

李锴

年龄：27 岁 · 个人主页：<https://likaiucas.github.io/>
微信：lkbb1213 · 邮箱：kevinlikai@foxmail.com
电话：13258200037



核心亮点

研究方向：多模态大模型（MLLM / VLM）· 视频世界模型（Video World Model）· 强化学习 · 3D 空间理解。两段腾讯研究实习经历，当前在腾讯 Omni Research Team 从事 视频世界模型的记忆机制研究，并同期推进强化学习驱动的稀疏视角 3D 室内场景重建。

学术产出：一作发表 2 篇 IEEE Trans 顶刊（中科院一区 top, CCF-B），4 篇一作在投；累计发表 19 篇，Google Scholar 引用 152 次，h-index 7。

竞赛成绩：WWW 2025（CCF-A）多模态对话系统意图识别挑战赛**决赛第一名**；天智杯 2023 变化检测赛道决赛第二名（30 万奖金）；航天宏图杯 2022 决赛三等奖。

荣誉：IEEE GRSS 分会年度优秀交流奖提名（共 10 人，唯一华人）；全国大学生数学建模大赛国家二等奖；国家级奖学金、香港城市大学全额奖学金。

工程落地：所提算法已在国家多地区测绘局落地应用。

实习与工作经历

腾讯 - Omni Research Team - 研究实习生（多模态大模型 / 视频世界模型） 2025.10 - 今

方向：视频世界模型的记忆机制；强化学习 + 多模态大模型驱动的稀疏视角 3D 场景生成与重建

- **视频世界模型的记忆机制（当前重心，在研）：**视频世界模型能够预测场景如何演化，却普遍缺少对「已经看到过什么」的持久表征。研究如何为生成式视频模型引入场景状态的 **记忆** 表征，使长时序推演保持自治、镜头回到旧视角时与先前生成内容一致，并控制记忆开销不随序列长度无约束增长。
- **强化学习提升跨视角一致性（在研）：**稀疏、无标定视角输入下，重建结果在不同视角间常出现几何与语义不一致。将「可执行场景程序」的执行结果作为可验证的奖励信号，用 **强化学习** 目标替代纯监督训练，直接优化跨视角一致性，而非逐视角单独拟合。
- **Scenix（一作，在投）：**提出 **可执行场景程序（Executable Scene Programs）** 范式，摆脱传统方法对相机参数与密集视图的依赖，让 VLM 直接从少量 RGB 观测中联合预测房间结构、物体属性与语义描述，并实例化为可编辑的 3D 场景。
- **数据与训练：**构造了一批含 6DoF 位姿标注的 3D layout 数据集，训练 LayoutVLM 并配合两个 agentic workflow，实现全流程解耦的 3D 室内场景重建；相比基于 2D 检测 + 深度估计的传统重建链路，具备真正的空间结构理解能力。

腾讯科技 - 视觉算法工程师实习 2021.05 - 2021.09

本科毕业至博士入学期间实习，调研 **3D 目标检测** 与 **3D 重建** 相关算法，为后续在 3D 空间理解方向的研究打下基础。

中国科学院空天信息创新研究院 - 博士研究（偏移摄影建筑物提取） 2021.09 - 今

独立构建 Offset Token 研究主线（详见「论文发表」），完成从概念提出、矢量级提取、问题重构到高效计算的完整闭环，算法已在国家多地区测绘局落地应用。

教育背景

香港城市大学 - 数据科学 - 博士 - 导师：赵翔宇教授 2024.09 - 2027.06

中国科学院大学 - 信号与信息处理 - 博士一导师：赵忠明（中科院遥感所党委书记、副所长）
2021.09 - 2027.06

美国德州大学奥斯汀分校 - 人工智能 - 暑期学校交换 2019.07 - 2019.08

日本上智大学 - 短期访学 2018.01 - 2018.02

电子科技大学（985） - 空间信息与数字技术 - 本科 2017.09 - 2021.06

平均绩点 3.92/4，专业排名 2/53，获推荐免试直接攻读博士学位。

论文发表

以一作发表 2 篇领域 (CCF-B 类) 顶刊 Trans; 另有 4 篇一作在投。我的博士课题研究集中在倾斜相机视角下 (远离星下点), 对于含高度物体 (建筑物) 的提取投影差校正。我的研究围绕着我所提出的概念, Offset Token (偏移量词元), 展开。完成了从概念提出、矢量级别提取、新问题重构、高效计算等方面的研究。我的算法已在国家各地区测绘局得到落地应用。

已发表论文:

[1] Li, Kai, et al. "Prompt-Driven Building Footprint Extraction in Aerial Images with Offset-Building Model." *IEEE Transactions on Geoscience and Remote Sensing* (2024).

论文核心: 首次提出偏移量词元概念。

(1) 我们首次提出并验证 Offset Token 的概念, 用于描述建筑物在倾斜视角下的形变。为领域实现了从 CNN 小模型, 到大模型研究的理论跨越。(2) 我们发现了建筑物投影差校正过程中的普遍规律: 矮小的建筑物, 更容易预测投影形变的长度, 但投影形变方向难预测; 高层建筑物, 形变方向明显易于预测, 但形变长度难以估计。基于此现象, 我们结合 NMS 的算法思想, 提出了 D-NMS 后处理算法, 使用高层建筑物的方向去校正低层建筑物的方向。

[2] Li, Kai, et al. "PolyFootNet: Extracting Polygonal Building Footprints in Off-Nadir Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* (2025).

论文核心: 从偏移量词元出发, 首次实现了偏移摄影下的建筑物结构矢量级提取。(1) 我们首次实现了在远离星下点影像中, 自动化的矢量级建筑物轮廓勾勒。(2) 依据建筑物校正的普遍规律, 我们结合 Nadaraya-Watson Regression 提出了 self-offset attention (SOFA) 机制去提升形变的预测能力。(3) 我们分析了建筑物轮廓预测在泛化性实验上的普遍失败案例, 提出使用建筑物本身的多解性去解决建筑物轮廓的语义混淆。

在投论文:

[3] Li, Kai, et al. "Scenix: Sparse-View 3D Scene Reconstruction via Executable Scene Programs"

论文核心: 腾讯期间工作: 提出新的范式, 让 VLM 直接从房间的 3D 布局中学习, 从而构建了一种只需要几张拍照就可以完成从 3D layout 到 3D 场景的全自动重建。核心: 训练学习 layoutvlm+ 两个 agentic 的 workflow

传统的基于 2D 图像分析的 detection、深度估计等算法进行的 3D 重建无法真实地理解空间, 我们直接构造了一批包含 6DoF 的 3D layout 数据集, 实现了全程解耦的 3D 室内场景重建。

[4] Li, Kai, et al. "DragOSM: Extract Building Roofs and Footprints from Aerial Images by Aligning Historical Labels."

论文核心: 从偏移量词元出发, 以去噪的视角重构了历史标签与更新影响的不匹配问题。我们对于建筑物相关问题提出了新的思路, 从开放世界地图角度, 重新思考建筑物相关问题。我们可能不需要每次都从头开始预测建筑物的结构: 首先, 开源地图中的各类型标注囊括了世界大部分的地区; 其次, 这些标签虽然质量不统一 (普遍存在位置偏差错误), 但提供了人工标注的清晰轮廓。我们可以通过校正这些位置偏差错误, 获得建筑物的轮廓标签。(1) 结合去噪模型的思想, 我们将由于复杂的系统性位置错误偏差理解为一个时间序列的高斯过程。(2) 我们提出 Alignment Token 的概念去引导和校正这些位置偏差的错误, 并提出 DragOSM 模型, 一个标签级自回归的模型。(3) 结合当下对于大语言模型自回归模型 Test-time Scaling Law 的思潮, 为 DragOSM 设计了类似的 TTA 策略, 提高推理时模型的预测能力。

[5] Li, Kai, et al. "ObliCity: A Benchmark and Baseline for Roof-to-Ground Projection Displacement Correction."

论文核心: 从偏移量词元出发叠加数学推理, 倡议将偏移量学习做成新任务。

针对论文 [1] 中发现的普遍规律, 针对长度错误部分, 我分析了实验上的结论: 对于高层建筑物的形变预测, 多数情况下的长度错误为相比于预测真值更短; 而低层建筑物的预测值普遍偏长。于是, 受启发于当下生成模型 Flow Matching 的思想, 将建筑物的投影差校正问题建模为一个 ODE 偏微分方程组, 并使用条件归一化流 (Conditional Normalizing Flow) 的思想, 训练模型, 以模拟整个人工标注的过程, 获得建筑物更加精准的长度形变校正。

[6] Li, Kai, et al. "LoDEOT: Low-Dimensional and Efficient Offset Tokens for Building Footprint Extraction from Off-Nadir Imagery"

论文核心: 之前的偏移量词元设定都是按照 CV 中的惯例设计, 本文从建筑物偏移摄影三维成像的数学原理与直觉出发, 重新定义了一种高效率的偏移量词元, 并自然引入去噪的思想进行训练。

其他参与发表论文:

- Zhou, Z., **Li, K.**, & Deng, Y. (2026). Remote-Sensing City Layout Extraction with MLLM. arXiv:2608.16484. 【指导学生一作，本人通讯/二作】
- Meng, Y., Deng, L., Xi, Z., Chen, J., Chen, J., Yue, A., Liu, D., **Li, K.**,... & Sun, X. (2025). IRSAMap: Towards Large-Scale, High-Resolution Land Cover Map Vectorization. IEEE Transactions on Geoscience and Remote Sensing. 【中科院一区 top, CCF-B, IF:8.6】
- Wang, C., Xi, Z., Liu, D., Feng, Y., Deng, Y., **Li, K.**, Chen, J., Chen, J., & Meng, Y. (2025). PCP: A Prompt-Based Cartographic-Level Polygonal Vector Extraction Framework for Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing. 【中科院一区 top, CCF-B, IF:8.6】
- Li, Z., Wu, B., Zhang, Y., Li, X., **Li, K.**, & Chen, W. (2025, May). CuSMer: Multimodal Intent Recognition in Customer Service via Data Augment and LLM Merge. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 3058-3062). 【CCF-A，比赛决赛第一】
- Wang, C., Chen, J., Meng, Y., Deng, Y., **Li, K.**, & Kong, Y. (2024). Sampolybuild: Adapting the segment anything model for polygonal building extraction. ISPRS Journal of Photogrammetry and Remote Sensing, 218, 707-720. 【中科院一区 top, IF:12.2】
- Chen, W., Deng, Y., Jin, W., Chen, J., Chen, J., Feng, Y., Xi, Z., Liu, D., **Li, K.**, & Meng, Y. (2025). DGTRSD and DGTRSClip: A Dual-Granularity Remote Sensing Image-Text Dataset and Vision-Language Foundation Model for Alignment. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
- Wang, H., Li, Z., Zhang, Y., He, Z., Jiang, L., **Li, K.**, et al. (2026). ConsistNav: Closing the Action Consistency Gap in Zero-Shot Object Navigation with Semantic Executive Control. arXiv:2605.09869.
- Ma, J., Wang, S., Yang, J., Hu, J., Liang, J., Lin, G., **Li, K.**, & Meng, Y. (2025). SayAnything: Audio-Driven Lip Synchronization with Conditional Video Diffusion. arXiv:2502.11515.
- Chen, W., Chen, J., Deng, Y., Chen, J., Feng, Y., Xi, Z., Liu, D., **Li, K.**, & Meng, Y. (2025). LRSCLIP: A Vision-Language Foundation Model for Aligning Remote Sensing Image with Longer Text. arXiv:2503.19311.
- 本科期间另参与 IGARSS 2019 会议论文 3 篇 (Land Price Assessment Based on Deep Neural Network; Urban Functional Regions Discovering Based on Deep Learning; A WordNet-Based Geospatial Web Services Search Method Supporting Quality of Service Constraints)。

竞赛与荣誉

博士阶段

学术竞赛：

[1] WWW2025 (CCF-A)，多模态对话系统意图识别挑战赛，决赛第一名。 (1) 我发现在多模态大语言的训练和传统的计算机视觉模型一样，尤其是在训练中针对图像进行多尺度和旋转等变化可以显著提升模型的预测能力。(2) 结合队友的模型融合策略，我们的模型在最终的多模态电商场景任务中获得了第一名。

[2] 天智杯 2023，变化检测赛道 (30 万奖金)，决赛第二名。 该比赛工程性较强，策略使用常见的计算机视觉增强，以及余弦退火训练策略。主要难度为现场限时 2 天，要把一台工作站，首先完成在离线状态下，从操作系统到运行环境的配置，再完成算法类比赛。

[3] 航天宏图杯 2022，道路交叉口及道路提取赛道 (3 万奖金)，决赛三等奖我们发现道路的提取存在本身标注的主观性，*i.e.*，标注人员对于不同粗细的道路定义不同，有的细小道路被标注，而有的却没有。我们创新地提出使用图像压缩策略，去模糊化道路的不同粗细，通过实验测试，找到了最佳下采样策略。由于图像的压缩，我们在性能赛道获得了第一名，综合成绩决赛第四。

[4] IGARSS 2025 三分钟论文竞赛(3MT),决赛 Top 10(以 OBM 参赛),并获 IGARSS 2025 Travel Grant 资助。

荣誉与奖学金： IEEE GRSS 分会年度优秀交流奖提名 (共 10 人，唯一华人)、2025 年三好学生、企业奖学金 (共 2 名)、香港城市大学全额奖学金。

本科阶段

学业成绩： 平均绩点：3.92/4，专业排名：2/53，获评：成电留学精英 (每年全校仅 10 人) 获得学校与新东方各类资助。

学术竞赛： 全国大学生数学建模大赛国家二等奖；在校内获得学术类比赛三等奖以上 12 次。

奖学金： 国家级奖学金 (1 次)、优秀学生奖学金 (3 次)。

学生指导与教学

学生指导（与香港城市大学赵翔宇教授联合指导）

- Jiajun Zhang, 香港城市大学硕士，下一站：北京大学博士生。
- Zigan Zhou, 香港城市大学硕士，曾于中国科学院实习；指导其完成一作论文 *Remote-Sensing City Layout Extraction with MLLM* (arXiv:2608.16484)。

学生指导（与中国科学院大学陈静波教授联合指导）

- Weizhi Chen, 中国科学院大学（空天信息创新研究院）；研究方向为 CLIP 图文对齐，发表 2 篇 **IEEE JSTARS** 论文；现任中公教育大模型算法研究员。

助教经历

中国科学院大学《最优化方法》助教（赵振江教授）；《学术英语》助教（Prof. James）。

其他技能

英语：雅思 6.5（口语 6.5），六级 547。