孙络祎

📞 电话: 18621064042 ✉邮箱: loie@zju.edu.cn 🎓 主页: loiesun.github.io

教育背景

浙江大学（博士研究生） 2024.09 - 2028.03

- 计算机科学与技术专业; 研究方向: 音频-视频-语言多模态理解 导师: 谢伟迪

上海大学（硕士） 2020.09 - 2023.06

- 数字媒体创意工程专业, 上海电影学院; 研究方向: 视觉引导的音效生成

云南大学（学士） 2015.09 - 2019.06

- 数字媒体技术专业, 软件学院

工作经历

上海人工智能实验室（见习研究员） 2024.09 - 至今

上海人工智能实验室（科研助理） 2023.10 - 2024.08

项目经历

1. 音频-文本表征学习及音频描述生成

第一作者 接收至 ACM MM 2024（**CCF-A**）

- 概述: (i) 基于视听数据集, 构建自动化流程, 利用计算机视觉与音频理解工具自动生成大规模、高质量的**音频-文本表征学习数据集**; (ii) 构建一个由音频编码模块和文本编码模块构成的**音频-文本对比学习框架**, 并提出一个由音频编码模块、映射网络模块、文本生成模块构成的**轻量级音频自动描述网络**; (iii) 实验结果表明, 利用新数据集训练的模型, 在检索、音频自动描述及音频分类等任务的性能上均有大幅提升。

2. 基于音频-语言模型的细粒度音频时序定位

第一作者 投稿至 ACM MM 2026（**CCF-A**）

- 概述: (i) 提出 SpotSound 模型, 通过**时间戳与音频特征交错输入**大语言模型实现细粒度时间定位, 使模型能够精准捕捉声学事件的起止时间; (ii)针对负样本引入判别式训练策略, 有效**抑制模型对不存在声音事件的幻觉**, 提升复杂场景下的鲁棒性; (iii) 开发自动化流程合成长音频时序定位数据, 构建 SpotSound-Bench 测试基准, 聚焦“大海捞针”式**短时事件定位任务**; (iv) 实验结果表明, SpotSound 在多个音频时序定位基准上达到**SOTA**, 在声音事件检测等下游任务中保持良好泛化能力。

3. 音频感知驱动的长视频全模态理解方法

第二作者 投稿至 ECCV 2026（**CCF-B**）

- 概述: (i) 基于视频密集描述数据集, 构建自动化流程, 利用多种音频模型与大语言模型, 生成大规模**音视频解耦描述数据集AVDC** 及带思维链推理的**问答数据集 AVDC-QA-CoT**; (ii) 提出两阶段全模态训练策略: 先通过描述生成预训练对齐音视频表征, 再经指令微调提升全模态理解与推理能力及音频理解鲁棒性; (iii) 实验表明,模型在音视频描述、全模态问答及纯音频理解等任务上效果均显著提升, 尤其在识别**不可见声音和减少幻觉**方面, 有效缩小了开源与闭源模型的差距。

4. 时序对齐视觉特征映射的音效生成方法

学生第一作者 接收至 CADCG（**CCF-A期刊**）

- 概述: (i) 针对**无声视频片段**, 构建视觉特征引导模型, **生成时序匹配、内容一致的声音效果**; (ii) 提出一个由视觉特征聚合模块、视音频跨模态特征映射模块、音频解码模块组成的端到端的时序对齐特征映射网络; (iii) 实验结果表明, 对比目前最先进的方法, 模型输出结果在**保真度**和**时序对齐效果**方面, 均有显著提升。

5. 面向癌症诊断的知识增强型病理视觉-语言基础模型

第二作者 接收至 Cancer Cell（**IF=44.5**）

- 概述: (i) 整合疾病本体论与统一医学语言系统等权威医学知识库, 构建覆盖 11,454 种人类疾病及139,143 个疾病属性的**疾病知识图谱**; (ii) 设计知识增强的对比学习机制,构建基于知识图谱层级的病理图像-文本语义组, 将**领域知识显式注入**视觉-语言对齐过程, 提升模型的理解能力; (iii) 实验结果表明, 模型在**区域分割、癌症检测和亚型分类**等任务中达到SOTA, 在**罕见肿瘤诊断**任务中, 性能显著优于现有模型。

论文及专利发表情况

- Sun L, Xu X, Wu M, Xie W. A Large-scale Dataset for Audio-Language Representation Learning[C]. ACM Multimedia 2024. 2024.
- Sun L, Zhou X, Li Z, et al. SpotSound: Enhancing Large Audio-Language Models with Fine-Grained Temporal Grounding[C]. Submitting to ACM MM 2026.
- 谢志峰, 孙络祎, 孙郁洲, 余椿鹏, 马利庄. 时序对齐视觉特征映射的音效生成方法 [J]. 计算机辅助设计与图形学学报, 2022, 34(10):1506-1514.
- Yan K, Sun L, Zhou X, et al. From Sensing to Reasoning: Empowering Long-form Omni-modal Understanding with Robust Audio Perception[C]. Submitting to ECCV 2026.
- Zhou X, Sun L, He D, et al. Knowledge-enhanced Pretraining for Vision-language Pathology Foundation Model on Cancer Diagnosis[J]. Cancer Cell, 2026.
- Che W, Sun L, Xie Z, et al. Multi-Scale Graph Convolutional Interaction Network For Salient Object Detection[C]//2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, 679-683.
- 谢伟迪, 孙络祎, 严恺颖, 周晓. 音频感知驱动的长视频全模态理解方法及系统.
- 郑迦恒, 夏世宇, 孙络祎, 谢志峰. 一种语音驱动的可编辑人脸重演方法、装置及存储介质.
- 廖毅慧, 蒋智文, 孙络祎, 谢志峰. 一种多模态语义交互的视频场景分割方法.

其他合作项目

1. 基于时序对比学习的视音频同步模型 2022.11 – 2023.09

项目合作单位: 牛津大学计算几何组(VGG), 亚马逊公司

- 概述: (i) 构建包含**2,000+**个高光片段的高质量网球比赛数据集, 涵盖多场职业赛事的多视角视音频素材; (ii) 实现基于时序的视音频对比学习预训练, 完成视音频同步模型的领域微调, 使模型能够预测击球瞬间与声音间的偏移量; (iii) 在**真实测试集**上的评估结果显示, 模型同步预测准确率达63.6%。

2. 基于对比学习的心音诊断模型 2025.04 – 2025.11

项目合作单位: 盖茨基金会, 上海交通大学附属新华医院, 上海交通大学人工智能学院

- 概述: (i) 处理并清洗**10,000+**条心音信号, 通过自动化脚本从心超报告中提取诊断标签, 构建大规模多模态数据集; (ii) 设计基于对比学习的预训练框架, 并结合多分类头, 使模型在学习心音信号有效表征的同时, 构建与心超金标准之间的映射关系; (iii)实验结果表明, 模型对心脏功能异常的判别准确率达到67.8%, **超过主治医师人工判读**的平均准确率。

获奖情况

上海市优秀毕业生	2023 年
国家奖学金	2022 年
第十八届中国研究生数学建模竞赛二等奖	2021 年
上海大学一等奖学金	2020, 2021, 2022 年
云南大学二等奖学金	2017, 2018 年
云南大学优秀学生	2017, 2018 年
云南大学优秀学生干部	2017, 2018 年

专业技能

- 编程技能: 精通 Python 语言; 熟练使用 PyTorch 框架
- 科研技能: 熟悉音频理解、多模态感知任务; 具有构建数据集的丰富经验
- 语言能力: 英语 (CET-6), “LSCAT 中国翻译协会”S100 级认证
- 数字媒体创作技能: 精通 Adobe Photoshop, Premiere, Audition 等软件