

黄冠杰

多模态大模型算法研究员 | 语音与音视频理解 | MLLM / Agent / 模型可靠性

(+86) 18981003311 | ghuang565@connect.hkust-gz.edu.cn | 香港科技大学 (广州) 人工智能博士在读

个人简介

香港科技大学 (广州) 人工智能博士在读，具备 5 年以上多模态算法研发与科研经历，方向覆盖音视频学习、低资源语音识别、MLLM、多智能体、音频伪造检测与大模型不确定性。论文及成果发表于 TPAMI、NeurIPS、IEEE/ACM TASLP、ACM Multimedia、ICASSP 等；兼具学术研究、数据集/评测基准构建和工业视觉算法经验。

核心能力

研究方向：多模态表示与对齐、音视频理解、语音/ASR、MLLM 与多智能体、模型不确定性、音频伪造检测

算法方法：Transformer、Adapter/参数高效学习、注意力与原型学习、自监督学习、CNN、U-Net/GAN、多任务与证据学习

工程实践：Python；数据处理与半自动标注流水线；模型训练、评测与可视化；算法接口与自动化测试

教育经历

香港科技大学 (广州) | 人工智能博士在读 | 导师：刘李教授、曾宪国教授 2023.09 - 至今

研究方向：高效音视频学习、自动线索语识别、音视频伪造检测、大模型不确定性分析

澳大利亚国立大学 | 人工智能硕士 2018.07 - 2020.12

电子科技大学 | 软件工程学士 | 优秀毕业生 2014.09 - 2018.06

科研与工作经历

香港科技大学 (广州) | 博士生研究员 2023.05 - 至今

面向听障者发声改善的线索语多模态智能评估研究

项目背景：针对听障语音数据稀缺、语音与口形等多模态信息难以精确对齐，以及通用 ASR/MLLM 难以识别非典型发音并给出细粒度反馈等问题，依托听障人群参与的线索语培训前后测试实验，开展发声改善评估、低资源识别与个性化反馈研究。

- 多维实验评估与数据构建：**同步采集语音、口部视频、文字记忆与 fNIRS 数据，从发音清晰度、口形变化、认知表现和脑功能活动等维度量化线索语训练带来的发声改善。
- 低资源 ASR 与细粒度发音评价：**面向通用语音模型在听障语音上的识别偏差，研究低资源适配方法，并在字、音素和声调层面定位具体发音错误。
- 视听对齐与可解释建模：**联合建模目标文本、语音信号与口部运动，学习线索语手势、口形和实际发音之间的跨模态关联，识别影响发声改善的关键特征。
- 个性化发声练习工具：**将 ASR 错误定位与多模态分析结果转化为针对辅音、元音、声调及口形的可操作反馈，并依据个人历史表现动态调整练习内容。

基于语义调制提示的小样本音视频感知理解框架 | 第一作者，IEEE/ACM TASLP 2026

项目背景：针对小样本音视频分类中模型易过拟合、音视频时序异步导致跨模态融合困难，以及极少样本下模态不平衡等问题，提出参数高效的语义调制提示框架。

- 提示优化的视听高效学习：**基于 Adapter 设计参数高效学习模块，引入提示引导的潜在注意力机制，以文本语义提示定位冗长且异步的音视频序列中的关键 token，实现低资源条件下的跨模态对齐与融合。
- 提示调整原型正则化：**将单模态原型校准构建为动态过程，在训练迭代中利用语义提示持续拉近音频/视觉原型与语义中心，并依据各模态学习状态进行重平衡，缓解极少样本下的模态偏置。

半免训练自动线索语识别系统 | 第一作者，ICASSP 2026

项目背景：针对自动线索语识别中数据稀缺导致复杂手唇建模与跨模态融合网络易过拟合的问题，提出半免训练识别范式 STF-ACSR。

- 免训练手部提示模块：**将连续手部建模转化为基于 MLLM 的零样本分类任务；通过手部运动速度构建无参数关键帧筛选机制，并设计包含背景、多模态上下文、对比与推理提示的专家提示模板，激活 MLLM 的手势识别能力。

- **极简跨模态融合**：将手部识别结果转换为嵌入矩阵，仅使用单一线性层与唇部特征对齐，再以特征相加方式将手部信息作为提示注入预训练唇读模型，降低融合参数量与训练复杂度。

基于多智能体协作的自动线索语识别系统 | 第一作者, ACM MM 2025 Oral

项目背景：针对手唇模态异步、训练数据不足以及传统方法只能输出音素而难以生成自然语言句子的限制，提出首个协作式多智能体自动线索语识别系统 Cued-Agent。

- **四智能体协同架构**：集成 MLLM 手部识别智能体、Transformer 唇部识别智能体、手部提示解码智能体和 LLM 音素到词自校正智能体，形成从多模态感知、融合到自然语言输出的完整链路。
- **免训练手部解析与零参数融合**：设计关键帧筛选器与专家提示策略，将手部运动解码简化为免训练分类；推理时用手部提示动态加权唇读解码分数，实现无需新增训练参数的手唇融合。
- **自校正语言翻译**：引入线索语专家规则提示，由音素到词智能体对预测音素序列进行多轮语法和语义修正，实现从音素序列到自然语言句子的纠错转换。
- **听障场景数据构建**：采集并构建首个包含多名听障者的普通话线索语数据，支持更贴近真实应用场景的识别评测；成果获 ACM MM Oral 与 Student Travel Award。

基于物理约束的文本到音视频生成细粒度评测基准 | 核心作者, ACM MM 2026 Oral

项目背景：针对文本到音视频生成模型难以遵循声学物理规律、现有基准又多集中于音视频同步而忽略底层物理机制的问题，构建面向 T2AV 模型音频物理感知能力的评测基准 PhyAVBench。

- **物理敏感度评估范式**：设计 1000 组仅改变单一受控物理变量的成对提示词，通过对比物理响应分数评估模型对声学条件变化的敏感度及物理规律理解能力。
- **多维高质量数据构建**：覆盖日常声学场景及数十个细粒度测试点，包括声音衍射、共振等复杂现象；每个测试组包含至少 20 个经人工校对的新录制或收集真实视频，以降低预训练数据泄露风险。

基于全局-局部自适应的音频伪造检测 | ICME ESDD2 2026 第 4 名

项目背景：针对现有音频深度伪造检测方法在未知攻击和跨域场景下泛化能力不足，以及 Transformer/SSL 模型容易忽略细粒度信号伪造痕迹的问题，提出全局-局部自适应检测框架。

- **特征忽略与泛化分析**：通过显著性可视化和跨域层权重分析，验证 SSL 检测器易忽略局部伪造痕迹，且静态层选择难以适应未知攻击下的判别线索迁移。
- **层次化全局-局部编码**：设计语义-声学交叉注意力模块，协同建模双流自监督编码器特征；引入多粒度融合模块联合时序全局特征和 CNN 局部特征，弥合高级语义与底层伪造痕迹之间的粒度差异。
- **样本级自适应门控**：在多个 SSL 层和异构骨干网络上执行样本级动态重加权，根据攻击类型调整模型关注点，抑制域相关干扰并放大关键伪造证据。
- **SaniBoost 数据增强**：结合噪声净化与信号归一化，降低模型对环境噪声的捷径学习，将环境偏差与内在伪造痕迹解耦；获 ICME ESDD2 2026 环境音频伪造检测挑战赛第 4 名。

基于证据学习的大语言模型上下文学习不确定性量化 | 研究项目

项目背景：针对固定标签上下文学习中的不确定性估计依赖提示措辞、示例采样及重复生成，导致分数易受上下文变化干扰且计算成本高的问题，提出查询级证据不确定性框架 E-UQ。

- **查询级证据建模**：冻结大语言模型并仅提取测试查询的隐藏表示，将其映射为类别概率上的狄利克雷分布，统一获得分类结果、数据不确定性与模型不确定性；推理无需上下文示例、重复生成、束搜索或语义聚类。
- **双模式证据量化**：设计轻量训练版本 E-UQ-T 与免训练少样本版本 E-UQ-TF：前者仅训练轻量证据头，并联合预测误差、狄利克雷方差与 KL 正则；后者利用少量样本构建类别原型，通过相似度和类别间隔感知门控生成证据。
- **可靠性与效率评估**：在 5 个自然语言分类数据集及 LLaMA、Qwen 系列不同规模模型上验证误分类检测、提示/示例扰动稳定性和分布外检测能力；查询级单次前向将平均输入 Token 成本降低约 96.7%。

苏州 INSNE 图像软件技术有限公司 | 深度学习算法工程师

2022.05 - 2023.05

基于生成对抗网络的缺陷数据生成算法 | 国家发明专利相关研发

项目背景：针对少量缺陷样本训练易导致检测模型稳定性和泛化能力不足的问题，研发基于生成对抗网络的缺陷数据生成方法。

- **掩膜引导 U-Net 生成器**：对原始缺陷掩膜进行随机变换，并将其与随机噪声共同输入 U-Net 生成器，利用掩膜约束缺陷形态和生成区域，使模型聚焦局部缺陷结构。

- **多损失对抗训练**: 构建由原始背景、缺陷、掩膜及生成图像组成的增强序列, 联合相似度损失、结构损失和对抗损失优化生成器与判别器, 提高少样本条件下生成结果的稳定性与合理性。

缺陷检测模型生成方法与系统 | 国家发明专利相关研发

项目背景: 针对工业缺陷形态差异大、少样本条件下检测与定位性能受限的问题, 构建数据增广与多任务联合学习框架。

- **掩膜引导的数据增广**: 结合分时图像采集与质检先验筛选缺陷样本, 提取缺陷与背景掩膜; 通过模板匹配寻找最佳融合位置, 并结合几何变换与复制粘贴实现缺陷特征的自然融合。
- **多任务协同训练**: 联合提取真实目标图像与增广图像的深层特征, 端到端优化分类和分割子模型, 提升复杂形态缺陷的类别识别与像素级定位能力。

深圳市大数据研究院 | 算法工程师

2021.04 - 2022.05

百万级无人机目标跟踪基准 WebUAV-3M | 共同一作, IEEE TPAMI 2023

项目背景: 针对现有 UAV 跟踪数据集规模小、模态单一、类别与场景多样性不足, 难以充分训练和评估深度跟踪算法的问题, 构建百万级多模态无人机跟踪基准 WebUAV-3M。

- **超大规模基准与高效标注**: 构建包含 330 万帧、4500 个视频序列和 223 个目标类别的数据集; 设计可扩展的半自动标注流水线, 完成密集边界框标注。
- **文本与音频多模态扩展**: 为视频序列增加自然语言规范和音频描述, 以高层语义缓解复杂运动、外观变化和长时遮挡下的边界框歧义, 为多模态 UAV 跟踪提供数据基础。
- **细粒度评估协议**: 构建 7 个场景约束子测试集, 并对 40+ 种代表性跟踪算法进行系统评测与分析; 成果获深圳市优秀科技学术论文。

线性规划求解器研发

项目背景: 针对传统 MILP 求解器依赖启发式分支定界策略、处理大规模问题时计算耗时长的的问题, 探索深度学习方法对求解过程进行加速。

- **深度学习求解加速**: 引入 Solution Mapping 与 Neural Diving 等方法, 优化分支定界过程中的节点和候选解选择策略, 探索提升求解速度与收敛效率的可行路径。
- **接口与性能评估工具**: 主导开发求解器 Python 交互接口和全链路自动化测试模块; 设计时间分析可视化工具, 用于细化评估算法耗时与定位运行瓶颈。

澳大利亚国立大学 | 硕士研究

2020.03 - 2020.12

统一风格迁移算法

项目背景: 针对图像风格定义模糊、客观评估方法不足, 以及基于图像和基于域两类迁移范式联系不明确的问题, 提出统一风格迁移框架与统计风格分析方法。

- **统一风格迁移架构**: 引入生成式内部风格表征, 并对人工定义和人类视觉定义下的风格分布进行建模, 在形式上连接两类迁移范式; 该模块可作为外部风格迁移模型的即插即用组件。
- **统计风格分析**: 以不同风格域之间的统计距离量化相似性, 用于检验内部风格表征与人类艺术感知的一致性, 为风格迁移模型提供区别于主观视觉对比的评估路径。

澳大利亚联邦科学与工业研究组织 (CSIRO) | 研究助理

2019.07 - 2019.11

3D 岩层性质可视化工具

- **核心渲染与插值**: 基于 PyVista 开发三维岩性图像生成模块, 构建自动化数据处理管线, 对钻井数据进行空间插值与解析, 实现复杂地质结构的三维可视化。
- **云端部署与开源集成**: 基于 Binder 搭建云端运行环境, 降低本地依赖配置门槛; 成果集成至 GitHub 开源项目 csiro-hydrogeology/pyela。

早期工程经历

卫士通信息产业股份有限公司 | 实习软件工程师

2018.02 - 2018.05

- **算法预研**: 开展基于机器学习的犯罪画像特征分类研究; 研发无源图像清晰度识别模块, 用于在上传环节识别模糊图像并减少低质量信息干扰。

- **工具开发与系统测试：**负责通信设备测试工具的需求分析、程序结构与界面设计、实时状态采集模块开发，并参与 Ad-Hoc 网络设备的外场与软件系统测试。

代表论文与成果

- **第一作者：** Huang, G. et al. Semantic Modulated Prompting for Few-Shot Audio-Visual Classification. IEEE/ACM TASLP, 2026.
- **第一作者：** Huang, G. et al. Lend a Hand: Semi Training-Free Cued Speech Recognition via MLLM-Driven Hand Modeling. ICASSP, 2026.
- **第一作者；Oral；Student Travel Award：** Huang, G. et al. Cued-Agent: A Collaborative Multi-Agent System for Automatic Cued Speech Recognition. ACM MM, 2025.
- **核心作者；Oral：** Xie, T. et al. PhyAVBench: A Challenging Audio Physics-Sensitivity Benchmark for Physically Grounded Text-to-Audio-Video Generation. ACM MM, 2026.
- **共同一作：** Zhang, C. *, Huang, G. * et al. WebUAV-3M: A Benchmark for Unveiling the Power of Million-Scale Deep UAV Tracking. IEEE TPAMI, 2023.
- **共同作者：** Zhang, C. et al. WebUOT-1M: Advancing Deep Underwater Object Tracking with A Million Scale Benchmark. NeurIPS, 2024.
- **共同作者；CV4Animals 杰出论文：** Zhang, C. et al. Underwater Camouflaged Object Tracking Meets Vision-Language SAM2. CVPRW, 2025.
- **第一作者；最佳学生论文提名：** Huang, G., Lin, W. & Liu, L. Content-Aware Efficient Learner for Audio-Visual Emotion Recognition. ICSR, 2024.

奖项与证书

- **2026：** ICME ESDD2 环境音频伪造检测挑战赛第 4 名
- **2025：** ACM Multimedia Student Travel Award； CVPR Workshop CV4Animals 杰出论文
- **2024：** ICSR 最佳学生论文提名
- **2023：** 深圳市优秀科技学术论文
- **2018：** 电子科技大学优秀毕业生
- **2017：** ISTQB 软件测试工程师认证