牌蒸馏策略, 缓解跨模态数据间的鸿沟问题.

值得关注的是, 无监督与弱监督方法为应对数据稀缺提供了新思路. Zhao 等[58] 通过跨模态对比损失降低标注依赖; Liang 等[59] 和 Oorloff 等[60] 分别通过自监督语音表征迁移与两阶段跨模态学习机制, 挖掘真实视频的语义一致性特征, 增强模型对未知伪造类型的适应能力. Li 等[61] 则提出一种无监督的伪造检测方法, 其通过预训练模型提取语义文本序列, 利用编辑距离量化视听断层.

此外, 多项研究致力于提升模型的实用性与泛化能力. Muppalla 等[62] 将细粒度深度伪造识别与二值分类相结合, 精准检测视频中的视听伪影, 显著强化域内及跨域检测性能. Yang 等[63] 设计跨模态分类器, 精准检测模态内不一致性, 并发布基准数据集 DefakeAVMiT. 面对模态缺失问题, Yu 等[64] 通过单独预测各模态真实性, 结合时间聚合模块与假成分检测器, 显著提升鉴别性能.

2.1.6 音频增强的视觉修复

视觉修复[65] 作为一项重要的 CV 任务, 通过对低质视觉内容进行增强修复与重建, 致力于提升其质量、清晰度和流畅度, 或尽可能还原因各种因素而损失的信息. 图像修复侧重于恢复静态图像的细节和提升其分辨率, 视频修复则聚焦于还原连续帧序列中的动态内容. 音频信号同样蕴含着丰富的信息, 例如, 声音信号能传递说话人身份特征 (如性别、年龄) 和场景环境线索, 音乐则包含不同乐器的独特音质和特征. 近年来, 将音频信息与视觉信息相结合, 已成为视觉修复研究的一个新兴方向.

然而, 在视觉修复任务中引入异构的音频信息, 将引发以下挑战: 音频信号的高采样率 (通常 16 kHz) 与视频帧率 (通常 30/60 fps) 存在量级差异, 导致跨模态对齐时产生高达 ±3 帧的时序偏差; 复杂场景中音频仅与局部视觉元素存在间接关联, 如多声源场景中部分声音对应特定物体, 显著增加了从音频提取有效视觉重建信息的难度; 相同声学事件 (如掌声) 可能对应多种视觉场景 (会议室/音乐厅), 这

种一对多映射关系阻碍了音频到视觉场景的确定性推理.

针对这些关键挑战, 研究者提出了多种创新方法. Sanguineti 等[66] 通过构建音频空间中像素的概率分布模型来生成低分辨率图像, 并借助生成对抗网络 (generative adversarial networks, GAN) 进一步优化图像质量. 此外, 研究者不仅关注从单帧图像中恢复丢失的细节, 还积极探索利用视频相邻帧之间的时空关联性来提升视频分辨率. 例如, Lu 等[67] 将时空差值引入视频超分中, 使得模型能够从时空坐标、图像帧以及视频中不同事件的特征中学习更深层次的表达. Chen 等[68] 在多个粒度层级上综合考虑目标帧的语义特征, 并采用自上而下的多尺度融合机制, 将高层抽象的语义信息逐步反馈至底层全局视觉特征中, 以此为视频恢复模型提供有效的音频隐式语义引导. 为提高实时性, Xiao 等[69] 设计一种轻量级网络结构, 实现在不依赖未来帧作为输入的情况下, 有效降低在线处理延迟.

2.2 视觉增强的音频任务

AP 作为研究机器“听”的科学, 其核心目标在于实现对音频信号的处理、分析和理解. 然而, 在实际应用中, 仅依赖单一听觉模态进行环境感知存在诸多固有局限:

- 采集阶段易受干扰: 在嘈杂环境中 (如鸡尾酒会场景), 背景噪声极易混入目标音频信号, 导致其被掩盖;
- 传输过程信号失真: 可能受到电磁干扰, 引发杂音或失真;
- 设备因素导致频响失衡: 采集设备的频率响应不均衡或滤波器设置不当, 可能造成频率失真, 进而丢失原始特征.

相比之下, 视觉信息如嘴唇运动、肢体动作或视觉场景等, 通常不易受此类因素干扰. 因此, 在 AP 任务中引入视觉信息, 有助于任务从多维度捕捉和解析音频信号, 更有效地应对复杂多变的现实场景, 显著提升音频处理能力. 如图 5(b) 所示, 视觉增强的典型音频任务包括音素识别、语音分离、音频增强、语音识别和说话人识别. 此类视听多模态学习任务充分结合视觉信息在不同场景下的独特优势, 增强音频信息的表达能力和理解深度.

2.2.1 视觉增强的音素识别

音素是音频的最小单元, 音素识别旨在将语音信号转换为对应的音素序列[70], 揭示语音的内在结构和语言学规律, 为相关下游任务如语音合成、身份识别以及情感识别等任务提供更准确的信息. 传统音素识别方法包括隐马尔科夫模型和循环神经网络 (recurrent neural network, RNN) 等. 尽管这些已经取得较大进展, 但在嘈杂环境下, 其识别性能仍面临严峻挑战. 为应对这一挑战, 研究者开始探索视觉模态的互补作用. 嘴唇运动等视觉信息因不受声学噪声干扰, 可作为增强噪声环境下音素识别鲁棒性的关键线索.

引入视觉模态虽然能提升噪声环境下的音素识别鲁棒性, 但也带来了以下关键挑战: 视觉与音频模态的时序需严格同步, 但实际数据常存在同步偏差, 如音频流与视频流的硬件延迟差异. 视觉输入易受遮挡 (如口罩)、低分辨率或头部姿态变化干扰, 导致关键唇部特征丢失, 噪声视觉特征可能比纯音频更不可靠; 模型在训练过程中可能出现单模态偏好, 如低噪环境下视觉模态主导决策. 这种偏差源于模态间信息密度差异及训练目标的不均衡优化.

为了解决视频和音频在受到污染时如何提高模型的鲁棒性, Fang 等[71] 提出一种不确定性驱动的混合融合方案, 通过在决策中纳入模态不确定性, 使模型能够自适应地决定是否融合多种模态及其依赖程度, 从而减轻了不可靠视觉输入的影响. Richter 等[72] 分别采用 CNN 和 ResNet 提取嘴唇运动的视觉特征和声学特征, 并通过门控循环单元 (gated recurrent unit, GRU) 建模跨模态时序依赖关系, 有效提升了噪声场景下的识别精度. Biswas 等[73] 利用颌骨和唇部区域提取的形状等外观信息来增强噪声环境下音素识别系统的鲁棒性.

Hu 等[74] 提出视素–音素分布平衡模型, 通过对抗性互信息估计器, 构建噪声不变的跨模态关联, 有效缓解了同音异形带来的语义歧义问题. 该方法与 Kim 等[75] 的离散化视觉语音表征形成互补, 后者采用自监督学习压缩视觉数据维度, 实现了高效的知识蒸馏. 针对发音变异等特殊场景, Pai 等[76] 通过深度可分离卷积实现音素级特征对齐. 这些方法通过对抗互信息估计、自监督离散化或深度可分离对齐等策略, 抑制发音变异与同音异形带来的语义歧义, 实现高效、噪声不变的视听语音表征.

2.2.2 视觉增强的语音分离

语音分离[77] 的核心目标是从复杂声学混合信号中解析出特定目标声源. 人类语音产生过程具有固有的视听关联特性: 发声行为始终伴随唇部运动与面部动态变化. 典型案例如鸡尾酒会场景, 听者可通过视觉线索快速定位目标说话人, 并借助唇动轨迹增强语音理解能力. 上述机制表明, 视觉信息可构成强引导性跨模态先验知识. 通过建模语音信号与唇部运动的时空耦合关系, 视觉模态能有效提升复杂声学环境下的分离精度, 尤其在信噪比显著

劣化的场景中强化目标声源提取能力.

在语音分离任务中引入视觉信息, 虽然能够显著提升分离精度, 但也带来了以下核心挑战: 视觉表征需对唇动等发音内容高度判别, 同时抵御身份、光照、遮挡等干扰, 而音频表征需精准区分目标语音与噪声, 可能结合空间信息; 当目标说话人处于视域外时, 视觉线索完全缺失或不可靠; 精准对齐的视听–语音数据集构建成本高, 需挖掘未标记视频的内在监督信号.

Li 等[78] 构建语音分离、去噪与识别三阶段框架, 通过联合优化视听模态提升分离性能. Tan 等[79] 提出一种基于自然语言查询的自监督学习方法, 通过利用视觉–语言模型实现发声对象描述、视觉特征及音频特征的跨模态语义对齐. Gao 等[80] 探索无监督的语音分离, 借助人脸信息从未标记视频中同步完成语音分离与跨模态嵌入学习. 为从音频中提取视频内特定说话者语音, Yoshinaga 等[81] 通过融合唇动时序嵌入与预录声纹嵌入, 分别建模视域内与视域外线索. Pan 等[82] 利用面部信息建立声学差异感知模块, 使模型识别目标语音与噪声差异以有效提取受损特征. Chen 等[83] 通过跨模态注意力机制, 去除估计波形中潜在声源干扰, 实现显式声源解耦. Chatterjee 等[84] 通过结合 RNN 与多头注意力来生成视觉正交嵌入, 并以此引导频谱分离过程. 上述的研究主要集中在单声道视听分离, 而在沉浸式虚拟环境中, 听众需区分不同方位声源. 为此, Ye 等[85] 利用双耳音频相位差蕴含的空间线索, 并将发声体位置作为额外引导因子, 建立频谱–空间特征映射以实现方位感知分离.

在真实复杂环境中, 模型还需具备对多说话人、视觉缺失及持续学习等更具挑战性场景的适应能力. Pan 等[86] 提出一种说话人级的音视频交互机制, 在单次分离过程中实现多个说话人协同分离. Li 等[87] 通过模仿丘脑到皮层的层级结构, 并在音频和视觉子网络中使用递归连接, 增强信息在不同层级之间的流动与融合. Liu 等[88] 提出一种无需视觉输入的两阶段音视频语音分离模型, 在训练阶段利用视觉特征优化分离网络, 而测试阶段仅依赖音频输入完成分离. Pian 等[89] 通过跨模态相似性蒸馏约束有效缓解灾难性遗忘, 实现新旧声源类别间的连续分离与知识保持. 针对嘈杂环境下说话人无关的语音分离问题, Kalkhorani 等[90] 在训练中引入信噪比调度机制以强化视觉流的利用, 从而捕获频率特定的时序依赖与长程多模态关联. Fan 等[91] 利用预训练的音视频模型作为教师, 通过知识蒸馏提升仅依赖音频输入的学生模型性能, 并通过设计多阶段自适应特征融合机制, 从全局与局部视角充分挖掘音视频深层关联.

2.2.3 视觉增强的音频增强

音频增强旨在采用多种技术手段来消除音频中的杂质, 进而提高音频信号的质量、清晰度以及可理解度等. 传统的音频增强方法如: 谱减法[92] 和深度聚类[93] 等, 主要聚焦于音频特征的提取与分析, 但忽略了面部等视觉信息对音频理解的潜在作用. 心理学研究表明[21], 人类在嘈杂环境中能够通过观察说话人的唇部运动显著提高语音理解能力. 因此, 融合视觉信息的音频增强方法有望更有效地提升语音清晰度. 然而, 引入视觉信息也带来了一些关键问题: 语音与唇部运动涉及音节、音素等多时间尺度的分层动态 (快慢模式交织), 模型需协同捕捉跨尺度时序依赖; 纯净/轻度噪声下侧重音频主导的互补融合, 而极低信噪比或复杂场景 (如鸡尾酒会) 下强化视觉线索贡献.

一些研究通过精细的特征提取和创新的融合策略, 提升语音增强性能. Zhu 等[94] 利用嘴角像素作为视觉线索, 通过多阶段门控求和模块融合视听特征, 并采用密集连接缓解性能下降. Balasubramanian 等[95] 采用多通道 CNN 提取动态视听特征, 集成上下文信息强化噪声环境语音理解. Li 等[96] 首创细粒度 3D 唇部特征表示, 结合扩张卷积与注意力机制提取多尺度特征并抑制信息损失. Hussain 等[97] 将情感特征作为上下文, 显式地加入模型中, 同时利用时序、谱域与跨模态同步关系作为融合策略以改善特征表示. Zheng 等[98] 引入教师模型向仅基于音频与唇部视频的学生模型传递舌部表征知识, 从而更全面地刻画发音过程.

在追求模型性能的同时, 实时性与低延迟处理也成为实际应用中的关键需求. Gogate 等[99] 通过 GAN 进行视觉特征增强, 并与实时语音增强算法相结合, 提出一种低延迟音视频语音增强方法. Chen 等[100] 提出一种创新的逐帧实时处理框架, 通过设计基于 Emformer 的因果型音视频融合架构, 实现端到端的低延迟语音增强. Jung 等[101] 提出一种基于条件流匹配的音视频语音增强方法, 在保持语音质量的同时显著提升推理效率与模型轻量化.

此外, 针对复杂环境下的鲁棒性问题和个性化需求, 研究者还开发了多种优化策略与自适应框架. Passos 等[102] 融合图神经网络与典型相关分析, 增强通道内视图关联及音视频跨模态对齐以优化特征. Morrone 等[103] 探究多时间尺度视觉信息对音频修复的影响, 基于双向 LSTM 重建丢失语音. Chen 等[104] 提出一种多层次失真度量, 通过序列化加权的方式综合多种优化目标, 使模型能够同时提升语音

质量、可懂度和语音识别性能. Chen 等[105] 通过将音频与视觉数据、语音增强算法及模糊推理机制深度融合, 并引入用户偏好与环境上下文自适应调节机制, 实现语音增强策略的个性化动态优化.

2.2.4 视觉增强的语音识别

语音识别[106] 旨在通过分析和处理声音信号, 将其解码为可理解的自然语言, 从而实现人与机器之间语言交互的一种重要技术. 然而, 该领域仍面临诸多挑战, 例如声学通道易受环境噪声干扰, 以及说话者语音风格多样性对识别准确率的负面影响. 相较于声学模态, 视觉信息, 特别是与语音内容高度同步且语义一致的唇部运动, 受声学环境影响较小. 因此, 融合视觉信息可有效提升语音解析精度.

动态与自适应融合策略是提升模型在复杂场景下语音识别性能的关键途径. Hong 等[107] 通过量化输入质量、动态评估各模态可靠性, 并在预测阶段自适应地加权高可靠性模态, 有效缓解模态缺失或受损场景下的性能下降问题. Wang 等[108] 构建音视频双模态网络, 利用卷积块、注意力模块和时序注意力机制, 提取关键特征并过滤冗余噪声. Wang 等[109] 通过在音视频编码器的多个中间层进行特征融合, 使各模态在表示学习阶段就能相互获取低层到高层的互补上下文信息. Wang 等[110] 通过生成视觉嵌入, 并利用跨模态 Conformer 将其与音频融合, 从而利用视觉提示引导模型提前关注语音关键信息.

此外, 知识迁移与数据利用方法聚焦于通过知识蒸馏、伪标签等技术解决数据稀缺和模态依赖问题. Ma 等[111] 使用预训练的自动语音识别模型来自动生成视听数据的标签, 从而大幅扩充训练样本规模, 使得模型学习更泛化的视听表征. Yeo 等[112] 设计音频记忆单元存储量化音频信息, 通过过滤非语言相关成分, 实现大规模音频预训练模型知识向视听任务的迁移. Lian 等[113] 通过指导学生模型逼近教师模型提取的多层次上下文特征表示, 从而捕获不同粒度的视听特征信息. Dai 等[114] 通过均衡教师模型引导学生模型, 减轻对音频的过度依赖. Rouditchenko 等[115] 利用门控跨注意力机制将视频信息注入 Whisper 的解码器, 增强模型对训练数据稀缺和噪声环境的鲁棒性.

然而, 实际应用场景充满挑战, 针对遮挡、噪声和异步等问题设计的鲁棒性增强技术至关重要. Wang 等[116] 通过一种音频驱动的唇部修复策略, 在合成完整唇形的基础上匹配未遮挡区域进行优化, 有效提升唇部遮挡场景下的语音识别鲁棒性. Li 等[117] 引入不同步感知损失与统一编码框架, 实现了多模态信息的自适应融合, 有效提升模型在音频缺失、视觉损坏及多说话人干扰等复杂场景下的鲁棒性. Fu 等[118] 通过定制化提示机制与对比分解策略, 实现预训练知识向低质量数据的迁移与纯净特征学习, 增强了模型对多模态失真的鲁棒性. Burchi 等[119] 通过自动生成多语言转录数据以扩充训练集, 提升在噪声环境及非英语语料上的识别性能.

2.2.5 视觉增强的说话人识别

说话人识别[120] 作为一项经典的 AP 任务, 其核心在于利用人类的语音特征来精准鉴别个体身份. 然而, 在嘈杂环境中, 仅靠音频信息难以准确识别说话人. 此时, 引入视觉模态信息可构建多模态互补框架, 通过融合面部口型运动轨迹、表情动态等视觉线索与语音韵律特征, 形成时空域联合表征, 从而提升复杂场景下的识别鲁棒性. 然而, 视听双模态的异构性也为该任务带来了新的挑战: 视觉特征多呈现高维、局部细节丰富的多尺度结构, 而音频特征则以时序动态信息为主, 其统计分布和表示形式存在显著差异; 视觉特征易受遮挡与光照变化等干扰, 音频特征则易受环境噪声等干扰, 如何有效抑制视觉和音频噪声的影响, 同时避免过拟合, 是该任务的关键.

针对上述挑战, 研究者提出多种解决方案. Chelali 等[121] 通过对比不同融合策略, 证实视觉信息对说话人识别的有效性. Tang 等[122] 借助音频信号定位视频的关键帧, 以选取具有显著表情变化的帧进行人脸识别. Gong 等[123] 在随机子空间上实施子空间分析, 以缓解面部遮挡或语音失真等问题. Tao 等[124] 提出一种两阶段深度去噪框架, 消除表征学习中噪声标签的影响. 为了同时处理多个语音信号, Gebru 等[125] 利用有监督的视听对准技术, 将音频特征映射到图像上, 再利用半监督聚类方法将多个音频特征分配给不同的目标说话人.

Lin 等[126] 提出开放场景下的说话人识别任务, 并发布 VoxBlink2 基准数据集, 实现对输入语音属于已知说话人或未知个体的判断. Clarke 等[127] 专注于头戴式视角下的说话人检测, 通过参考语音与候选音频的帧级比对, 在低质量视频场景中仍能有效辨识说话人身份. Tao 等[128] 提出一种选择性听觉注意力机制, 用于抑制干扰说话人和非语音信号.

3 跨模态交互

听到海浪声, 大脑会浮现出海涛翻涌的画面; 看到树叶随风摇曳, 仿佛也能听见沙沙声响. 这种单一感官刺激触发跨模态联觉的现象, 在认知心理学中被定义为“交叉模式知觉”[8]. 认知神经科学研究进一步揭示, 人类大脑具有高度可塑的跨模态联

想能力[6]. 在 AI 领域, 模仿甚至超越人类的跨模态联想与创造力, 始终是核心研究目标之一.

视听跨模态交互的核心目标在于模拟人脑的跨模态可塑能力. 其核心是通过建立视听数据间的语义映射, 构建双向的信息通路, 实现双向模态转换与信息补偿. 根据数据转换形式的差异性, 本文将其细分为:

- 视听内容生成: 旨在生成跨模态、创造性的视听内容, 具体包括音频驱动的视觉生成 (如图 6(a) 所示) 和视觉驱动的音频生成 (如图 6(b) 所示);
- 视听语义检索: 通过分析度量视听数据间的相关性, 实现跨模态的语义检索 (如图 6(c) 所示).

视听跨模态交互的技术突破, 不仅引领了多模态认知计算的新方向, 也极大地丰富了多媒体内容的创作, 推动个性化和定制化体验的发展.

3.1　音频驱动的视觉生成

音频驱动的视觉生成致力于通过多层次解析与结构化表征声学信号, 实现从音频模态到视觉内容的语义级映射与转换, 如生成与音乐旋律相匹配的舞蹈动作. 此类技术不仅体现了 AI 在多媒体内容理解与生成领域的前沿探索, 更展现出 AI 在推动娱乐、教育和艺术等领域变革的潜力. 如图 6(a) 所示, 音频驱动的视觉生成具体包括说话人脸生成、手势生成、动作生成和音频可视化.

3.1.1　音频驱动的说话人脸生成

给定语音输入及目标个体的视觉身份特征 (静态图像或参考视频片段), 说话人脸生成[129] 的核心目标是面部动画, 特别是实现高保真音唇同步. 该任务基于人类视听感知的一致性特性, 即 McGurk 效应, 建立从声学特征到面部肌肉运动的跨模态映射关系. 在视听跨模态生成中, 音唇同步精度与身份保持能力是衡量生成质量的核心指标.

针对说话人面部生成任务中唇形同步不稳定问题, 研究者提出了多种基于约束与精细化建模的解决方案. Agarwal 等[130] 通过模态间对抗训练增强唇形–语音的时序对齐能力, 缓解了合成视频中音画异步问题. Wang 等[131] 进一步在身份一致性约束下, 生成符合个性化风格的唇动序列, 提升同步自然度. Jang 等[132] 通过构建正交约束的双空间, 兼顾身份保持与唇部同步精度. 为实现更细粒度的控制, Park 等[133] 引入音素感知唇部动力学建模, 将音频分解为音素单元以预测对应唇动. Wang 等[134] 聚焦唇部空间结构与语音相关通道, 优化嘴型拓扑准确性. Zhang 等[135] 使用时频变换提取语音语义相关特征, 增强唇动物理合理性.

随着研究的深入, 生成任务的目标从唇形同步扩展至整体面部运动的自然性与多样性, 涵盖表情、眨眼、头部姿态等多方面运动建模. Liu 等[136] 采用两阶段模型, 生成具有目标身份特征且表情生动的

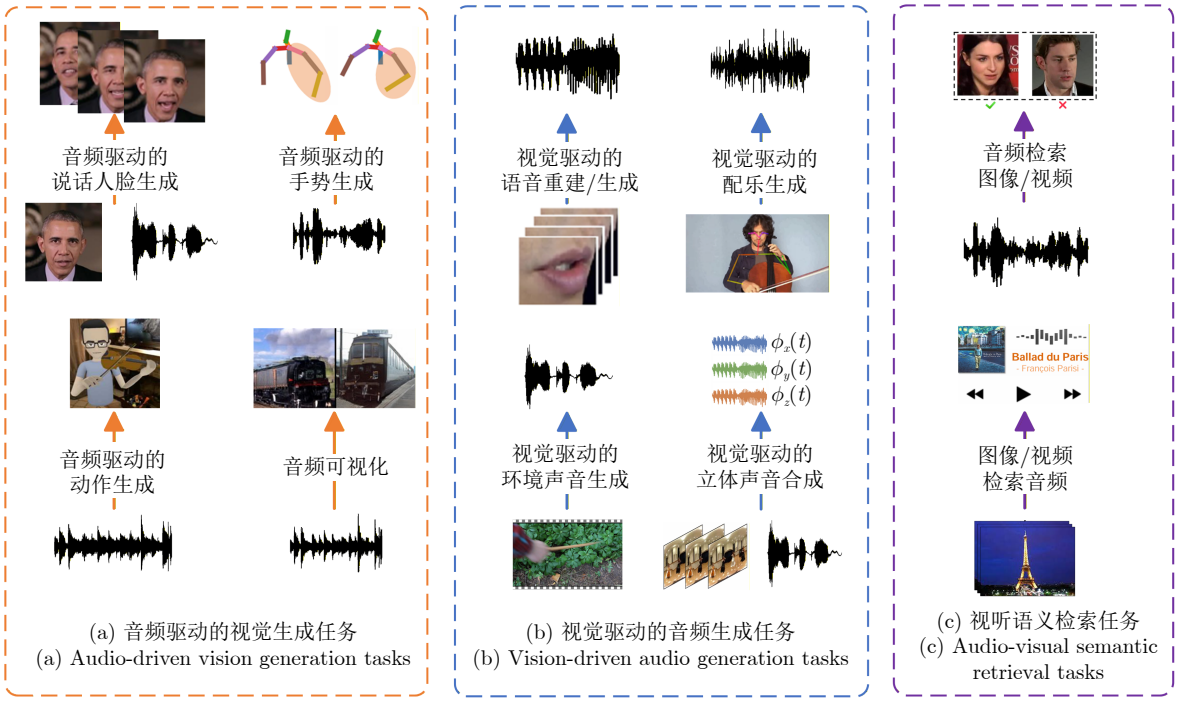


图 6　跨模态交互分类及其任务划分示意图

Fig. 6　Diagram of cross-modal interaction categorization and its task division

说话人脸. Yaman 等[137] 引入唇同步损失, 并设计三种新评估指标, 以提升同步稳定性与测评鲁棒性. Jang 等[138] 利用最优传输条件流匹配增强头部姿态与表情的多样性与连贯性, 并结合语音建模机制保障身份一致性. Xu 等[139] 将唇动、表情、眨眼与头部运动统一为潜变量, 设计基于 Transformer 的扩散模型架构实现整体自然生成. Zhang 等[140] 引入自发眨眼生成模块来模拟真实眨眼行为, 通过音频到关键点再到人脸的两阶段生成过程, 增强面部表情的自然度.

3.1.2 音频驱动的手势生成

手势作为人类非语言交流的核心载体, 手势生成[141] 旨在合成与语音内容在语义、节奏及情感维度协同一致的身体动作. 其方法可分为两大类: 基于规则的方法[142] 利用启发式方法, 根据音频数据选择相应的手势动作, 优势在于动作可控性强, 但泛化能力受限; 数据驱动的方法[143] 则侧重于从大量数据中学习和理解音频与手势动作之间的隐式关联, 可生成更自然的动作. 音频驱动的手势生成的结果需同时满足以下约束: 语义一致性, 语音中的语义单元必须精确诱导对应的手势模式, 确保语音–动作的强关联性; 节奏同步性, 手势运动的相位转折点需与语音重音音节的时间偏差 (通常 $\leq$ 200 ms); 运动自然性, 手势动作需满足生物力学可行性, 包括关节角速度限制、运动平滑性等.

在手势生成的同步性与多样性优化方面, 研究者提出了多种基于运动先验与特征解耦的建模方法. Ferstl 等[144] 通过预训练运动动力学先验, 实现未来帧序列的自回归预测, 显著提升长时动作连贯性. Liang 等[145] 将音频中的节奏与语义信息解耦, 分别驱动节拍性手势与语义性手势生成, 降低无关线索干扰. Hasegawa 等[146] 通过隐状态门控机制增强音轨同步性, 增强音频与手势之间的同步性, 从而有效降低时序误差. Yoon 等[147] 结合对抗多模态融合策略, 在联合优化内容匹配与节奏对齐的同时, 借助风格编码器实现说话人身份的自适应生成.

随着任务复杂度的提升, 研究重点逐渐扩展至情感融合、质量评估与语境理解等语义层面. Qi 等[148] 通过情感–节奏挖掘模块, 联合建模情感与语音节奏特征, 实现情感丰富且节奏对齐的手势生成. Liu 等[149] 结合时间序列对比学习与知识蒸馏技术, 将语音上下文信息融入手势表示, 增强生成动作的语境契合度. 在评估体系方面, Gao 等[150] 构建了 Ges-QA 数据集, 并提出一种基于三分支 Transformer 架构的自动化客观评分模型, 提供包含手势质量、一致性及情感匹配的评价.

为进一步提升生成动作的真实感与完整性, 部分研究开始探索全身动作建模与高保真视频合成. Liu 等[151] 从部分初始化手势和音频中解码出完整的动作序列, 实现音频同步、部位一致的全身动作生成. Chen 等[152] 则引入潜在偏差学习机制, 对前景与背景运动进行像素级建模, 并结合弱监督训练策略, 生成手势与口型同步的高保真共语视频.

3.1.3 音频驱动的动作生成

除了对手势动作的模拟与生成, 研究者还致力于探索更广泛的动态人体运动序列建模, 涵盖舞蹈动作生成、交流动作合成以及乐器演奏动作生成等方向. 动作生成技术的核心在于对音频信息的深度解析与特征提取, 进而实现对人体运动的精确建模与重建. 这一研究方向在虚拟现实、人机交互和数字艺术表演等领域具有重要的应用价值, 为智能动画生成、虚拟角色驱动和沉浸式交互体验提供了关键技术支撑.

在舞蹈动作生成领域, 早期研究主要将其视为相似性检索问题[153]. 然而, 这类基于检索的方法在生成多样性方面存在明显局限. 为突破这一限制, 后续研究转向探索音乐特征到动作特征的端到端映射机制. 例如, Huang 等[154] 通过建立舞蹈动作特征与音乐特征的跨模态关联, 实现了帧级动作生成. Wang 等[155] 通过离散化建模人体前景、背景和姿势等要素, 显著提升了模型对人体外观与行为模式的理解能力.

进一步地, 研究重点逐渐转向动作控制、节奏对齐与风格化生成等精细化方向. Tseng 等[156] 支持关节级条件控制和动作插值操作, 生成结果兼具节奏同步性、运动真实性与物理合理性. Yang 等[157] 则利用归一化流技术学习音乐与关键姿势联合条件下的动作分布, 有效增强了生成多样性. 在节奏同步方面, Li 等[158] 提出节拍对齐奖励函数, 实现动作与音乐的亚帧级对齐, 并通过优化人体旋转域表示提升动作自然度. 针对风格转换需求, Yin 等[159] 扩展 CycleGAN 架构以捕捉时序依赖, 实现了跨舞蹈风格的动作转换.

在交流动作合成中, 研究者致力于生成与语音内容协同的面部表情、身体姿态和手势序列. Habibie 等[160] 采用 CNN 建模面部–手部动作关联, 基于 GAN 实现对面部、身体和手势运动序列的联合生成. Yi 等[161] 则提出分区生成策略, 为不同身体部位设计独立生成器并建模其与语音的特异性关联. 在乐器演奏动作生成中, 研究者探索了从音乐到演奏者身体动作的生成. Zhu 等[162] 进一步提出可微分空间转换器, 将生成的关键点转化为热图表示, 结合

音乐时序特征合成高保真演奏视频.

近年来, 研究者开始探索更具结构感知与时空适应性的生成范式. Xu 等[163] 开创性地提出空间音频驱动人体动作生成任务, 基于扩散模型融合空间与时间特征, 实现空间音频与人体动作的高效对齐与建模. Zhang 等[164] 引入注意力引导的掩码建模机制, 在时空维度自适应关注关键动作, 显著提升控制精细度与生成质量. Zhang 等[165] 通过人体部位划分实现细粒度动作控制, 并结合变帧率与随机掩码策略进行预训练优化.

3.1.4 音频可视化

人类与生俱来的跨模态感知能力, 使其能将声学信号映射为具象化的视觉场景, 该认知现象在神经科学中被称为“联觉机制”[8]. 听觉信息的视觉表征技术通过计算模型来模拟这一生物机制, 旨在将时序声学特征转化为结构化视觉元素. 相较于传统的声纹图等低阶特征可视化, 数据驱动的音频可视化技术能够建立更高维度的语义级跨模态映射, 如根据鸟类声学指纹生成具有物种特异性的视觉形象, 不仅包含形态特征, 还能反映栖息环境等丰富语义的视觉表征, 极大提升了音频可视化的表达能力. 音频可视化研究的核心在于建立准确反映高层语义关联的跨模态映射, 同时精细捕捉声学特征的微观差异, 从而生成既符合人类认知预期又具备丰富多样性的视觉表征.

该领域通过深度学习模型模拟人类联觉机制, 致力于实现从抽象声学信号到具象视觉元素的可解释、细粒度转换, 为生物多样性监测、无障碍辅助等应用提供技术支持. 例如, Shim 等[166] 基于 c-GAN 实现从鸟类声学特征到对应物种图像的端到端生成. Song 等[167] 引入动态注意力约束, 使非配对音视频数据中关键声学特征得以在生成视频中保持视觉可辨识性, 为听障人士构建新型声学辅助系统. 针对模态对齐难题, 文献 [168] 通过预训练生成模型的潜在空间优化, 建立视听特征的几何一致性约束. Lee 等[169] 则将声学嵌入与 CLIP 多模态空间对齐, 实现声音引导的图像语义编辑. Zhuang 等[170] 通过将声景编码为语义向量并利用级联扩散模型, 实现从音频到街景图像的跨模态生成.

3.2 视觉驱动的音频生成

视觉驱动的音频生成技术致力于通过解析视觉信息中的空间结构、语义内容和上下文特征, 构建从视觉到听觉的精准映射关系, 从而合成与视觉内容高度匹配的音频信号. 此类技术已从对无声电影的语音重建等具体应用, 迈向为构建沉浸式视听一体化体验奠定技术基石的新阶段. 如图 6(b) 所示, 根据生成内容和应用场景的差异, 视觉驱动的音频生成任务主要涵盖语音重建/生成、配乐生成、环境声音生成以及立体声音合成.

3.2.1 视觉驱动的语音重建/生成

给定一段无声视频, 语音重建/生成的目标是合成与之匹配且同步的语音信号. 这一技术的应用前景广阔, 例如, 在电影行业, 通过对演员口型的精准分析, 可以高效生成电影对白, 甚至能够将早期的无声电影转化为有声电影, 为经典作品注入新的生命力. 然而, 基于视觉信息重建语音信号的技术在实现过程中面临多重相互关联的核心挑战: 需建立视频帧与语音波形的精确时间对应关系; 需区分并控制语音内容、说话人特征和发音风格等不同属性. 这些挑战直接影响语音生成的准确性和自然度.

在早期研究中, 研究重点集中于从视频中直接提取发音特征并重建语音波形. Ephrat 等[171] 利用 CNN 提取音素特征, 结合相邻帧的时间信息合成语音波形. Kumar 等[172] 通过整合不同摄像机角度的视觉信息, 提升声学特征重建的全面性, 尤其在发音部位遮挡场景下效果显著. Dong 等[173] 将低维运动特征映射为语音参数, 简化静默视频的语音重建流程. Kefalas 等[174] 则引入对抗性损失和三重对比损失, 强化输入视频与生成音频的对应关系.

随后, 研究重点逐渐转向语音内容和说话人身份分离与精细化控制. Hong 等[175] 将无声视频分解为内容 (音素) 与说话人身份独立成分, 通过双分支网络分别建模, 实现身份可控的语音重建. Kameoka 等[176] 利用人脸编码器生成潜在风格特征, 作为语音生成器的条件输入, 引导模型输出与特定人脸匹配的音色. Weng 等[177] 则利用帧级面部特征生成多样化语音, 并借助多尺度判别器与自注意力模块增强对内容特征的聚焦能力. Wang 等[178] 通过向量量化对比预测编码推导离散音素单元, 并将其传递给唇读网络, 实现语音内容与说话人身份的有效解耦与控制. Lu 等[179] 采用三阶段训练策略从人脸信息构建说话人相关分布, 以学习更精确的说话风格.

此外, 一些研究利用多源信息提升复杂场景下的语音重建质量. Mira 等[180] 采用预训练的视频–语音模型生成缺失语音信号, 再结合原始无声视频共同训练语音合成模型, 有效提升合成语音的自然度与准确性. Chen 等[181] 引入唇部运动信息辅助异常发音重建, 通过多模态编码器、方差调节器和说话人编码器的协同工作, 生成高质量语音. Pham 等[182] 同时利用视觉特征和唇语文本预测语言信息, 并通过声学路径预测韵律, 有效重建说话人的语音内容、

口音和韵律特征. Rong 等[183] 通过自适应面部提示查询器提取身份相关特征, 并利用对比学习优化音频匹配, 生成与图像一致的音频内容.

3.2.2 视觉驱动的配乐生成

随着短视频平台的兴起, 背景音乐 (配乐) 在提升用户体验方面发挥着至关重要的作用. 在此背景下, 人工智能辅助的配乐生成技术应运而生. 该技术通过对视频中的视觉元素进行分析, 创作出与之匹配的背景音乐, 从而增强视频的整体效果和用户沉浸感. 然而, 视觉驱动的配乐生成技术也面临诸多挑战. 一方面, 视觉与音乐模态存在显著差异, 需要构建统一的表征以实现语义对齐; 另一方面, 音乐结构复杂, 需要同时捕捉风格特征并控制细节要素.

Zhu 等[184] 通过舞蹈动作等视觉信息量化音频表示, 使模型具备符号化与连续性特征提取能力, 从而生成风格复杂且与舞蹈高度匹配的音乐. Su 等[185] 采用多阶段自回归模型, 从对齐的视频帧与音乐波形中学习通用视觉–音频对应关系, 缓解场景依赖问题. Liu 等[186] 利用光流提取逐帧动态特征, 结合自注意力进行帧内聚合, 并将动态信息映射为音频标记, 实现跨视频场景的时间对齐音乐生成. Lin 等[187] 提出视频节拍对齐方案与高效时间编码器, 捕捉细粒度动态信息, 使生成音乐在高层语义与低层动作上均与视觉精确同步, 克服了因依赖符号化标注导致的表达力弱、音质差与泛化能力不足的问题. Wang 等[188] 借助多模态音乐描述模型将视觉信息转为文本描述, 再通过全局主题检索与针对性属性检索模块, 提供音乐桥梁并增强控制, 最终在显式条件生成框架下输出高质量、可控且与输入高度一致的音乐. Tian 等[189] 设计长短期视觉模块分别提取全局与局部动态信息, 结合 Transformer 音乐令牌解码器, 实现端到端的音乐生成.

3.2.3 视觉驱动的环境声音生成

环境声音, 如物体被击中的声响、动物的吼叫以及烟花声等, 对于人类理解和感知周围环境具有至关重要的作用. 这些声音不仅丰富了我们的感官体验, 还为我们的认知提供了重要的线索. 在此背景下, 环境声音生成技术应运而生. 该技术以特定的视觉场景为输入, 旨在构建与之匹配的声音信息. 通过这种方式, 环境声音生成技术能够为视障人士提供一种全新的感知和识别场景中物体、动作和事件的方式, 从而显著增强视障人士的独立生活能力.

然而, 实现高保真且物理可信的环境声音生成仍面临诸多挑战. 在复杂场景中, 多种声源和时序事件链 (例如玻璃破碎后碎片落地) 交织在一起, 模型需要解耦空间声源定位、时间事件同步以及多声源干涉, 从而构建出符合物理场景的连贯声音. 此外, 环境声音涵盖了从低频 (如脚步声、风声) 到高频 (如玻璃破碎、烟花爆炸) 的多种频率, 因此需要从视频中提取多层次的时空特征, 以构建跨尺度的视听表征, 生成空间上真实且一致的声场.

早期研究主要聚焦传统时序建模与特征优化. 例如, Owens 等[190] 通过 RNN 从视觉数据中提取音频特征并转换为声波, 实现静默视频的动作/材质声音预测. Zhou 等[191] 结合多视觉编码器提取特征, 利用 LSTM 实现视频到音频长序列生成. Du 等[192] 以视听片段为条件, 自回归预测频谱图序列, 生成与视频同步且保留音色的声音.

后续研究主要聚焦生成质量提升与跨模态对齐. Iashin 等[193] 引入频谱图感知损失与基于窗口的 GAN, 加速波形生成, 显著提升音频生成的速度和质量. Sheffer 等[194] 提出一种两阶段框架: 第一阶段生成低维音频表示, 第二阶段上采样为高保真音频; 同时引入预训练 CLIP 模型确保跨模态特征对齐. Chen 等[195] 提出从视频帧解耦外观与运动特征, 通过对比真实/生成声音特征过滤噪声, 解决视听信息不匹配问题.

随着多模态大模型兴起, 研究转向语义化与端到端生成范式. Liu 等[196] 利用多模态大模型生成语义对齐的思维链, 引导音频基础模型合成高保真音频, 提升在捕捉视频细节、视觉动态及声学环境方面的真实度. Jeong 等[197] 将视频作为条件控制注入文本到音频模型, 结合文本提示实现对音频结构的动态调控. Xie 等[198] 引入视觉–语言模型, 将视频到音频的生成问题分解为视频到文本再到音频两阶段任务, 提升语义准确性与时间同步性. Chen 等[199] 联合文本–图像与文本–谱图扩散模型, 在共享潜在空间中参考双模态噪声估计, 生成视听一致的样本并通过预训练解码器转换为波形.

3.2.4 视觉驱动的立体声音合成

当声源发出的声波到达双耳时, 会出现微小的时延和声级差异. 这种现象被称为双耳听觉效应[200], 它使人类能够辨别声源的方向、距离和空间位置. 立体声音合成, 也称为双声道音频合成, 旨在模拟生成立体声音体验[201]. 早期的立体声音合成主要依赖于信号处理技术, 常用的方法包括头部相关传输函数[202] 和脉冲效应[203]. 近年来, 研究者开始尝试将视觉数据作为辅助信息加入音频双耳化的过程中. 然而, 基于视觉信息生成立体声的核心挑战在于如何有效地学习全景声场表征. 具体而言, 模型需要在复杂声学场景中解耦混合声源并定位其三维空间坐标. 这一任务要求模型依据多视角视觉线索感知

声源的方位分布, 从而生成物理一致的沉浸式空间音频.

早期研究聚焦几何信息与空间感知建模. Morgado 等[204] 基于音频与 360° 视频帧的多模态分析, 分离单个声源并定位, 生成沉浸式的空间音频. Parida 等[205] 利用场景深度图结合分层注意力机制, 强化声源距离感知, 实现精准音频双耳化. Li 等[206] 提出视觉引导的生成对抗方法, 通过共享时空信息协同优化模型, 提升立体声空间真实感.

随着生成模型的发展, 新一代方法逐渐转向隐式场景建模与端到端的空间音频合成. Chen 等[207] 提出基于点云的音视频渲染框架, 通过稀疏锚点构建场景表示, 并解码为听者视角的空间传输函数, 将单声道转为视角相关立体声. Xie 等[208] 则构建 4D 场景音频生成框架, 借助多模态大模型进行像素级定位, 将每帧声源映射为三维点云以追踪轨迹, 再结合物理模拟与脉冲响应生成动态双耳音频. Marinoni 等[209] 利用深度图与目标检测框提取空间线索, 作为扩散模型中的跨模态注意力条件, 生成与场景几何和运动动态一致的空间化音频. Li 等[210] 通过分别编码左右声道并设计新型音频标记序列, 使单阶段语言模型能够有效建模立体声信号, 在保持结构简洁的同时提升生成质量. Kim 等[211] 通过融合 CLIP 视觉特征与方向感知的神经音频编解码, 实现从视频直接合成时序一致、空间连贯的多声道音频.

3.3 视听语义检索

“未见其人先闻其声”的典型场景, 生动诠释了人类大脑能够快速建立视听信息的关联能力. 这种精准高效的跨模态匹配机制, 显著提升了人类对动态环境的感知与理解效能. 在多媒体数据中, 视听数据内在的语义相关性和时间同步性, 为实现视听语义检索任务提供了坚实基础. 如图 6(c) 所示, 视听语义检索旨在建模视听跨模态数据间的语义关联性, 实现视听数据间的双向检索.

3.3.1 音频检索图像/视频

音频检索图像/视频任务旨在通过音频中的语义特征, 实现音频与视觉信息之间的语义级关联, 从而利用音频信息检索相关的图像或视频. 这一技术在多个实际应用场景中展现出显著的价值, 在智能安防领域, 音频检索图像/视频技术可通过监控音频快速定位嫌疑人影像, 显著提高案件侦破效率. 例如, 系统能利用犯罪现场录音, 从海量视频中找出匹配的说话人影像, 为侦破提供关键线索. 在多媒体内容管理中, 广播电视台利用该技术建立智能媒资库, 通过输入解说词或背景音乐, 快速检索相关画面或片段, 提升内容制作效率. 这不仅优化了资源管理, 还为创作者提供了高效工作流程.

为建立音频与图像/视频间的语义关联, 现有研究主要围绕跨模态表示学习、关联对齐以及特定挑战驱动的机制设计展开. Nagrani 等[212] 利用 CNN 挖掘人脸图像与音频频谱图之间的潜在关联, 实现跨模态人脸–音频匹配. Fang 等[213] 通过设计情感约束的目标损失函数, 在优化类间距离的同时保持情感一致性, 提升了语义层面的匹配精度. 为进一步构建共享表示空间, Zeng 等[214] 基于深度典型相关分析与 LSTM 的联合建模方法进行关键片段选择. Guo 等[215] 利用图像–语音语义关联作为监督信号, 将该范式拓展至遥感图像检索. Chen 等[216] 则融合聚类学习与度量学习, 以无监督方式捕捉声音与面部之间的深层关联特征. 面向实际应用中存在的音频描述缺失与噪声干扰等问题, Oncescu 等[217] 借助视觉–音频双标注数据引导大语言模型 (large language model, LLM) 生成音频描述, 从而缓解现有视频文本数据在音频检索任务中音频信息弱化问题. Lin 等[218] 提出重要性感知的多粒度融合模型, 通过音频重要性预测器抑制无效或噪声片段, 改善视频片段检索中的语义完整性与鲁棒性.

3.3.2 图像/视频检索音频

与音频检索图像/视频不同, 图像/视频检索音频是利用图像或视频信息来检索相关的音频, 该技术在现实场景中具有广泛的应用价值. 例如, 在影视后期制作中, 剪辑师可利用关键画面快速匹配契合的背景音乐或音效, 如战争场景匹配爆炸声、枪击声, 浪漫镜头匹配抒情音乐, 从而提升声音设计效率. 在社交媒体平台, 该技术能为短视频自动推荐风格匹配的背景音乐, 优化内容创作体验. 在博物馆等文化场所, 展品图像可触发相关解说音频, 提供沉浸式导览服务. 这些应用提升了人机交互自然度, 推动了多媒体产业智能化升级.

早期研究聚焦跨模态关联基础架构与嵌入空间优化. Li 等[219] 首次提出 image2song 任务, 通过 CNN 提取图像区域语义标签, 结合 RNN 建模歌词特征, 利用多层感知机构建跨模态关联. Hong 等[220] 设计跨模态排序损失, 缩小相关视频–音乐嵌入距离并保留模态特性. Nakatsuka 等[221] 采用深度度量学习, 约束相关对接近/无关对远离以学习潜在关系. Cheng 等[222] 提出半监督学习策略, 通过跨片段信息整合提取显著特征, 缓解标签噪声影响. 随着对语义精度需求的提升, 研究转向细粒度对齐与外部知识注入. Era 等[223] 实现人体动作–音乐特征的

细粒度对齐, 在共享空间中精准匹配节拍与音乐类型. McKee 等[224] 引入 LLM 生成音乐描述, 并应用文本丢弃正则化技术, 确保检索结果与视频风格一致.

为突破语义对齐瓶颈, 新型自适应交互网络成为研究焦点. Chen 等[225] 利用对称四元注意力机制, 自适应融合通道–空间信息以突出关键语义, 并设计基于指令的跨学习模块, 实现视听特征的高效交互. Huang 等[226] 开发区域增强学习注意力模块, 联合建模图像局部–全局特征来提升视觉表示力, 进而通过双向排序加权重排模块优化模态间相似度矩阵, 显著提升跨模态对齐精度. Wu 等[227] 针对长视频无关片段干扰, 通过基于异常先验增强关键片段权重, 实现精细对齐下的精准异常匹配.

综上所述, 视听数据间的双向检索任务在实际应用中主要面临以下挑战: 音频与视觉模态在时序连续性、空间结构表达及语义编码形式上存在异质性差异; 类内样本的高变异性 (如不同场景下的同类声音) 与类间样本的强相似性 (如不同乐器的同类音高) 易引发语义模糊. 针对上述挑战, 当前研究通常从以下两个方面入手: 基于注意力机制的共享子空间学习方法, 通过模态交互增强特征判别性; 在全局场景匹配基础上建模局部语义单元 (如特定声源与视觉对象) 的细粒度关联.

4 视听协作

人类视听感知系统通过多感官整合机制实现对环境感知的优化, 这种多模态协同效应的“联合作用”可产生超加性 (“1 + 1 > 2”) 的感知强化效果[3]. 受此启发, 视听协作研究致力于探索视听信息的综合处理手段和协同作用, 实现更高效的环境感知建模. 如图 7 所示, 基于认知层次和处理目标的不同, 视听协作任务可系统性地划分为:

- 视听实例感知: 致力于对视听实例 (事件、声源和物体) 及其行为状态 (情感) 进行多维度解析;
- 视听场景理解: 实现对二维视听数据到三维空间场景的深入理解;
- 视听推理与交互: 赋予机器对视听信息的推理能力, 支撑复杂环境下的智能决策与人机交互;
- 非传统的视听学习: 探索自监督学习、增量学习 (incremental learning)、零样本学习 (zero-shot learning) 及迁移学习等前沿机器学习范式在视听多模态学习中的应用和推广, 解决标注数据稀缺场景下的模型泛化难题.

4.1 视听实例感知

在视听多模态学习中, 视听协同的核心要义并非简单的模态互补, 而是由认知机制、信息特性和任务本质共同决定的必然选择. 其不可替代性主要体现在:

- 神经进化机制: 神经科学研究表明, 人类大脑的颞上沟区域专门执行视听整合, 其反应速度比单模态处理快 30% 以上[1–3];
- 信息的固有耦合: 物理事件的发生必然同时产生特定的视觉模式与声学特征, 视听信号在物理世界具有内在同步性;
- 任务驱动的不可分性: 关键感知任务本质上要求视听信号的联合推理, 如必须同时解析声学语义特征与视觉空间关系才能实现精准定位声源, 单模态输入无法独立支撑决策.

如图 7(a) 所示, 视听实例感知致力于对视听实

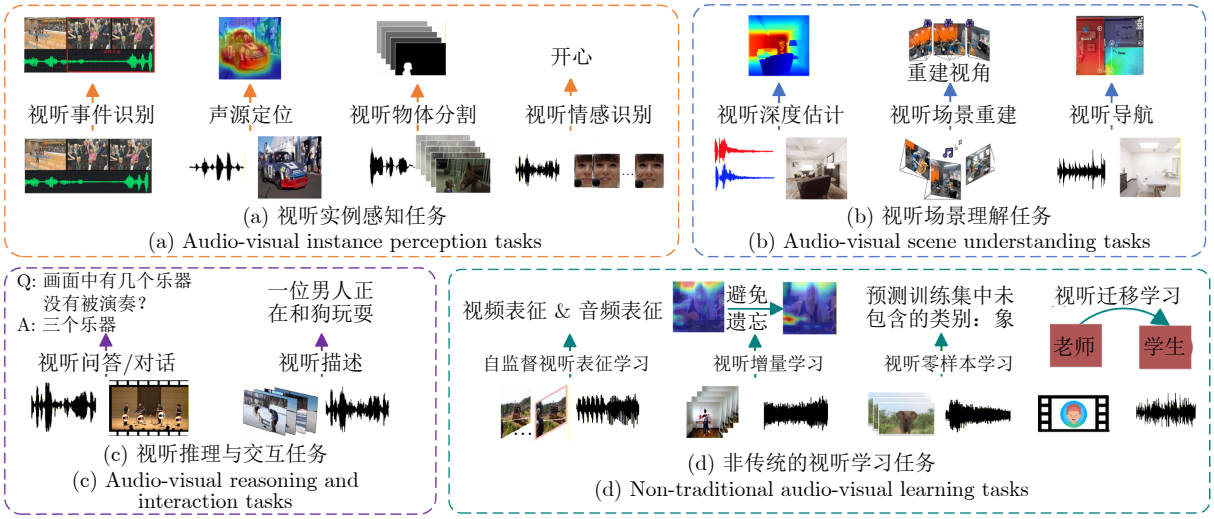


图 7 视听协作分类及其任务划分示意图

Fig. 7 Diagram of audio-visual collaboration categorization and its task division

例及其行为状态进行多维度解析, 具体包括视听事件识别、声源定位、视听物体分割和视听情感识别.

4.1.1 视听事件识别

视听事件识别致力于解决两个关键问题: 事件存在性判定与事件属性解析. 根据任务目标的差异, 该任务可进一步细分为: 视听事件定位旨在精确检测事件在连续时间轴上的发生区间; 视听事件解析则在视听事件定位的基础上, 进一步标注视听事件的模态属性 (即可听、可见或视听共存).

视听事件识别面临的核心挑战主要表现为: 在多尺度时序建模中, 模型需捕捉音视频片段在时间维度上的动态变化, 实现从局部动态细节到全局语义信息的跨粒度协同; 在跨模态整合中, 则需克服视听信号固有的时间异步性与环境噪声干扰, 实现鲁棒性融合; 在弱监督噪声抑制中, 需克服伪标签偏差与标注不确定性, 提升监督信号的可靠性; 在特定任务优化中, 则应解决模态竞争与语义干扰问题.

Tian 等[228] 首次定义了视听事件定位任务, 通过音频引导的视觉注意力机制建模跨模态关联, 并设计双模态残差网络以融合异构特征, 增强模型表征能力. Xuan 等[229] 针对视听信号时间异步问题, 引入三重自适应注意力机制, 动态筛选关键空间区域、高响应片段及主导模态, 实现有效跨模态融合. Mahmud 等[230] 则通过跨数据集集成与多阶段训练框架, 对多尺度时序信息进行建模, 并引入跨域注意力机制, 协同局部动态与全局语义, 提升长视频事件定位精度. Bao 等[231] 首次探索无监督视听事件定位范式, 利用多模态自监督信号生成伪标签, 并通过期望最大化算法优化标注质量.

在此基础上, 更多视听事件方法聚焦于结构优化与噪声抑制. Zhou 等[232] 引入跨模态一致性协作与多时序粒度协作机制, 显著提升对长视频中密集、重叠视听事件的定位能力. Sun 等[233] 通过自约束双模态交互模块、前景事件增强模块与弱监督对比损失, 缓解多模态融合机制中的噪声干扰问题. Liu 等[234] 通过引入共享文本语义标签构建统一语义空间, 并利用跨模态对齐与图交互模块来弥合音视频语义鸿沟. Zhang 等[235] 通过音频引导的视觉关系挖掘、轻量化的单流跨模态关系编码和时序连续性损失函数, 综合学习视频内的多种关系.

对于视听事件解析任务, 其研究重点主要聚焦在语义关联与模态平衡. Lin 等[236] 通过挖掘不同视频间共享的事件语义模式与跨模态事件共现规律, 增强模型对相似事件的区分能力. Fan 等[237] 通过计算语言提示与视觉/音频片段的余弦相似度, 将最大相似度对应的文本标签作为片段级伪标签, 有效过滤无关片段噪声. 针对特征学习中模态竞争失衡问题, Fu 等[238] 引入可学习的模态不确定性度量函数, 以自适应的方式平衡多模态特征的优化方向. 针对密集重叠事件带来的定位挑战, Geng 等[239] 将视听事件解析任务重构为分类–回归联合学习范式, 通过显式建模事件间时空依赖关系, 实现对不同持续时间、高度重叠事件的精准定位. Yu 等[240] 通过引入多分支特征金字塔模块和门控自适应融合模块, 动态整合多尺度语义信息.

近年来, 弱监督学习与语义解耦成为视听解析任务的重要发展方向. 针对伪标签生成中的段间依赖缺失与预测偏差问题, Lai 等[241] 引入不确定性加权的伪标签和特征混合训练正则化, 以增强标签的可靠性与模型泛化能力. Gao 等[242] 通过将标签去噪过程与视频解析模型进行协同学习, 并借助验证反馈机制直接引导去噪策略的更新. 为应对弱监督与数据本身带来的不确定性, Gao 等[243] 设计一种证据收集机制, 综合利用模态的存在证据与互补模态的缺失证据, 并通过联合模态互学习过程实现证据的动态校准. Zhao 等[244] 通过将片段特征解耦为类别特定特征, 使模型能选择性聚合跨片段的匹配语义, 从而抑制无关事件类别的干扰.

4.1.2 声源定位

声源定位是人类感知系统中一个至关重要的功能, 使我们能够在复杂环境中快速准确地判断声音的来源方向和位置. 鸡尾酒会效应[3] 凸显了人类在该能力上的卓越表现: 即使处于嘈杂环境中, 个体仍能选择性关注并定位出特定对话对象的语音信号, 从而实现更有效地交流与互动. 受此启发, 声源定位任务致力于从复杂的视听场景中, 精准定位发声源所对应的视觉区域.

然而, 复杂环境下的声源定位技术仍面临多重核心挑战: 首先, 精确的位置标注成本高昂且难以获取, 模型只能依靠粗粒度标签 (图像级或视频片段级的标签); 其次, 实际场景中常存在多声源混叠、环境动态变化及视听干扰等复杂因素.

在声源定位任务中, 部分工作尝试减少对精细标注的依赖. Senocak 等[245] 通过注意力机制处理视听输入, 证明模型在无监督条件下仍可有效定位多类别声源. Xuan 等[246] 利用全局响应图作为一种无监督的空间约束, 提出一种无监督的物体级别的声源定位框架. Liu 等[247] 强调组合增强策略 (如图像与声音的变换不变性) 与几何一致性 (即声源位置随视频帧变换同步变化) 在表征学习中的关键作用, 并通过自监督框架探索数据转换的有效性.

针对多声源混叠与复杂场景建模, Qian 等[248]

提出一种两阶段学习框架, 以从粗到细的方式分离并对齐不同类别的跨模态表征. Mo 等[249] 则侧重于从混合信号中学习各声源的语义特征, 并以此引导视觉定位. Ryu 等[250] 使模型在混合音频信号上进行训练, 并借助音视频对齐目标, 使模型同时学习不同声音的对应关系与分离能力.

在语义理解优化方面, Senocak 等[251] 引入语义对齐与对象先验知识, 以增强模型的跨模态语义理解能力. Senocak 等[252] 通过结合检索式策略和手工数据增强技术, 增强模型对音视频事件关联性的判别能力, 使其能够更准确地识别静默物体及多物体场景中的真实声源. Kim 等[253] 将图像和音频特征分解为目标表示与非目标表示, 并结合跨模态注意力匹配进一步对齐局部特征, 有效缓解噪声及无关背景的干扰. Liu 等[254] 提出一种迭代式多模态参数融合策略, 将双流网络压缩为单流结构, 在保持性能的同时将推理参数量减半.

4.1.3 视听物体分割

声源定位任务通常难以实现发声物体的精确轮廓分割. 与传统视觉物体分割不同, 视听物体分割是在音频线索的引导下, 对视频帧中的发声物体进行像素级掩码预测, 从而实现声源的精准定位与轮廓提取. 然而, 听觉模态的引入, 也为视听物体分割带来了以下挑战: 视觉模态固有的空间密集特性, 易导致音频线索被压制, 引发模态表征偏差; 物体形态与运动的多样性, 叠加声学信号的宽频特性, 要求模型具备层次化融合能力以保证分割结果的连贯性; 像素级掩码标注成本高昂, 如何在弱/无监督条件下实现精确分割.

为了缓解上述问题, 研究者提出了多种技术路径. Zhou 等[255] 首次提出视听物体分割问题并构建基准数据集, 同时引入时序像素级视听交互模块, 以音频语义引导视觉分割过程. 针对标注稀缺问题, 弱监督与无监督方法成为研究热点. Mo 等[256] 提出一种弱监督视听物体分割框架, 借助多尺度多实例对比学习的方法实现视听对齐. Bhosale 等[257] 进一步探索无监督范式, 基于预训练的多模态基础模型实现重叠区域的精细分割. Liu 等[258] 探索利用现有图像数据与音频语料合成训练样本, 以降低标注负担.

为了解决视觉模态占主导地位、音频线索利用不足导致的模态失衡问题, Chen 等[259] 通过增强音频线索, 缓解音频和视觉模态之间的模态失衡问题. Mao 等[260] 提出一种显式条件多模态变分自编码器 (variational autoencoder, VAE), 将模态内的表征分解为模态共享和模态特定的表征, 以有效评估各模态在联合表征中的独立贡献. Liu 等[261] 提出音频引导的模态对齐模块与不确定性估计模块, 缓解分割过程中因对应关系模糊与物体发声状态频繁变化所导致的过分割与欠分割问题.

随着基础模型的兴起, 多项研究探索了借助大模型先验知识提升分割性能与效率. Bhosale 等[262] 通过整合 DINO、SAM 和 ImageBind 等基础模型, 并引入像素匹配聚合策略, 在无需精细标注的情况下实现像素级音视频关联. Zhu 等[263] 则通过帧间一致性模块和音频引导的提示编码器, 将 SAM 的先验知识有效迁移至视听物体分割任务, 缓解音频信息利用不足与时序关系建模困难的问题. Xuan 等[264] 通过引入跨模态适配器、跨模态提示器和跨模态自监督机制三个核心组件, 提升 SAM 模型先验知识的利用效率.

4.1.4 视听情感识别

人类情感感知本质上是多感官整合的过程: 笑声中升高的基频与面部颧大肌收缩, 共同构成愉悦的双模态证据链; 哭声的颤抖韵律常与皱眉肌激活在 200 ms 内同步出现, 构成悲伤的时空耦合特征. 视听情感识别旨在通过发掘视听信息的内在联系与交互协作, 实现对人类情感状态的精准识别. 然而, 该任务仍面临以下固有挑战: 视听模态间常存在表达不一致性 (如抑制性表情与哽咽声) 及时间非同步现象; 情感状态具有连续动态演变特性, 呈现积累、峰值与消退的完整时序过程.

为应对上述挑战, 研究者提出了多种技术路线. Lei 等[265] 通过同时分析意图标签与感知标签在不同模态下的差异性与一致性, 并构建三组多尺度排序规则来生成相关分数, 使得模型能够有效利用不同标签提供的互补信息. Goncalves 等[266] 提出统一的多任务学习框架, 可同时处理单模态和多模态输入, 实现对情绪属性的预测.

在特征融合机制设计上, Hsu 等[267] 依据模态间隐含的情感一致性估计注意力权重, 动态融合跨模态特征以增强上下文感知. Mocanu 等[268] 则使用时–空–通道三重注意力机制, 分别捕捉模态内与跨模态的判别性特征. Praveen 等[269] 则引入联合交叉注意力机制, 显式建模模态间与模态内依赖关系, 以生成更具判别力的融合表示. Praveen 等[270] 引入门控机制, 根据音视频模态间的互补性强度自适应调节信息流动: 互补性强时强化跨模态注意力特征, 反之保留原始特征, 从而灵活捕捉跨模态协同关系.

在特征增强与表征学习方面, Pan 等[271] 提出基于统计几何的关键情感帧学习方法, 并借助进化算法优化视听判别特征提取. Shi 等[272] 通过引入光流

信息增强视频表征以捕捉面部细节变化, 并借助注意力机制同步强化模态内外特征, 提升特征的判别能力. 针对效率与实时性需求, Ding 等[273] 则采用 1D CNN 处理音频 MFCC (mel-frequency cipstal coefficients) 特征, 2D CNN 处理面部图像, 在显著降低计算复杂度的同时兼顾情感识别性能, 为实时紧凑型系统的部署提供了可行路径. Sharafi 等[274] 同时利用 CNN 和双向 LSTM 并行处理视听信息, 增强模型对双模态情绪特征的捕获能力.

4.2 视听场景理解

研究表明, 大脑颞上沟区域存在专门的神经网络, 用于强化视听同步信号的整合[2]. 该机制主要表现在: 在嘈杂环境中, 观察说话者唇动可使目标语音信噪比提升 15 ∼ 20 dB[6]; 视听联合定位使方位判断误差从纯听觉的 30° 降至 5° 以内; 视听反射弧仅需 120 ∼ 150 ms, 较纯视觉搜索快 200 ms. 这种视听整合的核心优势主要体现在以下三个层面:

- 功能互补性: 视觉信息提供空间分辨能力, 音频信息增强时间敏感度, 而跨模态校验则进一步提升视听多模态学习的鲁棒性;
- 认知增强效应: 通过模拟大脑颞上沟区域的神经整合机制, 多模态模型能够展现出更强的认知能力, 显著提升感知精度和效率;
- 语义一致性: 视听协作确保了多模态信息在语义层面的一致性, 使得模型能够更自然地理解和解释复杂环境.

这些优势直接推动了视听场景理解技术的进步和发展, 如图 7(b) 所示, 视听场景理解的目标在于实现对二维视听数据到三维空间场景的深入理解, 其核心任务包括视听深度估计、视听场景重建以及视听导航. 此外, 如图 4 所示, 视听场景理解任务通过二维数据推理出三维空间信息的过程, 也体现出对时序视听数据“增强”效应. 因此, 视听场景理解同时体现出视听协作与视听增强的双重特性.

4.2.1 视听深度估计

自然界中, 海豚和蝙蝠等生物通过回声定位系统能够精确感知环境的深度信息, 其空间分辨率可达到厘米级甚至毫米级. 受此启发, 视听深度估计通过提取并融合视觉几何线索与声学传播特性, 构建从二维视听信号到深度维度的映射关系, 为视听导航等下游任务提供精确的距离参照基准. 然而, 在复杂开放环境中, 视听深度估计面临如下挑战: 视觉模态在透明或镜面表面易产生深度估计偏差, 声学模态则易受混响与噪声干扰; 需保证宏观场景布局 (米级) 与微观物体细节 (厘米级) 的跨尺度几何一致性; 高精度深度真值标注成本高昂.

为解决上述挑战, 研究者提出多种视听深度估计框架. 在多尺度与上下文建模方面, Wang 等[275] 利用金字塔池化模块提取多尺度特征, 通过聚合不同区域内的上下文信息, 增强模型捕获全局信息的能力. 在自监督学习与物理属性利用上, Gao 等[276] 通过捕获 3D 室内场景回声响应, 并利用交互式学习预测其视觉特征, 以自监督的方式从回声信号中提取空间相关特征. Liu 等[277] 利用回声固有的物理速度属性, 为视觉深度估计提供了可靠的尺度基准, 有效缓解深度估计中的尺度模糊问题.

针对单一传感器存在的不确定性与噪声问题, Karaoguz 等[278] 提出基于奖励信号的学习策略, 联合音频、视觉及材料属性预测响应奖励, 从而指导深度估计模型的优化. Zhang 等[279] 引入特征迁移模块和相对深度不确定性估计模块, 在输出层面实现像素级融合. 在时序信息整合与材料感知建模中, Sun 等[280] 通过由粗到细的流程, 将时间定位精度从事件级提升至毫秒级, 并融合光流、视觉时间戳、音频与物体特征以回归深度, 在长距离测距中展现出优越性能. Parida 等[281] 通过显式引入物体材料属性, 将 RGB 图像与双耳回声有效融合, 利用了不同场景元素与回声深度之间的内在关联性.

4.2.2 视听场景重建

人类大脑通过多级皮层网络的协同计算, 实现了视觉和听觉信号在时空与语义维度的自适应融合, 从而完成高精度的三维场景重建. 与仅关注距离信息的视听深度估计不同, 视听场景重建旨在对三维环境进行全面建模, 涵盖几何结构、材质属性和声学特性等多维特征. 这种全息化的环境重建技术为虚拟现实和增强现实系统提供了高度逼真的数字场景, 显著提升用户的沉浸式体验. 然而, 由于实际场景的复杂性与视听模态的异构性, 该任务面临如下挑战: 视觉与听觉数据在时空对齐方面存在固有差异, 传感器标定误差和传输延迟进一步加剧了跨模态配准的难度; 同时, 动态场景中物体运动引发的声光特性变化, 也大幅增加了物理一致性建模的复杂度.

为了解决上述挑战, 研究人员提出了多种解决方案: Liang 等[282] 提出了首个基于 NeRF 的真实世界音视频场景合成方法, 通过声学感知的音频生成模块将音频生成与 3D 场景几何和材质属性隐式关联, 并利用以声源为中心的坐标变换模块学习声学场. Purushwalkam 等[283] 利用音频信息感知相机视野外几何结构, 通过视听联合推理预测室内布局与语义标签, 从而在有限的视角中快速重建完整的环

境三维平面图. 针对窗户、镜子等具有反射特性的表面导致的深度不连续问题, Wilson 等[284] 则采用了深度滤波与平面填充技术进行有效修复, 显著提升了复杂场景与物体三维结构的重建质量. 为克服透明或反光表面导致的视觉重建失真, Kim 等[285] 结合 360° 立体视觉深度与声学反射器位置估计, 有效改善该类场景下的重建精度. Alawadh 等[286] 通过构建语义化 3D 体素模型, 结合语义场景补全技术与声学渲染, 实现了沉浸式 VR 环境的自动构建. Konno 等[287] 分别重建运动物体与静态背景的三维结构, 并将其融合为完整动态场景, 实现对包含运动物体场景的三维重建.

4.2.3 视听导航

人类通过听觉和视觉系统的协同作用实现精准声源定位: 听觉系统解析双耳时间差和强度差获取方位信息, 视觉系统则提供空间参照线索. 视听导航技术旨在通过深度挖掘音频信号中的声波传播特性 (如混响时间、频谱特征) 和视觉场景的几何约束 (如透视关系、遮挡信息), 实现对发声源的亚米级定位与智能路径规划. 该技术在多领域展现出广阔前景, 如家庭服务机器人响应呼叫、医院护理机器人定位求助声等.

然而, 受限于动态环境变化与开放场景的语义复杂性, 该技术面临的核心挑战主要体现在: 跨模态空间关联建模的难题, 需建立视觉场景结构与声学特征之间的精确映射; 开放语义泛化能力不足, 需有效解耦声源语义与空间特征以识别未知类别; 多模态协同决策复杂, 涉及目标定位消歧、动态环境下的强化学习优化以及多目标路径规划等多个子任务的协同处理.

为此, 研究人员提出多种关键解决路径. 在跨模态空间关联方面, Younes 等[288] 通过融合视觉和音频模态的空间特征, 学习环境地图结构与音频几何属性之间的关联, 实现在多源音频环境下的有效导航. Shi 等[289] 通过声音事件描述符提取目标声源的空间与语义特征, 并利用多尺度场景记忆 Transformer, 实现对目标的稳定跟踪.

针对语义泛化与消歧问题, Wang 等[290] 提出一种语义与空间注意力解耦机制, 提升模型在未知声源类别上的采样效率与零样本泛化能力. Huang 等[291] 构建多模态语义消歧框架, 通过融合视觉、音频与语言线索来降低目标定位过程中的不确定性. Chen 等[292] 提出语义音视频导航新任务, 突破传统持续发声假设, 引入符合语义的偶发声音 (如冲马桶声). 他们通过推断融合空间与语义属性的目标描述符, 并利用持久多模态记忆模块, 实现在声音停止后仍能导航至目标位置.

在多模态协同决策与动态环境适应方面, Chen 等[293] 采用多模态强化学习技术, 优化了复杂动态场景下的发声目标检测与追踪策略. Kondoh 等[294] 则设计面向多目标的决策模型, 结合声源分离与路径记忆机制, 实现对多声源场景的高效导航规划. Yu 等[295] 通过构建听觉复杂环境模拟真实世界的声音干扰, 模拟对抗性攻击者, 并采用中心化评论家与分布式执行者协同训练, 使导航智能体在零和博弈中学会规避干扰. Liu 等[296] 通过双向自然语言对话与人类交互, 并借助轨迹预测和问答推理网络, 应对音频目标稀疏或环境嘈杂的挑战, 在噪声场景下显著提升导航成功率.

4.3 视听推理与交互

逻辑推理和抽象总结能力是人类智能的核心特征, 它们共同支撑着人类的高级认知活动, 如问题解决、决策制定和复杂交流. 现代神经科学研究揭示, 人类高级思维活动 90% 以上的认知负载依赖于视听通道的协同处理[3]. 这种进化机制所形成的多感官整合效应, 具有以下优势:

- 通过视觉与听觉通路的异步并行处理, 实现高效的多模态特征绑定, 极大降低单模态认知负荷;
- 当某一模态受损 (如视频静音或噪声干扰), 另一模态可提供互补信息;
- 通过跨模态类比实现创造性思维, 基于视听线索进行高效的因果推理.

为了让机器具备类似甚至超越人类的这种多模态信息推理交互能力, 视听推理与交互任务应运而生. 其核心目标是赋予机器对视听信息的综合分析与推理能力, 从而实现更自然、高效的人机交互. 如图 7(c) 所示, 视听推理与交互任务包括视听问答/对话和视听描述. 此外, 如图 4 所示, 视听推理与交互任务的结果往往以文本的形式呈现. 因此, 这类视听多模态学习任务也体现出“视 & 听 → 文本”的跨模态属性.

4.3.1 视听问答/对话

视听问答[297] 旨在基于视频中的视觉内容与音频信息, 结合给定的自然语言问题, 生成准确答案. 其核心挑战在于: 多模态特征提取需有效过滤噪声, 克服视听片段无关性; 跨模态对齐需解决语义与时空表征差异; 数据集的语言与联合偏见易导致伪相关推理; 动态场景下的对象追踪与事件因果关联则对时序建模提出更高要求.

为应对上述挑战, 研究者提出多种技术路径. Yang 等[298] 通过分层融合模块建模音频–视觉–文

本三模态间的多层次语义关联, 并发布了一个视听问答基准数据集. 在时空筛选与对象关联机制上, Zhao 等[299] 将视觉、音频与问题语义统一建模为统一多模态图, 并利用跨模态特征对齐模块弥合语义鸿沟, 从而显式刻画跨时序、跨模态的动态事件关联. Li 等[300] 提出一种分级时空筛选机制, 即先基于问题相关性选择关键时间段, 再进一步聚焦空间区域来回答问题, 有效降低视频中无关噪声对问答过程的干扰.

Li 等[301] 通过运动驱动、声音驱动和问题驱动三个并行模块协同工作, 精准追踪与问题相关的发声对象. Li 等[302] 通过问题引导的线索发现与模态引导的线索收集机制, 结合对象感知的自适应正样本对比学习策略, 精准关联问题相关对象与发声物体. Li 等[303] 提出了一种特征对比约束的多粒度关联注意方法, 通过对每个问题均给予同等重要性, 将问题与相应的视频/音频片段进行关联, 从而有效地捕获问题的局部顺序表达. Jiang 等[304] 通过融合视听特征, 实现对事件时间和目标对象的精准定位与感知, 从而显式地捕捉问题所涉及的时空语义信息.

在偏见缓解与鲁棒性增强方面, Ma 等[305] 构建带有分布偏移的诊断性数据集, 并提出多角度循环协作去偏架构, 有效缓解音视频问答模型中的数据集偏差问题. Lao 等[306] 基于因果分析揭示视听问答模型存在语言偏见与跨模态联合偏见, 借助偏见图谱干预与多路径去偏机制动态调整不同样本的去偏策略.

在推理增强与表示学习层面, Li 等[307] 将问题转化为声明式提示来增强时序感知能力, 并设计空间感知模块融合关键视觉片段与音频进行跨模态交互, 从而精准捕捉与问题相关的音视频线索. Pei 等[308] 通过问题引导变换器, 学习多个问题间的集体推理信息, 并将其编码为可学习令牌, 使得推理时即使仅输入单个问题也能利用这些信息, 在显著减少训练时间的同时保持最优性能.

视听对话作为视听问答的扩展任务, 着重于在动态变化的视听环境中与用户进行连续的多轮对话. 该任务除继承视听问答中的核心挑战外, 还面临以下独特难点: 多轮对话中上下文信息的长期依赖与主题漂移问题; 语音内容、情感语调与面部表情等非语言线索的融合与理解; 生成响应在语义、情感与上下文层面的一致性保持.

为推进视听对话研究, Alamri 等[309] 首次提出了视听问答任务, 并为此构建了多模态视听对话基准数据集. Heo 等[310] 构建语言驱动的多模态表示框架, 通过将视听特征转换为自然语言描述, 以进一步增强模型理解相关事件的多模态特征表示. 在多轮对话建模与上下文感知方面, Chen 等[311] 提出一个跨模态对齐特征的视听记忆网络模块, 通过上下文建模实现不同模态间对话的深度融合, 从而增强模型对追踪对话历史主题信息的动态变化能力. Li 等[312] 采用多任务学习策略, 联合优化多模态表征与响应生成过程, 成功将预训练语言模型扩展并适配于多模态对话任务, 实现了有效的跨模态语义融合. Ye 等[313] 依据问题优先级引导模型对关键特征的选择并优化答案生成路径, 从而使网络集中关注与问题相关的关键视听信息.

在多线索融合与情感生成上, Park 等[314] 从音频与视觉信号中提取并融合语音内容、语音情感语调及细粒度面部表情等多重非语言线索, 以端到端方式生成在情感与上下文层面均更恰当、更具共情能力的对话响应. 在知识迁移与推理优化方面, Shah 等[315] 分别基于注意力机制与时间域区域提议网络, 结合注意力多模态融合、联合师生学习与模型组合技术, 实现对视频中答案证据的定位与支持. Ye 等[316] 通过知识蒸馏对齐跨模态语义以弥合异质模态间的语义鸿沟, 并进一步解耦视听依赖关系, 避免因深层特征导致的过拟合, 在提升模型性能的同时增强其对多样化问题的泛化能力.

视听问答/对话技术经历了从静态分析到动态建模的转变, 早期研究主要关注单帧特征提取, 而最新进展则聚焦于连续时空建模. 与此同时, 特征处理方法也从粗放式融合逐步演化为精细化交互机制, 显著提升了关键信息的捕获能力. 系统功能范围已从最初的单轮问答扩展至支持持续多轮对话, 通过引入记忆机制实现了更接近人类认知的交互模式.

4.3.2 视听描述

视听描述[317] 通过深度挖掘视频中的视觉场景与听觉线索的协同关系, 利用自然语言生成技术实现对视频内容的语义化表征. 正是这种需要融合多模态信息并最终转化为文本描述的特性, 使得该任务在实现过程中面临一系列关键挑战: 视觉与听觉信号的异步性导致时序对齐困难; 生成的文本需在细粒度描述与整体简洁性之间取得平衡; 视觉信息常占据主导, 导致音频线索未被充分挖掘; 长视频内容的关键信息提取与概括能力不足.

针对上述挑战, 研究者提出了多种解决方案. 在视听融合方面, Shen 等[318] 在视频描述模型中添加独立的音频分支, 从而捕获视频中每个对象的外观与空间位置信息, 生成更全面的文本描述. Xie 等[319]

采用音频引导的视觉注意力机制以提高事件定位精度, 并结合概念检测技术识别关键语义, 生成内容更丰富的视频描述. 面对长视频中冗余信息过多的问题, Çayli 等[320] 采用池化前端和下采样策略对音频和视觉属性实施知识蒸馏, 减少冗余视听帧数量的同时有效降低模型复杂度.

此外, 部分研究引入文本模态作为额外信息补充. 例如, Xie 等[321] 通过在浅层融合音频和视觉特征, 并在高层融入全局共享文本表示, 从而生成蕴含多模态互补上下文信息的特征; Yang 等[322] 将未标注视频中的语音信号及其对应的时间戳作为生成密集视频描述的弱监督信号源, 从而有效地利用未标注数据.

针对事件重叠导致起始点定位不准的问题, Han 等[323] 结合跨模态注意力机制和知识增强的无偏场景图以提高事件定位精度及描述的逻辑合理性. Kim 等[324] 从系统层面出发, 将音视频特征作为文本令牌输入, 围绕编码器架构探索、预训练模型适配与模态融合效能三个维度, 对模型进行系统设计. 随着研究深入, Shen 等[318] 提出细粒度视听描述, 构建首个多语言视听描述基准数据集, 并设计实体完整性与音频描述能力的专用评估指标, 实现音频线索与段落级生成的有机结合.

4.4 非传统的视听学习

传统机器学习 (如监督学习) 虽然在许多任务上取得了瞩目成绩, 但其核心范式依赖于大规模标注数据和静态任务设定, 导致其在开放环境中面临以下关键困境:

- 模型性能对高质量人工标注呈刚性依赖, 而专业领域标签的获取往往代价高昂甚至不可得;
- 训练数据与测试数据需满足同分布假设;
- 模型训练后固定, 无法适应新任务或新类别;
- 学习新任务时, 会覆盖旧任务的知识.

相较于传统基于监督学习的视听多模态学习方法, 非传统的视听学习更侧重于在有限标注条件下, 如何保障模型的稳健学习能力. 其核心目标在于通过借助新兴机器学习范式, 提升模型的学习效能与泛化能力. 为此, 如图 7(d) 所示, 研究者将自监督学习、增量学习、零样本学习及迁移学习等新型机器学习范式扩展到视听多模态学习领域. 这些探索为视听多模态学习领域提供了更为灵活高效的知识获取与模型学习思路.

4.4.1 自监督视听表征学习

随着互联网多媒体数据的爆炸式增长, 海量无标注的视听数据蕴含丰富的知识信息, 如何从海量无标签的视听数据中学习有效的特征表示, 已成为视听多模态学习领域的关键科学问题. 自监督视听表征学习[325] 通过挖掘视听数据间的内在关联性, 构建了不依赖人工标注的特征学习范式, 为下游任务提供了高质量、强泛化的特征表示基础.

然而, 这一研究方向仍面临若干关键性的挑战. 一方面, 视觉信号与听觉信号在原始特征空间存在结构性差异, 同一语义在不同模态的表达粒度不同 (如“狗吠”在视觉中表现为狗张嘴, 在听觉中为特定声波模式), 导致跨模态映射存在模糊性; 另一方面, 多模态对比学习方法面临两个关键问题, 即传统均匀采样易产生假负例 (false negative) 和简单负例对模型泛化能力提升有限.

自监督视听表征学习可被划分为两类: 基于二值化验证的方法和基于对比学习的方法. Arandjelović等[326] 通过定义一个视听对应任务, 即预测音视频片段是否来自于同一视频, 使模型间接地学习隐藏在大规模音视频中的高级语义信息. 与之不同的是, Owens 等[327] 在视频编码器中引入 3D 卷积以更好地捕获视频中的运动信息. Korbar 等[328] 通过预测音视频片段在时间上是否具有一致性, 来学习更有效的视听表征, 并引入课程式学习进一步提升视听表征学习的效果.

然而上述方法在处理大规模音视频数据集时, 难以充分揭示数据的整体分布特性. Huang 等[329] 利用视听事件的时间同步性作为监督信号, 通过对比学习对单个音频和视觉模态刻画不同层次的特征表示. 为了解决自监督学习中假阴性样本问题, Sun 等[330] 通过采用邻接矩阵和正则化手段对视听数据进行分析和处理以减轻假阴性样本对学习过程的影响. 此外, 理论研究表明, 通过增加负样本数量, 可以实现更严格的互信息下界. Morgado 等[331] 通过设计动态的表征存储机制, 在训练过程中实时存储和更新音视频表征, 并通过随机抽取负样本以打破负样本数量与批次大小之间的耦合. 然而, 随机采样的策略可能会选取与目标样本语义相近的负样本, Ma 等[332] 借鉴 MoCo 思想, 提出一种多模态版本的 MoCo 框架, 解决上述问题. 此外, 随机采样策略可能无法保证采样样本在语义上的一致性, Xuan 等[333–334] 进一步提出一种主动负样本挖掘方法, 缓解了随机采样可能会得到与目标样本相同或极其相似的语义信息负样本.

近年来, 研究进一步拓展至时序动态建模与跨模态知识迁移. Li 等[335] 提出一种融合语义、时间与空间三层对应的音视频多表征自监督学习方法. Sarkar 等[336] 引入“异步”跨模态关系学习机制, 通

过放松音视频模态间的时间同步约束, 使模型能够学习泛化能力更强的鲁棒表征. Jenni 等[337] 将时序自监督任务 (如播放速度识别、时序排序) 扩展至音视频双模态, 该目标在动态特征空间中构建正负样本对. Zhang 等[338] 以语音基础模型作为教师网络, 通过多教师集成与表征知识蒸馏损失, 指导学生网络从音视频数据中学习高质量表征.

4.4.2 视听增量学习

人类认知系统具有显著的持续学习能力, 这一特性在技能习得过程中表现得尤为突出. 以乒乓球学习为例, 初学者首先通过反复练习建立基础动作的自动化 (如正手挥拍) 和规则认知, 随后逐步习得高阶技能 (如旋转发球、网前截击). 整个过程呈现出知识结构的渐进式完善, 这种学习模式具有两个关键特征: 新旧知识的协同整合, 确保已有技能的稳定性; 增量式能力扩展, 实现认知边界的持续拓展.

受此启发, 视听增量学习通过计算建模模拟人类的持续学习机制, 其核心的挑战性问题是“灾难性遗忘”现象, 即模型在学习新任务时, 因参数更新而对已学任务的性能严重退化. 具体表现为: 任务间的语义干扰与模态间表征差异加剧了遗忘程度; 严格的存储限制使得历史样本的保存与复用面临困难; 不同任务中视听信号的贡献度差异也增加了持续学习的不确定性.

为缓解灾难性遗忘问题, Zhu 等[339] 提出在严格存储限制下, 采用基于相关性的样本选择与自监督蒸馏方法. Zuo 等[340] 通过分层增强模块和分层蒸馏模块, 系统利用数据与模型的内在层次化结构进行知识保护. Cui 等[341] 针对水产养殖中鱼类摄食强度评估任务, 首次引入音频–视觉类增量学习范式, 并提出基于原型的特征表示以兼顾存储效率与知识保持. Mo 等[342] 通过利用可学习的视听类标记和分组机制, 持续地聚合类感知特征, 从而有效地提升模型对判别性视听类别的捕获能力.

新增任务可能破坏已学习的跨模态语义关联或任务间依赖关系, 需保证新旧知识的逻辑一致性. Pian 等[343] 利用视听相似性约束, 以维持跨模态特征的实例与类感知语义相似性, 确保所学知识的连贯与完整. Yue 等[344] 通过递归最小二乘法将模态融合问题重构为闭式求解, 并设计模态特定知识补偿模块, 在保护数据隐私的同时保证增量分类精度. 针对模态不平衡问题, Cui 等[345] 通过时序相关提示在有限数据下高效微调基础模型并捕获音视频关联, 同时引入时间提示正则化, 从多视角利用时序知识以巩固旧类记忆.

4.4.3 视听零样本学习

人类在面对新事物时, 常依托先验经验体系与认知类比机制实现知识迁移: 即使面对全新领域, 也能通过已有知识结构进行有效推理与泛化. 例如, 即使从未见过“蜂鸟”的人, 也能通过“鸟类”和“快速振翅”等特征组合进行准确识别. 受此认知特性启发, 视听零样本学习[346] 旨在解决如何对由音频–视频序列组成的未见过样本进行分类. 其中, 测试类别在训练阶段完全不可见, 需通过跨模态语义迁移实现判别. 然而, 这一过程面临多重挑战: 视听模态间的表征差异使得共享语义空间对齐困难; 模型易偏向已见类别, 存在严重的域偏移问题; 语义信息利用不充分导致类别间区分度不足; 模态缺失与数据稀疏情境下的泛化能力也有待提升.

为实现跨模态知识迁移, 研究者首先致力于构建有效的共享语义空间. Parida 等[347] 将音频、视频和文本三种模态映射到共享的语义空间中, 并施加类别与跨模态相似性约束的双模态三元组损失进行零样本学习, 同时发布基准数据集 AudioSetZSL. Li 等[348] 提出一种基于信息瓶颈理论的变分学习方法, 学习视听及文本标签的联合表示, 帮助模型提升判别能力并缓解零样本学习中的强偏置问题.

为增强未知类别的表征能力, Zheng 等[349] 采用生成模型预测未知类别的视听特征, 并融合对比学习与判别学习策略以增强不同模态之间的语义区分度. Mo 等[350] 摒弃传统跨模态特征重构方法, 采用单一文本–音视频对比损失学习模态对齐, 并通过差分优化将类别嵌入转换为更具区分度的空间, 在保留语义结构的同时提升了零样本学习性能.

针对模态缺失与数据稀疏问题, Dong 等[351] 通过跨模态特征增强子网整合对象感知信息以学习更鲁棒的多模态表征, 并利用缺失模态生成子网为不完整输入生成虚拟特征, 有效提升模型在资源受限环境下的性能与泛化能力. Li 等[352] 引入时间步因子动态合成推理信息, 结合全局–局部池化机制优化脉冲信号, 并采用时序–语义融合模块在保持全二阶交互的同时实现多尺度特征融合.

在网络架构与推理机制方面, Yang 等[353] 结合脉冲神经网络与张量回归网络, 通过时空–语义融合模块, 整合视听信息以保留高维结构; Li 等[354] 则侧重利用其记忆机制, 通过交叉注意力融合时间与语义特征来捕获模态关联. Zhang 等[355] 通过时序信息推理模块, 增强音频时序特征捕获能力, 结合跨模态推理模块构建鲁棒联合嵌入表征. Li 等[356] 通过事件流转换、递归联合学习及差异分析等集成模块, 有效捕捉动态信息并建模跨模态关联.

4.4.4 视听迁移学习

人类在特定情境中习得的技能与知识, 可通过认知类比、神经可塑性重组和情境关联机制迁移至新情境, 显著提升学习效能[357]. 受这种机制启发, 视听迁移学习通过跨模态知识迁移, 将源域 (如大规模视觉数据集) 中学习到的特征表示、模型参数或结构先验, 迁移至目标域 (如音频任务或视听联合任务), 以减少目标域对标注数据的依赖, 并提升模型泛化能力. 因此, 视听迁移学习同时也体现了视听信息间的跨模态学习特性.

然而, 由于视听数据间存在的固有差异, 视听迁移学习在实际应用中面临如下挑战: 解决源域到目标域的模态异构性, 实现语义一致性知识迁移; 基于源域知识在目标域极低标注下高效迁移, 避免过拟合; 通过动态特征适配应对分布偏移/任务差异, 提升未知环境鲁棒性. 为了应对上述挑战, 研究者主要从跨模态迁移和跨域迁移两个层面开展工作, 旨在充分挖掘视听数据之间的互补潜力并提升模型在目标域的泛化能力.

在跨模态迁移方面, 其核心思路是利用视觉或音频的互补信息, 借助知识蒸馏、特征对齐或对比学习等方法, 将源域中知识丰富模态的特征迁移至目标模态. 针对跨模态知识迁移挑战: 研究者主要通过知识迁移与对齐机制缩小模态鸿沟. 例如, Hajavi 等[358] 基于特权信息学习的框架, 将视频作为训练阶段的特权信息, 通过知识蒸馏提升纯音频推理场景下的表征学习能力; Yun 等[359] 提出空间对齐蒸馏框架, 通过添加可学习空间模块将视觉模态的知识有效迁移至音频模态, 从而缓解视觉与音频在空间上的几何不一致性; Kim 等[360] 结合组对比学习与知识蒸馏, 通过约束音频特征以逼近共享空间表示, 并基于重要性加权策略抑制低质量模态, 实现仅依靠音频输入的鲁棒车辆检测.

在跨域迁移层面, 研究重点则放在如何应对源域和目标域之间的分布漂移、环境动态变化以及标签空间不一致等难题. Chen 等[361] 通过教师策略的置信度自适应加权蒸馏机制与全向信息采集机制解决动态环境中视听导航的样本效率与鲁棒性问题. Chen 等[362] 提出一种测试时选择性自适应方案, 使用模态专属的适配器和路由器, 根据输入自动判定哪个模态需要适配, 从而缓解跨数据集迁移时数据分布差异和标签空间不一致的问题. Duan 等[363] 提出一种跨模态适配机制, 通过将音频和视觉模态作为“软提示”, 通过可训练的跨模态空间–通道–时序注意力层动态调整特征提取过程, 从而提升模型在跨任务场景中的有效性. Mo 等[364] 采用共享的双流编码器和各任务专属解码器, 通过学习模态间的共同表征实现任务间正迁移, 验证了不同音视频任务间的互助效应.

5 核心问题分析与与前沿方向

尽管视听多模态学习在应用场景、任务目标和技术方法上呈现出显著的多样性, 涵盖分类、分割、定位、生成和推理等多个方向, 其在场景、目标与方法上差异显著. 然而, 这些多样性背后所涉及的核心科学问题, 可进一步凝练为以下五个共同挑战:

- 视听表征: 构建融合时空信息的联合嵌入空间, 解决多源异构信息的统一编码问题;
- 视听对齐: 建模跨模态语义关联, 核心在于建立跨模态时序对齐与语义一致性, 解决跨模态时间同步与语义对应关系;
- 视听转换: 实现跨模态内容生成与迁移, 完成模态间映射;
- 视听融合: 通过多源信息协同决策机制, 整合隐含线索以实现深度理解;
- 视听共同学习: 设计互补性知识蒸馏与迁移范式, 优化多模态协同训练.

如图 8 所示, 这五个共性问题共同构成了视听多模态学习领域的理论基石, 各研究方向之间的差异性, 本质上体现为对这些基本挑战所采取的差异化解决路径和创新视角的差异化探索与创新.

5.1 核心问题分析

5.1.1 视听表征

视觉数据和音频数据作为两种广泛存在的多模态信息载体, 通过对客观事物的形象化与动态化表

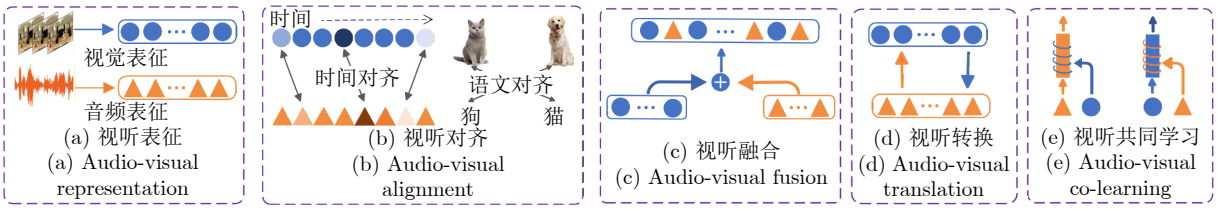


图 8 核心问题/挑战示意图

Fig. 8 Schematic diagram of the core problems/challenges

征, 实现对物理世界的数字化建模. 然而, 这两种数据形式在信号本质与信息表征方式上存在显著差异, 具体表现为:

● 在信号形式与采样特性上, 视觉数据表现为二维或三维的离散像素矩阵, 其采样过程基于空间分辨率; 音频数据则呈现为一维连续波形信号, 其采样过程基于时间频率特性.

● 在空间特性上, 视觉数据包含显性的空间结构信息, 可直接反映物体的几何特征与空间关系; 音频仅隐含空间信息.

● 在语义密度上, 视觉数据具有高语义密度特性, 单帧图像可同时包含多个对象的完整语义信息; 音频数据则表现出时序依赖性, 需通过时间累积才能形成完整的语义表达.

● 在特征尺度上, 视觉特征具有层次性, 从低级的边缘特征到中级的纹理特征, 再到高级的物体级特征; 音频特征呈现时频二象性, 同时在时间域和频率域展现其特性.

视听表征学习的目标是通过自监督或有监督的学习范式, 从原始视听信号 $X=\{x_a,\ x_v\}$ 中提取具有高度语义一致性的特征表示 $\mathcal{Z}$. 这样的特征表示能够有效捕捉跨模态数据的内在关联与结构特性, 为下游 CV 或 AP 任务 $\mathcal{L}_{\text{task}}$ 提供强健且可迁移的特征基础.

$$\begin{aligned} &X \xrightarrow{\phi(\cdot)} \mathcal{Z}=\{\boldsymbol{z}_i\}_{i=1}^{N} \\ &\phi^*(\cdot)=\arg\min_{\phi\in\Phi} \mathrm{E}_{X\sim\mathcal{D}}\left[\mathcal{L}_{\text{task}}\left(\psi(\mathcal{Z})\right)\right] \end{aligned} \tag{4}$$

其中, Φ 表示表征函数; $\mathcal{D}$ 表示数据分布; $\psi(\cdot)$ 表示任务驱动的映射函数. 最优表征函数 $\phi^*(\cdot)$ 需满足: 能够捕捉不同模态间的语义一致性; 保持同类样本在特征空间的聚集性; 有效分离语义相关因素与模态特定因素.

视听表征学习作为多模态信息处理的核心环节, 所有视听多模态学习任务均需通过表征学习, 构建模态内基础特征. 其网络架构直接制约着特征提取的效能, 不同架构在特征抽象层级、时空建模方式和计算效率上存在显著差异. 对于视觉表征学习, 常用的网络架构包括:

● CNN[14, 39, 171–173, 190, 204, 259, 275, 294, 330]

● ResNet 系列[16, 22, 37, 192, 283, 329]

● ViT[193, 205, 255]

● RNN[191] 和 LSTM[24, 195]

对于音频表征学习, 常见的特征提取技术包括梅尔频率倒谱系数 (MFCC) 和短时傅里叶变换 (short-time Fourier transform, STFT); 随后, 进一步利用

● CNN[16, 37, 39, 171, 259, 275, 294, 333]

● LSTM[24, 195]

● VGGish[26, 193, 255, 260, 328]

等网络架构提取音频表征. 这些网络架构提取音频信号的局部特征 (如频率和时长), 并通过预训练的方式, 捕捉音频中的高级语义信息 (如音高、音色和节奏).

不同网络架构通过差异化机制, 提取视听数据的跨模态、多尺度特征, 形成互补性优势. 例如, CNN 与 VGGish 通过卷积核的空间局部感知特性, 优先捕获基础视觉模式 (如边缘纹理) 与声谱局部模式 (如共振峰频带); ResNet 系列引入残差连接破解深度网络退化难题; Transformer 借助自注意力实现对全局上下文的无偏建模; RNN/LSTM 则通过门控机制, 赋予模型动态记忆时序信息的能力.

5.1.2 视听对齐

大脑通过跨 θ-γ 神经振荡耦合, 在百毫秒级时间尺度内协调多感官信息, 构建统一的视听体验[1–3]. 其中, θ 波段协调颞叶与枕叶的长程语义关联, γ 精准对齐视听信号时序. 作为视听多模态学习的核心步骤之一, 视听对齐需在异构模态间建立时空与语义的对应关系, 其复杂性主要源于:

● 模态异构性: 视觉表征为离散像素阵列, 而听觉表征为连续波形序列, 二者在维度、稀疏性与数据结构上存在显著差异;

● 时序异步性: 动作–声音物理延迟 (平均 ±150 ms) 与采样率不匹配 (视频 30 fps, 音频 44.1 kHz);

● 语义鸿沟: 同一概念在不同模态中存在表达差异, 且同一语义在不同模态的表达粒度也不同, 如“狗吠”在视觉中表现为狗张嘴, 在音频中则表现为特定声波模式.

在视听多模态学习中, 数据处理的核心目标在于构建视听模态间“语义–时空”联合嵌入空间:

$$\mathcal{T}:\mathcal{A}\xrightarrow{\phi_a}\mathcal{Z}\xleftarrow{\phi_v}\mathcal{V} \tag{5}$$

其中, $\mathcal{A}$ 和 $\mathcal{V}$ 分别表示音视频特征空间; ϕ_a 和 ϕ_v 为语义编码函数.

如图 8(b) 所示, 在作用维度层面, 视听对齐可以分为两个方面:

● 时间对齐: 强调音频与视觉信号在时序上的同步关系, 如视听学习任务中的声画同步要求;

● 语义对齐: 强调跨模态内容在语义概念层面上的对应关系, 如视听协同任务中听觉与视觉事件匹配.

在实现方法层面, 视听对齐主要分为两大类:

● 显式对齐: 基于数据统计特性的方法, 如典

型相关分析 (canonical correlation analysis, CCA); 基于监督学习的对齐方法, 利用额外标注信息指导对齐过程.

● 隐式对齐: 通过模型内置机制 (如注意力机制) 实现跨模态表征对齐; 基于图神经网络等架构的隐式关联建模.

在视听多模态学习的研究中, 现有工作通常基于视听数据固有的一致性假设开展. 具体来说, 在视觉增强的音频任务[84, 95] 和视听场景理解[276, 284−285] 中, 研究者普遍假设视听模态已经实现时空对齐, 并直接利用对齐特征执行下游任务. 然而, 视听数据可能存在画外音等异步现象. 视听事件/物体感知[236−237, 239−240] 专门研究模态不一致问题, 这类工作通过显式建模事件同步/异步特性, 实现细粒度事件感知. 此外, 部分视听多模态学习任务则侧重于更高层次的语义对齐, 例如, 文献 [121, 124] 通过音频信息识别说话人的身份, 文献 [122−123] 通过面部信息进行身份信息匹配. 在音频驱动的视觉生成任务中[135−136, 145−146, 157−158], 需要将音频信息与视觉信息进行语义对齐, 以生成生动逼真的视频. 在视听事件/物体感知任务中[256−257], 通过将音频信息与发声物体进行匹配, 实现更深层次的语义理解.

5.1.3 视听转换

人类大脑具有与生俱来的跨模态感知能力, 这一机制在认知神经科学中被称为“交叉模式知觉”[8]. 当接收单一感官刺激时, 大脑能自动激活跨模态神经表征, 例如, 听觉皮层的海浪声信号会同步激活视觉联合皮层的波浪意象; 视觉皮层的树叶运动信号会引发听觉联合皮层的沙沙作响声. 这一视听转换机制赋予人类显著的进化优势:

● 鲁棒感知: 当某一感官信号衰减时, 另一感官可即时补偿;

● 快速语义绑定: 听见“狗吠”即刻联想到视觉中“狗”的形象, 减少认知延迟, 提高反应速度;

● 记忆与联想: 声音即刻唤醒画面, 跨模态线索加固情景记忆, 有助于实现高效知识迁移.

视听转换是指视觉和听觉信息之间的互相转换, 实现感官信息从一种模态到另一种模态的映射. 在计算层面, 视听转换可形式化为下列双射函数:

$$\mathcal{T}: \mathcal{A} \underset{\phi_a}{\overset{\phi_v}{\longleftrightarrow}} \mathcal{V} \tag{6}$$

这种映射需满足:

● 模态不变性: $\forall\, a \in \mathcal{A},\ \exists\, v \in \mathcal{V}$;

● 语义保持性: $|\phi_a(a) - \phi_v(v)|_2 < \epsilon$, 其中 ϵ 表示预定义的阈值.

在音频驱动的视觉生成任务中[130, 147, 159−160], 通过 GAN 从音频信号中提取结构化特征, 并驱动与之同步的舞蹈或手势序列的生成. 文献 [158, 161] 则基于 VAE 显式地学习数据分布, 确保生成的视觉数据具备良好的稳定性和可解释性. 在视觉驱动的音频生成任务中, 研究者根据视觉信息中的空间、语义、上下文等特征, 构建从视觉到听觉的有效映射. 例如, 文献 [172−173, 175] 基于 Transformer 和 LSTM 来捕获全局的依赖关系, 以生成高质量的长序列音频数据. 文献 [174, 180, 195] 基于 GAN 及其变种, 直接利用视频帧生成波形或 MFCC, 实现对语音或环境声音的重建, 完成视觉到听觉的跨模态转换. 此外, 在跨模态检索任务中, 研究者的目标是减少视频与音频内容间的高级语义差异, 如 Cheng 等[222] 采用半监督训练策略, 提升关键视频片段的语义捕获能力和模型的鲁棒性.

视听转换作为视听多模态学习的关键挑战之一, 其核心是实现跨模态内容的实时生成与同步, 以解决模态缺失或质量不均的问题. 具体而言, 它既能重构缺失模态以保障信息完整, 也能增强弱模态以提升感知质量. 这项技术因此成为构建虚拟现实、增强现实与混合现实等沉浸式体验的基石.

5.1.4 视听融合

人类通过视觉与听觉感官持续捕捉环境中的多模态信息, 这些原始感知数据经视网膜与耳蜗初步处理后, 被传递至颞叶、顶叶及枕叶等高级脑区内进行深度整合. 认知神经科学研究表明[3, 8], 人脑的视听融合并非简单的信号叠加, 而是通过精密的时空对齐与语义关联, 实现“$1+1>2$”协同增效. 这一高效的生物整合机制, 为视听多模态融合奠定了坚实的理论基础. 视听融合旨在有效整合视听模态中蕴含的互补线索, 以突破单一模态的感知局限, 实现对环境更全面、更鲁棒的理解. 其必要性主要体现在三方面:

● 鲁棒性增强: 单一模态易受环境干扰, 如视觉在夜间失效、听觉在嘈杂环境中失真, 视听融合可利用多模态信息间的冗余性补偿干扰;

● 语义表征丰富: 视觉信息提供空间结构与运动轨迹, 听觉信息解析时域波形与频谱特征, 通过视听融合, 二者共同解码单一模态无法表达的复杂语义;

● 认知效率提升: 多模态融合显著提升信息传递效率, 并同步压缩决策延迟, 实现“感知–决策”闭环的加速与稳健化.

视听融合在计算模型中被定义为: 将视觉特征 $\mathcal{A}$ 与听觉特征 $\mathcal{V}$ 映射至联合的“时空–语义”统一空间, 以获得联合表征 $\mathcal{F}$. 该过程的通用公式化表达

可写为:

$$\mathcal{F} = F_\theta(\mathcal{A},\ \mathcal{V}) \tag{7}$$

其中, F 为融合算子; θ 为对应的可学习参数.

根据融合策略的不同, 视听融合可大致分为以下两类:

● 模型无关的融合方法: 此类方法不依赖特定模型结构, 直接操作特征或决策结果. 例如, 特征级融合[25, 266] 直接在原始特征或浅层特征层面进行拼接或加权组合, 形成联合特征向量; 决策级融合[46, 271] 则独立训练视觉与听觉模型后, 融合各模型的决策结果. 然而, 此类方法可能无法充分捕捉模态间的复杂关联, 在复杂任务中性能受限.

● 模型驱动的融合方法: 通过可学习的神经网络模型实现跨模态信息交互, 自适应地捕捉模态间复杂的关联性. 此类融合策略大致可细分成以下两种: 其一基于注意力机制的特征融合[26, 237–239, 268, 283], 通过动态权重分配实现特征选择; 其二则依赖 U-Net、GRU 等特定网络架构的融合策略[267, 275–276, 293–294], 以端到端方式学习多模态时序表征. 虽然模型驱动的融合方法需要更多的数据和计算资源, 但这类基于学习的模型能够捕捉更复杂的交互模式, 在复杂任务中表现更优.

5.1.5 视听共同学习

知识迁移指个体或系统将已习得的认知模式和技能等应用于新领域的过程, 其核心在于突破情境壁垒实现知识复用[365]. 从认知科学视角可分为:

● 正向迁移: 已有知识促进新任务学习;

● 负向迁移: 原有经验干扰新知识获取;

● 垂直迁移: 不同抽象层级的知识传递.

视听共同学习是一种通过模态间深度交互, 实现双向知识迁移的计算范式. 与传统的单向跨模态迁移不同, 该过程强调模态间的协同进化与知识互补, 其核心在于构建动态平衡的知识交换机制. 这一机制的复杂性主要源于:

● 模态差异性: 视听觉数据在分布、特征空间及噪声模式等方面存在显著差异;

● 表征异构性: 音视频表征在抽象层级上存在固有差异与语义鸿沟;

● 预测模型适配性: 源域 (知识丰富模态) 的预测模型需适应目标域 (知识匮乏模态) 的任务需求与数据分布.

视听共同学习主要利用非传统的机器学习范式 (如无/弱监督学习), 通过模态间知识迁移来增强系统的感知与协作能力. 面对标注数据缺失、数据噪声干扰等资源受限的场景, 视听迁移学习常被作为一种有效手段, 广泛应用于各类视听多模态学习任务. 例如, 文献 [358, 360–361] 运用知识蒸馏技术, 将知识传输至特定模态以优化目标任务的性能. 文献 [347–349, 354] 在训练阶段利用资源丰富的模态辅助资源匮乏模态学习更优表示, 而在测试阶段仅需依赖资源匮乏模态即可识别新类别. 通过视听共同学习, 可以促进不同模态之间的知识融合与迁移, 提高模型在跨域跨任务中的泛化能力.

5.2 实证分析

尽管视听多模态学习领域的相关研究在应用场景、任务目标和技术方法上呈现高度多样性, 本文根据人类多感官整合过程在视听认知加工中的重要性和视听多模态学习任务间的内在关联性, 将现有视听多模态学习的相关研究, 系统归纳为视听“增强–跨模态–协作”三类. 此外, 本文也对视听多模态学习背后所涉及的视听“表征–对齐–融合–转换–共同学习”五个核心科学问题进行归纳和总结. 在本节中, 本文将进一步剖析上述的三维分类框架与五大核心问题之间的映射关系, 具体包括:

● 不同类别的学习任务对核心问题的侧重差异;

● 各类任务中关键科学问题的具体表现形式.

如第 2 节所述, 视听增强学习通过在初始的单模态 CV 或 AP 任务中引入另一种模态的互补信息, 实现对初始任务的“增强效应”. 根据初始任务的模态主导性差异, 视听增强学习又可被细分为: 音频增强的视觉任务和视觉增强的音频任务. 具体而言, 如第 2.1 节和第 2.2 节所述, 二者旨在初始 CV 任务 (动作识别、显著性检测、人群计数、航空场景识别、深度伪造检测和视觉修复) 和初始 AP 任务 (音素识别、语音分离、音频增强、语音识别以及说话人识别) 中引入另一种模态, 通过跨模态信息的协同验证与互补增强, 突破传统单模态方法的性能瓶颈和鲁棒性.

为实现这一目标, 视听增强学习中的关键技术路径主要涵盖: 构建时空关联的联合嵌入空间; 挖掘视听模态间的深层语义相关性; 通过多源信息融合实现隐含线索的有效整合. 这些关键技术从根本上“强化”了单模态特征的表征能力与语义理解深度. 因此, 如表 2 所示, 视听增强学习中所涉及的核心科学问题可被归纳为视听表征、视听对齐和视听融合.

如第 3 节所述, 视听跨模态学习通过建模声学与视觉信号的关联性机制, 模拟人脑的跨模态可塑能力. 其核心目标在于构建音频与视觉数据间的语义映射关系, 实现双向模态转换与知识迁移. 基于模态转换过程中映射关系的差异性, 视听跨模态学

表 2 视听学习任务涉及的核心问题/挑战

Table 2 Core issues/challenges involved in audio-visual learning tasks

分类	子类	视听表征	视听转换	视听对齐	视听融合	视听共同学习
视听增强学习	音频增强的视觉任务	√		√	√	
	视觉增强的音频任务	√		√	√	
视听跨模态学习	视听生成任务	√	√	√		
	视听检索任务	√	√	√		
视听协作学习	视听实例感知任务	√		√	√	
	视听场景理解任务	√		√	√	
	视听推理与交互任务	√		√	√	
	非传统的视听学习任务	√		√	√	√

习可被细分为: 视听生成任务和视听检索任务. 如第 3.1 节和第 3.2 节所述, 前者聚焦于创造性跨模态内容合成, 包括音频驱动的视觉生成和视觉驱动的音频生成; 如第 3.3 节所述, 后者则通过设计跨模态相关性度量准则, 实现音频–视觉数据的高效互检索. 视听跨模态学习中的关键技术路径可归纳为: 构建时空关联的联合嵌入空间; 挖掘视听模态间的高阶语义对应关系; 设计跨模态迁移框架. 因此, 如表 2 所示, 其核心科学问题可被总结为视听表征、视听对齐和视听转换.

如第 4 节所述, 视听协作学习旨在构建视觉与听觉信息融合的认知计算框架, 通过探索多模态信息的协同作用机制, 实现环境感知效率的显著提升及类人认知模型的构建. 基于认知层级与处理目标的差异性, 视听协作学习可被细分为: 视听实例感知、视听场景理解、视听推理与交互和非传统的视听学习. 如第 4.1 节所述, 视听实例感知致力于对音视频中的实体 (视听事件、声源和视听物体) 及其行为状态 (情感表达) 进行多维度解析. 如第 4.2 节所述, 视听场景理解旨在突破二维视听数据的表征局限, 致力于构建对视听场景的语义化空间认知, 包括视听深度估计、视听场景重建以及视听导航. 如第 4.3 节所述, 视听推理与交互旨在赋予机器对视听信息的综合分析与推理能力, 实现更自然、高效的人机协同, 包括视听问答/对话和视听描述. 上述三者的关键技术路径主要涵盖: 时空关联的联合嵌入空间构建、深层语义相关性挖掘、多源信息融合机制. 因此, 其核心科学问题可归纳为视听表征、视听对齐、视听融合. 如第 4.4 节所述, 非传统视听学习探索自监督学习、增量学习、零样本学习、迁移学习等新型机器学习范式, 解决标注数据稀缺下的泛化挑战, 扩展视听多模态模型的适应边界. 其进一步引入互补性知识蒸馏与迁移, 通过优化多模态协同训练架构, 实现模型在有限样本下的鲁棒性和泛华性的提升. 因此, 如表 2 所示, 除了上述三个核心科学问题, 非传统视听学习还涉及视听共同学习.

5.3 大模型背景下的视听多模态学习

随着 GPT-3 的推出, AI 正式迈入以预训练与 LLM[366] 为核心范式的新时代. 在 AP 领域, 通用音频大模型如 AudioCraft[367]、Whisper[368]、AudioLM[369] 和 FunAudioLLM[370] 等, 通过在海量未标注音频数据集 (如 LibriSpeech、VoxCeleb、AudioSet) 上或联合 LLM 进行自监督预训练, 学习高层级音频特征表示, 显著提升语音识别、合成等下游 AP 任务的性能. 在 CV 领域, 大规模预训练模型如 Sora[371]、Video Diffusion[372]、T2I[373] 和 ViT[374] 等, 依托 ImageNet、LAION 等超大规模图像数据集进行预训练, 构建通用视觉感知能力, 并成功迁移到图像分类、目标检测、语义分割、图像/视频生成等多种 CV 任务中.

尽管单模态大模型成就显著, 其训练范式存在若干根本性局限[375]:

- 模态孤岛: 仅依赖单一模态的无约束数据集进行训练;
- 融合障碍: 模态异质性导致特征空间难以对齐;
- 性能瓶颈: 在跨模态推理任务中表现欠佳.

为克服上述问题, 多模态 LLM 应运而生. 例如, AnyGPT[376] 通过离散化的 Token 来统一所有模态的输入输出, 实现“任意模态到任意模态”的生成与交互; NextGPT[377] 利用 ImageBind 实现统一感知, 并依赖扩散模型进行高质量生成, 构建了强大的模块化多模态系统; X-LLM[378] 通过 Q-Former 等对齐模块将多模态信息高效“翻译”给 LLM, 专注于增强模型的多模态深度理解能力; VITA[379] 通过参数庞大的深度网络对音频等进行精细编码, 体现了 “重编码、精理解” 的设计哲学. 这些多模态 LLM 模型均致力于将文本、图像、音频等多种模态信息与 LLM 相融合, 以构建统一的理解与生成框架.

在此背景下, 视听多模态学习领域受到了一定程度的波及, 但影响程度相对有限. 在视听多模态学习领域, 其研究方法主要呈现以下几种典型范式:

1) 统一序列建模

将音频和视觉数据转换为统一的离散语义 Token, 并输入同一模型中进行联合建模. 例如, Zhan 等[376] 利用模态专用分词器将多模态数据压缩为离散 Token, 再使用统一的 Transformer, 对不同任务进行联合建模. Akbari 等[380] 则利用统一的 Transformer 架构, 结合多模态对比损失, 在无监督条件下同时学习视频、音频和文本的表征. Lin 等[381] 提出一种共享权重的 Siamese 网络, 使用统一的 ViT, 同时处理视频帧片段和音频谱片段, 通过对比学习实现跨模态匹配, 并引入掩码机制以增强模型泛化能力. Huang 等[382] 进一步融合音视频掩码重建、跨模态对比学习以及基于上下文特征的自训练策略, 来构建联合自监督预训练模型.

2) LLM 驱动的融合

保持各模态的独立编码器结构, 同时引入 LLM 作为核心语义处理中枢, 负责完成跨模态推理与内容生成任务. 例如, Chowdhury 等[383] 引入基于最优传输的模态对齐模块与跨模态注意力机制, 实现音视频在时空维度上的细粒度对齐. Chen 等[384] 构建一种端到端的情感对话视听 LLM, 并提出一种语义–声学解耦分词器, 以分离语音内容与声学风格, 从而精确建模带情感的语音对话. Sun 等[385] 采用多分辨率 Q-Former 结构, 将预训练的音频与视觉编码器接入 LLM, 并引入多样性损失和混合训练策略以缓解模态主导问题, 在音视频问答、定位等任务上取得显著性能提升. Chu 等[386] 将编码后的音视频与文本信息输入 LLM, 依次驱动情感识别与对话生成等任务, 从而理解用户状态并生成共情响应. Sun 等[387] 提出一种推理增强型的音视频 LLM, 通过构建推理密集型数据集, 并设计过程直接偏好优化方法, 对中间推理步骤进行奖励建模, 显著提升模型推理能力. Cappellazzo 等[388] 则基于预训练的音视频编码器与 LLM 协同, 以自回归方式生成唇读响应.

3) 下游任务适配

采用参数高效的适配技术 (如 LoRA 及其变体)[389] 结合面向目标任务的训练策略, 在保持模型轻量化的同时, 实现多任务、多粒度的高效协同与性能提升. 例如, Du 等[390] 提出一种音视频统一学习框架, 通过构建细粒度指令数据集与交互感知的 LoRA 结构, 在数据与模型层面实现多任务显式协作, 提升多项音视频理解任务的性能. Jin 等[391] 提出利用多模态 LLM 生成蕴含跨模态上下文的语义令牌作为提示, 驱动 SAM 模型实现声源目标分割, 并设计目标一致性语义对齐损失以增强视听语义学习的稳定性. Cappellazzo 等[392] 引入分层 LoRA 微调策略, 结合全局与尺度特定的适配模块, 在保证视觉增强的语音识别精度的同时, 显著提升模型的计算效率与跨场景适应能力. Tang 等[393] 引入多轮直接偏好优化方法, 通过周期性更新参考策略以避免策略过时, 在视听问答任务上取得显著性能提升. Huang 等[394] 基于预训练 Transformer 的适配器网络和内容感知注意力机制, 以极低的计算成本缓解预训练大模型在视听情感识别中的模态冲突问题.

6 结束语

作为 AI 领域的前沿方向, 对视听多模态学习领域仍然存在“任务驱动型研究”与“理论体系缺失”之间的结构性矛盾. 本文基于多感官整合机制在人类视听认知中的重要性, 以及视听多模态学习任务间的内在关联性, 将现有研究归纳为视听增强、跨模态交互和视听协作三大类. 其中, 视听增强被细分为音频增强的视觉任务和视觉增强的音频任务; 跨模态交互被细分为视听生成和视听检索; 视听协作被细分为视听实例感知、视听场景理解、视听推理与交互和非传统视听学习. 在此基础上, 本文对视听多模态学习中所涉及的 30 余个细分任务, 进行了系统的总结和分析, 广度与深度并重.

此外, 尽管视听多模态学习在应用场景、任务目标和技术方法上呈现出显著的多样性, 涵盖分类、分割、定位、生成和推理等多个任务, 本文对其背后所涉及的五个核心科学问题进行全面的归纳和总结, 即视听表征、对齐、转换、融合和共同学习. 同时, 本文也进一步剖析了所提出的三维分类框架与五大核心问题之间的映射关系, 并总结了大模型背景下视听多模态学习的发展现状. 从而构建了一个从“任务”到“方法”、从“核心问题”到“前沿动态”的完整闭环参考体系, 为国内研究者提供了可复制、可扩展的视听多模态学习研究路径.

视听多模态学习的未来发展, 将沿着“场景–技术–理论–伦理”四条主线纵深推进. 首先, 在场景层面, 视听多模态学习需突破实验室局限, 将技术嵌入医疗、教育、安防、自动驾驶等多元现实场景, 打造真正可落地的轻量级行业解决方案; 其次, 在技术层面, 需致力于提升生成内容的“物理一致性”与“交互可控性”, 追求亚帧级音唇同步、音乐节拍级动作编排等精细控制, 并支持用户通过自然语言实时引导生成过程; 再次, 在理论层面, 关键突破在于

融合认知神经科学与 LLM, 构建统一的时空–语义联合表征框架, 以攻克跨模态对齐、灾难性遗忘及零样本泛化等核心难题; 最后, 在伦理层面, 需构建兼顾隐私保护的无障碍系统与文化包容体系, 助力听障者“看见”声音、视障者“听见”环境, 同时借助联邦学习与可解释机制保障数据安全并促进社会公平.

参考文献

1 Treichler D G. Are you missing the boat in training aids? *Film and Audio-Visual Communication*, 1967, **1**(48): 14–30

2 Wen Xiao-Hui, Liu Qiang, Sun Hong-Jin, Zhang Qing-Lin, Yin Qin-Qing, Hao Ming-Jie, et al. Theoretical models of multisensory cues integration. *Advances in Psychological Science*, 2009, **17**(4): 659–666
(文小辉, 刘强, 孙弘进, 张庆林, 尹秦清, 郝明洁, 等. 多感官线索整合的理论模型. 心理科学进展, 2009, **17**(4): 659–666)

3 Calvert G A, Brammer M J, Bullmore E T, Campbell R, Iversen S D, David A S. Response amplification in sensory-specific cortices during crossmodal binding. *NeuroReport*, 1999, **10**(12): 2619–2623

4 Zhu H, Luo M D, Wang R, Zheng A H, He R. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 2021, **18**(3): 351–376

5 Wei Y K, Hu D, Tian Y P, Li X L. Learning in audio-visual context: A review, analysis, and new perspective. arXiv preprint arXiv: 2208.09579, 2022.

6 Bedny M. Evidence from blindness for a cognitively pluripotent cortex. *Trends in Cognitive Sciences*, 2017, **21**(9): 637–648

7 Calvert G A, Bullmore E T, Brammer M J, Campbell R, Williams S C R, Mcguire P K, et al. Activation of auditory cortex during silent lipreading. *Science*, 1997, **276**(5312): 593–596

8 Michael G A, Jacquot L, Millot J L, Brand G. Ambient odors modulate visual attentional capture. *Neuroscience Letters*, 2003, **352**(3): 221–225

9 Shannon C E. A mathematical theory of communication. *The Bell System Technical Journal*, 1948, **27**(3): 379–423

10 Sedaghati N, Ardebili S, Ghaffari A. Application of human activity/action recognition: A review. *Multimedia Tools and Applications*, 2025, **84**(28): 33475–33504

11 Liu Y, Tan Y, Lan H Y. Self-supervised contrastive learning for audio-visual action recognition. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). Kuala Lumpur, Malaysia: IEEE, 2023. 1000–1004

12 Shaikh M B, Chai D, Islam S M S, Akhtar N. MAiVAR-T: Multimodal audio-image and video action recognizer using Transformers. In: Proceedings of the 11th European Workshop on Visual Information Processing. Gjovik, Norway: IEEE, 2023. 1–6

13 Shahabinejad M, Kezele I, Nabavi S S, Liu W T, Patel S, Yu Y H, et al. Video action recognition with adaptive zooming using motion residuals. In: Proceedings of the International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE, 2023. 1206–1215

14 Shaikh M B, Chai D, Islam S M S, Akhtar N. Multimodal fusion for audio-image and video action recognition. *Neural Computing and Applications*, 2024, **36**(10): 5499–5513

15 Gao R H, Oh T H, Grauman K, Torresani L. Listen to look: Action recognition by previewing audio. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 10454–10464

16 Song L C, Bi J, Huang C, Xu C L. Audio-visual action prediction with soft-boundary in egocentric videos. In: Proceedings of the International Conference on Computer Vision Workshops. New York, USA: IEEE, 2024. 1–5

17 Chalk J, Huh J, Kazakos E, Zisserman A, Damen D. TIM: A time interval machine for audio-visual action recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 18153–18163

18 Wang K, Hatzinakos D. MOMA: Mixture-of-modality-adaptations for transferring knowledge from image models towards efficient audio-visual action recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 8055–8059

19 Han H C, Zheng Q H, Luo M N, Miao K Y, Tian F, Chen Y. Noise-tolerant learning for audio-visual action recognition. *IEEE Transactions on Multimedia*, 2024, **26**: 7761–7774

20 Wang W G, Shen J B, Guo F, Cheng M M, Borji A. Revisiting video saliency: A large-scale benchmark and a new model. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4894–4903

21 Luo Xiao-Xiao, Kang Guan-Lan, Zhou Xiao-Lin. The influential factors and neural mechanisms of McGurk effect. *Advances in Psychological Science*, 2018, **26**(11): 1935–1951
(罗霄骁, 康冠兰, 周晓林. McGurk 效应的影响因素与神经基础. 心理科学进展, 2018, **26**(11): 1935–1951)

22 Tavakoli H R, Borji A, Rahtu E, Kannala J. DAVE: A deep audio-visual embedding for dynamic saliency prediction. arXiv preprint arXiv: 1905.10693, 2019.

23 Min X K, Zhai G T, Gu K, Yang X K. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2017, **13**(1): Article No. 6

24 Jiang L, Xu M, Liu T, Qiao M L, Wang Z L. DeepVS: A deep learning based video saliency prediction approach. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 625–642

25 Jain S, Yarlagadda P, Jyoti S, Karthik S, Subramanian R, Gandhi V. ViNet: Pushing the limits of visual modality for audio-visual saliency prediction. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021. 3520–3527

26 Xiong J W, Wang G L, Zhang P, Huang W, Zha Y F, Zhai G T. CASP-Net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 6441–6450

27 Chang Q Y, Zhu S P. Human vision attention mechanism-inspired temporal-spatial feature pyramid for video saliency detection. *Cognitive Computation*, 2023, **15**(3): 856–868

28 Xie J W, Liu Z, Li G Y, Song Y J. Audio-visual saliency prediction with multisensory perception and integration. *Image and Vision Computing*, 2024, **143**: Article No. 104955

29 Chen Z, Zhang K, Cai H, Ding X Y, Jiang C X, Chen Z Z. Audio-visual saliency prediction for movie viewing in immersive environments: Dataset and benchmarks. *Journal of Visual Communication and Image Representation*, 2024, **100**: Article No. 104095

30 Zhu D D, Zhu K, Ding W P, Zhang N N, Min X K, Zhai G T, et al. MTCAM: A novel weakly-supervised audio-visual saliency prediction model with multi-modal Transformer. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024, **8**(2): 1756–1771

31 Qiao M L, Liu Y F, Xu M, Deng X, Li B, Hu W M, et al. Joint learning of audio-visual saliency prediction and sound source localization on multi-face videos. *International Journal of Computer Vision*, 2024, **132**(6): 2003–2025

32 Xiong J W, Zhang P, You T, Li C Y, Huang W, Zha Y F. DiffSal: Joint audio and video learning for diffusion saliency prediction. In: Proceedings of the Conference on Computer Vision

and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 27263−27273

33 Khan M A, Menouar H, Hamila R. Revisiting crowd counting: State-of-the-art, trends, and future perspectives. *Image and Vision Computing*, 2023, **129**: Article No. 104597

34 Hu D, Mou L C, Wang Q Z, Gao J Y, Hua Y S, Dou D J, et al. Ambient sound helps: Audiovisual crowd counting in extreme conditions. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops. New York, USA: IEEE, 2020. 1−4

35 Zou Y, Min W D, Zhao H Y, Han Q. A novel framework for crowd counting using video and audio. *Computers and Electrical Engineering*, 2023, **109**: Article No. 108754

36 Hu R H, Mo Q L, Xie Y F, Xu Y Q, Chen J Q, Yang Y L, et al. AVMSN: An audio-visual two stream crowd counting framework under low-quality conditions. *IEEE Access*, 2021, **9**: 80500−80510

37 Sajid U, Chen X Y, Sajid H, Kim T, Wang G H. Audio-visual Transformer based crowd counting. In: Proceedings of the International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE, 2021. 2249−2259

38 Hu D, Li X H, Mou L C, Jin P, Chen D, Jing L P, et al. Cross-task transfer for geotagged audiovisual aerial scene recognition. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 68−84

39 Heidler K, Mou L C, Hu D, Jin P, Li G Y, Gan C, et al. Self-supervised audiovisual representation learning for remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 2023, **116**: Article No. 103130

40 Sun X, Gao J Y, Yuan Y. Alignment and fusion using distinct sensor data for multimodal aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, **62**: Article No. 5626811

41 Han F Z, Yu T Y, Zhang L M, Si L Y, Zhang Y Q. SlotFusion: Object-centric audiovisual feature fusion with slot attention for remote sensing scene recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE, 2025. 1−5

42 Khanal S, Xing E, Sastry S, Dhakal A, Xiong Z X, Ahmad A, et al. PSM: Learning probabilistic embeddings for multi-scale zero-shot soundscape mapping. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 1361−1369

43 Corley I, Robinson C, Dodhia R, Ferres J M L, Najafirad P. Revisiting pre-trained remote sensing model benchmarks: Resizing and normalization matters. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2024. 3162−3172

44 Liu X L, Yu Y, Li X L, Zhao Y. MCL: Multimodal contrastive learning for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, **34**(4): 2803−2813

45 Rana S, Nobi M N, Murali B, Sung A H. Deepfake detection: A systematic literature review. *IEEE Access*, 2022, **10**: 25494−25513

46 Hashmi A, Shahzad S A, Lin C W, Tsao Y, Wang H M. AVTENet: Audio-visual Transformer-based ensemble network exploiting multiple experts for video deepfake detection. arXiv preprint arXiv: 2310.13103, 2023.

47 Zhang Y B, Lin W G, Xu J F. Joint audio-visual attention with contrastive learning for more general deepfake detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2024, **20**(5): Article No. 137

48 Wang R, Ye D P, Tang L, Zhang Y M, Deng J C. AVT.2-DWF: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies. *IEEE Signal Processing Letters*, 2024, **31**: 1960−1964

49 Wang Y F, Wu X C, Zhang J, Jing M H, Lu K D, Yu J, et al. Building robust video-level deepfake detection via audio-visual local-global interactions. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 11370−11376

50 Liu W F, She T Y, Liu J W, Li B H, Yao D Y, Liang Z Y, et al. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. 91131−91155

51 Astrid M, Ghorbel E, Aouada D. Audio-visual deepfake detection with local temporal inconsistencies. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE, 2025. 1−5

52 Koutlis C, Papadopoulos S. DiMoDif: Discourse modality-information differentiation for audio-visual deepfake detection and localization. arXiv preprint arXiv: 2411.10193, 2024.

53 Shahzad S A, Hashmi A, Peng Y T, Tsao Y, Wang H M. AV-Lip-Sync+: Leveraging AV-HuBERT to exploit multimodal inconsistency for deepfake detection of frontal face videos. *IEEE Transactions on Human-Machine Systems*, 2025, **55**(6): 973−982

54 Chugh K, Gupta P, Dhall A, Subramanian R. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM, 2020. 439−447

55 Zou H Q, Shen M, Hu Y C, Chen C, Chng E S, Rajan D. Cross-modality and within-modality regularization for audio-visual deepfake detection. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 4900−4904

56 Bekheet A A, Ghoneim A, Khoriba G. A comprehensive comparative analysis of deepfake detection techniques in visual, audio, and audio-visual domains. In: Proceedings of the Intelligent Methods, Systems, and Applications (IMSA). Giza, Egypt: IEEE, 2024. 122−129

57 Nie F, Ni J Q, Zhang J, Zhang B, Zhang W Z. FRADE: Forgery-aware audio-distilled multimodal learning for deepfake detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 6297−6306

58 Zhao H Q, Zhou W B, Chen D D, Zhang W M, Guo Y, Cheng Z, et al. Audio-visual contrastive pre-train for face forgery detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025, **21**(2): Article No. 45

59 Liang Y C, Yu M, Li G, Jiang J G, Li B Q, Yu F, et al. SpeechForensics: Audio-visual speech representation learning for face forgery detection. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 2735

60 Oorloff T, Koppisetti S, Bonettini N, Solanki D, Colman B, Yacoob Y, et al. AVFF: Audio-visual feature fusion for video deepfake detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 27092−27102

61 Li X L, Liu Z H, Chen C, Li L T, Guo L, Wang D. Zero-shot fake video detection by audio-visual consistency. In: Proceedings of the 25th Annual Conference of the International Speech Communication Association. Kos, Greece: ISCA, 2024.

62 Muppalla S, Jia S, Lyu S W. Integrating audio-visual features for multimodal deepfake detection. In: Proceedings of the MIT Undergraduate Research Technology Conference (URTC). Cambridge, USA: IEEE, 2023. 1−5

63 Yang W Y, Zhou X Y, Chen Z K, Guo B F, Ba Z J, Xia Z H, et al. AVoiD-DF: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 2023, **18**: 2015−2029

64 Yu C, Chen P, Tian J H, Liu J, Dai J, Wang X, et al. Modal-

ity-agnostic audio-visual deepfake detection. arXiv preprint arXiv: 2307.14491, 2023.

65 Li Jin-Xin, Huang Zhi-Yong, Li Wen-Bin, Zhou Deng-Wen. Image super-resolution based on multi-hierarchical features fusion network. *Acta Automatica Sinica*, 2023, **49**(1): 161−171 (李金新, 黄志勇, 李文斌, 周登文. 基于多层次特征融合的图像超分辨率重建. 自动化学报, 2023, **49**(1): 161−171)

66 Sanguineti V, Thakur S, Morerio P, del Bue A, Murino A. Audio-visual inpainting: Reconstructing missing visual information with sound. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

67 Lu Y F, Wang Z P, Liu M J, Wang H J, Wang L. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 1557−1567

68 Chen Y X, Zhao P C, Qi M B, Zhao Y, Jia W, Wang R G. Audio matters in video super-resolution by implicit semantic guidance. *IEEE Transactions on Multimedia*, 2022, **24**: 4128−4142

69 Xiao J, Jiang X Y, Zheng N X, Yang H, Yang Y F, Yang Y Q, et al. Online video super-resolution with convolutional kernel bypass grafts. *IEEE Transactions on Multimedia*, 2023, **25**: 8972−8987

70 Chakraborty C, Talukdar P H. Issues and limitations of HMM in speech processing: A survey. *International Journal of Computer Applications*, 2016, **141**(7): 13−17

71 Fang H J, Frintrop S, Gerkmann T. Uncertainty-driven hybrid fusion for audio-visual phoneme recognition. In: Proceedings of the 15th ITG Conference on Speech Communication. Aachen, Germany: VDE, 2023. 255−259

72 Richter J, Liebold J, Gerkmann T. Continuous phoneme recognition based on audio-visual modality fusion. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Padua, Italy: IEEE, 2022. 1−8

73 Biswas A, Sahu P K, Bhowmick A, Chandra M. VidTIMIT audio visual phoneme recognition using AAM visual features and human auditory motivated acoustic wavelet features. In: Proceedings of the 2nd International Conference on Recent Trends in Information Systems (ReTIS). Kolkata, India: IEEE, 2015. 428−433

74 Hu Y C, Li R Z, Chen C, Qin C W, Zhu Q S, Chng E S. Hearing lips in noise: Universal viseme-phoneme mapping and transfer for robust audio-visual speech recognition. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023. 15213−15232

75 Kim M, Yeo J, Park S J, Rha H, Ro Y M. Efficient training for multilingual visual speech recognition: Pre-training with discretized visual speech representation. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 1311−1320

76 Pai L T, Wang Y C, Yan B C, Wang H W, Lu J L, Lin C H, et al. An effective contextualized automatic speech recognition approach leveraging self-supervised phoneme features. In: Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Macao, China: IEEE, 2024. 1−6

77 Vincent E, Virtanen T, Gannot S. *Audio Source Separation and Speech Enhancement*. Hoboken: John Wiley & Sons, 2018. 110−234

78 Li G N, Deng J J, Geng M Z, Jin Z R, Wang T Z, Hu S J, et al. Audio-visual end-to-end multi-channel speech separation, dereverberation and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, **31**: 2707−2723

79 Tan R B, Ray A, Burns A, Plummer B A, Salamon J, Nieto O, et al. Language-guided audio-visual source separation via trimodal consistency. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 10575−10584

80 Gao R H, Grauman K. VisualVoice: Audio-visual speech separation with cross-modal consistency. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 15490−15500

81 Yoshinaga T, Tanaka K, Morishima S. Audio-visual speech enhancement with selective off-screen speech extraction. In: Proceedings of the 31st European Signal Processing Conference (EUSIPCO). Helsinki, Finland: IEEE, 2023. 595−599

82 Pan Z X, Wichern G, Masuyama Y, Germain F G, Khurana S, Hori C, et al. Scenario-aware audio-visual TF-Gridnet for target speech extraction. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU). Taipei, China: IEEE, 2023. 1−8

83 Chen J B, Zhang R R, Lian D Z, Yang J Q, Zeng Z Y, Shi J B. iQuery: Instruments as queries for audio-visual sound separation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 14675−14686

84 Chatterjee M, le Roux J, Ahuja N, Cherian A. Visual scene graphs for audio source separation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 1184−1193

85 Ye Y X, Yang W M, Tian Y P. LAVSS: Location-guided audio-visual spatial audio separation. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 5496−5507

86 Pan T R, Liu J, Wang B H, Tang J, Wu G S. RAVSS: Robust audio-visual speech separation in multi-speaker scenarios with missing visual cues. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 4748−4756

87 Li K, Xie F H, Chen H, Yuan K X, Hu X L. An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(10): 6637−6651

88 Liu Y G, Deng Y J, Wei Y. A two-stage audio-visual speech separation method without visual signals for testing and tuples loss with dynamic margin. *IEEE Journal of Selected Topics in Signal Processing*, 2024, **18**(3): 459−472

89 Pian W G, Nan Y Y, Deng S J, Mo S T, Guo Y H, Tian Y P. Continual audio-visual sound separation. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 2421

90 Kalkhorani V A, Kumar A, Tan K, Xu B Y, Wang D L. Audiovisual speaker separation with full-and sub-band modeling in the time-frequency domain. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 12001−12005

91 Fan C H, Xiang W, Tao J H, Yi J Y, Lv Z. Cross-modal knowledge distillation with multi-stage adaptive feature fusion for speech separation. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, **33**: 935−948

92 Boll S. Suppression of acoustic noise in speech using spectral subtraction. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Piscataway, USA: IEEE, 2020. 7314−7318

93 Isik Y, le Roux J, Chen Z, Watanabe S, Hershey J R. Single-channel multi-speaker separation using deep clustering. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco, USA: ISCA, 2016. 545−549

94 Zhu Z R, Yang H M, Tang M, Yang Z Y, Eskimez S E, Wang H M. Real-time audio-visual end-to-end speech enhancement. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

95 Balasubramanian S, Rajavel R, Kar A. Ideal ratio mask estimation based on cochleagram for audio-visual monaural speech

enhancement. *Applied Acoustics*, 2023, **211**: Article No. 109524

96 Li Y K, Zhang X M. Lip landmark-based audio-visual speech enhancement with multimodal feature fusion network. *Neurocomputing*, 2023, **549**: Article No. 126432

97 Hussain T, Dashtipour K, Tsao Y, Hussain A. Audio-visual speech enhancement in noisy environments via emotion-based contextual cues. arXiv preprint arXiv: 2402.16394, 2024.

98 Zheng R C, Ai Y, Ling Z H. Incorporating ultrasound tongue images for audio-visual speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, **32**: 1430−1444

99 Gogate M, Dashtipour K, Hussain A. Robust real-time audio-visual speech enhancement based on DNN and GAN. *IEEE Transactions on Artificial Intelligence*, 2025, **6**(11): 2860−2869

100 Chen H L, Mira R, Petridis S, Pantic M. RT-LA-VocE: Real-time low-SNR audio-visual speech enhancement. arXiv preprint arXiv: 2407.07825, 2024.

101 Jung C, Lee S, Kim J H, Chung J S. FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching. arXiv preprint arXiv: 2406.09286, 2024.

102 Passos L A, Papa J P, del Ser J, Hussain A, Adeel A. Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement. *Information Fusion*, 2023, **90**: 1−11

103 Morrone G, Michelsanti D, Tan Z H, Jensen J. Audio-visual speech inpainting with deep learning. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE, 2021. 6653−6657

104 Chen H, Wang Q, Du J, Yin B C, Pan J, Lee C H. Optimizing audio-visual speech enhancement using multi-level distortion measures for audio-visual speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, **32**: 2508−2521

105 Chen S, Kirton-Wingate J, Doctor F, Arshad U, Dashtipour K, Gogate M, et al. Context-aware audio-visual speech enhancement based on neuro-fuzzy modeling and user preference learning. *IEEE Transactions on Fuzzy Systems*, 2024, **32**(10): 5400−5412

106 Ahlawat H, Aggarwal N, Gupta D. Automatic speech recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 2025, **6**: 201−237

107 Hong J, Kim M, Choi J, Ro Y M. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 18783−18794

108 Wang X M, Mi J C, Li B Q, Zhao Y X, Meng J X. CATNet: Cross-modal fusion for audio-visual speech recognition. *Pattern Recognition Letters*, 2024, **178**: 216−222

109 Wang H, Guo P C, Zhou P, Xie L. MLCA-AVSR: Multi-layer cross attention fusion based audio-visual speech recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 8150−8154

110 Wang J D, Qian X Y, Li H Z. Predict-and-update network: Audio-visual speech recognition inspired by human speech perception. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, **33**: 11−22

111 Ma P C, Haliassos A, Fernandez-Lopez A, Chen H L, Petridis S, Pantic M. Auto-AVSR: Audio-visual speech recognition with automatic labels. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

112 Yeo J H, Kim M, Choi J, Kim D H, Ro Y M. AKVSR: Audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model. *IEEE Transactions on Multimedia*, 2024, **26**: 6462−6474

113 Lian J C, Baevski A, Hsu W N, Auli M. Av-Data2Vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU). Taipei, China: IEEE, 2023. 1−8

114 Dai Y S, Chen H, Du J, Wang R Y, Chen S H, Wang H T, et al. A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 27435−27445

115 Rouditchenko A, Gong Y, Thomas S, Karlinsky L, Kuehne H, Feris R, et al. Whisper-Flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. In: Proceedings of the 25th Annual Conference of the International Speech Communication Association. Kos, Greece: ISCA, 2024. 2420−2424

116 Wang J D, Pan Z X, Zhang M L, Tan R T, Li H Z. Restoring speaking lips from occlusion for audio-visual speech recognition. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 19144−19152

117 Li J H, Li C D, Wu Y F, Qian Y M. Unified cross-modal attention: Robust audio-visual speech recognition and beyond. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, **32**: 1941−1953

118 Fu D J, Cheng X Z, Yang X D, Wang H T, Zhao Z, Jin T. Boosting speech recognition robustness to modality-distortion with contrast-augmented prompts. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 3838−3847

119 Burchi M, Puvvada K C, Balam J, Ginsburg B, Timofte R. Multilingual audio-visual speech recognition with hybrid CTC/RNN-T fast conformer. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 10211−10215

120 Kabir M M, Mridha M F, Shin J, Jahan I, Ohi A Q. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *IEEE Access*, 2021, **9**: 79236−79263

121 Chelali F Z. Bimodal fusion of visual and speech data for audiovisual speaker recognition in noisy environment. *International Journal of Information Technology*, 2023, **15**(6): 3135−3145

122 Tang X O, Li Z F. Audio-guided video-based face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2009, **19**(7): 955−964

123 Gong D H, Li N, Li Z F, Qiao Y. Multi-feature subspace analysis for audio-vidoe based multi-modal person recognition. In: Proceedings of the 4th IEEE International Conference on Information Science and Technology. Shenzhen, China: IEEE, 2014. 776−779

124 Tao R J, Lee K A, Shi Z, Li H Z. Speaker recognition with two-step multi-modal deep cleansing. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

125 Gebru I D, Ba S, Li X F, Horaud R. Audio-visual speaker diarization based on spatiotemporal Bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(5): 1086−1099

126 Lin Y K, Cheng M, Zhang F L, Gao Y Y, Zhang S L, Li M. VoxBlink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark. arXiv preprint arXiv: 2407.11510, 2024.

127 Clarke J, Gotoh Y, Goetze S. Speaker embedding informed audiovisual active speaker detection for egocentric recordings. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE, 2025. 1−5

128 Tao R J, Qian X Y, Jiang Y D, Li J J, Wang J D, Li H Z. Audio-visual target speaker extraction with selective auditory attention. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, **33**: 797−811

129 Li Han-Chao, Shen Cheng-Ze, Liu Xin-Guo. Facial blendshape generation method with virtual edge constraints. *Chinese Journal of Computers*, 2023, **46**(11): 2453−2462
(李韩超, 沈成泽, 刘新国. 带虚拟边约束的面部表情基生成方法. 计算机学报, 2023, **46**(11): 2453−2462)

130 Agarwal M, Mukhopadhyay R, Namboodiri V P, Jawahar C V. Audio-visual face reenactment. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 5167−5176

131 Wang S Z, Li L C, Ding Y, Yu X. One-shot talking face generation from single-speaker audio-visual correlation learning. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 2531−2539

132 Jang Y, Rho K, Woo J, Lee H, Park J, Lim Y, et al. That′s what I said: Fully-controllable talking face generation. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 3827−3836

133 Park S J, Kim M, Hong J, Choi J, Ro Y M. SyncTalkFace: Talking face generation with precise lip-syncing via audio-lip memory. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 2062−2070

134 Wang G L, Zhang P, Xie L, Huang W, Zha Y F. Attention-based lip audio-visual synthesis for talking face generation in the wild. arXiv preprint arXiv: 2203.03984, 2022.

135 Zhang J Y, Liu Y Z, Li X, Li W, Tang Y. Talking face generation driven by time-frequency domain features of speech audio. *Displays*, 2023, **80**: Article No. 102558

136 Liu S G, Wang H X. Talking face generation via facial anatomy. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023, **19**(3): Article No. 125

137 Yaman D, Eyiokur F I, Bärmann L, Aktı S, Ekenel H K, Waibel A. Audio-visual speech representation expert for enhanced talking face video generation and evaluation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2024. 6003−6013.

138 Jang Y, Kim J H, Ahn J, Kwak D, Yang H S, Ju Y C, et al. Faces that speak: Jointly synthesising talking face and speech from text. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 8818−8828

139 Xu S C, Chen G J, Guo Y X, Yang J L, Li C, Zang Z Y, et al. VASA-1: Lifelike audio-driven talking faces generated in real time. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 21

140 Zhang Z J, Zhang J, Mai W J. VPT: Video portraits Transformer for realistic talking face generation. *Neural Networks*, 2025, **184**: Article No. 107122

141 Nyatsanga S, Kucherenko T, Ahuja C, Henter G E, Neff M. A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum*, 2023, **42**(2): 569−596

142 Cassell J, Vilhjálmsson H H, Bickmore T. BEAT: The behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: ACM, 2001. 477−486

143 Neff M, Kipp M, Albrecht I, Seidel H P. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 2008, **27**(1): Article No. 5

144 Ferstl Y, McDonnell R. Investigating the use of recurrent motion modelling for speech gesture generation. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. Sydney, Australia: ACM, 2018. 93−98

145 Liang Y Z, Feng Q Y, Zhu L C, Hu L, Pan P, Yang Y. SEEG: Semantic energized co-speech gesture generation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 10463−10472

146 Hasegawa D, Kaneko N, Shirakawa S, Sakuta H, Sumi K. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. Sydney, Australia: ACM, 2018. 79−86

147 Yoon Y, Cha B, Lee J H, Jang M, Lee J, Kim J, et al. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 2020, **39**(6): Article No. 222

148 Qi X Q, Liu C, Li L C, Hou J, Xin H R, Yu X. EmotionGesture: Audio-driven diverse emotional co-speech 3D gesture generation. *IEEE Transactions on Multimedia*, 2024, **26**: 10420−10430

149 Liu P X, Zhang P F, Kim H, Garrido P, Shapiro A, Olszewski K. Contextual gesture: Co-speech gesture video generation through context-aware gesture representation. In: Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM, 2025. 9803−9812

150 Gao Z L, Li Y H, Wu S J, Cao Y Q, Duan H Y, Zhai G T. GES-QA: A multidimensional quality assessment dataset for audio-to-3D gesture generation. In: Proceedings of the International Conference on Visual Communications and Image Processing (VCIP). Klagenfurt, Austria: IEEE, 2025. 1−5

151 Liu H Y, Zhu Z H, Becherini G, Peng Y C, Su M Y, Zhou Y, et al. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 1144−1154

152 Chen J H, Yang H, Shi R H, Ding C F, Mo X Q, Xiong S Y, et al. Audio-driven gesture generation via deviation feature in the latent space. In: Proceedings of the IEEE International Conference on Multimedia and Expo. Nantes, France: IEEE, 2025. 1−6

153 Lee M, Lee K, Park J. Music similarity-based approach to generating dance motion sequence. *Multimedia Tools and Applications*, 2013, **62**(3): 895−912

154 Huang Y H, Zhang J J, Liu S Y, Bao Q, Zeng D, Chen Z N, et al. Genre-conditioned long-term 3D dance generation driven by music. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 4858−4862

155 Wang T, Li L J, Lin K, Zhai Y H, Lin C C, Yang Z Y, et al. Disco: Disentangled control for realistic human dance generation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 9326−9336

156 Tseng J, Castellon R, Liu C K. EDGE: Editable dance generation from music. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 448−458

157 Yang Z P, Wen Y H, Chen S Y, Liu X, Gao Y, Liu Y J, et al. Keyframe control of music-driven 3D dance generation. *IEEE Transactions on Visualization and Computer Graphics*, 2024, **30**(7): 3474−3486

158 Li S Y, Yu W J, Gu T P, Lin C Z, Wang Q, Qian C, et al. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(12): 14192−14207

159 Yin W J, Yin H, Baraka K, Kragic D, Björkman M. Dance style transfer with cross-modal Transformer. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 5047−5056

160 Habibie I, Xu W P, Mehta D, Liu L J, Seidel H P, Pons-Moll G, et al. Learning speech-driven 3D conversational gestures from video. In: Proceedings of the 21st ACM International

Conference on Intelligent Virtual Agents. Virtual Event: ACM, 2021. 101–108

161 Yi H W, Liang H L, Liu Y F, Cao Q, Wen Y D, Bolkart T, et al. Generating holistic 3D human motion from speech. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 469–480

162 Zhu H, Li Y, Zhu F X, Zheng A H, He R. Let′s play music: Audio-driven performance video generation. In: Proceedings of the International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021. 3574–3581

163 Xu S Y, Dou Z Y, Shi M Y, Pan L, Ho L, Wang J B, et al. MOSPA: Human motion generation driven by spatial audio. In: Proceedings of the 39th Annual Conference on Neural Information Processing Systems. San Diego, USA: OpenReview.net, 2025.

164 Zhang Z Y, Wang Y R, Mao W, Li D N, Zhao R, Wu B, et al. Motion anything: Any to motion generation. arXiv preprint arXiv: 2503.06955, 2025.

165 Zhang M Y, Jin D S, Gu C Y, Hong F Z, Cai Z G, Huang J F, et al. Large motion model for unified multi-modal motion generation. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2024. 397–421

166 Shim J Y, Kim J, Kim J K. S2I-Bird: Sound-to-image generation of bird species using generative adversarial networks. In: Proceedings of the International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021. 2226–2232

167 Song C J, Zhang Y C, Peng W, Mohaghegh P, Wandt B, Rhodin H. AudioViewer: Learning to visualize sounds. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 2205–2215

168 Sung-Bin K, Senocak A, Ha H, Owens A, Oh T H. Sound to visual scene generation by audio-to-visual latent alignment. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 6430–6440

169 Lee S H, Roh W, Byeon W, Yoon S H, Kim C, Kim J, et al. Sound-guided semantic image manipulation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 3367–3376

170 Zhuang Y G, Kang Y H, Fei T, Bian M, Du Y Y. From hearing to seeing: Linking auditory and visual place perceptions with soundscape-to-image generative artificial intelligence. *Computers, Environment and Urban Systems*, 2024, **110**: Article No. 102122

171 Ephrat A, Peleg S. Vid2speech: Speech reconstruction from silent video. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017. 5095–5099

172 Kumar Y, Aggarwal M, Nawal P, Satoh S, Shah R R, Zimmermann R. Harnessing AI for speech reconstruction using multi-view silent video feed. In: Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Republic of Korea: ACM, 2018. 1976–1983

173 Dong Z P, Xu Y, Abel A, Wang D. Lip2Speech: Lightweight multi-speaker speech reconstruction with Gabor features. *Applied Sciences*, 2024, **14**(2): Article No. 798

174 Kefalas T, Panagakis Y, Pantic M. Audio-visual video-to-speech synthesis with synthesized input audio. arXiv preprint arXiv: 2307.16584, 2023.

175 Hong J, Kim M, Ro Y M. VisageSynTalk: Unseen speaker video-to-speech synthesis via speech-visage feature selection. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 452–468

176 Kameoka H, Tanaka K, Puche A V, Ohishi Y, Kaneko T. Crossmodal voice conversion. arXiv preprint arXiv: 1904.04540, 2019.

177 Weng S E, Shuai H H, Cheng W H. Zero-shot face-based voice conversion: Bottleneck-free speech disentanglement in the real-world scenario. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 13718–13726

178 Wang D S, Yang S, Su D, Liu X Y, Yu D, Meng H. VCVTS: Multi-speaker video-to-speech synthesis via cross-modal knowledge transfer from voice conversion. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 7252–7256

179 Lu H H, Weng S E, Yen Y F, Shuai H H, Cheng W H. Face-based voice conversion: Learning the voice behind a face. In: Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event: ACM, 2021. 496–505

180 Mira R, Vougioukas K, Ma P C, Petridis S, Schuller B W, Pantic M. End-to-end video-to-speech synthesis using generative adversarial networks. *IEEE Transactions on Cybernetics*, 2023, **53**(6): 3454–3466

181 Chen X Y, Wang Y J, Wu X X, Wang D S, Wu Z Y, Liu X Y, et al. Exploiting audio-visual features with pretrained AV-HuBERT for multi-modal dysarthric speech reconstruction. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 12341–12345

182 Pham L K, Tran T V T, Pham M T, Nguyen V. RESOUND: Speech reconstruction from silent videos via acoustic-semantic decomposed modeling. arXiv preprint arXiv: 2505.22024, 2025.

183 Rong Y, Liu L. Seeing your speech style: A novel zero-shot identity-disentanglement face-based voice conversion. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI Press, 2025. 25092–25100

184 Zhu Y, Olszewski K, Wu Y, Achlioptas P, Chai M L, Yan Y, et al. Quantized GAN for complex music generation from dance videos. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 182–199

185 Su K, Li J Y, Huang Q Q, Kuzmin D, Lee J, Donahue C, et al. V2Meow: Meowing to the visual beat via video-to-music generation. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 4952–4960

186 Liu X H, Tu T, Ma Y S, Chua T S. Extending visual dynamics for video-to-music generation. arXiv preprint arXiv: 2504.07594, 2025.

187 Lin Y B, Tian Y, Yang L J, Bertasius G, Wang H. VMAs: Video-to-music generation via semantic alignment in web music videos. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Tucson, USA: IEEE, 2025. 1155–1165

188 Wang B S, Zhuo L, Wang Z K, Bao C X, Wu C J, Nie X C, et al. Multimodal music generation with explicit bridges and retrieval augmentation. arXiv preprint arXiv: 2412.09428. 2024.

189 Tian Z Y, Liu Z Y, Yuan R B, Pan J H, Liu Q F, Tan X, et al. VidMuse: A simple video-to-music generation framework with long-short-term modeling. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 18782–18793

190 Owens A, Isola P, McDermott J, Torralba A, Adelson E H, Freeman W T. Visually indicated sounds. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 2405–2413

191 Zhou Y P, Wang Z W, Fang C, Bui T, Berg T L. Visual to sound: Generating natural sound for videos in the wild. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3550–3558

192 Du Y X, Chen Z Y, Salamon J, Russell B, Owens A. Conditional generation of audio from video via Foley analogies. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 2426–2436

193 Iashin V, Rahtu E. Taming visually guided sound generation.

In: Proceedings of the British Machine Vision Conference. London, UK: BMVA, 2021. 1−15

194 Sheffer R, Adi Y. I hear your true colors: Image guided audio generation. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

195 Chen P H, Zhang Y, Tan M K, Xiao H D, Huang D, Gan C. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 2020, **29**: 8292−8302

196 Liu H D, Luo K C, Wang J L, Wang W, Chen Q, Zhao Z, et al. ThinkSound: Chain-of-thought reasoning in multimodal LLMs for audio generation and editing. In: Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS). San Diego, USA: OpenReview.net, 2025.

197 Jeong Y, Kim Y, Chun S, Lee J. Read, watch and scream! Sound generation from text and video. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 17590−17598

198 Xie Z F, Yu S Y, He Q L, Li M T. SonicVisionLM: Playing sound with vision language models. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 26856−26865

199 Chen Z Y, Geng D, Owens A. Images that sound: Composing images and sounds on a single canvas. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 2700

200 Hawley M L, Litovsky R Y, Culling J F. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 2004, **115**(2): 833−843

201 Wang Rui-Qi, Cheng Hao-Nan, Ye Long. Visually guided binaural audio generation method based on hierarchical feature encoding and decoding. *Journal of Software*, 2024, **35**(5): 2165−2175
(王睿琦, 程皓楠, 叶龙. 分层特征编解码驱动的视觉引导立体声生成方法. 软件学报, 2024, **35**(5): 2165−2175)

202 Cheng C I, Wakefield G H. Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. *Journal of the Audio Engineering Society*, 2001, **49**(4): 5026−5035

203 Lin Y Q, Lee D D. Bayesian regularization and nonnegative deconvolution for room impulse response estimation. *IEEE Transactions on Signal Processing*, 2006, **54**(3): 839−847

204 Morgado P, Vasconcelos N, Langlois T, Wang O. Self-supervised generation of spatial audio for 360° video. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc., 2018. 360−370

205 Parida K K, Srivastava S, Sharma G. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2022. 2151−2160

206 Li Z J, Zhao B, Yuan Y. Cross-modal generative model for visual-guided binaural stereo generation. *Knowledge-Based Systems*, 2024, **296**: Article No. 111814

207 Chen M F, Shlizerman E. AV-Cloud: Spatial audio rendering through audio-visual cloud splatting. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 4477

208 Xie S Y, Zhu H X, Chen X Y, He T Y, Li X, Chen Z B. Sonic4D: Spatial audio generation for immersive 4D scene exploration. arXiv preprint arXiv: 2506.15759, 2025.

209 Marinoni C, Gramaccioni R F, Shimada K, Shibuya T, Mitsufuji Y, Comminiello D. StereoSync: Spatially-aware stereo audio generation from video. In: Proceedings of the International Joint Conference on Neural Networks. Rome, Italy: IEEE, 2025. 1−8

210 Li X D, Zhuo F, Luo D, Chen J, Kang S Y, Wu Z Y, et al. Generating stereophonic music with single-stage language models. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 1471−1475

211 Kim J, Yun H, Kim G. ViSAGe: Video-to-spatial audio generation. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025.

212 Nagrani A, Albanie S, Zisserman A. Seeing voices and hearing faces: Cross-modal biometric matching. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 8427−8436

213 Fang Z, Liu Z, Hung C C, Sekhavat Y A, Liu T T, Wang X. Learning coordinated emotion representation between voice and face. *Applied Intelligence*, 2023, **53**(11): 14470−14492

214 Zeng D H, Yu Y, Oyama K. Audio-visual embedding for cross-modal music video retrieval through supervised deep CCA. In: Proceedings of the International Symposium on Multimedia (ISM). Taichung, China: IEEE, 2018. 143−150

215 Guo M, Zhou C H, Liu J H. Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, **12**(11): 4644−4654

216 Chen G Y, Zhang D Y, Liu T, Du X Y. Self-lifting: A novel framework for unsupervised voice-face association learning. In: Proceedings of the International Conference on Multimedia Retrieval. Newark, USA: ACM, 2022. 527−535

217 Oncescu A M, Henriques J F, Zisserman A, Albanie S, Koepke A S. A sound approach: Using large language models to generate audio descriptions for egocentric text-audio retrieval. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024. 7300−7304

218 Lin J A, Liu D Z, Chen X K, Qu X Y, Yang X, Zhu J X, et al. Audio does matter: Importance-aware multi-granularity fusion for video moment retrieval. In: Proceedings of the 33rd ACM International Conference on Multimedia. Dublin Ireland: ACM, 2025. 6027−6036

219 Li X L, Hu D, Lu X Q. Image2song: Song retrieval via bridging image content and lyric words. In: Proceedings of the International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 5650−5650

220 Hong S, Im W, Yang H S. CBVMR: Content-based video-music retrieval using soft intra-modal structure constraint. In: Proceedings of the ACM International Conference on Multimedia Retrieval. Yokohama, Japan: ACM, 2018. 353−361

221 Nakatsuka T, Hamasaki M, Goto M. Content-based music-image retrieval using self-and cross-modal feature embedding memory. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 2173−2183

222 Cheng X X, Zhu Z H, Li H X, Li Y W, Zou Y X. SSVMR: Saliency-based self-training for video-music retrieval. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

223 Era Y, Togo R, Maeda K, Ogawa T, Haseyama M. Video-music retrieval with fine-grained cross-modal alignment. In: Proceedings of the International Conference on Image Processing (ICIP). Kuala Lumpur, Malaysia: IEEE, 2023. 2005−2009

224 McKee D, Salamon J, Sivic J, Russell B. Language-guided music recommendation for video via prompt analogies. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 14784−14793

225 Chen Y X, Du C, Zi Y F, Xiong S W, Lu X Q. Scale-aware adaptive refinement and cross-interaction for remote sensing audio-visual cross-modal retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, **62**: Article No. 4706914

226 Huang J H, Chen Y X, Xiong S W, Lu X Q. Cross-modal remote sensing image-audio retrieval with adaptive learning for aligning correlation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, **62**: Article No. 4705213

227 Wu P, Su W S, He X T, Wang P, Zhang Y N. VarCMP: Adapting cross-modal pre-training models for video anomaly retrieval. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 8423−8431

228 Tian Y P, Shi J, Li B C, Duan Z Y, Xu C L. Audio-visual event localization in unconstrained videos. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 252−268

229 Xuan H Y, Zhang Z Y, Chen S, Yang J, Yan Y. Cross-modal attention network for temporal inconsistent audio-visual event localization. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 279−286

230 Mahmud T, Marculescu D. AVE-CLIP: AudioCLIP-based multi-window temporal Transformer for audio visual event localization. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 5147−5156

231 Bao P J, Yang W H, Ng B P, Er M H, Kot A C. Cross-modal label contrastive learning for unsupervised audio-visual event localization. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 215−222

232 Zhou Z H, Zhou J X, Qian W, Tang S G, Chang X J, Guo D. Dense audio-visual event localization under cross-modal consistency and multi-temporal granularity collaboration. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 10905−10913

233 Sun C, Chen M, Zhu C B, Zhang S, Lu P, Chen J C. Listen with seeing: Cross-modal contrastive learning for audio-visual event localization. *IEEE Transactions on Multimedia*, 2025, **27**: 2650−2665

234 Liu L, Li S Y, Zhu Y Q. Audio-visual semantic graph network for audio-visual event localization. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 23957−23966

235 Zhang P F, Wang J X, Wan M, Chang S J, Ding L H, Shi P. Multi-relation learning network for audio-visual event localization. *Knowledge-Based Systems*, 2025, **310**: Article No. 112925

236 Lin Y B, Tseng H Y, Lee H Y, Lin Y Y, Yang M H. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 875

237 Fan Y Y, Wu Y, Du B, Lin Y T. Revisit weakly-supervised audio-visual video parsing from the language perspective. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1767

238 Fu J, Gao J Y, Bao B K, Xu C S. Multimodal imbalance-aware gradient modulation for weakly-supervised audio-visual video parsing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, **34**(6): 4843−4856

239 Geng T T, Wang T, Duan J M, Cong R M, Zheng F. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 22942−22951

240 Yu J S, Cheng Y, Zhao R W, Feng R, Zhang Y J. MM-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In: Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM, 2022. 6241−6249

241 Lai Y H, Ebbers J, Wang Y C F, Germain F O, Jones M J, Chatterjee M. UWAV: Uncertainty-weighted weakly-supervised audio-visual video parsing. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 13561−13570

242 Gao Y B, Sun X C, Lv G H, Yu D, Niu S J. Reinforced label denoising for weakly-supervised audio-visual video parsing. In: Proceedings of the 13th International Conference on Computational Visual Media. Hong Kong, China: Springer, 2025. 107−124

243 Gao J Y, Chen M Y, Xu C S. Learning probabilistic presence-absence evidence for weakly-supervised audio-visual event perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(6): 4787−4802

244 Zhao P C, Zhou J X, Zhao Y, Guo D, Chen Y X. Multimodal class-aware semantic enhancement network for audio-visual video parsing. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 10448−10456

245 Senocak A, Oh T H, Kim J, Yang M H, Kweon I S. Learning to localize sound source in visual scenes. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4358−4366

246 Xuan H Y, Wu Z L, Yang J, Yan Y, Alameda-Pineda X. A proposal-based paradigm for self-supervised sound source localization in videos. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1019−1028

247 Liu J X, Ju C, Xie W D, Zhang Y. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In: Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM, 2022. 3742−3753

248 Qian R, Hu D, Dinkel H, Wu M Y, Xu N, Lin W Y. Multiple sound sources localization from coarse to fine. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 292−308

249 Mo S T, Tian Y P. Audio-visual grouping network for sound localization from mixtures. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 10565−10574

250 Ryu H, Kim S, Chung J S, Senocak A. Seeing speech and sound: Distinguishing and locating audio sources in visual scenes. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 13540−13549

251 Senocak A, Ryu H, Kim J, Oh T H, Pfister H, Chung J S. Sound source localization is all about cross-modal alignment. In: Proceedings of the International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 7743−7753

252 Senocak A, Ryu H, Kim J, Oh T H, Pfister H, Chung J S. Toward interactive sound source localization: Better align sight and sound! *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(9): 7643−7659

253 Kim I, Song Y, Park J, Kim W H, Kwak S. Improving sound source localization with joint slot attention on image and audio. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 3121−3130

254 Liu T Y, Zhang P, Xiong J W, Li C Y, Huo Y, Huang W, et al. Less means more: Single stream audio-visual sound source localization via shared-parameter network. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, **33**: 4350−4360

255 Zhou J X, Shen X Y, Wang J Y, Zhang J Y, Sun W X, Zhang J, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, 2025, **133**(4): 1644−1664

256 Mo S T, Raj B. Weakly-supervised audio-visual segmentation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 753

257 Bhosale S, Yang H S, Kanojia D, Zhu X T. Leveraging foundation models for unsupervised audio-visual segmentation. arXiv preprint arXiv: 2309.06728, 2023.

258 Liu J X, Wang Y, Ju C, Ma C F, Zhang Y, Xie W D. Annotation-free audio-visual segmentation. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 5592−5602

259 Chen T X, Tan Z T, Gong T, Chu Q, Wu Y, Liu B, et al. Bootstrapping audio-visual video segmentation by strengthening audio cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, **35**(3): 2398−2409

260 Mao Y X, Zhang J, Xiang M C, Zhong Y R, Dai Y C. Multimodal variational auto-encoder based audio-visual segmentation. In: Proceedings of the International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 954−965

261 Liu C, Li P K, Yang L Y, Wang D D, Li L C, Yu X. Robust audio-visual segmentation via audio-guided visual convergent alignment. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 28922−28931

262 Bhosale S, Yang H S, Kanojia D, Deng J K, Zhu X T. Unsupervised audio-visual segmentation with modality alignment. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 15567−15575

263 Zhu Y, Li K, Yang Z X. Exploiting EfficientSAM and temporal coherence for audio-visual segmentation. *IEEE Transactions on Multimedia*, 2025, **27**: 2999−3008

264 Xuan H Y, Liu T X, Dong W X, Li Z H, Chen S. X-STA: Cross-modal spatial-temporal alignment network for unified audio-visual segmentation. *IEEE Signal Processing Letters*, 2025, **32**: 2883−2887

265 Lei Y Y, Cao H W. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 2023, **14**(4): 2954−2969

266 Goncalves L, Leem S G, Lin W C, Sisman B, Busso C. Versatile audio-visual learning for emotion recognition. *IEEE Transactions on Affective Computing*, 2025, **16**(1): 306−318

267 Hsu J H, Wu C H. Applying segment-level attention on bimodal Transformer encoder for audio-visual emotion recognition. *IEEE Transactions on Affective Computing*, 2023, **14**(4): 3231−3243

268 Mocanu B, Tapu R, Zaharia T. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing*, 2023, **133**: Article No. 104676

269 Praveen R G, Cardinal P, Granger E. Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023, **5**(3): 360−373

270 Praveen R G, Alam J, Charton E. United we stand, divided we fall: Handling weak complementarity for audio-visual emotion recognition in valence-arousal space. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE, 2025. 5741−5751

271 Pan B, Hirota K, Jia Z Y, Zhao L H, Jin X M, Dai Y P. Multimodal emotion recognition based on feature selection and extreme learning machine in video clips. *Journal of Ambient Intelligence and Humanized Computing*, 2023, **14**(3): 1903−1917

272 Shi T, Ge X R, Jose J M, Pugeault N, Henderson P. Detail-enhanced intra-and inter-modal interaction for audio-visual emotion recognition. In: Proceedings of the 27th International Conference on Pattern Recognition. Kolkata, India: Springer, 2024. 451−465

273 Ding S Y, Tang T B, Lu C K. Lightweight spatio-temporal convolutional neural network for audio-visual emotion recognition. *IEEE Transactions on Affective Computing*, 2025, **16**(4): 2721−2734

274 Sharafi M, Yazdchi M, Rasti J. Audio-visual emotion recognition using K-means clustering and spatio-temporal CNN. In: Proceedings of the 6th International Conference on Pattern Recognition and Image Analysis (IPRIA). Qom, Islamic Republic of Iran: IEEE, 2023. 1−6

275 Wang A J, Fang Z J, Jiang X Y, Gao Y B, Cao G F, Ma S W. Depth estimation of multi-modal scene based on multi-scale modulation. In: Proceedings of the International Conference on Image Processing (ICIP). Kuala Lumpur, Malaysia: IEEE, 2023. 2795−2799

276 Gao R H, Chen C G, Al-Halah Z, Schissler C, Grauman K. VISUALECHOES: Spatial image representation learning through echolocation. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 658−676

277 Liu X H, Hornauer S, Moutarde F, Lu J L. AVS-Net: Audio-visual scale net for self-supervised monocular metric depth estimation. arXiv preprint arXiv: 2412.01637, 2024.

278 Karaoguz C, Weisswange T H, Rodemann T, Wrede B, Rothkopf C A. Reward-based learning of optimal cue integration in audio and visual depth estimation. In: Proceedings of the International Conference on Advanced Robotics (ICAR). Tallinn, Estonia: IEEE, 2011. 389−395

279 Zhang C H, Tian K, Ni B L, Meng G F, Fan B, Zhang Z X, et al. Stereo depth estimation with echoes. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 496−513

280 Sun W, Qiu L L. Visual timing for sound source depth estimation in the wild. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 12348−12355

281 Parida K K, Srivastava S, Sharma G. Beyond image to depth: Improving depth prediction using echoes. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 8264−8273

282 Liang S S, Huang C, Tian Y P, Kumar A, Xu C L. AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1629

283 Purushwalkam S, Garí S V A, Ithapu V K, Schissler C, Robinson P, Gupta A, et al. Audio-visual floorplan reconstruction. In: Proceedings of the International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 1163−1172

284 Wilson J, Rewkowski N, Lin M C, Fuchs H. Echo-reconstruction: Audio-augmented 3D scene reconstruction. arXiv preprint arXiv: 2110.02405, 2021.

285 Kim H, Remaggi L, Jackson P J B, Fazi F M, Hilton A. 3D room geometry reconstruction using audio-visual sensors. In: Proceedings of the International Conference on 3D Vision (3DV). Qingdao, China: IEEE, 2017. 621−629

286 Alawadh M. 3D Audio-visual Indoor Scene Reconstruction and Completion for Virtual Reality From a Single Image [Ph.D. dissertation], University of Southampton, UK, 2025.

287 Konno T, Nishida K, Itoyama K, Nakadai K. Audio-visual 3D reconstruction framework for dynamic scenes. In: Proceedings of the International Symposium on System Integration. Honolulu, USA: IEEE, 2020. 802−807

288 Younes A, Honerkamp D, Welschehold T, Valada A. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *IEEE Robotics and Automation Letters*, 2023, **8**(2): 928−935

289 Shi Z B, Zhang L, Li L F, Shen Y. Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 14673−14680

290 Wang H C, Wang Y X, Zhong F W, Wu M D, Zhang J W, Wang Y Z, et al. Learning semantic-agnostic and spatial-aware representation for generalizable visual-audio navigation. *IEEE Robotics and Automation Letters*, 2023, **8**(6): 3900−3907

291 Huang C G, Mees O, Zeng A, Burgard W. Audio visual language maps for robot navigation. In: Proceedings of the 18th International Symposium on Experimental Robotics (ISER). Chiang Mai, Thailand: Springer, 2023. 105−117

292 Chen C G, Al-Halah Z, Grauman K. Semantic audio-visual navigation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 15511−15520

293 Chen C G, Jain U, Schissler C, Gari S V A, Al-Halah Z, Ithapu V K, et al. SoundSpaces: Audio-visual navigation in 3D environments. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 17−36

294 Kondoh H, Kanezaki A. Multi-goal audio-visual navigation using sound direction map. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS). Detroit, USA: IEEE, 2023. 5219−5226

295 Yu Y F, Huang W B, Sun F C, Chen C G, Wang Y K, Liu X H. Sound adversarial audio-visual navigation. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022.

296 Liu X L, Paul S, Chatterjee M, Cherian A. CAVEN: An embodied conversational agent for efficient audio-visual navigation in noisy environments. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 3765−3773

297 Bao Xi-Gang, Zhou Chun-Lai, Xiao Ke-Jing, Qin Biao. Survey on visual question answering. *Journal of Software*, 2021, **32**(8): 2522−2544
(包希港, 周春来, 肖克晶, 覃飙. 视觉问答研究综述. 软件学报, 2021, **32**(8): 2522−2544)

298 Yang P C, Wang X, Duan X G, Chen H, Hou R Z, Jin C, et al. AVQA: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM, 2022. 3480−3491

299 Zhao Y H, Xi W, Bai G R, Liu X H, Zhao J Z. Heterogeneous interactive graph network for audio-visual question answering. *Knowledge-Based Systems*, 2024, **300**: Article No. 112165

300 Li G Y, Hou W X, Hu D. Progressive spatio-temporal perception for audio-visual question answering. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 7808−7816

301 Li Z B, Zhou J X, Zhang J, Tang S G, Li K, Guo D. Patch-level sounding object tracking for audio-visual question answering. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 5075−5083

302 Li Z B, Guo D, Zhou J X, Zhang J, Wang M. Object-aware adaptive-positivity learning for audio-visual question answering. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 3306−3314

303 Li L J, Jin T, Lin W, Jiang H, Pan W W, Wang J, et al. Multi-granularity relational attention network for audio-visual question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, **34**(8): 7080−7094

304 Jiang Y Y, Yin J Q. Target-aware spatio-temporal reasoning via answering questions in dynamic audio-visual scenarios. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 9399−9409

305 Ma J, Hu M, Wang P H, Sun W C, Song L Y, Pei H B, et al. Look, listen, and answer: Overcoming biases for audio-visual question answering. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 302

306 Lao M R, Pu N, Liu Y, He K, Bakker E M, Lew M S. COCA: Collaborative causal regularization for audio-visual question answering. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 12995−13003

307 Li G Y, Du H H, Hu D. Boosting audio visual question answering via key semantic-aware cues. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 5997−6005

308 Pei B Q, Huang Y F, Chen G, Xu J L, Wang Y L, Wang L M, et al. Guiding audio-visual question answering with collective question reasoning. *International Journal of Computer Vision*, 2025, **133**(10): 6912−6929

309 Alamri H, Cartillier V, Das A, Wang J, Cherian A, Essa I, et al. Audio visual scene-aware dialog. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 7550−7559

310 Heo Y, Kang S, Seo J. Natural-language-driven multimodal representation learning for audio-visual scene-aware dialog system. *Sensors*, 2023, **23**(18): Article No. 7875

311 Chen Z, Liu H C, Wang Y. DialogMCF: Multimodal context flow for audio visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, **32**: 753−764

312 Li Z K, Li Z J, Zhang J C, Feng Y, Zhou J. Bridging text and video: A universal multimodal Transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, **29**: 2476−2483

313 Ye M C, You Q Z, Ma F L. QUALIFIER: Question-guided self-attentive multimodal fusion network for audio visual scene-aware dialog. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2022. 2503−2511

314 Park S J, Kim Y, Rha H, Godiva B, Ro Y M. AV-EmoDialog: Chat with audio-visual users leveraging emotional cues. arXiv preprint arXiv: 2412.17292, 2024.

315 Shah A, Geng S J, Gao P, Cherian A, Hori T, Marks T K, et al. Audio-visual scene-aware dialog and reasoning using audio-visual Transformers with joint student-teacher learning. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 7732−7736

316 Ye Q L, Yu Z T, Liu X. Answering diverse questions via text attached with key audio-visual clues. arXiv preprint arXiv: 2403.06679, 2024.

317 Hou Jing-Yi, Qi Ya-Yun, Wu Xin-Xiao, Jia Yun-De. Cross-lingual knowledge distillation for Chinese video captioning. *Chinese Journal of Computers*, 2021, **44**(9): 1907−1921
(侯静怡, 齐雅昀, 吴心筱, 贾云得. 跨语言知识蒸馏的视频中文字幕生成. 计算机学报, 2021, **44**(9): 1907−1921)

318 Shen X Y, Li D, Zhou J X, Qin Z, He B W, Han X D, et al. Fine-grained audible video description. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 10585−10596

319 Xie Z Y, Yang Y, Yu Y K, Liu Y. Exploring audio-visual con-

cepts for dense video captioning. In: Proceedings of the International Conference on Digital Society and Intelligent Systems (DSInS). Sydney, Australia: IEEE, 2024: 334−338

320 Çayli Ö, Liu X B, Kiliç V, Wang W W. Knowledge distillation for efficient audio-visual video captioning. In: Proceedings of the European Signal Processing Conference (EUSIPCO). Helsinki, Finland: IEEE, 2023. 745−749

321 Xie Y L, Niu J J, Zhang Y, Ren F. Global-shared text representation based multi-stage fusion Transformer network for multimodal dense video captioning. *IEEE Transactions on Multimedia*, 2024, **26**: 3164−3179

322 Yang A, Nagrani A, Seo P H, Miech A, Pont-Tuset J, Laptev I, et al. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 10714−10726

323 Han S X, Liu J, Zhang J Y M, Gong P Z, Zhang X L, He H H. Lightweight dense video captioning with cross-modal attention and knowledge enhanced unbiased scene graph. *Complex & Intelligent Systems*, 2023, **9**(5): 4995−5012

324 Kim J, Shin J, Kim J. AVCap: Leveraging audio-visual features as text tokens for captioning. In: Proceedings of the 25th Annual Conference of the International Speech Communication Association. Kos, Greece: ISCA, 2024.

325 AlSuwat M, Al-Shareef S, AlGhamdi M. Audio-visual self-supervised representation learning: A survey. *Neurocomputing*, 2025, **634**: Article No. 129750

326 Arandjelović R, Zisserman A. Objects that sound. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 451−466

327 Owens A, Efros A A. Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 639−658

328 Korbar B, Tran D, Torresani L. Cooperative learning of audio and video models from self-supervised synchronization. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc., 2018. 7774−7785

329 Huang C, Tian Y P, Kumar A, Xu C L. Egocentric audio-visual object localization. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 22910−22921

330 Sun W X, Zhang J Y, Wang J Y, Liu Z Y, Zhong Y R, Feng T P, et al. Learning audio-visual source localization via false negative aware contrastive learning. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 6420−6429

331 Morgado P, Vasconcelos N, Misra I. Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 12470−12481

332 Ma S, Zeng Z Y, McDuff D, Song Y L. Active contrastive learning of audio-visual video representations. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview, 2021. 1−19

333 Xuan H Y, Xu Y H, Chen S, Wu Z L, Yang J, Yan Y, et al. Active contrastive set mining for robust audio-visual instance discrimination. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI, 2022. 3643−3649

334 Xuan H Y, Wu Z L, Yang J, Jiang B, Luo L, Alameda-Pineda X, et al. Robust audio-visual contrastive learning for proposal-based self-supervised sound source localization in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(7): 4896−4907

335 Li Z J, Zhao B, Yuan Y. Bio-inspired audiovisual multi-representation integration via self-supervised learning. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 3755−3764

336 Sarkar P, Etemad A. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 9723−9732

337 Jenni S, Black A, Collomosse J. Audio-visual contrastive learning with temporal self-supervision. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 7996−8004

338 Zhang J X, Wan G S, Gao J Q, Ling Z H. Audio-visual representation learning via knowledge distillation from speech foundation models. *Pattern Recognition*, 2025, **162**: Article No. 111432

339 Zhu B Q, Wang C J, Xu K L, Feng D W, Zhou Z M, Zhu X Q. Learning incremental audio-visual representation for continual multimodal understanding. *Knowledge-Based Systems*, 2024, **304**: Article No. 112513

340 Zuo Y K, Yao H T, Zhuang L S, Xu C S. Hierarchical augmentation and distillation for class incremental audio-visual video recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(11): 7348−7362

341 Cui M, Yue X H, Qian X Y, Zhao J Z, Liu H H, Liu X B, et al. Audio-visual class-incremental learning for fish feeding intensity assessment in aquaculture. arXiv preprint arXiv: 2504.15171, 2025.

342 Mo S T, Pian W G, Tian Y P. Class-incremental grouping network for continual audio-visual learning. In: Proceedings of the International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 7754−7764

343 Pian W G, Mo S T, Guo Y H, Tian Y P. Audio-visual class-incremental learning. In: Proceedings of the International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 7765−7777

344 Yue X H, Zhang X Y, Chen Y M, Zhang C W, Lao M R, Zhuang H P, et al. MMAL: Multi-modal analytic learning for exemplar-free audio-visual class incremental tasks. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 2428−2437

345 Cui Y W, Liu L, Yu Z T, Huang G J, Hong X P. Few-shot audio-visual class-incremental learning with temporal prompting and regularization. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025. 16118−16126

346 Zhang Lu-Ning, Zuo Xin, Liu Jian-Wei. Research and development on zero-shot learning. *Acta Automatica Sinica*, 2020, **46**(1): 1−23
(张鲁宁, 左信, 刘建伟. 零样本学习研究进展. 自动化学报, 2020, **46**(1): 1−23)

347 Parida K K, Matiyali N, Guha T, Sharma G. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). Snowmass, USA: IEEE, 2020. 3240−3249

348 Li Y P, Luo Y, Du B. Audio-visual generalized zero-shot learning based on variational information bottleneck. In: Proceedings of the International Conference on Multimedia and Expo (ICME). Brisbane, Australia: IEEE, 2023. 450−455

349 Zheng Q C, Hong J, Farazi M. A generative approach to audio-visual generalized zero-shot learning: Combining contrastive and discriminative techniques. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Gold Coast, Australia: IEEE, 2023. 1−8

350 Mo S T, Morgado P. Audio-visual generalized zero-shot learning the easy way. In: Proceedings of the 18th European Confer-

ence on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 377−395

351 Dong Y J, Chen S M, Duan B W, Ding W P, Wang Y S, You X G. Object-aware image augmentation for audio-visual zero-shot learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025, **9**(6): 4106−4118

352 Li W R, Wang P H, Xiong R Q, Fan X P. Spiking tucker fusion Transformer for audio-visual zero-shot learning. *IEEE Transactions on Image Processing*, 2024, **33**: 4840−4852

353 Yang Z, Li W R, Hou J X, Cheng G H. Multi-modal spiking tensor regression network for audio-visual zero-shot learning. *Neurocomputing*, 2025, **629**: Article No. 129636

354 Li W R, Ma Z Y, Deng L J, Man H Y, Fan X P. Modality-fusion spiking Transformer network for audio-visual zero-shot learning. In: Proceedings of the International Conference on Multimedia and Expo (ICME). Brisbane, Australia: IEEE, 2023. 426−431

355 Zhang K W, Zhao K C, Tian Y N. Temporal-semantic aligning and reasoning Transformer for audio-visual zero-shot learning. *Mathematics*, 2024, **12**(14): Article No. 2200

356 Li W R, Wang P H, Wang X T, Zuo W M, Fan X P, Tian Y H. Multi-timescale motion-decoupled spiking Transformer for audio-visual zero-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, **35**(11): 10772−10786

357 Larsen-Freeman D. Transfer of learning transformed. *Language Learning*, 2013, **63**(S1): 107−129

358 Hajavi A, Etemad A. Audio representation learning by distilling video as privileged information. *IEEE Transactions on Artificial Intelligence*, 2024, **5**(1): 446−456

359 Yun H, Na J, Kim G. Dense 2D-3D indoor prediction with sound via aligned cross-modal distillation. In: Proceedings of the International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 7829−7838

360 Kim J U, Kim S T. Towards robust audio-based vehicle detection via importance-aware audio-visual learning. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

361 Chen J Y, Wang W G, Liu S, Li H S, Yang Y. Omnidirectional information gathering for knowledge transfer-based audio-visual navigation. In: Proceedings of the International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 10959−10969

362 Chen M C, Zhang B M, Han Z B, Jiang W Y, Wang Y M, Feng S, et al. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: OpenReview.net, 2025. 1−10

363 Duan H Y, Xia Y, Zhou M Z, Tang L, Zhu J M, Zhao Z. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2445

364 Mo S T, Morgado P. A unified audio-visual learning framework for localization, separation, and recognition. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: Curran Associates Inc., 2023. Article No.1041

365 Cai Chao-Yang, Zhou Li-Jing. Study on learning transfer and ability generation from the perspective of cognitive psychology. *Advances in Education*, 2025, **15**(1): 1226−1237
(蔡朝阳, 周黎婧. 认知心理学视角下学习迁移与能力生成研究. 教育进展, 2025, **15**(1): 1226−1237)

366 Chen Guang, Guo Jun. Artificial intelligence in the era of large language models: Technical significance, industry applications, and challenges. *Journal of Beijing University of Posts and Telecommunications*, 2024, **47**(4): 20−28
(陈光, 郭军. 大语言模型时代的人工智能: 技术内涵、行业应用与挑战. 北京邮电大学学报, 2024, **47**(4): 20−28)

367 Peng P Y, Huang P Y, Li S W, Mohamed A, Harwath D. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: ACL, 2024. 12442−12462

368 Andreyev A. Quantization for OpenAI′s whisper models: A comparative analysis. arXiv preprint arXiv: 2503.09905, 2025.

369 Su Y, Bai J S, Xu Q S, Xu K L, Dou Y. Audio-language models for audio-centric tasks: A systematic survey. arXiv preprint arXiv: 2501.15177, 2025.

370 Jiang F, Lin Z Y, Bu F, Du Y H, Wang B Y, Li H Z. S2S-arena, evaluating speech2speech protocols on instruction following with paralinguistic information. arXiv preprint arXiv: 2503.05085, 2025.

371 He H R, Zhang Y, Lin L, Xu Z W, Pan L. Pre-trained video generative models as world simulators. arXiv preprint arXiv: 2502.07825, 2025.

372 Wang Y M, Liu X Y, Pang W, Ma L, Yuan S, Debevec P, et al. Survey of video diffusion models: Foundations, implementations, and applications. arXiv preprint arXiv: 2504.16081, 2025.

373 Zhang Y B, Wei Y X, Lin X H, Hui Z, Ren P R, Xie X S, et al. VideoElevator: Elevating video generation quality with versatile text-to-image diffusion models. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI Press, 2025. 10266−10274

374 Wang Y L, Deng Y J, Zheng Y J, Chattopadhyay P, Wang L P. Vision Transformers for image classification: A comparative survey. *Technologies*, 2025, **13**(1): Article No. 32

375 Yu Z P, Ananiadou S. Understanding multimodal LLMs: The mechanistic interpretability of Llava in visual question answering. arXiv preprint arXiv: 2411.10950, 2024.

376 Zhan J, Dai J Q, Ye J S, Zhou Y H, Zhang D, Liu Z G, et al. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: ACL, 2024. 9637−9662

377 Wu S Q, Fei H, Qu L G, Ji W, Chua T S. NExT-GPT: Any-to-any multimodal LLM. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: JMLR.org, 2024. Article No. 2187

378 Chen F L, Han M L, Zhao H Z, Zhang Q Y, Shi J, Xu S, et al. X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv: 2305.04160, 2023.

379 Fu C Y, Lin H J, Long Z W, Shen Y H, Dai Y H, Zhao M, et al. VITA: Towards open-source interactive Omni multimodal LLM. arXiv preprint arXiv: 2408.05211, 2024.

380 Akbari H, Yuan L Z, Qian R, Chuang W H, Chang S F, Cui Y, et al. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 1853

381 Lin Y B, Bertasius G. Siamese vision Transformers are scalable audio-visual learners. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 303−321

382 Huang P Y, Sharma V, Xu H, Ryali C, Fan H Q, Li Y H, et al. MAViL: Masked audio-video learners. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 894

383 Chowdhury S, Nag S, Dasgupta S, Chen J, Elhoseiny M, Gao

R H, et al. MEERKAT: Audio-visual large language model for grounding in space and time. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 52−70

384 Chen K, Gou Y H, Huang R H, Liu Z L, Tan D X, Xu J, et al. EMOVA: Empowering language models to see, hear and speak with vivid emotions. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 5455−5466

385 Sun G Z, Yu W Y, Tang C L, Chen X Z, Tan T, Li W, et al. video-SALMONN: Speech-enhanced audio-visual large language models. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: Curran Associates Inc., 2024. Article No. 1921

386 Chu Y Q, Liao L Z, Zhou Z Y, Ngo C W, Hong R H. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*, 2025, **27**: 8276−8287

387 Sun G Z, Yang Y D, Zhuang J M, Tang C L, Li Y X, Li W, et al. video-SALMONN-o1: Reasoning-enhanced audio-visual large language model. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Vancouver, Canada: OpenReview.net, 2025. 1−10

388 Cappellazzo U, Kim M, Chen H J, Ma P C, Petridis S, Falavigna D, et al. Large language models are strong audio-visual speech recognition learners. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE, 2025. 1−5

389 Mao Y R, Ge Y H, Fan Y J, Xu W Y, Mi Y, Hu Z H, et al. A survey on LoRA of large language models. *Frontiers of Computer Science*, 2025, **19**(7): Article No. 197605

390 Du H H, Li G Y, Zhou C, Zhang C J, Zhao A, Hu D. Crab: A unified audio-visual scene understanding model with explicit cooperation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 18804−18814

391 Jin D, Zhou Y H, Zhou J X, Ma J Q, Guo R H, Guo D. SimToken: A simple baseline for referring audio-visual segmentation. arXiv preprint arXiv: 2509.17537, 2025.

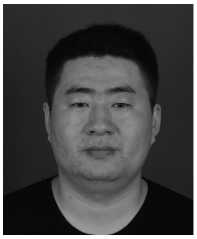
392 Cappellazzo U, Kim M, Petridis S. Adaptive audio-visual speech recognition via Matryoshka-based multimodal LLMs. arXiv preprint arXiv: 2503.06362, 2025.

393 Tang C J, Li Y X, Yang Y D, Zhuang J M, Sun G Z, Li W, et al. video-SALMONN 2: Captioning-enhanced audio-visual large language models. In: Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: OpenReview.net, 2025.

394 Huang G J, Lin W L, Liu L. Content-aware efficient learner for audio-visual emotion recognition. In: Proceedings of the 16th International Conference on Social Robotics. Shenzhen, China: Springer, 2024. 31−40

宣寒宇 安徽大学大数据与统计学院副教授. 主要研究方向为计算机视觉和多模态学习.
E-mail: 22176@ahu.edu.cn
(**XUAN Han-Yu** Associate professor at the School of Big Data and Statistics, Anhui University. His research interests include computer vision and multimodal learning.)

陈 强 安徽大学计算机科学与技术学院博士研究生. 主要研究方向为计算机视觉.
E-mail: e125111017@stu.ahu.edu.cn
(**CHEN Qiang** Ph.D. candidate at the School of Computer Science and Technology, Anhui University. His main research interest is computer vision.)



吴之亮 浙江大学计算机科学与技术学院博士后. 主要研究方向为计算机视觉和多模态技术.
E-mail: wu_zhiliang@zju.edu.cn
(**WU Zhi-Liang** Postdoctor at the School of Computer Science and Technology, Zhejiang University. His research interests include computer vision and multimodal technology.)

韩向敏 清华大学软件学院博士后. 主要研究方向为医学超图计算, 尤其是高阶关联启动的脑网络及病理图像分析.
E-mail: simon.xmhan@gmail.com
(**HAN Xiang-Min** Postdoctor at the School of Software, Tsinghua University. His main research interest is medical hypergraph computation, with a particular emphasis on brain networks and pathology image analysis driven by high-order correlations.)

董文祥 合肥综合性国家科学中心数据空间研究院研究员. 主要研究方向为网络空间安全, 网络社会认知, 数据智能计算.
E-mail: javin0304@foxmail.com
(**DONG Wen-Xiang** Researcher at the Institute of Dataspace, Hefei Comprehensive National Science Center. His research interests include cyberspace security, network social cognition, and data intelligent computing.)

马 楠 北京工业大学信息学院教授. 主要研究方向为交互认知, 具身智能和机器视觉. 本文通信作者.
E-mail: manan123@bjut.edu.cn
(**MA Nan** Professor at the School of Information Science and Technology, Beijing University of Technology. Her research interests include interactive cognition, embodied intelligence, and machine vision. Corresponding author of this paper.)