杨豪 Yang Hao

模型量化与部署负责人 · 机器学习基础架构 · 自动驾驶与端侧 AI

邮箱 howellyang@buaa.edu.cn
主页 howellyang.com/zh
GitHub github.com/Howell-Yang

7 年以上把深度学习模型从训练落地到量产车规芯片的经验。在轻舟智航负责约 19 个量产行车、泊车模型（BEV 感知、Occupancy、矢量地图、时序跟踪、预测规划与端到端）的量化与端侧部署，平台覆盖地平线 J5/J6M/J6P、英伟达 Orin、高通和爱芯。技术上深入到定点算子的 LSB 级误差、量化感知蒸馏和板端耗时建模；工程上把经验沉淀为工具链、发版门禁和团队流程：带领 6 人团队，并担任新芯片量产项目技术负责人。

≈19	2 → 0.5	⩽ 2 LSB	0.7%	4,273	2 周 → 3–5 天
个量产模型的量化与部署 行车 + 泊车	**个点的量化掉点** PTQ → QAT 5 个感知模型	**查表算子最大误差** 600 个模块 表数 −85%	**板端耗时预测误差** 预测 19.38 实测 19.25 ms	**次量产模型转换** 单次编译 30–60 → 5 分钟	**单模型量化适配周期** 三层解耦 + AI 自动适配

2012 – 2019 北航 本硕
自动化
模式识别

2019.04 依图科技
人脸识别引擎
转向芯片量化

2021.07 腾讯地图
TensorRT INT8
端侧感知 SDK

2022.10 加入轻舟智航
QNN Web
量产发版

2024 晋升 L5
J6M 量化工具链
AutoQuant

2025 量化方向负责人
带团队
蒸馏 QAT

2026 实线带队新芯片技术负责人
大规模 QAD
平台化框架

工作经历与代表项目

轻舟智航 QCraft　模型量化与部署负责人（感知 ML-Infra）

2022.10 – 至今

入职为机器学习基础架构工程师；2024 年晋升 L5；2025 年起负责量化方向并带团队（2026 年起实线）。

代表项目在部署链路上的位置

浮点训练 ⑤ 训练提速 共享 backbone → 量化 ①② PTQ / QAT / QAD 与浮点逐 case 一致 → 导出 IR ③ 定点查表 算子对齐 → 多芯片编译 ④ QNN Web 四个芯片平台 → 对齐与门禁 ③ 逐跳比对 证据分级 → 板端与实车 ⑤⑥ 耗时预测 新芯片适配

① 与浮点模型逐 case 一致的量化：四代算法演进

2023 – 2026

问题　量产对回退近乎零容忍，AEB 目标晚一帧检出即失败；而只用任务损失的 QAT 会改变 precision/recall 工作点，交付的实际上是一个“重训模型”。

- **Scale-search PTQ**（2023，已申请专利）：以与浮点输出的 cosine 相似度为目标，分级搜索量化参数：校准方式 × 参数组合 → 逐层前/后向搜索 → 逐层量化。红绿灯检测模型上，默认校准的 mAP 为 0.747，逐层量化恢复到 **0.803（浮点 0.800）**。
- **LSQ AutoQuant**（2024）：面向 CNN + Transformer 混合结构（矢量地图、时序跟踪），学习量化步长，用 AdaRound 式取整约束权重漂移，并以浮点中间特征监督；训练前做跨层均衡（CLE）与 LayerNorm–Linear 融合。
- **蒸馏 QAT → 大规模 QAD**（2025–2026）：浮点模型做教师、PTQ 模型为起点，任务损失 + 特征/输出蒸馏。在端到端规划模型上以单变量对照实验（25 项关键指标、388 个公共标量）确认：场景边界重置时序状态、教师独立历史、训练后期权重锚定是必要条件。

结果　5 个量产感知模型量化掉点由约 2 个点降至约 **0.5 个点**；端到端规划模型的量化版本进入实车测试；模型 owner 已开始自助完成 auto-QAT 发版。

② 预测规划模型的量化误差累积攻坚（QuestNet v3）

2025 – 2026

问题　模型量化后轨迹输出异常，PTQ 与常规 QAT 均失效，量产交付受阻。

- 用逐层误差与累积误差分析定位出误差在深层网络**逐级累积**的机制，对敏感层与鲁棒层做差异化处理；用实验否定“冻结部分层再微调”的折中方案，改为**两段式量化 + 全量 QAT**。
- 基于对芯片算子特性的理解，推动全链路 **INT16** 量化，为误差累积留出数值余量。

结果　轨迹异常消除，方案合入主线并完成发版验证；“中融合模型量化必须全量训练”成为团队方法论。

③ 定点数值一致性：自适应查表与分层对齐

2024 – 2026

- **自适应查表**：rsqrt、reciprocal、exp、sigmoid 等非线性算子在 NPU 上以定点分段查表实现。算法以芯片真实定点输出码（而非浮点近似）度量误差，保留原始 6 表边界，对误差最大的段递归二分，预算用尽仍超限则阻断导出。4 个量产计算图共 600 个查表模块：平均 **0.217 LSB**、最大 **2 LSB**（均匀 126 表方案最大 11,381 LSB），表数量减少 **85%**。
- **分层对齐**：Float → QAT → 内部 IR → LibTorch → HBIR → HBM → 板端后处理，每个阶段只与前一阶段在相同输入下比对，让每个不一致只可能来自一处；定位了归一化层 rsqrt 查表分段点、upsample align_corners 不一致、位姿精度等问题。
- **发版门禁**：结论按证据分级（源码 / 编译 / 仿真 / 板端 / 实车），新特性默认关闭（feature flag），历史量化问题回归集每次发版前重跑；在 3 个车企项目中把量化问题拦截在发版之前。

4 工具链与平台：从一键转换到平台化量化部署 2022 – 2026

- **QNN Web** 一键多芯片模型转换服务（gRPC 前后端分离、按 ONNX 计算图自动生成配置、TensorRT timing cache；后续加入任务数据库、队列和面向 AI Agent 的 MCP 接口）：累计 **4,273** 次量产转换，单次编译 30–60 分钟 → 约 5 分钟，人工耗时减少约 80%。
- 把多平台量化拆为**通用量化层 / AI 自动适配层 / 平台适配层**：单模型量化适配周期由约 2 周缩短到 **3–5 天**。带领团队打通高通（模型拆分与耗时优化，推理耗时降低 50% 以上）、TensorRT QDQ INT8 和地平线 J6P 上的 VLM 量化链路。
- 2026.11 起牵头**平台化量化部署框架**的目标架构（设计稿，路线图至 2027 Q2）：以定点仿真结果为唯一真值，用七份带版本的契约（产物身份、IR 元信息、IO 契约、真实 IO 包、精度方案、平台能力、门限）连接训练、转换、车端三个代码库，逐跳门禁默认阻断，并以“敏感度 × 板端时延代价”的闭环自动规划混合精度。

5 耗时建模、模型架构与训练效率 2025 – 2026

- **结构级耗时模型**：按模块结构预测板端耗时，Occupancy 模型预测 19.38 ms、实测 19.25 ms（误差 **0.7%**）；在写代码之前就否定了“扩大范围 + 10 Hz”直接上线的方案，并验证 MapFormer query 合并可节省 2.1–3.3 ms。
- **OneModel 二期**（进行中）：5 个感知任务、9 路相机共享图像 backbone 与 neck，按帧率折算预计每帧节省 5–9 ms；对 5 个任务 checkpoint 的特征做 PCA 冗余分析确定 backbone 结构，并以通道对齐 + 闭式线性求解合成预训练权重；采用 ReLU6 + 代理梯度的量化友好激活，修复 backbone 感受野中心偏移。
- BEV Occupancy 训练迁移到 PyTorch 2.x：单 iter 1.56 s → 0.71 s（**−54%**，V100、batch 16、AMP），显存降低 9%，完成 1–112 卡扩展性压测。
- **端到端驾驶模型架构设计**（提案阶段，尚未训练）：camera-first 的 token 化方案，9 路相机经共享编码器 token 化，流式 Transformer + 位姿感知记忆，以曲率/加速度节点表示的多模态规划头，并从设计之初给出蒸馏 + INT8 量化的落地路径。

6 团队管理与项目交付 2025 – 2026

- 带领 **6 人**量化部署团队（2025 年起虚线、2026 年起实线）：约 19 个量产模型各设 owner 与 backup，季度目标 + P0/P1 优先级，按“紧急度 × 技术难度”分配任务。2026 年 Q2 四个量产项目并行、约两周一版的节奏下，未发生因量化导致的发版延期。
- 2026.06 起任**爱芯芯片项目技术负责人**（先在 M97 700 TOPS 开发板验证，目标芯片 M95 350 TOPS）：分阶段适配计划、模型适配与量化、路测问题聚类分发；城市 NOA v4 方案已在测试车上运行，关键模型无掉帧。
- **AI 提效**：飞书机器人 + 编程 Agent Skill 自动完成新模型特性的量化适配（60 人天项目）；维护 33 篇文档的知识库，组织月度量化分享会，培养成员独立负责模型。

此前经历

腾讯地图 Tencent Maps　算法工程师（CSIG 地图平台部） 2021.07 – 2022.10

- **TensorRT INT8 量化部署**：搭建云端 GPU 模型的量化部署流程与转换工具（偏置校正、权重均衡、逐层校准），与浮点 cosine 相似度高于 0.999，推理耗时降低 20%，GPU 占用与显存各降低 10%，节省约 300 万元算力成本；校准工具开源为 onnx2trt。
- **端侧感知 SDK**：AR 导航多流程并行框架 + 自动 profile，CPU 占用降至 46%、单帧端到端 80 ms，在量产车型上线；采集脱敏使用 38 MFLOPs 人脸/车牌检测模型（recall 74.0、precision 90.6）+ DeepSORT/光流跟踪。

依图科技 Yitu Technology　算法工程师 2019.04 – 2021.07

- 开发 C++ 人脸识别引擎，组合多个模型并优化组合策略，兼顾吞吐与识别率，应用于地铁刷脸通行等项目；公司自研 AI 芯片后转向模型量化，负责量化工具、精度评测与 model zoo。

专业技能

量化	PTQ、QAT、QAD 与蒸馏、LSQ / AutoQuant、AdaRound、CLE、scale search、INT8/INT16 混合精度、定点查表算子、敏感度分析、逐层误差定位
部署平台	地平线 J5 / J6M / J6P（Horizon plugin、HBIR/HBM、BPU）、英伟达 Orin（TensorRT、自定义 plugin、QDQ）、高通（QNN）、爱芯 M95 / M97
模型类型	BEV 检测、Occupancy、矢量地图、稀疏时序检测与跟踪、预测规划、端到端、多任务感知、VLM
训练与工程	PyTorch 1.x/2.x、DDP、AMP/BF16、多机训练（英伟达 V100/H20、阿里 PPU）；Python、C++、ONNX、TorchScript/LibTorch、gRPC、MCP 工具服务、AI 编程 Agent
语言	中文（母语），英文（可用于技术汇报和文档）

教育背景

北京航空航天大学 2016 – 2019
模式识别 · 硕士

北京航空航天大学 2012 – 2016
自动化 · 本科

专利、分享与客户交流

- **专利（已申请）**：自动搜索量化参数框架。
- **客户交流**：为海外车企做中英双语的量化方法技术介绍；参与与地平线、爱芯的技术交流会。
- **技术分享**：组织月度量化分享会，在公司 tech weekly 分享模型量化。