

张益新 (Yixin Zhang)

✉ zyxcambridge@gmail.com

☎ 17521398109

📍 上海

算法工程师

工作经历

VLN 部署算法工程师 | 某上市公司机器人子公司 | 2025.9 - 至今

- 项目名称：视觉语言导航（VLN）算法部署与自动跟随机器人系统
- 核心算法能力：
 - 长距离规划：支持超过 150 米的长距离导航，实现复杂环境下的路径规划
 - 零样本泛化：实现陌生环境自主适配能力，无需预训练即可在新场景中稳定运行
 - 密集障碍穿行：突破密集障碍物挑战，在障碍间距小于 50cm 的极端场景下仍可安全通过
 - 动态避障：实时感知并避开移动障碍物，保障机器人安全运行
- 模型部署与优化：
 - 在 NVIDIA Thor 和 Orin 平台上成功部署 VLN 模型，实现端到端推理优化
 - 通过模型量化、算子融合、内存优化等技术手段，显著降低时间延迟
 - 完成自动跟随机器人功能开发，实现稳定可靠的跟随性能
 - 成功完成 42 次关键测试验证，系统稳定性和可靠性达到量产标准
 - 具身智能算法研究部署落地：在 NVIDIA Thor 和 Orin 平台上实现具身智能算法的工程化部署
- 项目战果：
 - 在几乎“零基础”的条件下，独立完成软硬件全链路打通，包括 Thor 硬件配置、算法视频调优等全部工作
 - 成功实现跟随功能稳定运行，攻克了项目中最难啃的技术难题
 - 建立了完整的部署流程和验证标准，为后续规模化应用奠定基础
- 12.18 演示救场：
 - 在演示现场语音网络全面崩溃的紧急情况下，凭借 Plan B 预案和完整的控制环境备份
 - 现场统筹协调团队，稳定团队心态，确保第二次演示顺利通过
 - 在高压环境下突破心理阈值，展现了出色的项目把控和应急处理能力
- 项目规划与团队建设：
 - 在 demo 成功运行的基础上，制定 Thor 项目规模化扩展计划
 - 申请团队扩充，规划基础工作分工，聚焦核心技术攻坚
 - 从“单兵作战”向“集团军冲锋”转变，争取更大的技术突破和商业价值

智能体算法研发（竞赛期间） | 自由职业 | 2025.2 - 2025.9

- 竞赛成就：NeurIPS 2025 智能体工具增强推理 Workshop - CureBench 国际智能体评测竞赛全球第二名 (Top 2)
- 论文发表：作为共同第一作者发表论文“CureAgent: A Training-Free Executor-Analyst Framework for Clinical Reasoning” (arXiv:2512.05576)
- 项目背景：参与 NeurIPS 2025 顶级会议 Workshop，构建生物医学领域专业 Agent 系统。针对小规模 LLM 临床智能体的上下文利用失败问题，提出 Executor-Analyst 框架

- 技术架构: CureBench + TxAgent RL Framework - ART Training
- 核心工作:
 - 设计并实现 Executor-Analyst 模块化架构, 将工具执行与临床推理解耦, 缓解单体模型的推理缺陷
 - 提出分层集成策略 (Stratified Ensemble), 通过保留证据多样性解决信息瓶颈问题
 - 发现上下文-性能悖论和动作空间维度诅咒等关键缩放洞察
 - 利用 Agent 和 Test-time Scaling 技术构建用药助手智能体
 - 调用 FDA、OpenTargets、PubMed 等生物医学工具进行智能体工具增强推理
 - 完成《自进化智能体-动态记忆与持续运行的架构实践》一书初稿撰写

深度学习算法工程师 | 安波福中央电气 (上海) 有限公司 | 2024.8 - 2025.2

- 项目名称: 通用障碍物感知算法研发与端到端量产预研
- 模型部署工作分层:
 - 部署一个网络
 - 部署多个网络, 在一个芯片上进行性能优化:
 - * 2 个 backbone + 3 个 head, 多任务流水线
 - * 共享内存和队列, instance bank
 - 统一适配多个芯片框架 (GPU+ASIC+FPGA):
 - * 多芯片平台调度框架
 - * 一套代码适配多个芯片
- 参与 L2++ 无图端到端网络设计方案, 主导通用障碍物 OCC 分支的网络结构设计和优化
- 负责芯片选型, 完成三种芯片的 benchmark 测试, 设计端到端网络多任务调度框架
- 建立部署工作标准 SOP 和 work flow, 独立完成 OD 和 OCC 分支在 2 种芯片的 AI 模型部署
- 异步调度: 设计异步流水线, 实现 CPU 和多个 AI 加速核心 (NPU/VP) 最大限度重叠工作, 缩短总体完成时间
- 奔向空闲: 将 1000ms 任务压缩到 28ms (14ms*2), 使芯片更快进入低功耗 idle 状态
- 系统级优化 - 多任务调度框架: 设计串行、并行或流水线并行的 Pipeline 编排多个模型执行顺序
 - 车间 A (Stage 1): 特征提取中心, 配备两台并行机器分别计算 Backbone_1 (BEV) 和 Backbone_2 (Temporal)
 - 车间 B (Stage 2): 感知分析部, 配备两台并行机器分别计算 Head_1 (OD) 和 Head_2 (Map)
 - 车间 C (Stage 3): 决策融合站, 负责最后的 Head_3 (Predict) 计算
- 部署与验证工作流 (量产规范):
 - 模型导出与分段: 将 PyTorch 模型导出为 ONNX 并按 DAG 节点分割为独立文件 (backbone1.onnx, head1.onnx 等)
 - 模型编译: 使用硬件厂商工具链将 ONNX 编译为二进制 bin 文件, 完成算子融合与量化优化
 - 分阶段验证:
 - * IO 对齐: 确保部署程序与仿真脚本输入逐比特一致
 - * 单节点验证: 对比 bin 文件输出与 PC 端 ONNX 结果
 - * 多帧端到端对齐: 验证完整 APP 业务指标 (mAP, IoU) 与黄金参考模型一致性
 - 最终集成与交付: 封装为符合规范的 5 个核心 API (Init/Run/Release/GetResult/GetStatus)
 - 核心优化指标: 时间延迟、吞吐量、内存带宽 (减少占用)、功耗 (W4A4)、算力占用 (软硬一体)
 - 适配车机芯片: 英伟达 Orin、地平线 J5/J6、CV3

模型部署负责人/高级软件架构 | 图达通智能科技 (上海) 有限公司 | 2022.6-2024.04

- 招聘组建高效的深度学习部署团队, 完善算法工程落地流程
- 基于 Jetson 平台实现 AI LiDAR 模型实时推理, 时间延迟降低 4 倍
- 开发 ADAS 代码补全 VSCode 插件, 提升自动驾驶算法开发效率
- 大模型推理部署探索, 成功部署 Ollama、llama.cpp、vllm、TensorRT-LLM、mlc-llm 等框架
- 在 Mac M1 上成功运行 yi-34B 模型, 在 Android 平台使用 mlc-llm 运行 llama3-8b 模型
- 基于 NVIDIA Drive Sim 完善 MLOps 流程, 产生 10W 帧仿真数据集

- 实现模型仿真数据和真实数据的联合训练，精度提高 10 个百分点
- **【模型部署】**
 - 性能分析工具应用：使用 Nsight Systems 和 Nsight Compute 定位部署瓶颈，解决关键性能卡点
 - lidar 车端框架构建：实现激光雷达点云检测与语义分割网络部署迭代，包括 rangeimage 和 pillar voxel 网络实时推理，设计端到端统一检测分割框架
 - NVIDIA 技术协作：与 Jetson 团队建立直接沟通渠道，分析模型瓶颈并获取优化支持，通过任务依赖关系优化将整体延迟降低 75%（四分之一）
 - 地平线生态对接：为地平线 J5 平台争取硬件和软件开发支持，完成感知算法全量移植
 - 技术难题突破：与 SPConv 核心开发团队协作，解决自定义稀疏卷积算子的部署挑战

高级算法工程师 | 上海雪湖科技有限公司 | 2019.05-2022.05

- 从 0 到 1 搭建算法团队，构建 V2X 感知算法工程化落地流程
- 设计 FPGA 混合量化方案，精细化到乘法和加法运算级别
- 成功量产 MEC 设备 100 台以上，软硬一体变现 500 多万，服务近 10 家客户
- **【难点与解决方案】**
 - 精度损失问题：针对量化导致的 30 个点精度损失，提出混合量化策略
 - 量化方案：实现 PTQ(Post-training Quantization)，参考 QAT(Quantization-aware Training) 全流程
 - 量化算法对比：系统评估 minmax、KL 散度和 MSE 等量化方法的误差分布特性
 - 工具开发：自主研发精度比对工具，可视化不同量化参数下的误差热力图
 - 训练推理一致性：从量化原理公式出发，解决训练/推理精度偏差问题
 - 客户信任构建：通过理论推导和实测数据双重验证，解决时间延迟质疑
- **【实现方法】**
 - 核心优化流程：构建哈希表 → 生成 Rulebook → Gather 输入特征成稠密矩阵 M_{in} → 执行 GEMM → Scatter 输出结果
- **【lidar MLOPS】**
 - 分层策略：关键特征提取层保留浮点精度，分类头和后处理层量化至 INT8
 - 量化算法：采用 KL 散度校准方法，使用校准集进行数据分布匹配
 - 网络分解：将复杂卷积操作分解为基本乘加运算，适配 FPGA 整数计算架构
 - 基于 ZCU104 IP 域控制器构建深度学习部署流程，实现 10 倍速度提升
- **【技术落地】**
 - 根据项目和现有公司解决方案创建 V2X 解决方案文档，提供现场解释，获得 10+ 客户
 - 在项目现场进行现场开发，在中国各地城市实施 10+ 试点项目
 - 推动 FPGA lidar 加速 IP 向速腾、禾赛、Ouster、Livox 等激光雷达公司的适配
- **【项目收入证明】**
 - 基于 AMD Xilinx ZCU104 大规模生产 100+MEC 设备，实现 200 万元 + 软硬件综合收入
 - 拥有近 10 个客户，建立长期合作关系

算法工程师 | 深兰科技(上海)有限公司 | 2018.05 - 2019.02

- 汇报对象：技术经理
- 基于关键点的目标检测算法工程落地：
 - 实现基于 OpenPose 的人体关键点检测算法，自主完成 C++ 推理流程
 - 算法改进包括高斯响应增强、Heatmap 半径调整和中继监督机制
 - 创新实现基于峰值的 NMS 方法替代传统 IoU 方式
- 模型部署与优化：
 - 在 P100 GPU 上部署模型，单卡支持 28 个模型并行推理
 - 将 OpenPose 推理时延优化至 500ms 以内
- 算法验证：

- 构建密集场景测试集，在复杂环境下达到 99% 关键点识别准确率
- 其他成果：
 - 获得 3 项发明专利授权
 - 开发基于关键点的云端商品识别服务

深度学习工程师 | 友邦网络科技有限公司 | 2017.12 - 2018.04

- 汇报对象：技术经理
- ADAS 系统开发：
 - 研发行人和车辆目标检测算法，结合图像分割技术实现车道线检测
 - 在 RK3399 等嵌入式平台进行网络优化，显著提升检测速度
 - 使用 TensorFlow Lite 进行 Android 端模型移植和量化优化
- 人体姿态识别：
 - 基于 MobileNet 架构实现轻量级人体关键点检测
 - 分别使用 Caffe2 和 TensorFlow 框架完成算法实现
 - 成功移植至 Android 设备，支持实时推理

教育背景

网络工程学士 | 北华航天工业学院 | 2010.09-2014.06

核心成就与著作

论文发表：

- **CureAgent: A Training-Free Executor-Analyst Framework for Clinical Reasoning** (arXiv:2512.05576). Ting-Ting Xie, **Yixin Zhang**. NeurIPS 2025 Workshop - CURE-Bench Competition 2nd Place Solution. [arXiv:2512.05576](https://arxiv.org/abs/2512.05576)

著作出版：

- 撰写《自进化智能体-动态记忆与持续运行的架构实践》一书，目前已完成初稿并进入出版社三审三校阶段，计划由电子工业出版社出版

竞赛成就：

- NeurIPS 2025 智能体工具增强推理 Workshop - CureBench 国际智能体评测竞赛全球第二名 (Top 2)

项目与成就

【项目名称：通用障碍物感知算法研发与端到端量产预研】

- 项目目标：研发通用障碍物检测算法，基于物理规律（深度分布和语义约束）实现开放场景泛化感知，预研车规级低功耗平台（16W）端到端部署方案
- 核心技术突破：

- 范式转变：从‘学习物体形状’转向‘验证物理规律’，通过深度分布建模与语义约束验证，实现‘不识形状却感知存在’的泛化能力
- 场景优化：强化建筑区域垂直边缘梯度约束，降低悬挂物误识别风险
- 负障碍物检测：利用高度突变和点密度异常等几何特征，实现坑洼、沟渠等负障碍物识别
- 核心挑战与解决方案：
 - 挑战：端到端感知（OD+Map+Predict+Plan）计算量与 16W 功耗限制的矛盾
 - 解决方案：
 - * 1) 模型轻量化：优化 OD 与 Map 模型参数结构
 - * 2) 混合量化策略：FP16+INT8 混合量化，结合离线初始化与在线校准，精度损失 <2%
 - * 3) 算子定制：开发体素化、可变形聚合等关键算子，提升计算效率
- 关键成果：
 - 端到端推理速度提升 10 倍，成功部署至目标平台
 - 混合量化方法申请技术专利
 - 验证基于物理规律感知范式在开放场景的可行性
- 前瞻性研究：
 - 混合精度训练梯度不匹配问题解决方案
 - 端侧内存带宽限制下的算子融合方案

【项目名称：lidar 推理框架】

- 难点与解决方案
 - 精度损失控制：针对量化导致的 30 个点精度损失，提出混合量化策略
 - 量化方案实现：搭建 PTQ/QAT 全流程，支持 minmax、KL 散度、MSE 三种校准算法
 - 误差分析工具：自主研发精度比对平台，可视化不同量化参数下的误差热力图
 - 训练推理一致性：从量化原理公式出发，解决浮点/整数计算偏差问题
 - 客户信任构建：通过理论推导 + 实测数据双重验证，将延迟降低至客户要求的 1/3
- 技术实现
 - 核心优化流程：构建哈希表 → 生成 Rulebook → Gather 输入特征 → 执行 GEMM → Scatter 输出结果
 - 量化算法对比：系统评估三种方法在不同层的误差分布，选择最优组合策略
- 商业价值
 - 基于 AMD Xilinx ZCU104 大规模生产 100+MEC 设备
 - 实现 200 万元 + 软硬件综合收入，服务近 10 家客户
 - 量产陕重汽

【项目名称：NeurIPS 2025 CureBench 医药 Agent 系统（CureAgent）】

- 项目角色：NeurIPS 2025 Workshop 项目负责人/核心开发者、论文共同第一作者
- 项目背景：参与 NeurIPS 2025 顶级会议 Workshop，构建生物医学领域专业 Agent 系统。针对当前基于小规模 LLM（如 TxAgent）的临床智能体存在的上下文利用失败（**Context Utilization Failure**）问题，提出 Executor-Analyst 框架，将工具执行的语法精确性与临床推理的语义鲁棒性解耦。
- 技术架构：CureBench + TxAgent RL Framework - ART Training
 - **Executor-Analyst** 框架：模块化架构，将专门的 TxAgent 执行器（Executors）与长上下文基础模型分析师（Analysts）协同工作，缓解单体模型的推理缺陷
 - 分层集成策略（**Stratified Ensemble**）：通过保留证据多样性，显著优于全局池化，有效解决信息瓶颈问题
 - 训练无关架构工程：无需昂贵的端到端微调，在 CURE-Bench 上达到最先进性能
- 核心技术突破：
 - 上下文-性能悖论（**Context-Performance Paradox**）：发现推理上下文超过 12k tokens 会引入噪声导致准确性下降
 - 动作空间维度诅咒：工具集扩展需要分层检索策略

- 调用 FDA、OpenTargets、PubMed 等生物医学工具进行智能体工具增强推理
- 实现 Test-time Scaling 技术，提升智能体在测试阶段的性能
- 基于强化学习框架 (TxAgent RL Framework) 进行智能体训练与优化
- 项目成果：
 - 在 CureBench 国际智能体评测竞赛中获得全球第二名 (Top 2)
 - 发表论文至 arXiv (arXiv:2512.05576)，代码已开源
 - 成功构建了可实际应用的生物医学领域智能体系统，为下一代可信 AI 驱动治疗学提供可扩展、敏捷的基础

大模型微调与部署项目

- 在 AWS 上部署 Baichuan2 模型为 OpenAI 兼容格式，实现 Lora 微调
 - 使用 LLaMA-Factory 在阿里云平台进行大模型微调与优化
 - 将 Start Code 模型部署至 HuggingFace 作为 API 服务
 - 优化 8B 模型性能达到相当于 671B 模型水平
-

学术与竞赛经历

- 2025 年: NeurIPS 2025 智能体工具增强推理 Workshop - CureBench 国际智能体评测竞赛全球第二名 (Top 2)
 - 2021 年 10 月: CVPR Workshop - 3D 目标检测算法 (第九名)
 - 多次黑客马拉松获奖：
 - 2023 年 10 月: 百川黑客马拉松 - 美国中医获客 Agent 项目 (特别奖)
 - 2023 年 9 月: Google I/O Hackathon - Stable Diffusion 计算共享 (三等奖)
 - 2023 年 7 月: 世界 AI 黑客马拉松 - 领导力追踪与训练 (二等奖)
 - 2019 年 9 月: AngelHack 上海 - 领导力追踪与训练 (一等奖)
 - 2017 年 7 月: 全球 AI 黑客马拉松 - 假新闻检测 (冠军)
-

技术社区参与

- Google 机器学习开发专家，连续 5 年每年举办 4 场技术讲座，影响 11,874 人