

昇腾 310B 实战

从入门到精通边缘计算与人工智能

周贤中

2025 年 9 月 21 日

前言

书名中的“实战”，核心是“在编程实践中学习”。本书作为昇腾 310B 芯片的入门指南，将跳出单纯的理论讲解，通过真实的 AI 推理部署案例，带读者直观理解昇腾 310B 的硬件架构特性、Atlas 工具链使用逻辑与端侧 AI 项目开发流程——从模型适配、量化优化到推理服务部署，每一个知识点都配套可落地的代码示例，让读者在动手编码的过程中，真正掌握昇腾 310B 的实战应用能力，实现从“了解芯片”到“能用芯片落地项目”的跨越。

本书定位与目标读者

本书面向以下三类读者：

- **高校学生 / 科研新人**: 希望通过一套系统化路径快速理解边缘 AI 硬件与部署流程。
- **嵌入式 / IoT 工程师**: 已有一定 Linux / C / Python 基础，希望把 AI 模型真正跑在边缘端并做性能调优。
- **AI 应用开发者 / 创客**: 已经能使用主流深度学习框架，希望将训练好的模型迁移到昇腾 310B 进行高效推理与产品化落地。

阅读预期：

- 零基础读者可依照“快速起步路径”完成第 1~3 章 + 精选案例；
- 进阶读者可继续深入算子优化、系统整合与复杂多模型协同部署；
- 有项目诉求的团队可参考“方法论 + 附录模板”直接搭建属于自己的边缘 AI 应用。

学习路径速览（建议路线）

1. 环境 + 工具链：硬件认知 → 开发环境装配 → CANN 工具初试
2. 基础模型部署：图像分类 → 目标检测 → 语义分割 / OCR / NLP
3. 性能优化：模型结构裁剪 → 精度-性能权衡 (FP16 / INT8) → 并行与 pipeline
4. 低级能力：自定义算子 → Profiling → ACL / GE 原理 → 内存与数据通路调优
5. 系统构建：多进程/多模型协同 → 任务调度 → 监控与日志体系
6. 实战案例：从需求分析 → 方案设计 → 模型适配 → 部署脚本 → 交付验收

全书结构（初版规划）

| 模块 | 章节标题 | 内容聚焦 | 读者收益 |
|--------|---------------|----------------------------------|-------------------|
| Part 0 | 导读与准备 | 芯片特性、开发形态、学习地图 | 建立整体心智模型 |
| Part 1 | 昇腾 310B 硬件与环境 | 硬件结构、固件、驱动、系统配置、容器化 | 能独立搭建可复现环境 |
| Part 2 | CANN 软件栈核心 | CANN 组件、ATC 模型转换、OM 模型结构、ACL 编程 | 掌握模型从框架到 OM 全过程 |
| Part 3 | 边缘计算基础 | 边缘计算价值、典型架构、数据流、协同模式 | 会做架构选型与资源拆分 |
| Part 4 | 模型部署实战 | 分类/检测/NLP/多模态部署、性能测试、批处理与流式 | 会把主流任务完整迁移上板 |
| Part 5 | 性能与算子优化 | Profiler 使用、数据对齐、算子融合、自定义算子开发 | 会定位瓶颈并提升帧率/延迟 |
| Part 6 | 系统工程方法 | 多模型编排、任务调度、异常恢复、日志与监控 | 会构建工程级可维护部署系统 |
| Part 7 | 项目实战方法论 | 需求拆解、Baseline 迭代、评测体系、交付模板 | 会组织团队快速交付边缘 AI 项目 |
| Part 8 | 典型综合案例 | 9 个端侧 AI 实战案例 (与 experiments 配套) | 迁移复用案例形成生产力 |
| Part 9 | 附录与工具 | FAQ、性能 Checklist、脚手架模板、术语表 | 快速检索与复用加速迭代 |

各模块核心要点概述

1. 昇腾 310B 硬件与环境
 - SoC 架构 (昇腾 AI Core、内存层次、带宽特性)
 - 开发板接口与外设 (摄像头/存储/网络)
 - 固件刷新 & 系统初始化
 - Docker / 容器化开发与远程调试

-
- 2. CANN 软件栈与模型转换
 - CANN 组件: Driver / Runtime / Compiler / Toolkit
 - ATC 模型转换参数详解 (shape、输入格式、最优算子选择)
 - OM 模型结构与可视化
 - ACL 编程流程 (初始化 → 内存 → 推理 → 释放)
 - 常见错误 (推理精度损失 / 内存不足 / 算子不支持) 定位
 - 3. 边缘计算原理
 - 边云协同模式: 云训练 + 边缘推理
 - 数据生命周期: 采集 → 预处理 → 推理 → 缓存 → 上报
 - 典型架构模式 (单板/多板/异构协同)
 - 边缘 QoS: 功耗、热设计、延迟、稳定性
 - 4. 模型部署与优化实践
 - 图像分类 (ResNet / MobileNet)
 - 目标检测 (YOLO / FasterRCNN)
 - OCR & NLP (文本检测 + 轻量文本识别 / 中文 BERT 推理)
 - 多模型串联 (检测 → 裁剪 → 分类) Pipeline 设计
 - 精度 vs. 性能: Batch、FP16、算子融合、降采样策略
 - 5. 低级算子与性能调优
 - Profiling 工具使用 (时间线 / 算子耗时 / 内存峰值)
 - 数据对齐与内存复用策略
 - 常用自定义算子开发模板 (算子描述 → 编译 → 集成)
 - 典型瓶颈案例: 数据搬运 > 计算、Host/Device 同步等待
 - 6. 系统集成与工程实践
 - 任务调度 (多进程、多线程、异步队列)
 - 资源隔离与监控 (显存 / Host 内存 / 温度 / 带宽)
 - 高可用设计: 看门狗、超时熔断、故障降级
 - 交付形态: 容器镜像 / 一键部署脚本 / OTA 升级
 - 7. 项目实战方法论
 - 需求澄清 & 场景指标设定 (Latency / FPS / Accuracy / Power)
 - Baseline 快速验证: 裁剪 vs. 重构 vs. 迁移
 - 评测体系: 功能、性能、稳定性、可维护性
 - SRE 视角的上线准备 Checklist
 - 8. 综合实战案例 (与 `src/experiment` 配套) 包含 9 大可复现案例: 人脸打卡机、实时跟踪、智能电子琴、掌纹识别、数据采集仪、智能小车、智能相册、手势识别、聊天机器人。每个案例均提供:

-
- 需求说明 & 功能结构图
 - 模型与数据选择依据
 - 转换 & 部署脚本
 - 性能测试报告 (延迟 / 吞吐 / 资源占用)
 - 可选 3D 打印结构件与装配说明
9. 附录与工具箱
- 常见报错速查表 (ATC / ACL / Runtime)
 - 模型转换与部署参数模板
 - 性能调优 Checklist (内存 / 数据流 / 并行 / 算子)
 - 术语表 / 推荐资料 / 贡献指南

实践驱动与开源协作

本书所有示例代码、脚本、案例与附录工具均开源托管于本仓库。欢迎通过 Issue / PR 反馈问题、提交改进、补充案例或翻译。我们鼓励：
- 增补新模型 / 新任务的部署范式 - 分享自定义算子优化经验 -
提交性能测试报告 (含硬件信息 + 指标) - 翻译与文档校对

如何使用本书

| 读者类型 | 推荐阅读路径 | 目标 | 补充建议 |
|----------|----------------------------------|---------|------------|
| 零基础学生 | Part1 → Part2 → Part4(入门任务) | 能跑通首个模型 | 结合案例做改动实验 |
| 嵌入式工程师 | Part1 → Part2 → Part5 → Part6 | 掌握部署与优化 | 关注资源与稳定性章节 |
| AI 应用开发者 | Part2 → Part4 → Part7 → Part8 | 快速场景落地 | 记录调参与性能差异 |
| 技术负责人 | Part0 → Part3 → Part6 → Part7 | 构建团队方法论 | 制定内部模板体系 |

更新与版本计划

- v0.1 (当前): 结构规划 + 前 3 章草稿 + 2 个示例案例
- v0.3: 补齐核心部署链路 + 性能调优初稿
- v0.6: 全案例上线 + 工程化章节完善
- v1.0: 补齐附录 + 全面审校 + PDF / LaTeX 发行

许可证与引用

本书内容采用 Apache 2.0 许可证。引用本书内容请注明：> 《昇腾 310B 实战：从入门到精通边缘计算与人工智能》(GitHub: zhouxzh/Ascend310)

欢迎加入共建，一起把“边缘 AI 实战”这件事做成！

Contents

| | |
|--|----------|
| 1 犀腾 310B 边缘计算基础 | 1 |
| 1.1 什么是云计算? | 1 |
| 1.2 什么是边缘计算? | 1 |
| 1.3 边缘计算 vs 云计算? | 2 |
| 1.4 何时采用边缘计算? | 3 |
| 1.4.1 工程实现关键差异速览 | 4 |
| 1.4.2 价值协同总结 | 4 |
| 1.5 章节小结 | 4 |
| 1.6 实践任务 | 5 |
| 2 CANN 软件栈核心与模型转换全流程 | 6 |
| 2.1 章节总览 | 6 |
| 2.2 CANN 软件栈分层与数据流 | 6 |
| 2.3 环境一致性与安装验证 | 7 |
| 2.4 模型准备与输入规范统一 | 7 |
| 2.5 ATC 模型转换详解 | 7 |
| 2.5.1 自定义算子加载 | 8 |
| 2.5.2 日志与告警 | 8 |
| 2.6 OM 文件结构解读 | 9 |
| 2.6.1 解析与统计脚本要点 | 9 |
| 2.7 ACL 推理编程模型 | 9 |
| 2.7.1 C 语言最小示例（核心片段） | 9 |
| 2.7.2 Python 封装思路 | 10 |
| 2.8 性能与初步调优策略 | 10 |
| 2.9 常见错误分类与排查路径 | 11 |
| 2.10 质量保障与自动化流水线 | 11 |
| 2.10.1 精度对齐示例指标 | 11 |
| 2.11 Dump / Profiling / 调试手段 | 12 |

| | |
|--|-----------|
| 2.12 动态 Shape 策略与内存规划 | 12 |
| 2.13 精度验证流程与脚本要点 | 12 |
| 2.14 安全与合规考量 | 12 |
| 2.15 章节小结 | 13 |
| 2.16 实践任务 | 13 |
| 3 异腾 310B 算子开发基础 | 14 |
| 3.1 算子开发概述 | 14 |
| 3.2 开发的理论基础 | 14 |
| 3.3 开发流程 (AI Core 路线) | 15 |
| 3.4 常见问题与排查 | 16 |
| 3.5 章节小结 | 17 |
| 3.6 实践任务 | 17 |
| 4 典型模型部署实践 | 18 |
| 4.1 章节总览 | 18 |
| 4.2 统一部署工作流与契约化 | 18 |
| 4.3 图像分类: ResNet / MobileNet | 18 |
| 4.3.1 模型导出 | 18 |
| 4.4 目标检测: YOLO / FasterRCNN | 19 |
| 4.4.1 输入尺寸与 Letterbox | 19 |
| 4.5 OCR: 文本检测 + 识别 Pipeline | 19 |
| 4.5.1 结构 | 19 |
| 4.6 NLP: BERT 推理优化 | 19 |
| 4.6.1 序列长度策略 | 19 |
| 4.7 多模型 Pipeline 串联 | 20 |
| 4.8 工程目录与脚本标准 | 20 |
| 4.9 性能基线方法与统计置信 | 21 |
| 4.10 常见问题诊断深度版 | 21 |
| 4.11 章节小结 | 21 |
| 4.12 实践任务 | 21 |
| 5 性能与算子优化初阶 | 22 |
| 5.1 章节总览 | 22 |
| 5.2 性能拆解与衡量框架 | 22 |
| 5.3 Profiling 工具与时间线解读 | 22 |

| | | |
|----------|-------------------------------|-----------|
| 5.4 | 瓶颈模式与处置策略矩阵 | 22 |
| 5.5 | Layout / 内存访问优化 | 23 |
| 5.6 | 精度与性能的层级折衷 | 23 |
| 5.7 | 内存管理专题 | 24 |
| 5.8 | 并行与流水线 | 24 |
| 5.9 | 自定义算子开发与评估 | 24 |
| 5.10 | 优化案例: Add + ReLU 融合 | 25 |
| 5.11 | 性能报告与回归模板 | 25 |
| 5.12 | 章节小结 | 25 |
| 5.13 | 实践任务 | 25 |
| 5.14 | 昇腾 310B 自定义算子开发全流程 | 26 |
| 5.14.1 | 开发概述 | 26 |
| 5.14.2 | 开发的理论基础 | 26 |
| 5.14.3 | 开发流程 (AI Core 为例) | 27 |
| 5.14.4 | AICPU 路线 (可选) | 28 |
| 5.14.5 | 常见问题与排错 | 28 |
| 5.14.6 | 本章小结 | 28 |
| 6 | 系统工程与高可用部署 | 29 |
| 6.1 | 章节总览 | 29 |
| 6.2 | 部署形态与演进路线 | 29 |
| 6.3 | 进程与线程模型设计 | 29 |
| 6.3.1 | 基本原理 | 29 |
| 6.3.2 | 线程池建议 | 29 |
| 6.4 | 任务调度与优先级控制 | 30 |
| 6.5 | 配置管理与热更新 | 30 |
| 6.6 | 日志体系与追踪 | 31 |
| 6.7 | 指标监控与探针 | 31 |
| 6.8 | 高可用与自愈机制 | 31 |
| 6.9 | 异常分类与处理矩阵 | 32 |
| 6.10 | 版本、灰度与回滚 | 32 |
| 6.11 | 安全与访问控制 | 32 |
| 6.12 | 审计与合规 | 32 |
| 6.13 | 示例: 两模型多进程结构 | 32 |
| 6.14 | 章节小结 | 33 |

| | |
|----------------------------------|-----------|
| 6.15 实践任务 | 33 |
| 7 项目实战方法论与交付模板 | 34 |
| 7.1 章节总览 | 34 |
| 7.2 需求澄清 Canvas | 34 |
| 7.3 指标分层与优先级 | 34 |
| 7.4 Baseline 策略与控制变量法 | 35 |
| 7.5 评测集设计原则 | 35 |
| 7.6 迭代计划与看板 | 36 |
| 7.7 资产沉淀文档体系 | 36 |
| 7.8 交付目录与不可变产物 | 37 |
| 7.9 上线前综合 Checklist | 37 |
| 7.10 验收、回归与漂移监测 | 38 |
| 7.11 风险管理与决策日志 | 38 |
| 7.12 章节小结 | 38 |
| 7.13 实践任务 | 38 |
| 8 合实战案例集 | 39 |
| 8.1 章节总览 | 39 |
| 8.2 案例统一模板（标准化规范） | 39 |
| 8.3 案例目录结构规范 | 39 |
| 8.4 例概览与重点 | 40 |
| 8.5 案例 1：人脸打卡机 | 40 |
| 8.5.1 场景 | 40 |
| 8.5.2 指标 | 41 |
| 8.5.3 模型链路 | 41 |
| 8.5.4 性能优化 | 41 |
| 8.5.5 metrics 示例 | 41 |
| 8.6 案例 2：实时跟踪（检测 + 关联） | 42 |
| 8.6.1 流程 | 42 |
| 8.6.2 难点 | 42 |
| 8.6.3 优化 | 42 |
| 8.6.4 评估指标 | 42 |
| 8.7 案例 3：智能电子琴（音频） | 42 |
| 8.7.1 流程 | 42 |
| 8.7.2 优化点 | 42 |

| | |
|--------------------------------------|-----------|
| 8.7.3 指标 | 42 |
| 8.8 结果记录与差异报告 | 42 |
| 8.9 自动化与复现保障 | 43 |
| 8.10 指标可视化建议 | 43 |
| 8.11 通用问题经验库 | 43 |
| 8.12 扩展方向 | 44 |
| 8.13 贡献工作流 | 44 |
| 8.14 章节小结 | 44 |
| 8.15 实践任务 | 44 |
| 9 附录与工具箱 | 45 |
| 9.1 章节总览 | 45 |
| 9.2 常见报错速查 | 45 |
| 9.3 模型转换参数模板合集 | 46 |
| 9.3.1 分类模型 (ResNet) | 46 |
| 9.3.2 YOLO 动态分辨率 | 46 |
| 9.3.3 INT8 量化 (示例) | 46 |
| 9.4 性能与质量 Checklist (执行勾项) | 47 |
| 9.5 术语表 (扩展) | 47 |
| 9.6 推荐资源与外部引用 | 48 |
| 9.7 贡献指南摘要 | 48 |
| 9.8 FAQ | 48 |
| 9.9 License 与引用 | 49 |
| 9.10 版本路线回顾 | 49 |
| 9.11 实践任务 | 49 |
| 10 导读与准备工作 | 50 |
| 10.1 章节总览 | 50 |
| 10.2 全书主线结构 | 50 |
| 10.3 读者路径矩阵 | 50 |
| 10.4 硬件准备与兼容性 | 50 |
| 10.5 软件与工具栈细化 | 51 |
| 10.6 仓库目录与命名约定 | 51 |
| 10.7 最小可行环境验证 (MVE) | 52 |
| 10.8 全局术语与约定 | 52 |
| 10.9 协作工作流与质量闸门 | 52 |

| | |
|-------------------------------------|-----------|
| 10.10 学习与实践建议 | 53 |
| 10.11 常见初学误区与规避 | 53 |
| 10.12 章节小结 | 53 |
| 10.13 实践任务 | 53 |
| 11 案例 0 | 54 |
| 11.1 昇腾 310B 开发板介绍 | 54 |
| 11.1.1 开发板详细视图 | 55 |
| 11.1.2 开发板硬件规格 | 58 |
| 11.1.3 所需配件 | 58 |
| 11.1.4 下载开发板的系统镜像 | 68 |
| 11.1.5 刷写系统到 TF 卡 | 75 |
| 11.1.6 启动开发板 (Ubuntu) | 82 |
| 11.1.7 WIFI 天线安装指南 | 86 |
| 11.1.8 Ubuntu Xfce 桌面使用说明 | 87 |
| 11.1.9 HDMI 口使用 | 88 |
| 11.1.10 USB 摄像头使用 | 89 |
| 11.1.11 音频使用 | 89 |
| 11.1.12 GPIO 口的引脚顺序 | 90 |

1 昇腾 310B 边缘计算基础

华为云等公共云计算平台允许企业以全国的云服务器作为其私有的数据与 AI 计算中心，将其基础设施扩展到任意地点，并根据需要向上或向下扩展计算资源。然而遍布全球实时运行的 AI 应用可能需要显著的本地处理能力，且往往位于距离集中式云服务器过远的偏远位置。而且出于低时延或数据驻留要求，一些工作负载需要保留在本地或特定地点，这就是为什么许多企业使用边缘计算来部署其 AI 应用。边缘计算指的是在数据产生的位置进行处理，边缘计算在边缘设备中本地处理与存储数据。由于边缘计算设备无需依赖互联网连接，可以作为独立的网络节点运行。

1.1 什么是云计算？

云计算 (Cloud Computing) 是一种计算风格，其可扩展与弹性能力通过互联网技术以服务形式交付。云计算依托集中化数据中心，通过互联网按需提供弹性、可扩展、托管式 IT 资源与服务（计算 / 存储 / 网络 / 数据 / AI 平台）。

云计算的好处是什么？ - 较低的前期成本：购买硬件、软件、IT 管理以及全天候电力与制冷的资本支出被消除。云计算使组织能够以较低的财务进入门槛快速将应用推向市场。灵活的定价 – 企业只为其使用的计算资源付费，从而对成本有更多控制并减少意外。 - 按需无限计算：云服务可以通过自动配置与撤销资源即时对不断变化的需求做出反应与适配。这可以降低成本并提升组织整体效率。 - 简化的 IT 管理：云提供商为其客户提供访问 IT 管理专家的渠道，使员工可以专注于企业的核心需求。 - 便捷更新：可一键访问最新的硬件、软件与服务。 - 可靠性：数据备份、灾难恢复与业务连续性更容易且更便宜，因为数据可以在云提供商网络的多个冗余站点镜像。 - 节省时间：企业可能在配置私有服务器与网络上耗费时间。借助按需云基础设施，它们可以在更短时间内部署应用并更快进入市场。

1.2 什么是边缘计算？

边缘计算 (Edge Computing) 是一种分布式计算框架，旨在将计算、存储和网络能力部署在靠近数据源或终端设备的网络边缘位置。通过在本地或近端节点处理数据，减少向远端数据中心/云的集中回传，获得更低的时延、更好的带宽利用与更高的数据主权与隐私保障。边缘计算是在距离数据产生源（传感器、摄像头、终端设备、工业控制点）物理更近的位置部署计算与存储，使数据在本地/近端被快速处理与筛选，减少长距离回传，保障低时延、带宽节省与数据主权。边缘计算是将计算能力在物理上靠近数

据生成位置（通常是物联网设备或传感器）的实践。因为计算能力被带到网络或设备的边缘，边缘计算能够实现更快的数据处理、增加的带宽以及确保的数据主权。

通过在网络边缘处理数据，边缘计算减少了大量数据在服务器、云与设备或边缘位置之间往返传输以被处理的需求。这对诸如数据科学与 AI 等现代应用尤为重要。许多高计算应用（如深度学习与推理、数据处理与分析、仿真与视频流）已成为现代生活的支柱。随着企业日益意识到这些应用由边缘计算驱动，生产中的边缘用例数量应会增加。

边缘计算的特点是什么？ - 靠近数据源：在物联网设备、网关、工业控制器或本地微型服务器附近直-成初级或核心推理逻辑，降低往返延迟。 - 分布式架构：区别于云的集中式调度，采用多节点协同与局部自治，适合-化、地理分散与对实时性敏感的任务。 - 实时响应：适配无人驾驶、工业控制、视频安防、AR/VR 等对毫秒级响应-的场景。 - 隐私与安全：敏感原始数据（面部特征、生产工艺、地理轨迹）在本地预-或结构化提取后再上云，降低泄露与合规风险。 - 带宽优化：仅上传事件/特征/聚合指标，显著降低原始全量视频/传感流量-路的占用。 - 弹性与容错：弱网/离线时保持关键功能脱网运行，网络恢复后再同步（延迟一致性）。

边缘计算的好处是什么？ - 更低时延：在边缘进行数据处理会消除或减少数据传输。这可为需要低时延的复杂 AI 模型用例（如完全自动驾驶车辆与增强现实）加速洞察。 - 降低成本：使用局域网进行数据处理相比云计算可为组织提供更高带宽与更低成本的存储。此外，由于处理发生在边缘，需要发送到云或数据中心进一步处理的数据更少。这导致需要传输的数据量减少，成本也降低。 - 模型精度：AI 依赖高精度模型，尤其是需要实时响应的边缘用例。当网络带宽过低时，通常通过降低输入模型的数据尺寸来缓解。这会导致图像尺寸缩小、视频跳帧、音频采样率降低。部署在边缘时，数据反馈回路可用于提升 AI 模型精度，并且可同时运行多个模型。 - 更广覆盖：传统云计算必须依赖互联网接入。而边缘计算可在本地处理数据，无需互联网接入。这将计算范围扩展到以前无法访问或偏远的位置。 - 数据主权：当数据在其采集位置被处理时，边缘计算允许组织将所有敏感数据与计算保留在局域网和公司防火墙内。这减少了暴露于云端网络安全攻击的风险，并在不断变化的严格数据法律下具备更好合规性。

1.3 边缘计算 vs 云计算？

| 维度 | 边缘计算 | 云计算 | 典型取舍 / 典型场景 |
|------|------------------|------------------|---------------------------------------|
| 处理位置 | 近端（本地/网关/边缘节点） | 远端数据中心 | 边缘降低时延并就近处理高带宽原始数据（如视频）；云用于集中算力与全局聚合。 |
| 时延特性 | 低（本地判决 20~几十 ms） | 受网络往返影响 (>100ms) | 实时/控制闭环优先边缘；非实时批处理、训练场景常见。 |
| 带宽占用 | 上行压缩（只传事件/特征/聚合） | 常上传原始数据到云 | 当传输成本高或链路受限，将预处理放到边缘；带宽充足且需保留原始数据则上云。 |

| 维度 | 边缘计算 | 云计算 | 典型取舍 / 典型场景 |
|---------------|-------------------------|--------------------|---|
| 部署/运维 | 分布式, 需节点管理 (异构、离线可用) | 集中化, 统一维护与弹性伸缩 | 节点数多时运维复杂度上升 (需探针/模板); 云侧适合动态弹性与大规模训练/批处理。 |
| 隐私合规 | 本地脱敏/保留数据 主权易控 | 数据集中存储 需合规审核 | 高度敏感或受监管数据优先边缘; 低合规风险且需统一治理优先云。 |
| 伸缩弹性 | 受制于本地硬件与现场成本 | 云端资源弹性丰富 | 现场扩展 (CAPEX) 与运维 (OPEX) 成本较高; 云适合突发/动态扩展工作负载。 |
| 典型适用场景 (示例对照) | 实时推理、低延迟控制、弱网/离线场所 | 非实时批处理、模型训练、集中化数据湖 | 云: 非时间敏感数据、可靠互联网、已在云存储的数据; 边缘: 实时数据处理、受限或无联网地点、传输成本过高的大规模本地数据、受严格法规约束的数据。 |

一个边缘计算优于云计算的示例是医疗机器人，外科医生需要实时数据访问。这些系统包含大量可在云中执行的软件，但手术室中日益出现的智能分析与机器人控制无法容忍时延、网络可靠性问题或带宽限制。在此示例中，边缘计算为患者提供了生死攸关的益处。

1.4 何时采用边缘计算？

边缘节点通常使用功耗、体积和成本折中硬件（如 Ascend 310B）。典型的应用场景如下：

| 场景 | 说明 | 边缘价值点 |
|-------|---------------|---------------------|
| 物联网网关 | 聚合海量传感数据 | 局部预处理 + 协议转换 + 降噪聚合 |
| 工业自动化 | 产线质量检测/能耗分析 | 毫秒级响应 + 数据本地闭环 |
| 智慧城市 | 交通流量/环境监测 | 低延时告警 + 带宽节省 |
| 安防监控 | 实时视频结构化 | 事件级上报 + 隐私保护 |
| 智能零售 | 客流/货架分析 | 设备自治 + 弱网容忍 |
| 车路协同 | 路侧单元 (RSU) 分析 | 超低时延 + 本地协同决策 |

边缘计算并非替代云，而是形成“端 边 云”分层协同：端侧产生原始数据，边缘做低时延智能决策与数据筛选，云端负责全局模型训练、长周期统计与跨区域调度。合理的切分策略直接影响系统的成本结构、响应性能与可持续演进能力。

| 典型更适合云 | 典型更适合边缘 |
|-------------------|-------------|
| 非实时批处理 / ETL / 训练 | 实时推理 / 控制闭环 |
| 动态弹性突发强 | 稳定持续低时延需求 |
| 数据已在云湖中 | 数据采集源密集分散 |
| 合规风险低 | 高敏感/受监管数据 |
| 带宽充足且廉价 | 带宽受限或成本高 |

1.4.1 工程实现关键差异速览

| 维度 | 云侧偏好 | 边缘侧偏好 | 说明 |
|------|--------------------|---------------|-------------|
| 日志策略 | 全量集中收集 | 采样 + 本地环形截断 | 带宽 & 存储控制 |
| 模型分发 | 大文件 CDN | 差分/分块 + 校验 | 断点续传/校验哈希 |
| 配置管理 | 中央配置中心 | 嵌入版本 + 增量下发 | 需要离线安全回滚 |
| 监控 | Prometheus/集中 TSDB | 轻量 Agent 本地缓存 | 冲突时丢弃低优先级指标 |
| 安全补丁 | 自动批量推送 | 计划窗口/手动确认 | 避免运行中断 |

1.4.2 价值协同总结

“边缘强化实时性 + 云强化全局优化”是主旋律：将需要毫秒级反馈、隐私受限、数据强局部性的处理前移；将需要大规模聚合、长周期分析、模型训练、跨区域调度的任务后移。设计时以“放在云端的必要性”反向审视每一段功能，并以可观测指标（时延、带宽、成本、精度、合规等级）量化切分边界。

依据画像做编排：- 高 I/O 密度任务与计算密集型错峰执行。- 热点算子分组，避免同一时间窗口内全部提交导致带宽抖动。热设计：读取温度曲线（如每 5s 采样），超过阈值 85°C 触发降频/任务降载。

1.5 章节小结

边缘系统设计的本质是多目标优化：时延、精度、稳定、成本、安全。通过资源画像、协同模式选择、分层缓存、任务编排与降级策略形成一套可演进体系。后续章节将把单模型部署扩展到多模型与工程化落地。

1.6 实践任务

1. 基于你的目标场景输出一份协同模式决策表（含放弃理由）。
2. 编写数据生命周期图（ASCII 或 Mermaid）。
3. 实现一个队列背压示例：当处理时延 $>$ 阈值时自动丢弃旧帧。
4. 采集 10 分钟温度与时延数据，绘制相关性（是否热导致抖动）。
5. 设计一份故障降级矩阵并评审可行性。

2 CANN 软件栈核心与模型转换全流程

2.1 章节总览

本章系统阐述 Ascend CANN 软件栈的分层结构、模型从框架格式到 OM 的转换原理、转换工具 ATC 的关键参数、OM 文件组织结构、AscendCL (ACL) 推理编程模型、精度与性能验证方法以及工程级质量保障流水线建设。阅读完成后应满足：1. 能解释 Driver / Runtime / Compiler / Toolkit / ACL 各组件职责及交互边界。2. 能为任意主流视觉模型编写一份无二义性的 ATC 转换命令并说明参数意义。3. 能通过脚本解析 OM 模型的输入输出信息、算子统计与内存占用估算。4. 能以 C 或 Python 写出健壮的最小推理程序（含异常处理与资源释放）。5. 能定位转换/推理常见错误，给出复现、分析与修复路径。6. 能构建“转换 → 精度对齐 → 性能基线 → 回归监测”的自动流水线。

2.2 CANN 软件栈分层与数据流

| 层级 | 组件 | 核心职责 | 典型交互 |
|-------|-------------------|--------------------------------------|-------------------|
| 硬件抽象 | Driver | 设备初始化、资源枚举、功耗/温度接口 | npu-smi / Runtime |
| 运行时 | Runtime | 上下文 (Context) 管理、Stream/Task 调度、内存分配 | ACL / Compiler |
| 编译优化 | Graph Compiler | 图解析、拓扑排序、算子匹配、内存复用、算子融合 | ATC / Runtime |
| 工具链 | Toolkit | ATC 转换、Profiling、Dump、可视化、日志 | 开发者 |
| API 层 | AscendCL | C 接口封装：模型管理 / 内存 / 数据传输 / 执行 | 应用 |

数据流（框架模型 → OM → 推理）核心阶段：1. 前端导出：PyTorch → ONNX（维度常量化、算子展开）。2. ATC 编译：图解析 → Shape Infer → 算子选择 → Kernel 排布 → 内存映

射 → 生成 OM (二进制 + 元数据段)。3. 运行加载: aclmdlLoadFromFile 读取 OM Header, 分配 Device 内存, 构建执行计划 (Task 列表)。4. 推理执行: Host 侧准备输入 → H2D 拷贝 → Runtime 提交 Task → 硬件执行 → D2H 拷贝 → 后处理。

2.3 环境一致性与安装验证

环境差异是隐性失败根源,建议形成“安装后自检”脚本,校验以下要点:1. 版本矩阵:固件/Driver/CANN/ATC 必须在官方 Release Note 支持组合内。2. 环境变量: ASCEND_INSTALL_PATH 指向安装根; LD_LIBRARY_PATH 中包含 driver 与 runtime/lib64;Python 绑定需在 PYTHONPATH 中。3. 设备可见: npu-smi info 返回芯片型号 Ascend310B 且状态正常,无 Fault 标记。4. 转换工具: atc --version 输出版本与期望匹配; atc --help 能正常列出参数。5. 运行权限: 当前用户具备访问 /dev/davinci* 设备节点读写权限 (若无,加入相应用户组或 udev 规则)。6. Python 依赖: numpy, onnx, onnxruntime (精度对齐), pyyaml, 自编写工具包。

2.4 模型准备与输入规范统一

| 项 | 说明 | 决策标准 |
|-----------|------------------|-----------------|
| 边界 Shape | 静态 or 动态 | 场景多尺寸/Batch 波动? |
| Layout | NCHW / NHWC | 上游预处理 & 算子最佳实现 |
| 颜色空间 | RGB / BGR / YUV | 原始采集格式 + 算子期望 |
| 归一化 | mean/std / scale | 训练环节定义必须完全对齐 |
| 精度策略 | FP16 / INT8 | 性能目标 & 可接受精度损失 |
| Quant 校准集 | 代表性样本 | 覆盖亮度/场景/尺寸多样性 |

核心风险: 训练与部署输入不一致 (尺寸拉伸方式、通道顺序、归一化顺序、色彩空间转换位置)。必须输出“输入契约文件”(JSON/YAML) 标注: shape、dtype、layout、color_space、mean/std、range、precision_mode。

2.5 ATC 模型转换详解

典型命令 (以 ResNet50 为例, 支持 FP16):

```
atc \
--model=resnet50.onnx \
```

```
--framework=5 \
--output=resnet50_fp16 \
--input_format=NCHW \
--input_shape="input:1,3,224,224" \
--soc_version=Ascend310B \
--precision_mode=allow_fp32_to_fp16 \
--op_select_IMPLmode=high_performance \
--log=info \
--insert_op_conf=aipp.cfg
```

关键参数说明： | 参数 | 作用 | 注意事项 | | — | — | — | | --framework | 输入框架类型 (5=ONNX) | 与实际导出一致，否则形状推理异常 | | --input_shape | 静态 shape 指定 | 多输入以逗号分隔 in1:1,3,224,224; in2:1,128 | | --dynamic_batch_size | 动态 Batch | 与 --input_shape 不能混用静态冲突 | | --dynamic_image_size | 动态分辨率 | YOLO 等多尺度部署 | | --precision_mode | 精度策略 | allow_mix_precision、allow_fp32_to_fp16 | | --soc_version | 硬件目标 | 与实际芯片匹配；310B 与 310P 不可混淆 | | --insert_op_conf | AIPP(预处理) | 可下沉色彩空间转换、均值/方差 | | --op_select_IMPLmode | 算子实现优先级 | high_precision vs high_performance | | --input_format | 模型输入排布 | 与 --input_shape 一致性检查 | | --output_type | 输出 dtype | 常用于 INT8 推理后转 FP32 便于后处理 | | --enable_small_channel | 小通道优化 | 某些轻量网络加速 |

2.5.1 自定义算子加载

1. 定义 JSON 描述 (输入输出、属性)。
2. 编写 Kernel 源码并使用官方编译脚本生成 .so。
3. ATC 阶段通过 --optypelist_for_impl 或 --soc_version + JSON 注册；运行时放置在 ASCEND_OPP_PATH 对应目录。

2.5.2 日志与告警

常见告警分类：
 - 未使用节点 (prune) → 确认是否为训练辅助算子 (e.g., Dropout)。
 - 算子降级 → 检查是否 fallback 到 Host；对性能敏感需重写/替换结构。
 - 精度截断 → 记录发生算子，评估对最终指标影响；必要时关闭相关优化策略。

2.6 OM 文件结构解读

OM 通常包含：1. Header：魔数、版本、输入输出 Tensor 数、DataType、Format。2. Graph Meta：节点拓扑、算子类型列表、权重偏移指针。3. Weights Segment：连续存放常量权重与常量张量。4. Task List：调度指令列表（Kernel Launch / MemCopy / Event）。5. AIPP 配置（可选）：预处理算子参数表。

2.6.1 解析与统计脚本要点

- 调用 aclmdlQuerySize 得到模型工作内存与权重内存需求。
- 利用 aclmdlGetInputIndexByName / aclmdlGetInputDims 获取 IO 维度与 dtype。
- 自建表格：{op_type: count} 用于识别热点类型（后续优化参考）。

2.7 ACL 推理编程模型

典型生命周期：1. 初始化：aclInit → aclrtSetDevice → aclrtCreateContext → （可选）创建 Stream。2. 模型：aclmdlLoadFromFile → 查询 IO 描述 → 预分配 Device Buffer。3. 数据准备：Host 侧申请内存（Pinned 优先）→ 格式/归一化 → H2D 拷贝。4. 执行：aclmdlExecute 或异步 aclmdlExecuteAsync + Stream 同步。5. 输出处理：D2H 拷贝 → 解码 / Softmax / NMS。6. 资源释放：aclmdlUnload → Free buffers → Destroy Context → aclFinalize。

2.7.1 C 语言最小示例（核心片段）

```
// 省略错误检查宏定义 ERR_CHK
aclInit(NULL);
aclrtSetDevice(0);
aclrtContext ctx; aclrtCreateContext(&ctx, 0);
uint32_t modelId; size_t wSize, rSize;
aclmdlLoadFromFile("resnet50_fp16.om", &modelId);
aclmdlDesc *desc = aclmdlCreateDesc();
aclmdlGetDesc(desc, modelId);
// 输入准备
void *hostIn = malloc(3*224*224*2); // FP16
void *devIn; aclrtMalloc(&devIn, 3*224*224*2, ACL_MEM_MALLOC_NORMAL_ONLY);
aclrtMemcpy(devIn, 3*224*224*2, hostIn, 3*224*224*2, ACL_MEMCPY_HOST_TO_DEVICE)
aclmdlDataset *input = aclmdlCreateDataset();
```

```

aclDataBuffer *inBuf = aclCreateDataBuffer(devIn, 3*224*224*2);
aclmdlAddDatasetBuffer(input, inBuf);
// 输出
size_t outSize = 1000 * 2; // FP16 logits
void *devOut; aclrtMalloc(&devOut, outSize, ACL_MEM_MALLOC_NORMAL_ONLY);
aclmdlDataset *output = aclmdlCreateDataset();
aclDataBuffer *outBuf = aclCreateDataBuffer(devOut, outSize);
aclmdlAddDatasetBuffer(output, outBuf);
aclmdlExecute(modelId, input, output);
// 回拷
void *hostOut = malloc(outSize);
aclrtMemcpy(hostOut, outSize, devOut, outSize, ACL_MEMCPY_DEVICE_TO_HOST);
// 解析 softmax ...
// 清理省略

```

2.7.2 Python 封装思路

官方 Python 包接口层次相似，建议封装 ModelSession 类：

```

class ModelSession:
    def __init__(self, om_path):
        self.model_id = load(om_path)
        self.desc = query(self.model_id)
        self._alloc_io_buffers()
    def infer(self, np_input: np.ndarray):
        # preprocess -> copy H2D -> execute -> copy D2H -> postprocess
        return logits
    def __del__(self):
        self._release()

```

2.8 性能与初步调优策略

| 问题 | 诊断信号 | 初级优化 | 进阶优化 |
|-------|-----------|---------------|--------------------|
| 时延波动大 | P95 » P50 | 固定 Batch / 预热 | Stream 并行 + Pin 内存 |

| 问题 | 诊断信号 | 初级优化 | 进阶优化 |
|------|-------------|--------|---------|
| 吞吐不足 | 利用率低 | FP16 | 多实例并行 |
| 拷贝过多 | H2D 大占比 | 合并预处理 | AIPP 下沉 |
| 算子退化 | 日志 Fallback | 替换模型结构 | 自定义算子 |

关键早期收集指标：平均时延、P95、H2D+Pre 占比、推理核心阶段占比、内存峰值。

2.9 常见错误分类与排查路径

| 场景 | 日志/现象 | 根因类型 | 排查步骤 | 修复 |
|---|--------------------|-----------------------------|----------------------------------|-------------------------------------|
| ATC Unsup- ported Op 动态 Shape OOM 精度下降 | E190xx Top1 -5% | 模型含新算子 最大分辨率超预算 归一化差异 | onnxsim → 拆解 统计输入分布 离线对齐脚本 | 替换/重写 分桶/裁剪 重新校准 修正预处 理 |
| 输出 NAN | logits 异常 | 上溢/量化尺度错误 | Dump 中间 Tensor | 重新校准 |
| 设备不可 见 | aclInit 失败 | Driver 未加载 | dmesg & npu-smi | 重装驱动 |

2.10 质量保障与自动化流水线

- 流水线阶段:
1. Export: 框架导出 + ONNX Simplify + 模型签名 (inputs/name/dtype/layout/mean/std).
 2. Convert: ATC 命令模板参数化 (YAML → 渲染)。
 3. Validate: ONNXRuntime vs OM 输出差异 (L1/L2/TopK 差异率 < 阈值)。
 4. Benchmark: Warmup N + Run M, 记录 JSON {avg, p50, p95, memory}。
 5. Archive: 产物归档(om, atc.log, metrics.json, signature.json)。
 6. Regression: 新提交对比基线差异, 超阈值报警。

2.10.1 精度对齐示例指标

| 指标 | 计算方式 | 推荐阈值 |
|---------|---|------|
| Top1 差异 | $\text{abs}(\text{top1_acc_onnx} - \text{top1_acc_om})$ | 0.2% |
| 平均 L1 | $\text{mean}(\text{y_onnx} - \text{y_om})$ | |
| 最大相对误差 | $\text{max}(\text{d})$ | |

2.11 Dump / Profiling / 调试手段

| 工具 | 使用时机 | 价值 | 代价 |
|----------------------|----------|--------|-----------|
| Dump 中间 Tensor | 精度异常 | 对齐中间层 | I/O 与存储占用 |
| Profiling Timeline | 性能不达标 | 定位瓶颈 | 额外开销 (W%) |
| 日志级别升高 (--log=debug) | 转换失败 | 细粒度错误码 | 噪声多 |
| 校准数据捕获 | INT8 偏差大 | 重新校准 | 需准备代表性样本 |

Dump 配置：通过环境变量或 JSON 指定层名称白名单，避免全量 Dump 导致性能与空间压力。

2.12 动态 Shape 策略与内存规划

多分辨率/Batch 场景建议：1. 分桶：统计历史尺寸 → 选 3~5 个“代表桶”→ ATC 生成多 OM；运行时按最近桶选择。2. Padding：对齐到 32/64 边界，减少算子内部分支；记录真实尺寸用于后处理。3. 内存预估：最大桶内存 + 安全冗余 15% 作为部署阈值，超出触发降级。

2.13 精度验证流程与脚本要点

流程：采样输入集（校准集或验证集子集）→ ONNXRuntime 前向 → Ascend 前向 → 指标聚合 → 报告。脚本关键：1. 随机种子固定；2. 输入预处理完全共用函数；3. 支持逐层 Dump 比对（差异 > 阈值输出层名）。

2.14 安全与合规考量

- 模型资产：带版权或敏感权重需加密存储（考虑文件系统权限 + 传输校验 hash）。
- 日志脱敏：避免输出用户数据路径/片段；开关化控制。
- Dump 数据：限定开发模式，生产禁用；数据自动过期删除策略（时间或数量）。

2.15 章节小结

本章从宏观分层、转换编译、OM 结构、ACL 编程、性能与精度保障、调试工具、自动化流水线到动态 Shape 与安全实践建立了闭环。掌握这些内容后即可进入后续“边缘系统架构与部署实践”章节，扩展到多模型、多进程及系统级优化。

2.16 实践任务

1. 任选一个公开 ONNX 分类模型（如 ResNet50）完成 ATC 转换，提交：命令 + atc.log。
2. 以 C 或 Python 实现最小推理程序，输出前 5 TopK 结果与 softmax 概率。
3. 编写对齐脚本比较 50 张图片 ONNX vs OM 输出差异（报告 L1/Top1 差异）。
4. 收集 Profiling Timeline，列出前 3 耗时算子类型及优化建议。
5. 输出 signature.json、metrics.json、conversion_meta.yaml 并归档。

3 昇腾 310B 算子开发基础

昇腾 310B 在通用算子覆盖广度上已能满足大多数推理任务，但在以下场景，自定义算子（Custom Op）能显著提升功能完备性与性能确定性：模型含未支持/半支持算子、复合算子频繁导致访存过多、需要业务特化（如阈值/形态学/后处理融合）、或内置实现对特定尺寸/布局性能欠佳。第三章将给出“为什么、怎么做、如何验证与上线”的完整路径。

3.1 算子开发概述

- 目标与收益：
 - 功能补齐：覆盖模型图中未支持或语义差异较大的算子；
 - 性能确定性：融合多算子、减少 GM<->UB 搬运与中间落地、利用向量化内核；
 - 工程可维护：以“算子契约”形式固化输入/输出/属性与边界行为，便于回归与复用。
- 执行形态：
 - AI Core（推荐）：基于 TBE/TE/TIK 运行于 NPU 核心，适合数值密集型；
 - AICPU（可选）：C/C++ 在 AICPU/Host 侧执行，适合控制流/轻量处理（注意 H2D/D2H 成本）。
- 产物要素：
 - 算子描述（op info/proto）：声明 op_type、inputs/outputs、dtype_format 组合、属性与形状推断；
 - 算子实现（Kernel）：TE/TIK 计算 + 调度或 AICPU C++ 实现；
 - 注册与打包：产物按规范放入 OPP 目录，ATC/Runtime 可发现与加载。

3.2 开发的理论基础

- 1) 硬件与存储层次：
 - GM (Global Memory)：容量大、带宽高；
 - UB (Unified Buffer)：片上高速缓存，容量有限；
 - DMA：GM UB 的数据搬运，偏好大块连续传输；
 - 向量/标量单元：支持 vadd/vmul/vmax 等，需数据对齐（常见 16/32）。

2) 计算表达与调度:

- TE (Tensor Expression) 描述计算公式; Schedule 负责 tile/并行/向量化/缓存;
- TIK 提供更贴近硬件的 DSL, 便于精细控制 DMA 与 UB 管理;
- 目标: 以较少的 GM 往返在 UB 内完成尽可能多的计算, 提升算子效率与吞吐。

3) 算子契约 (Operator Contract):

- 输入/输出张量的 shape、dtype、layout (NCHW/NC1HWC0 等)、属性 (如 alpha、mode);
- 广播与对齐规则、边界行为 (溢出/饱和/舍入)、精度策略 (FP16/FP32 混合);
- 动态 shape 与静态 shape: 实现需覆盖契约内的形状组合并保证 UB 不溢出。

4) 数值与精度:

- FP16 常用于 310B 推理通路; 必要时在关键步骤采用临时 FP32 计算再回写;
- 误差控制: 选择合适的舍入策略, 避免饱和/下溢导致 NAN/INF。

3.3 开发流程 (AI Core 路线)

1. 环境准备与约束

- 安装 CANN/Toolkit 并确认 atc --version 正常;
- 设置环境变量: ASCEND_INSTALL_PATH、ASCEND_OPP_PATH;
- 目标芯片: soc_version=Ascend310B; 优先使用 FP16 与硬件友好布局 (如 NC1HWC0)。

2. 定义算子信息 (op info/proto)

- 声明 op_type、inputs/outputs 名称与数量、可支持的 dtype_format 组合、属性与默认值;
- 提供形状推断规则 (静态或依据属性/输入维度计算)。

3. 编写算子实现 (TE/TBE/TIK)

- 计算表达 (示例: Add+ReLU 融合伪代码):

```
# y = relu(x1 + x2)
import te.lang.cce as tbe
from te import tvm

def add_relu_compute(x1, x2):
    y = tbe.vadd(x1, x2)
    z = tbe.vmaxs(y, tvm.const(0.0, x1.dtype))
    return z
```

- 调度要点:

- Tile 到 UB 容量可承载的块大小;
- 连续向量访问, 减少非对齐;
- 合并搬运, 避免频繁小块 DMA;
- 小尺寸路径避免调度开销超过计算开销。

4. 编译与注册

- 使用 Toolkit 提供的编译入口生成 kernel 与元数据;
- 将实现与描述文件放入 ASCEND_OPP_PATH 下 custom 目录(如 op_impl/custom/ai_core/tbe, op_proto/custom)。

5. 与 ATC 集成

- 转换模型时指定 --soc_version=Ascend310B;
- 确保 OPP 路径可被 ATC 读取, 必要时调整 --op_select_implmode;
- 转换日志中应能看到自定义算子被匹配与编译。

6. 运行时部署

- 目标环境包含同版本 OPP (含 custom 产物);
- 设置环境变量使 Runtime 能定位到自定义实现;
- 按常规 ACL 流程加载 OM 并执行推理。

7. 验证与度量

- 功能: 与 NumPy/ONNX 参考实现对齐, 随机多组张量比较 (平均绝对/相对误差、边界样本);
- 性能: Warmup 3 次, 采样 50 次, 统计 avg/p95/FPS;
- 资源: Profiling 检查 MemCopy 占比、Kernel 占比、Idle;
- 兼容: 覆盖不同 shape/dtype/layout 组合。

8. 打包与版本化

- 输出 op_contract.yaml (契约) 与 benchmark.json (性能);
- 目录建议:

```
op_pkg/<op_type>/<version>/
    op_proto/custom/
    op_impl/custom/ai_core/tbe/
    tests/
    docs/
```

3.4 常见问题与排查

- ATC 提示 Unsupported Op: 检查 op 描述是否生效、路径与 soc_version 是否匹配;

- 运行时回退 (fallback): 确认 dtype_format 覆盖到当前张量组合;
- 性能无提升: 检查是否出现额外 layout 转换、tile 过小造成 DMA 频繁;
- 精度异常: 核对归一化/广播规则、溢出与舍入策略, 必要时局部切 FP32;
- 动态 shape OOM: 缩小 tile 或分桶处理, 保证 UB 与工作区不溢出。

3.5 章节小结

自定义算子是 310B 场景下实现“功能补齐与性能确定性”的关键手段。遵循“明确契约 → 正确调度 → 可观测验证 → 规范打包”的路径, 选择计算/访存比例合适、出现频繁的目标起步, 先易后难、以基线与回归保障质量与收益的可持续。

3.6 实践任务

1. 选择你项目中的一个复合算子 (例如归一化 + 阈值), 写出算子契约草案 (IO/attr/d-type_format/边界)。
2. 基于 TE 写出该算子的计算表达伪代码, 并说明预期的 tile 与向量化策略。
3. 在开发环境完成编译注册, 将产物放入 OPP custom 目录并用一个最小模型验证 ATC 识别。
4. 设计功能与性能验证脚本: 随机张量对齐、Warmup/采样策略、输出 avg/p95 与资源占比。
5. 生成 op_contract.yaml 与 benchmark.json, 并归档到 op_pkg/<op_type>/<version>/。

4 典型模型部署实践

4.1 章节总览

本章以“统一流程 → 四类典型任务（分类/检测/OCR/NLP）→ 多模型 Pipeline → 工程化目录与脚本 → 性能基线采集 → 问题诊断”逻辑展开，强调“可复现、可量化、可演进”的部署范式。所有示例策略均可推广到后续复杂场景（多输入、多分辨率、流式/批式混合）。

4.2 统一部署工作流与契约化

标准六步：模型选择 → 框架导出 ONNX → ATC 转换（参数冻结）→ 推理引擎封装（I/O 契约）→ 运行形态编排 → 验证（精度 + 性能）。核心产物：
| 文件 | 作用 | — | — | — |
| export.py |
| 导出 & 简化 ONNX | | atc.sh |
| 标准化转换命令 | | config.yaml |
| 输入/归一化/颜色/阈值 | |
| signature.json |
| 模型输入输出字段与 dtype | | metrics.json |
| 性能统计 (avg/p95/memory) |

输入预处理必须模块化，业务层仅提供原始图像对象；可在 AIPP 中下沉部分（色彩空间、均值/方差），减少 Host 侧拷贝和转换。

4.3 图像分类：ResNet / MobileNet

4.3.1 模型导出

PyTorch → ONNX:torch.onnx.export(model, dummy, opset_version=13, dynamic_axes=None);
确保去掉训练专属层（Dropout, BN 置 eval）。### 预处理一致性
1. Resize: 保持短边 256 → CenterCrop 224。
2. Normalize: mean/std 与训练保持一致。
3. Layout: NCHW；若原始图像为 HWC(RGB) → 转 BGR/或保持一致并在 config 标记。
转换要点
—precision_mode=allow_fp32_to_fp16;
若需 INT8: 先做离线标定导出校准表，再加量化参数。
推理后处理
Softmax → ArgTopK → LabelMap。
为避免数值不稳定：FP16 logits 可先转 FP32 再 softmax。
性能采集
Warmup 5 次，采集 100 次：记录 avg, p50, p95, max；统计预处理耗时占比：pre_ms / total_ms，超过 25% 提示 AIPP 下沉或批处理优化。

4.4 目标检测：YOLO / FasterRCNN

4.4.1 输入尺寸与 Letterbox

Letterbox 使图像等比例缩放 + 填充，保持方形输入。部署需重现训练阶段相同逻辑，否则框坐标偏移。保存 scale 与 pad 用于反算原始坐标。### 多输出解析 YOLOv5s OM 输出通常包含一个或多个特征拼接张量：(num_boxes, attributes)；后处理：过滤 conf > 阈值 → 按类合并 → NMS。### NMS 实现决策 | 方案 | 优点 | 缺点 || — | — | — || CPU Python | 简单 | 高开销，多框场景慢 || CPU C++ SIMD | 中等复杂 | 仍需 D2H 拷贝 || Device Kernel | 减少拷贝 | 实现复杂 | 先评估 D2H + CPU NMS 占比，>15% 再考虑下沉。### 动态尺度支持转换阶段可生成多尺度 OM 或使用动态 shape；推荐：统计输入分辨率 → 选择 3 桶 (640/704/768) 提升命中率。

4.5 OCR: 文本检测 + 识别 Pipeline

4.5.1 结构

检测模型 (DB) → 文本框多边形 → 透视裁剪 → 识别模型 (CRNN / SVTR)。### 难点与策略 | 环节 | 风险 | 对策 | —— | —— | —— | 多边形裁剪 | 仿射失真 | 统一仿射矩阵 + padding | 长短文本差异 | 序列长度不均 | 动态 Batch 分组 (长度分桶) | 识别延迟 | 串行处理 | 检测与上一批识别并行 | 字典映射 | 乱码/对齐 | 固定 vocab + 版本号 | ### CTC 解码贪心：移除重复与 blank；大规模需 Beam Search (权衡性能)。

4.6 NLP: BERT 推理优化

4.6.1 序列长度策略

1. 静态最大长度（简单，浪费算力）。
 2. Bucketing：按输入长短分类（32/64/128/256），多 OM。
 3. 动态 shape：需评估内存分配抖动；提前预热各常见长度。### FP16 注意点 LayerNorm/Softmax 数值范围敏感；若发现精度下降：保持部分算子 FP32（通过混合精度策略或模型修改）。### 性能指标 tokens/s、avg_latency_ms（batch=1 与 batch>1）、内存占用；观察自注意力占比，必要时进行剪枝（去除冗余 head）或蒸馏。

4.7 多模型 Pipeline 串联

4.8 工程目录与脚本标准

```
deploy/
    classify/
        export.py
        atc.sh
        config.yaml
detect/
    export.py
    atc.sh
ocr/
    export_det.py
    export_rec.py
    atc_det.sh
    atc_rec.sh
runtime/
    core/acl_session.cpp
    preprocess/
    postprocess/
    pipelines/
tests/
    data/
    benchmark/
docs/
    model_cards/
```

版本归档要求: | 产物 | 检查点 | | — | — | | *.om | 与 atc.log hash 对应 | | signature.json | 与运行时动态查询一致 | | metrics.json | 包含时间戳/commit_sha | | model_card.md | 模型来源/License/精度 |

4.9 性能基线方法与统计置信

推荐: 1. Warmup 5~10 次; 2. 收集 200 次稳定样本; 3. 计算 avg, p50, p95, p99; 4. 计算置信区间: $\text{mean} \pm 1.96 * (\text{std}/\sqrt{n})$; 5. 记录环境: 芯片序列号/温度区间/电源模式/版本矩阵。差异判定: 新版本 avg 降低 >5% 或 p95 上升 >8% 触发报警分析。

4.10 常见问题诊断深度版

| 问题 | 表现 | 诊断步骤 | 修复 |
|-------------|--------------|------------------|------------------|
| 输出全 0 | logits 恒定 | Dump 中间 tensor | 校验预处理/权重损坏 |
| 检测框偏移 | 坐标不准 | 可视化缩放/Pad 参数 | 修正 letterbox 逆变换 |
| OCR 乱码 | 字符错位 | 对比 index→char 映射 | 统一 vocab & 排序 |
| BERT 性能差 | tokens/s 低 | 分析长度分布 | 分桶/裁剪长度 |
| Pipeline 堵塞 | 帧延迟增长 | 监控队列深度 | 降帧/扩线程池 |
| 内存持续上涨 | long run OOM | 内存快照/工具 | 释放缓存/池化 |

4.11 章节小结

本章提供四类典型任务部署详解，并抽象了跨任务可复用的脚手架与性能度量方法。重点在于“输入契约统一”、“阶段解耦”、“可观测性内建”。掌握后可进入性能与算子优化专题。

4.12 实践任务

1. 部署 ResNet50: 输出 Top5 及概率、提交 metrics.json。
2. 部署 YOLOv5s: 5 张测试图片生成可视化结果（描述框坐标与类别统计）。
3. 构建 OCR 双模型流水线: 统计单帧平均文本块数 + 平均识别耗时。
4. BERT: 对 3 组长度 (32/64/128) 测 tokens/s 与时延差异，生成对比表。
5. Pipeline 检测 → 分类: 实现批裁剪 + Buffer 池，比较优化前后平均时延下降百分比。

5 性能与算子优化初阶

5.1 章节总览

本章聚焦“定位 → 解释 → 改善”闭环：从性能分析模型、Profiling 工具、瓶颈模式分类、布局与精度策略、内存与并行调度、到自定义算子开发与验证标准，提供工程可落地方法。目标是让读者具备：
A) 定量证明问题；B) 选择低风险优化策略；C) 保证功能与性能回归一致性。

5.2 性能拆解与衡量框架

总时延公式: $T_{total} = T_{pre} + T_{h2d} + T_{infer} + T_{d2h} + T_{post} + T_{idle}$ 。吞吐上限受制于 $\max(T_{component})$ ；需收集：
- 平均/分位数 (p50/p95)；
- 波动系数 $CV = \text{std}/\text{mean} > 0.15$ 需进一步剖析；
- 稳定性：长跑 1h 是否存在漂移（内存泄漏或热降频）。对比优化前后必须保留固定随机种子和数据集，消除噪声。

5.3 Profiling 工具与时间线解读

关键观测元素：
| 轨迹 | 意义 | 异常信号 | | —— | —— | ——— | | Stream Timeline | 内核调度顺序
| 大量空洞 gap | | MemCopy | H2D/D2H 开销 | 频繁小块拷贝 | | Task Kernel | 算子执行
| 个别算子异常拖长 | | Sync/Wait | Host 等待 | Wait 占比高 |

使用策略：
1. 先全量 Profile → 定位热点范围；
2. 二次局部 Profile (过滤特定算子类型)；
3. 导出 JSON → 自动解析器归档：算子耗时 TOPK, Copy 占比, Idle 时间。

5.4 瓶颈模式与处置策略矩阵

| 模式 | 识别特征 | 定量指标 | 处置优先级 | 策略 |
|-------------|----------------|--------------|-------|---------------------|
| 调度空洞 | Timeline gap 多 | Idle > 10% | 高 | 合并小算子 / 预加载数据 |
| 访存受限 | 算子耗时与内存带宽正相关 | 算子内核利用率低 | 中 | Layout 变换 / Tile 分块 |
| H2D 瓶颈 | Memcpy 比例高 | H2D>20% | 高 | 合并/异步/Pin/AIPP 下沉 |
| 后处理拖慢 | Post>25% | NMS/Decode 长 | 中 | 并行化 / Device 化 |
| 量化退化 | INT8 未获收益 | 时延差 <10% | 低 | 重新校准/混合精度 |
| 单 Stream 阻塞 | 单流串行 | Stream=1 | 中 | 多流/流水线 |

优先处理“结构性收益”>“微优化”，避免局部手工 hack 影响可维护性。

5.5 Layout / 内存访问优化

常见格式：NCHW（框架常用）、NHWC（部分算子优化）、NC1HWC0（硬件友好对齐），转换策略：在数据首次落地时转换一次；若前后模型不同布局，以中间标准布局连接，减少重复重排。对齐：通道/宽高按 16/32 边界对齐可提升访存一致性；小通道 (<16) 可考虑 --enable_small_channel 以加载优化内核。缓存复用：多模型共享中间 Buffer（需尺寸与 dtype 一致），通过分配表管理生命周期。

5.6 精度与性能的层级折衷

| 精度层级 | 描述 | 性能收益 | 风险 |
|---------|-------|----------|----------|
| FP32 | 基准 | - | 内存带宽/算力高 |
| FP16 | 半精度 | 1.2~1.6x | 累积误差 |
| INT8 对称 | 量化整型 | 1.5~2.2x | 量化噪声 |
| 混合精度 | 局部高精度 | 中等 | 实现复杂 |

量化流程要点: 1. 收集代表性校准集 (覆盖光照/尺度/类别分布); 2. 校准统计 (MinMax / KL); 3. 评估 Top1/Top5 差异、关键指标差异 (mAP/F1)。误差定位: Dump 中间张量 (FP32 vs INT8) → 层级误差分布 → 定位失真层 (常见: 激活饱和/尺度不均衡)。

5.7 内存管理专题

策略: 1. 长期 Buffer: 模型 I/O、常量 Workspace; 2. 短期 Buffer: Batch 临时中间; 3. 建立内存池 (按 size class 分类 1KB/4KB/16KB/64KB/大块), 分配 → 归还; 4. 避免频繁 aclrtMalloc/Free: 使用池化接口封装; 5. 监控: 每 60s 记录一次池使用率与系统剩余内存, 突增后回收未引用对象; 6. 大对象对齐: 按 512B/4KB 对齐减少碎片。

5.8 并行与流水线

多 Stream: 将独立算子或多模型分离到不同 Stream 并行调度; 注意 Host 侧同步点过多会抵消收益。Pipeline: Pre → Infer → Post 分线程队列, 目标是 In-Flight 帧数达到平衡 (过多增加延迟, 过少利用率低)。自适应调度: 定期评估每阶段平均耗时, 动态调整线程池大小 (PID 控制思想)。

5.9 自定义算子开发与评估

决策条件: | 条件 | 必须满足至少一项 | — | — | — | 复合算子频繁出现 | 合并降低访存 | 内置实现回退 Host | 存在高额拷贝 | 内核模式不适配输入规模 | 小尺寸性能差 |

流程: 需求分析 → JSON 定义 (op_type, attr, inputs/outputs) → Kernel C++ 模板 (向量化 / Tile) → 编译注册 → ATC 识别 → 功能单测 (随机张量对比) → 性能对比 (3 次 Warmup + 50 次统计)。评估表: | 版本 | 输入规模 | 平均耗时 (us) | P95(us) | 访存次数 | 速度提升 | 备注 | — | — | — | — | — | — | — | — |

5.10 优化案例：Add + ReLU 融合

原始：Add → ReLU 两个算子各自读写内存；融合：单 Kernel 计算 $\text{out} = \text{relu}(a+b)$ ：减少一次读写；收益估算：内存带宽主导场景中延迟 ($T_{\text{add}} + T_{\text{relu}}$ - 重叠)，实际提升 10~25%。验证：随机输入 100 次 → 检查数值一致（允许 $1e-6$ FP16 差异）→ Benchmark 对比。

5.11 性能报告与回归模板

```
{
    "commit": "<git-sha>",
    "model": "resnet50_fp16",
    "batch": 1,
    "avg_latency_ms": 5.87,
    "p95_latency_ms": 6.24,
    "throughput_fps": 170.3,
    "h2d_ms_ratio": 0.11,
    "post_ms_ratio": 0.05,
    "memory_peak_mb": 486,
    "temperature_c_range": "54-58",
    "profiling_date": "2025-09-04T10:21:00Z"
}
```

自动化：CI 中若 avg_latency_ms 高于基线 5% → 标红注释。

5.12 章节小结

性能优化不等于盲调：应以数据驱动 + 分层定位为前提，先解决架构级与内存/拷贝问题，再考虑算子级微调与自定义算子开发。量化收益需伴随精度风险评估，内存与并行策略需要可观测支撑。

5.13 实践任务

1. 对一个部署模型收集 Profiling JSON，输出前 5 算子耗时与占比表。
2. 实现 H2D 合并：将 3 个连续小拷贝合并为单次，比较平均时延改善。
3. 尝试 INT8 量化：输出精度与性能对比 (Top1/Latency/FPS)。
4. 编写一个 Add+ReLU 融合算子伪代码 + 预期性能提升估算。
5. 生成基线性能报告，并设定 CI 回归阈值策略文本说明。

5.14 昇腾 310B 自定义算子开发全流程

本节面向 Ascend 310B 推理场景，给出“什么时候需要自定义算子、用什么方法开发、如何编译注册、怎样验证与上线”的系统指引。读完后，你应能独立完成一个简单自定义算子的端到端落地。

5.14.1 开发概述

- 目标：当模型中存在“内置算子不支持/性能欠佳/需要业务特化融合”的场景，通过自定义算子（Custom Op）补齐功能或获得确定性性能收益。
- 实现形态：
 - AI Core (TBE/TE/TIK，运行于 NPU 核心，适合数值密集型向量/矩阵计算)。
 - AICPU (C++/CPU 实现，在 Host/AICPU 执行，适合控制流或少量数据处理，注意 H2D/D2H 开销)。
- 产物：算子描述(op info/proto)、算子实现(AI Core: Python 实现并编译为内核; AICPU: C++ so)、注册与打包(放入 OPP 路径)，以及 ATC 与运行时可识别的元数据。
- 适配 310B：选择 soc_version=Ascend310B，优先 FP16 数据通路；对齐 NC1HWC0 等硬件友好布局；小通道/小尺寸注意 tile 策略。

5.14.2 开发的理论基础

1. 硬件/内存模型(简要)：
 - GM(Global Memory)：大容量全局显存，带宽高、时延高；
 - UB(Unified Buffer)：片上高速缓存，容量有限，需 tile 分块搬运；
 - Vector/Scalar 单元：提供 vadd/vmul/vmax 等向量指令，需保证数据对齐(通常以 16/32 对齐)。
 - DMA：GM 与 UB 之间的数据搬运，批量大块优于频繁小块。
2. 计算表达与调度：
 - TE (Tensor Expression)：描述计算公式与算子图(compute)；
 - Schedule：描述分块(tiling)、并行、缓存、向量化等执行计划；
 - TIK DSL：更接近硬件指令级的编程接口，适合精细控制。
3. 算子契约(Operator Contract)：
 - 输入/输出张量的 shape、dtype、format(如 NCHW/NC1HWC0)、属性(attr)；
 - 广播/对齐规则、边界行为(溢出/饱和/舍入)、精度策略(FP16/FP32 混合)。
4. 形状推断与动态 shape：
 - ATC 需要根据 op 描述完成 shape infer；
 - 动态尺寸需在实现中处理 tile 策略切换并保证 UB 不溢出。

5.14.3 开发流程 (AI Core 为例)

以下流程以一个“Add+ReLU 融合”示例说明，读者可据此扩展到实际业务算子。

1) 环境准备

- 确保 CANN/Toolkit 已安装，能使用 atc、Profiling 等工具；
- 设置环境变量：
 - ASCEND_INSTALL_PATH 指向 Toolkit 根；
 - ASCEND_OPP_PATH 指向 OPP 包路径 (custom 算子将被放置于此)；
 - soc_version=Ascend310B (ATC/编译时指定)。

2) 定义算子信息 (op info/proto)

- 指定: op_type、inputs/outputs 名称、dtype/format 组合、属性列表、融合类型等；
- 作用：
 - 供 ATC 做图解析、形状推断与算子选择；
 - 供运行时校验输入输出与 kernel 适配。

3) 编写算子实现 (TE/TBE)

- 计算表达：“`python # 伪代码: y = relu(x1 + x2) import te.lang.cce as tbe from te import tvm def add_relu_compute(x1, x2): y = tbe.vadd(x1, x2) z = tbe.vmaxs(y, tvm.const(0.0, x1.dtype)) return z`”
- 调度策略 (示例要点):
 - 选择合适的 tile 以满足 UB 容量；
 - 将连续内存访问向量化，减少非对齐访问；
 - 尽量合并搬运，减少 GM<->UB 往返；
 - 小尺寸场景避免过度拆分导致调度开销占比过高。

4) 编译与注册

- 使用官方提供的 TBE 编译入口生成内核与元数据 (具体命令因版本而异，遵循已安装 Toolkit 的说明)；
- 将生成的实现文件/元数据放入 ASCEND_OPP_PATH 下的 custom 目录(如 op_impl/custom/ai_core/tbe、op_proto/custom)。

5) 与 ATC 集成

- 在模型转换时指定 --soc_version=Ascend310B；
- 确保 ATC 能从 ASCEND_OPP_PATH 读取到你的 op 描述与实现信息；
- 若需要限制实现选择，可使用 --op_select_implmode 配合算子实现指示。

6) 运行时部署与加载

- 运行环境中需要包含同样的 OPP 目录 (含 custom 实现)；

- 应用进程启动时配置环境变量，使 Runtime 能定位自定义算子实现；
 - 按常规 ACL 流程加载 OM 并执行推理。
- 7) 验证与度量
- 功能正确性：与参考实现（NumPy/ONNXRuntime）对齐，随机多组张量比较（均值绝对误差、相对误差、边界样本）。
 - 性能评估：Warmup 3 次 + 采样 50 次，输出 avg/p95/FPS；对比内置算子或未融合版本；
 - 资源占用：Profiling 检查 MemCopy 占比、Kernel 占比、Idle；
 - 兼容性：不同 shape/dtype/format 组合覆盖测试。
- 8) 文档与产物归档
- 输出 op_contract.yaml (IO/Attr/格式/边界规则)；
 - 输出 benchmark.json (avg/p95、对比基线、硬件/版本信息)；
 - 产物目录：op_pkg/<op_type>/<version>/ {op_proto, op_impl, tests, docs}。

5.14.4 AICPU 路线（可选）

- 适用：控制流、轻量数据处理或暂不需在 NPU 上运行的功能性算子；
- 实现：C/C++ 编写，遵循 AICPU 接口，注册到相应目录生成动态库；
- 注意：Host 执行会引入 H2D/D2H；若在性能关键路径，优先 AI Core 版本。

5.14.5 常见问题与排错

- ATC 提示 Unsupported Op：检查 op info 是否被正确放置且生效；确认 soc_version 与路径；
- 运行时 Fallback：确认实现 dtype/format 与模型一致；必要时扩充 dtype_format 组合；
- 性能未达预期：增大 tile、减少小块 DMA、合并计算、检查是否出现额外 layout 转换；
- 精度差异：检查饱和/舍入策略、对齐与广播规则、数据范围（FP16 溢出）。

5.14.6 本章小结

自定义算子是 310B 场景下“功能补齐与性能确定性”的关键手段。核心抓手包括：明确契约 (IO/格式/属性)、用 TE/TIK 描述计算并设计合理调度、放在 OPP 中正确注册、生效于 ATC 与运行时、用可度量的基线进行功能/性能回归。建议从“融合与复合算子”起步，优先选择计算密集、访存友好的目标，循序渐进积累模板与脚手架，以降低维护成本。

6 系统工程与高可用部署

6.1 章节总览

从单机多模型到工程化高可用体系：进程与线程模型、调度与优先级、配置与热更新、日志指标监控、故障感知和自愈、版本交付与灰度回滚。核心目标：让推理系统具备“可观察、可控、可自愈、可演进”。

6.2 部署形态与演进路线

| 阶段 | 形态 | 特征 | 触发升级条件 |
|---------|-------------|-----------|-------------|
| POC | 单进程 | 简单，耦合高 | 模型增加/稳定性需求 |
| Beta | 多进程模块化 | 隔离故障 | 资源利用不均/需要扩展 |
| Prod 基础 | 本地 RPC 服务化 | 清晰 API 契约 | 多板协同/多客户端 |
| Prod 进阶 | 容器化 + 编排 | 可滚动更新 | 大规模交付/远程运维 |
| Edge 集群 | 中心调度 + 远程控制 | 全局负载均衡 | 弹性/集中监控 |

进程边界建议：capture、infer、postprocess、upload、monitor、watchdog。隔离崩溃影响并实现差异化资源限额（CPU 亲和 + 内存限制）。

6.3 进程与线程模型设计

6.3.1 基本原理

1. 最小可信核心：推理执行逻辑 + 输入输出队列；
2. 外围增强：监控、日志聚合、健康探针不影响核心路径。

6.3.2 线程池建议

| 线程组 | 职责 | 数量估算 |
|-------------|------------------|---|
| Capture | 采集与解码 | 摄像头数 (N) |
| Preprocess | Resize/Normalize | $\text{ceil}(N * \text{frame_rate} * \text{pre_time} / \text{CPU 核})$ |
| Inference | 调用 ACL | 通常 1~2 (避免过度上下文切换) |
| Postprocess | NMS/Decode | 与 Inference 分离防止阻塞 |
| Upload | 事件上报 | 1~2 |
| Monitor | 指标收集 | 1 |

CPU 亲和：将推理线程绑定至高性能核心，避免迁移污染缓存；预处理线程放置在剩余核心以平衡。

6.4 任务调度与优先级控制

多级队列：RealtimeQueue（最大长度 L1，满则丢弃旧帧）、NormalQueue（批处理）、BackgroundQueue（低优先日志/统计）。令牌桶限速：对外部请求（远程推理 API）采取令牌桶控制 QPS；令牌不足则延迟或返回限流错误码。超时策略：当帧在队列停留超过阈值（如 $2 \times$ 平均推理时延）标记过期，进入降级路径（丢弃或简化处理）。

6.5 配置管理与热更新

配置划分：

| 类别 | 内容 | 更新频率 | 是否热更新 |
|----|------------|------|---------|
| 资源 | 线程数、队列长度 | 低 | 是 |
| 模型 | 路径、版本、精度模式 | 中 | 滚动加载 |
| 策略 | 阈值、降级条件 | 中高 | 是 |
| 安全 | Token、公钥 | 低 | 非热（需重启） |

热更新流程：文件变更 → 校验 schema → 写入新 shadow 副本 → 原子指针切换（正在执行任务继续使用旧配置直至完成）。

6.6 日志体系与追踪

结构化字段: ts, level, module, thread, trace_id, latency_ms, event。Trace ID: 跨进程通过 IPC/RPC header 传递; 用于从采集到上报的全链路追踪。日志级别动态调整: 接收管理命令 (Unix Domain Socket / 本地控制端口) 将模块日志级别置 DEBUG 进行临时诊断。切割策略: 按大小 (100MB) 或按时间 (小时), 超限自动压缩归档; 保留策略 N 天 + 关键事件永久。

6.7 指标监控与探针

探针:

- Liveness: 进程是否在运行 (看门狗检查心跳文件更新时间)。
- Readiness: 模型是否加载完成 + 队列是否低压 (长度 < 阈值)。指标暴露格式: /metrics Prometheus 文本: model_latency_bucket{le="..."} 123。

核心指标分类:

| 分类 | 指标 | 说明 |
|-----|---------------------------------|--------|
| 性能 | model_latency_ms (histogram) | 推理时延分位 |
| 吞吐 | frames_processed_total | 每秒增量 |
| 背压 | queue_len / queue_wait_ms | 排队深度 |
| 资源 | npu_util / cpu_util / mem_bytes | 资源利用率 |
| 可靠性 | crash_count / restart_count | 重启频次 |
| 热 | temperature_c | 温度曲线 |
| 质量 | accuracy_drift | 精度回归差异 |

6.8 高可用与自愈机制

看门狗: 子进程每隔 T 秒写心跳文件; 超时 → 发送 SIGTERM → 宽限期 → SIGKILL → 重启并记录事件。

分级降级:

1. 软降级: 减小输入分辨率 / 降 FPS / 关闭次要模型;
2. 硬降级: 仅保留关键检测模型;
3. 熔断: 持续高温或资源不可用 → 暂停推理, 仅缓存数据。状态机: NORMAL → DEGRADED → CRITICAL → RECOVERY → NORMAL。

6.9 异常分类与处理矩阵

| 类别 | 触发信号 | 初步动作 | 深度动作 | 记录 |
|----|--------|--------------|---------|-------------|
| 输入 | 空帧/花屏 | 丢弃 + 计数 | 摄像头重置 | anomaly.log |
| 资源 | OOM 风险 | Dump 内存 | 重建上下文 | memory.log |
| 性能 | P95 飙升 | Profiling on | 降级策略 | perf.log |
| 硬件 | 温度高 | 降载 | 风扇策略/报警 | thermal.log |
| 数据 | 精度偏移 | Dump 样本 | 模型回滚 | quality.log |

6.10 版本、灰度与回滚

镜像标签: <model_version>—<git_sha>—<date>; 包含 manifest: 模型 hash、配置 hash、构建环境。灰度策略: 按设备集合 (Region/Batch) 逐步扩大; 监控关键指标偏差 (时延/Crash) 超过阈值立即回滚。回滚: 保留上一稳定版本镜像与配置快照; 执行原子 symbolic link 切换。

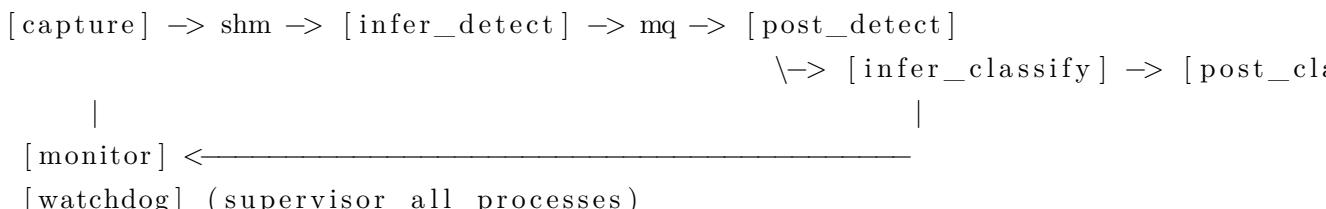
6.11 安全与访问控制

最小权限: 运行用户无 sudo; 只读挂载代码与模型目录, 写权限仅日志与缓存路径。配置签名: 管理端生成签名, 客户端部署时校验防篡改。远程指令: 白名单 + 签名校验; 禁止执行任意 shell。

6.12 审计与合规

记录: 运维操作、配置变更、模型替换、异常重启; 保存 JSON Line 格式, 便于集中检索。设定留存策略和脱敏规则 (剔除用户标识)。

6.13 示例: 两模型多进程结构



共享内存 (shm) 用于高带宽帧传输, 消息队列 (mq) 传递元数据 (指针、时间戳、追踪 ID)。

6.14 章节小结

通过模块化、可观察化与自动化自愈策略，边缘推理系统可以在资源约束与环境不稳定条件下提供接近云端的可靠性。重点：明确边界、度量驱动、降级可逆、版本可控。

6.15 实践任务

1. 设计多进程与队列拓扑图（ASCII）。
2. 编写队列监控小工具：输出队列长度与平均等待时长。
3. 实现一个看门狗脚本（检测心跳文件时间差 $>$ 阈值则重启模拟进程）。
4. 制作灰度发布计划（分三阶段 + 指标 + 回滚条件）。
5. 输出降级状态机定义（含转移条件）。

7 项目实战方法论与交付模板

7.1 章节总览

本章建立从需求澄清 → 指标体系 → 评测集 → 迭代节奏 → 资产沉淀 → 交付与回归的一套闭环方法论，让技术决策基于指标与风险敞口，而非经验臆测。核心理念：可量化、可比较、可复用、可追溯。

7.2 需求澄清 Canvas

| 维度 | 要素 | 问题提示 | 示例 |
|----|----------------|-------------|-------------------------|
| 场景 | 输入源/运行环境 | 摄像头？批处理？ | 室内 1080p30 低光 |
| 目标 | 功能/业务价值 | 用户希望看到什么结果？ | 实时检测 + 分析 |
| 指标 | Latency/FPS/精度 | 哪些分位数重要？ | <80ms / 25FPS / mAP 0.6 |
| 约束 | 能耗/内存/带宽 | 上限是多少？ | 功耗 15W 内存 3GB |
| 风险 | 数据/硬件/算法 | 失败模式有哪些？ | 低光/遮挡/抖动 |
| 合规 | 隐私/许可 | 是否需要脱敏？ | 仅上传事件元数据 |
| 成本 | 硬件/云 | ROI 衡量？ | 10 台板卡预算 |

输出：requirement.yaml（版本化），后续所有评审基于此文档。

7.3 指标分层与优先级

| 层级 | 类别 | 指标 | 说明 | 失败后果 |
|-------|----|-------------|-------|---------|
| SLO A | 体验 | p95 延迟 | 端到端 | 体验差/丢帧 |
| SLO A | 性能 | FPS | 稳态吞吐 | 处理拥堵 |
| SLO B | 质量 | mAP/F1/Top1 | 任务正确性 | 无法满足业务 |
| SLO B | 稳定 | Crash/小时 | 可靠性 | 运维成本高 |
| SLO C | 资源 | 内存峰值/功耗 | 成本约束 | 设备异常/降频 |
| SLO C | 带宽 | 上行 kbps | 成本/合规 | 费用/拥塞 |

优先级：先保障 A（体验 + 功能可用），再稳定 B（质量/稳定），最后优化 C（资源效率）。

7.4 Baseline 策略与控制变量法

Baseline 目标：建立“最小改动可运行”标尺。原则：

1. 不提前做微优化；
2. 记录所有关键参数：模型版本、输入尺寸、预处理策略、硬件温度范围；
3. 一次仅改变单个变量(batch、精度、线程数)。基线存档：baseline/<date>-<commit>/metrics.json；对比脚本生成差异报告。

7.5 评测集设计原则

| 原则 | 内容 |
|------|---------------|
| 代表性 | 涵盖主流场景/光照/角度 |
| 覆盖边界 | 极端尺寸、模糊、遮挡 |
| 可再现 | 文件命名规范 + 固定清单 |
| 可扩展 | 新增样本不破坏旧索引 |
| 标注一致 | 标注工具/规范/审校流程 |

目录示例：

```
dataset_eval/
images/
  day/*.jpg
  night/*.jpg
  occlusion/*.jpg
```

```

annotations/
instances_train.json
instances_val.json
meta/
README.md
version.txt

```

提供 Hash 列表，防止样本被替换而影响回归可信度。

7.6 迭代计划与看板

四阶段：

| Sprint | 目标 | 核心产出 | 风险控制 |
|--------|---------|------------------|-------------|
| 0 | 环境/基线 | baseline metrics | 依赖清单齐全 |
| 1 | 精度与功能稳固 | 精度报告 | 数据问题快速反馈 |
| 2 | 性能与稳定 | 性能对比表/监控上线 | Watchdog 验证 |
| 3 | 工程交付包装 | Release Notes/脚本 | 灰度计划制定 |

看板列：Backlog → Doing → Review → Bench → Done；性能/精度任务需进入 Bench 列执行对比脚本通过后才可 Done。

7.7 资产沉淀文档体系

| 文档 | 内容 | 更新频率 |
|--------------|---------------|-------|
| README | 快速启动 | 版本变化时 |
| ARCHITECTURE | 架构图/模块说明 | 结构调整 |
| MODEL_CARD | 模型来源/许可/精度/限制 | 模型更新 |
| EVAL_REPORT | 数据与评测方法/指标 | 每次发布 |
| PERF_REPORT | 基线/优化对比 | 优化后 |
| CHANGELOG | 可见版本差异 | 每次版本 |
| RISK_LOG | 已知风险列表 | 动态 |

MODEL_CARD 需包含：数据来源、训练超参摘要、输入契约、已知局限、许可（如 Apache-2.0）、安全与偏见说明（若涉及识别敏感属性声明避免用途）。

7.8 交付目录与不可变产物

```

release/
v1.0/
    manifest.json      # 产物 hash / 版本矩阵
    models/
        detect.om
        classify.om
        signature.json
    scripts/
        run.ps1
        run.sh
        watchdog.sh
    configs/
        default.yaml
    docs/
        model_card_detect.md
        model_card_classify.md
        QUICKSTART.md
    reports/
        perf.json
        accuracy.json

```

manifest.json 字段: {version, commit, build_time, model_hashes, dependencies}。

7.9 上线前综合 Checklist

| 类别 | 检查项 | 通过标准 |
|----|-------------------|-------------|
| 功能 | 核心用例 100% | 自动化用例通过 |
| 性能 | p95 < 目标 +5% | 连续 30min 稳定 |
| 精度 | mAP/Top1 回归差 < 阈值 | 与基线对比 |
| 资源 | 内存峰值 < 75% | 1h 稳态无泄漏 |
| 稳定 | Crash=0, 重启 =0 | 守护日志清洁 |
| 安全 | 日志无敏感泄露 | 关键字段脱敏 |

| 类别 | 检查项 | 通过标准 |
|----|----------|----------|
| 配置 | 签名校验一致 | Hash 匹配 |
| 回滚 | 验证上一版本可用 | 切换 < 30s |

7.10 验收、回归与漂移监测

交付后 7 天加密监控：记录时延、精度漂移（采样对比模型输出变化）。漂移检测：相同输入集合（Shadow Set）每天抽样跑一次 → 统计 logits KL 散度/TopK 变化率，高于阈值（如 $KL > 0.05$ ）触发报警（潜在数据分布变化或模型文件损坏）。回归集版本化：eval_set_vX；若需替换样本 → 新增版本，不覆盖旧数据。

7.11 风险管理与决策日志

风险登记表：risk_log.md 每条包含：ID、描述、影响、概率、缓解、当前状态。决策日志（Decision Record, ADR）：记录架构/模型/精度策略选择及备选方案放弃理由，以便新成员快速建立上下文。

7.12 章节小结

方法论的核心不是流程文档堆砌，而是“指标驱动 + 资产沉淀 + 可回滚”三支柱。通过契约化需求、标准化 Baseline、规范化评测与回归体系，使团队协作更高效、风险暴露更透明、交付结果更可信。

7.13 实践任务

1. 输出 requirement.yaml（含指标与约束）。
2. 构建 30 张代表图像的 mini 评测集并附 Hash 列表。
3. 生成 baseline_metrics.json 与后一次优化对比 diff 报告。
4. 制作一个 MODEL_CARD 模板并填写一个模型示例。
5. 编写上线 Checklist 并模拟一项未通过情形与处置方案。

8 合实战案例集

8.1 章节总览

本章通过九个真实应用场景串联前面章节的知识：模型选择、转换、部署、性能与稳定性验证、迭代优化。所有案例采用统一模板，支持快速复制与对比评估。强调“结构化指标 + 自动化脚本 + 可视化反馈”。

8.2 案例统一模板（标准化规范）

| 区块 | 内容要点 | 产出文件 |
|------|----------------------|--------------------|
| 场景描述 | 背景/输入/目标 | README.md #scene |
| 指标目标 | 延迟/FPS/精度/资源 | requirement.yaml |
| 模型选择 | 候选对比 + 取舍 | model_card*.md |
| 数据准备 | 采集/标注/增强 | data_prep.md |
| 转换部署 | 导出 → ATC 参数 | atc.sh / export.py |
| 运行脚本 | 启动/参数/日志路径 | run.sh / run.ps1 |
| 性能结果 | metrics.json (基线/优化) | metrics/*.json |
| 质量验证 | 精度/漂移检查 | accuracy.json |
| 风险改进 | 已知问题/迭代计划 | roadmap.md |

8.3 案例目录结构规范

```
experiments/caseX/
  README.md
  requirement.yaml
  models/          # onnx / om / signatures
  scripts/
```

```

export.py
atc.sh
run_infer.py
benchmark.py
data/          # 样本(或下载指令)
metrics/
    baseline.json
    optimized.json
logs/
eval/
    accuracy.json
    drift.json
assets/        # 截图/示意图

```

8.4 例概览与重点

| 序 | 名称 | 关键技术点 | 指标核心 | 风险要素 |
|---|-------|----------------|-------------|-------|
| 1 | 人脸打卡机 | 人脸检测 + 比对 + 活体 | 识别成功率/伪拒率 | 光照/遮挡 |
| 2 | 实时跟踪 | 检测 + 多目标关联 | 跟踪稳定度(IDF1) | 遮挡/抖动 |
| 3 | 智能电子琴 | 音频节拍识别 + 分类 | 识别延迟/准确率 | 噪声/延迟 |
| 4 | 掌纹识别 | ROI 提取 + 特征匹配 | 误识率/拒识率 | 采集姿态 |
| 5 | 数据采集仪 | 传感融合 + 缓存上传 | 数据丢失率 | 网络波动 |
| 6 | 智能小车 | 目标检测 + 路径策略 | 决策延迟 | 传感器同步 |
| 7 | 智能相册 | 分类 + 聚类 + 去重 | 聚类纯度 | 相似干扰 |
| 8 | 手势识别 | 时序建模(TSM) | 手势准确率/FPS | 动作模糊 |
| 9 | 聊天机器人 | NLP 推理 + 缓存 | 响应时延/意图准确 | 语料漂移 |

下列示例详细展开前三个具代表性的模式。

8.5 案例 1：人脸打卡机

8.5.1 场景

摄像头实时输入，人脸检测 → 关键点对齐 → 特征提取 → 特征库比对 → 授权决策 → 事件上报。

8.5.2 指标

| 指标 | 目标 | 说明 |
|-----------|---------|---------|
| 平均识别时延 | < 120ms | 从帧采集到结果 |
| 最大 P95 | < 150ms | 抖动控制 |
| 误识率 (FAR) | < 0.001 | 安全性 |
| 拒识率 (FRR) | < 0.02 | 体验 |

8.5.3 模型链路

1. 人脸检测 (RetinaFace);
2. 5 点关键点仿射对齐;
3. ArcFace 特征 512D;
4. 向量归一化 + 余弦相似度;
5. 阈值自适应 (基于滑动窗口均值校正)。

8.5.4 性能优化

- 批量特征比对：向量库转矩阵，使用 SIMD/BLAS;
- 缓存：最近识别通过用户特征缓存，减少重复比对；
- 光照增强：低光阈值触发 Gamma/直方图均衡。

8.5.5 metrics 示例

```
{
  "avg_latency_ms": 98.4,
  "p95_latency_ms": 121.3,
  "fps": 10.1,
  "face_detect_ms": 42.1,
  "feature_ms": 18.7,
  "match_ms": 5.2,
  "false_accept_rate": 0.0008,
  "false_reject_rate": 0.017
}
```

8.6 案例 2：实时跟踪（检测 + 关联）

8.6.1 流程

帧采集 → 目标检测 → 外观特征提取 → 卡尔曼预测 → 匈牙利匹配 → 轨迹输出。

8.6.2 难点

遮挡/丢失：轨迹生命周期管理（状态：Tentative → Confirmed → Lost → Removed）。

8.6.3 优化

1. 检测降频：每 N 帧做一次全检测，中间帧仅跟踪预测；
2. 多线程：检测与跟踪解耦；
3. ReID 模型轻量化（裁剪通道）。

8.6.4 评估指标

IDF1、MOTA、FP/FN、IDSW（身份切换）。

8.7 案例 3：智能电子琴（音频）

8.7.1 流程

音频采集 16kHz → 窗口分帧 FFT → 频谱/梅尔特征 → 分类模型（音符/节奏）→ 校准节拍输出。

8.7.2 优化点

FFT 批处理使用向量库；低延迟滑动窗口；模型输出置信度平滑（指数滑动平均）。

8.7.3 指标

节拍延迟 < 80ms；识别准确率 > 95%。

8.8 结果记录与差异报告

基线与优化版本差异自动生成：

| 指标 | baseline | optimized | 差异 | 状态 |
|-------------------|----------|-----------|--------|----|
| avg_latency_ms | 112.5 | 98.4 | -12.5% | |
| p95_latency_ms | 140.3 | 121.3 | -13.5% | |
| false_accept_rate | 0.0012 | 0.0008 | 改善 | |

8.9 自动化与复现保障

| 机制 | 说明 |
|-----------------|------------------------|
| Hash 校验 | onnx/om/脚本确保未篡改 |
| repeatable seed | 设定随机种子统一实验 |
| benchmark.py | 统一输出 metrics.json |
| drift 检测 | 周期性对比指标偏差 |
| 一键脚本 | run.sh + run.ps1 支持跨平台 |

8.10 指标可视化建议

- 时间序列: Latency / FPS / 温度。
- 箱线图: 不同优化阶段的时延分布。
- 堆叠条: 阶段占比 (检测/特征/比对)。
- 散点: 光照水平 vs 识别准确度。

8.11 通用问题经验库

| 问题 | 案例 | 根因 | 处理 |
|--------|-----|--------|-----------------|
| 相机丢帧 | 1/2 | 帧率不稳 | 缓冲 + 限速 |
| 模型加载慢 | 全部 | 冷启动未预热 | 预加载预热 10 次 |
| OCR 错字 | 新增 | 图像模糊 | 降噪/锐化 |
| 跟踪漂移 | 2 | 过度遮挡 | reinit + 短期外观缓存 |

8.12 扩展方向

- 多模态融合（视觉 + 语音指令）。
- 硬件加速协同（NPU + DSP 解码）。
- 大模型边缘裁剪（蒸馏 + 量化 + 分层推理）。

8.13 贡献工作流

1. Fork → 分支: case/<name>;
2. 新建目录遵循模板;
3. 提交包含: README、metrics、脚本、model_card;
4. CI 自动校验格式与 hash;
5. PR 模板填写: 动机/数据/指标/风险。

8.14 章节小结

案例是知识的验证与反哺：通过统一模板与自动化度量，形成可延展的案例库，帮助新模型与新任务快速落地并保障质量。

8.15 实践任务

1. 搭建 case1 目录，生成 baseline metrics。
2. 实现 face detection + feature 比对流程，并输出 FAR/FRR。
3. 将一次优化（裁剪/量化）前后差异写入 diff 表。
4. 编写 benchmark.py：支持 --repeat N --output metrics.json。
5. 增加 drift 检测脚本（比较两次 metrics 差异，阈值报警）。

9 附录与工具箱

9.1 章节总览

本附录聚焦“查得快、用得稳”：常见报错速查、转换参数模板、性能/质量 Checklist、术语字典、推荐资源与社区贡献规范。可作为日常开发随手翻阅的工具章节。

9.2 常见报错速查

| 分类 | 报错/现象 | 可能原因 | 排查步骤 | 解决建议 |
|---------|------------------------------|--------------|-------------------------|-----------------------|
| ATC | E19001: Op Not Supported | 新算子版本落后 | 确认 CANN 版本 + onnxsim 简化 | 升级/替换结构/自定义算子 |
| ATC | Shape 推断失败 | 动态维度不明确 | 检查 --input_shape/动态参数 | 固定关键维度或提供范围 |
| ACL | aclmdlLoadFromFile 权限问题/模型损坏 | 权限/模型损坏 | 校验文件 hash/权限 | 修正权限/重新生成 OM |
| Runtime | OOM / alloc 失败 | Batch 或分辨率过大 | 统计输入分布 | 降 batch/分桶/复用内存 |
| 运行 | 推理输出 NAN | 数值溢出/量化尺度错误 | Dump 中间 Tensor | 调整量化/保留 FP32 层 |
| 性能 | Timeline 大量 gap | Host 阻塞/小算子 | Profiling 分析 | 合并算子/异步预取 |
| 性能 | H2D 高占比 >25% | 多次小拷贝 | 合并缓冲 | AIPP 下沉/批量化 |
| 精度 | Top1 下降 >1% | 预处理不匹配 | 对比 ONNX 输出 | 统一 Normalize & Layout |
| 精度 | mAP 不稳定 | 阈值或 NMS 误差 | 调整阈值/比对中间框 | 校准 NMS 公式/尺度 |

| 分类 | 报错/现象 | 可能原因 | 排查步骤 | 解决建议 |
|----|----------|-----------|--------------|------------|
| 稳定 | 间歇 Crash | 悬空指针/并发访问 | 启用 ASAN/日志回溯 | 修订生命周期/加锁 |
| 部署 | 模型加载慢 | 冷启动/IO 慢 | 预热/缓存 | 预加载 + 固态存储 |
| 安全 | 日志泄露敏感路径 | 直接 print | grep 审计 | 结构化日志脱敏 |

9.3 模型转换参数模板合集

9.3.1 分类模型 (ResNet)

```
atc --model=resnet50.onnx \
--framework=5 \
--output=resnet50_fp16 \
--input_format=NCHW \
--input_shape="input:1,3,224,224" \
--soc_version=Ascend310B \
--precision_mode=allow_fp32_to_fp16 \
--log=info
```

9.3.2 YOLO 动态分辨率

```
atc --model=yolov5s.onnx \
--framework=5 \
--output=yolov5s_640_768 \
--dynamic_image_size="640,640;768,768" \
--input_format=NCHW \
--soc_version=Ascend310B \
--op_select_implmode=high_performance \
--precision_mode=allow_fp32_to_fp16
```

9.3.3 INT8 量化 (示例)

```
atc --model=resnet50.onnx \
--framework=5 \
```

```
--output=resnet50_int8 \
--input_format=NCHW \
--input_shape="input:1,3,224,224" \
--soc_version=Ascend310B \
--precision_mode=allow_mix_precision \
--insert_op_conf=aipp.cfg \
--enable_small_channel=true
```

9.4 性能与质量 Checklist (执行勾项)

性能:

- Profiling 无明显 Idle gap > 10%
- H2D + D2H 占比 < 25%
- Postprocess 占比 < 20%
- Stream 利用率平衡 (无单流饱和)
- 使用内存池减少频繁 alloc/free

精度:

- ONNX vs OM Top1 差异 < 0.2%
- L1 平均误差 < 1e-3 (FP16)
- NMS 输出框数量与基线差异 < 1 框/图 (平均)
- INT8 校准集覆盖多场景

稳定性:

- 1h 稳态无 Crash / OOM
- 温度在安全区间 < 85°C
- 看门狗重启次数 = 0

安全:

- 日志无明文密钥
- 模型文件 hash 校验通过

9.5 术语表 (扩展)

| 术语 | 说明 |
|----|------------------|
| OM | Ascend 离线模型二进制格式 |

| 术语 | 说明 |
|-------------------|-------------------|
| ACL | Ascend 计算语言 API 层 |
| ATC | 模型转换/编译工具 |
| AIPP | 自动图像预处理模块 |
| Stream | 异步任务调度通道 |
| Profiling | 性能采样分析工具体系 |
| Fallback | 算子未匹配优化实现退回通用实现 |
| Quant Calibration | 量化尺度统计过程 |
| Baseline | 初始标准对照性能/精度集 |
| Drift | 指标随时间未经预期的漂移 |

9.6 推荐资源与外部引用

- Ascend 官方文档入口（安装/算子列表/最佳实践）
- CANN Release Notes: 版本兼容与已知问题。
- ONNX Operator 列表与语义说明。
- Open Model Zoo / ModelScope: 获取预训练模型与许可信息。
- 学术资源：算子融合、低比特量化、蒸馏相关论文列表。

9.7 贡献指南摘要

流程: Fork → 新分支 → 修改/新增 → 本地 lint & 生成脚本 → PR (描述动机/影响面/验证方式)。PR 要求: | 要素 | 说明 | | — | — | | 标题 | 简明说明改动作用 | | 描述 | 背景 + 修改点 + 风险 | | 验证 | 性能/精度/功能截图或数据 | | 回滚 | 若失败如何恢复 | | 关联 Issue | 追踪链接 |

9.8 FAQ

| 问题 | 回答 |
|-----------|------------------------------|
| 模型转换慢怎么办？ | 使用 SSD，关闭调试日志，检查不必要动态 shape。 |
| 精度下降如何定位？ | 离线脚本层级 Dump 比对，逐层二分。 |
| 如何减少内存占用？ | 启用内存池 + 减少中间冗余张量 + 固定 batch。 |

| 问题 | 回答 |
|-----------|------------------------------------|
| 量化后收益不明显? | 检查是否 Compute-bound, 或激活分布集中导致尺度相近。 |
| NMS 很慢? | 合并小框批量处理/降低候选阈值/考虑 Device 版 NMS。 |

9.9 License 与引用

本书内容遵循 Apache 2.0 许可证。引用: > 《昇腾 310B 实战: 从入门到精通边缘计算与人工智能》(GitHub: zhouxzh/Ascend310)

9.10 版本路线回顾

| 版本 | 内容 | 目标 |
|------|--------|--------|
| v0.1 | 结构框架 | 验证框架可行 |
| v0.3 | 核心部署链路 | 形成可用主线 |
| v0.6 | 案例与工程化 | 贴近实战 |
| v1.0 | 全面审校发行 | 正式发布 |

9.11 实践任务

1. 为你的项目添加 1 条本地常见错误记录 (含根因与解决)。
2. 复制分类 ATC 模板并改写为检测模型版本 (含动态尺寸)。
3. 在术语表补充 3 个任务相关术语 (并验证唯一性)。
4. 选取 FAQ 一条, 写出更深入排查脚本思路。

10 导读与准备工作

10.1 章节总览

本章提供“鸟瞰 + 上手 + 约定 + 协作”四个维度：帮助读者在开始代码与实验前，建立清晰地图、完成环境自检、理解术语规范，并加入协作迭代。阅读后应能：A) 明确个人学习路径；B) 快速完成最小可行部署；C) 识别后续章节间的依赖关系。

10.2 全书主线结构

技术主线：硬件与环境 (1) → 软件栈与转换 (2) → 边缘系统视角 (3) → 典型部署实践 (4) → 性能与算子优化 (5) → 高可用工程体系 (6) → 方法论与交付 (7) → 综合案例 (8) → 工具与附录 (9)。
知识图谱建议：

硬件 / 板卡 → CANN 组件 → 模型转换 → 推理编程 → 多模型流水线 → 性能调优 → 系统可靠

10.3 读者路径矩阵

| 角色 | 起步路径 | 可跳过 | 深挖章节 | 目标里程碑 |
|-------|---------------|----------|------------|-----------|
| 零基础 | 1 → 2 → 4 | 5 深度优化细节 | 8 案例 | 跑通首个端到端推理 |
| 嵌入式 | 1 → 2 → 5 → 6 | 7 方法论部分 | 5/6 性能与可靠性 | 优化资源占比 |
| AI 应用 | 2 → 4 → 7 → 8 | 1 硬件细节 | 4/8 部署差异 | 多任务流水线 |
| 技术负责人 | 0 → 3 → 6 → 7 | 具体算子实现 | 7 评测体系 | 制定团队标准 |

10.4 硬件准备与兼容性

| 组件 | 推荐 | 说明 | 检查点 |
|-----|---------------------|--------|-----------------|
| 开发板 | OrangePi AIpro 310B | 标准平台 | npu-smi 识别型号 |
| 存储 | TF 64G+ / SSD | 加速 I/O | iostat 延迟 <10ms |
| 散热 | 风扇 + 鳍片 | 长时间稳定 | 温度 < 85°C |
| 摄像头 | USB UVC / MIPI | 即插即用 | v4l2-ctl 列设备 |
| 网络 | 千兆以太网 | 低抖动 | ping 丢包率 0 |
| 电源 | PD 65W | 稳定供电 | 无随机重启 |

准备完成后记录 hardware_inventory.md: 型号、序列号、固件版本、功耗模式。

10.5 软件与工具栈细化

| 层级 | 工具/组件 | 说明 |
|--------|-------------------------------|----------------------|
| OS | Ubuntu 22.04 / openEuler | 官方验证环境 |
| 驱动/固件 | 对应 CANN 版本 | 版本矩阵对齐 |
| CANN | Toolkit + Runtime | 提供 atc/acl/profiling |
| Python | 3.10+ | 脚本与评测 |
| 依赖 | numpy/onnx/onnxruntime/opencv | 模型与预处理 |
| 调试 | npu-smi/Profiler/日志系统 | 性能与稳定性分析 |

建议创建 requirements.txt 并使用 venv 或 Conda 隔离。

10.6 仓库目录与命名约定

| 目录 | 内容 | 约定 |
|-------------|------------|--------------|
| src/book | 文本章节 | 章节号前缀固定 |
| experiments | 案例 | caseX 模式 |
| models | 原始/导出中间模型 | 按模型名/版本 |
| scripts | 通用脚本 | 跨平台 .sh/.ps1 |
| tools | 辅助分析脚本 | 单一功能命令化 |
| docs | 生成 PDF / 图 | 不放大模型文件 |
| benchmarks | 性能记录 | 时间戳 + commit |

命名: <model>_<precision>_<shape>.om, 例如 yolov5s_fp16_1x3x640x640.om。

10.7 最小可行环境验证 (MVE)

执行脚本 scripts/verify_env.sh (建议添加):

1. npu-smi info: 输出芯片与状态;
2. atc --version: 版本号记录;
3. 运行随机张量推理 (内置简单 OM 或最小网络) 验证 ACL API;
4. Profiling 采集一次, 生成 timeline 文件;
5. 记录结果写入 env_report.json。

判定: 如某步骤失败阻断后续章节学习。

10.8 全局术语与约定

| 术语 | 约定 | 说明 |
|-----------|---------------|-------------------|
| FPS | frames/second | 统计处理输出帧数 |
| Latency | ms | 端到端完成时间 |
| Pxx | 分位数 | P95/P99 评估抖动 |
| Pipeline | 阶段组 | 多阶段并行结构 |
| Signature | 模型签名 | I/O 名称与形状/格式 json |
| Baseline | 初始基线 | 第一版性能/精度记录 |

所有时间单位默认 ms; 数据大小默认字节 (显式写 MB/GiB 时需指出换算基数)。

10.9 协作工作流与质量闸门

工作流: Issue (需求/缺陷) → 分支 feat | fix /<topic> → 提交 (含描述) → PR → 自动测试 (Lint + 精度/性能轻测) → Review → Merge。质量闸门:

| 闸门 | 说明 | 未通过处理 |
|--------------|---------|--------|
| Lint | 代码/文档格式 | 修复后再提交 |
| Spell | 关键术语拼写 | 更正 |
| Signature 验证 | 模型签名一致 | 拒绝合并 |

| | | |
|------|------------|-------|
| 闸门 | 说明 | 未通过处理 |
| 基线回归 | 性能/精度差异超阈值 | 标注需说明 |

PR 模板字段: Motivation / Changes / Test / Risk / Rollback Plan。

10.10 学习与实践建议

- 完成前 3 章后立即挑选一个轻量模型跑通部署（建立正反馈）。
- 每章输出“总结卡片”：知识点 → 应用场景 → 潜在风险。
- 建议建立个人实验日志：参数、结果、疑问与下一步假设。
- 失败样本收集：创建 failure_cases/ 目录存储误检/漏检图像用于持续改进。

10.11 常见初学误区与规避

| 误区 | 结果 | 规避 |
|---------|--------|-----------------|
| 直接优化无基线 | 无从评估收益 | 先建立 baseline |
| 混用不同预处理 | 精度随机波动 | 抽象统一函数 |
| 缺少签名文件 | 部署时出错 | 每次转换生成签名 |
| 未记录环境版本 | 难以复现 | env_report.json |
| 长日志未切割 | 磁盘占满 | 配置滚动策略 |

10.12 章节小结

通过环境、目录、术语、协作流程的标准化，后续学习聚焦问题本身，而不是环境与沟通摩擦。建议读者在继续前先完成“最小可行环境验证”并记录结果，以便后续调试时快速排除环境因素。

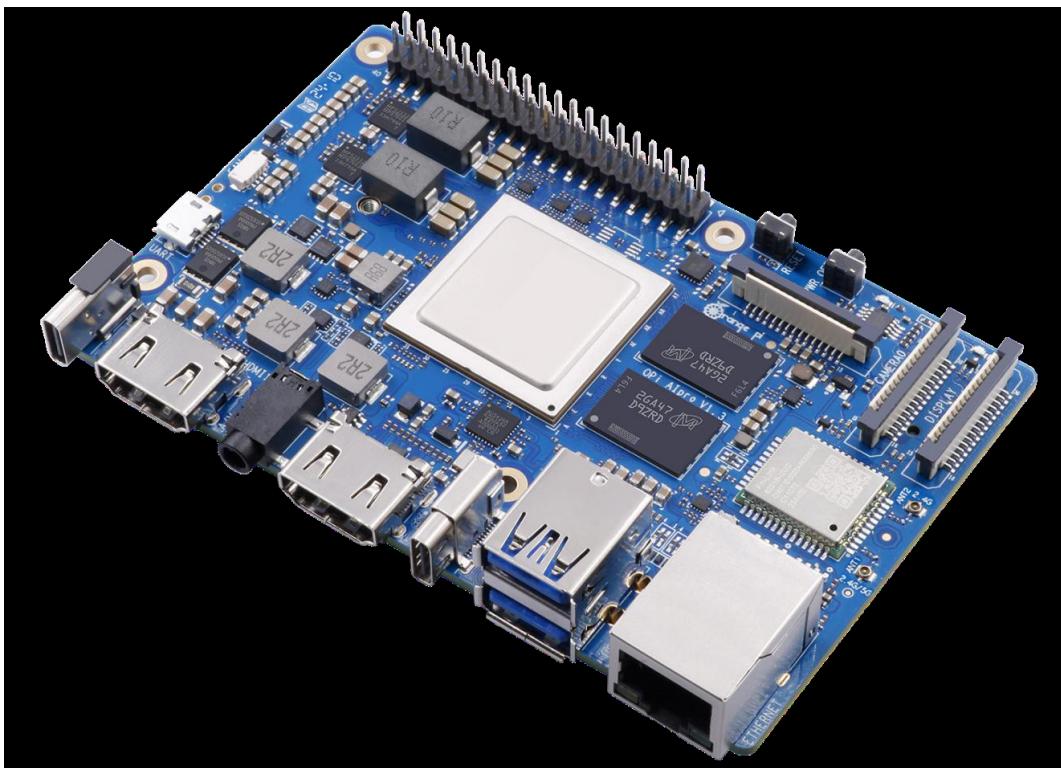
10.13 实践任务

- 撰写 hardware_inventory.md 与 env_report.json（可手动作草拟）。
- 建立 requirements.txt 并安装依赖，记录安装耗时。
- 创建一个最小随机张量 OM 推理脚本并输出结果摘要。
- 制定个人 4 周学习计划（章节 → 目标 → 产出）。

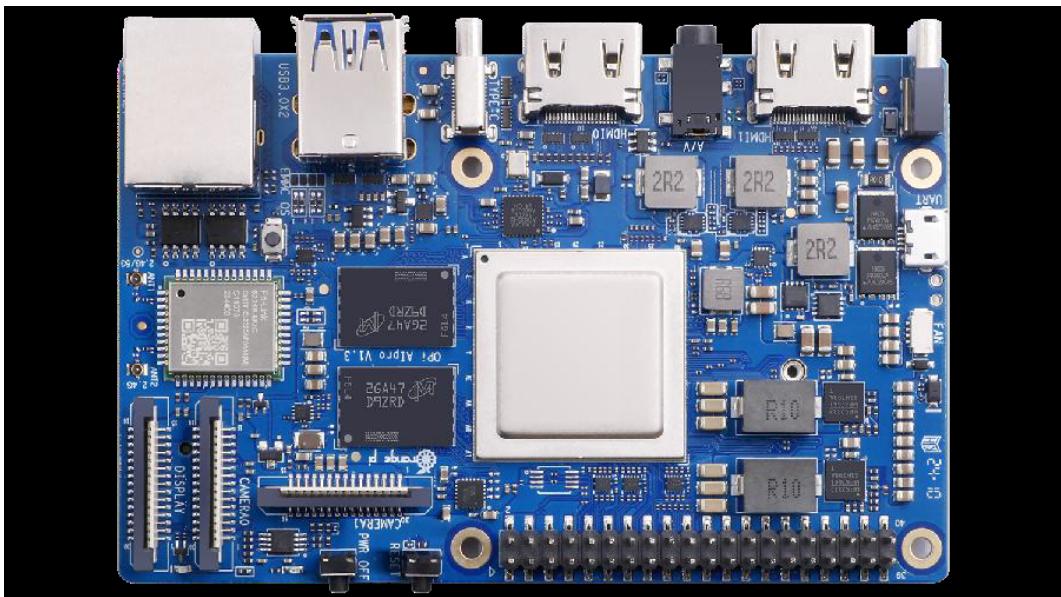
11 案例 0

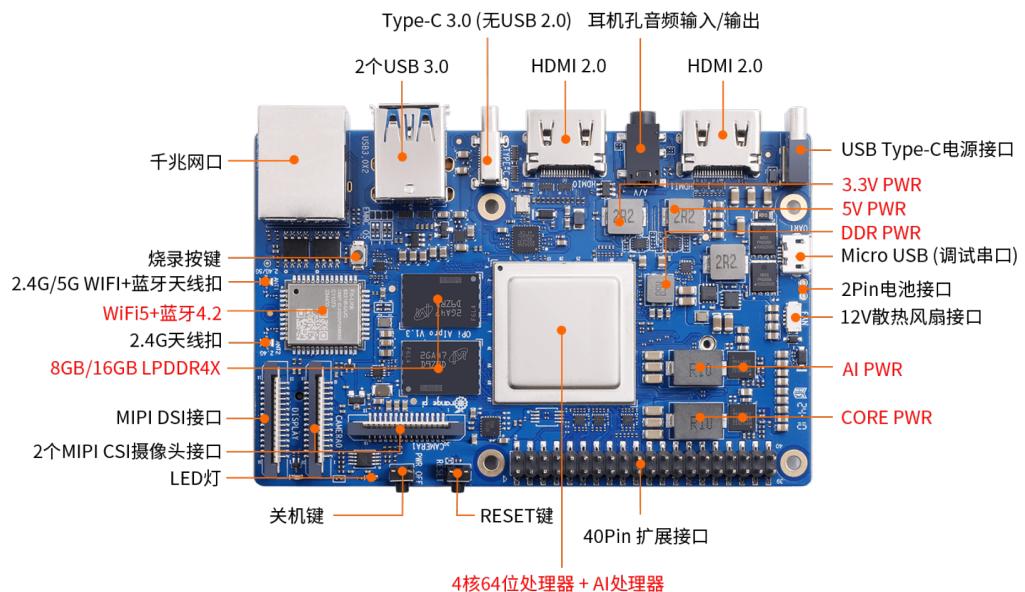
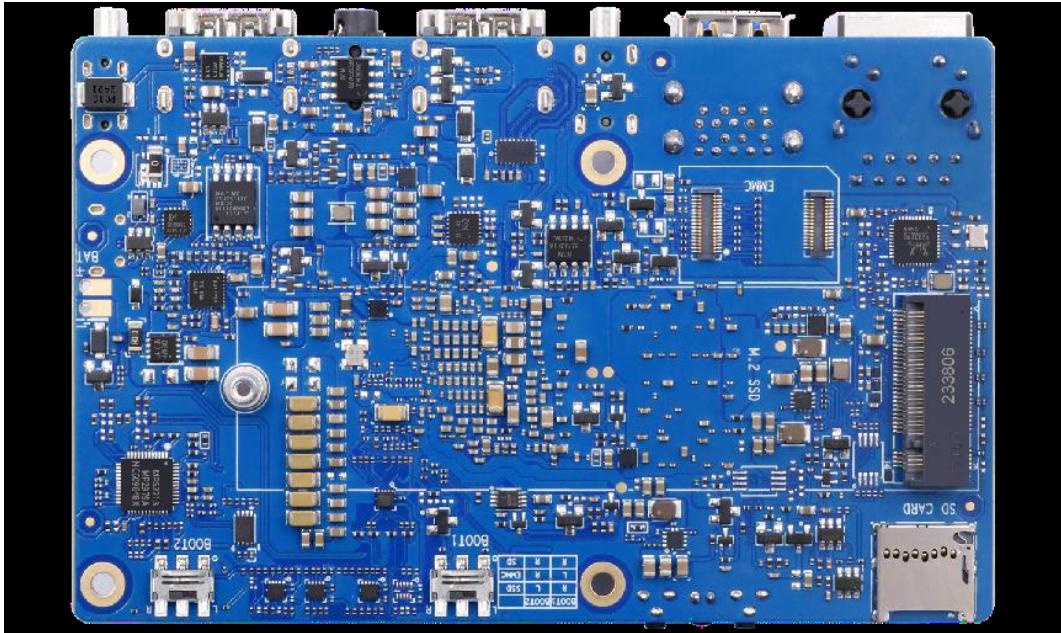
11.1 昇腾 310B 开发板介绍

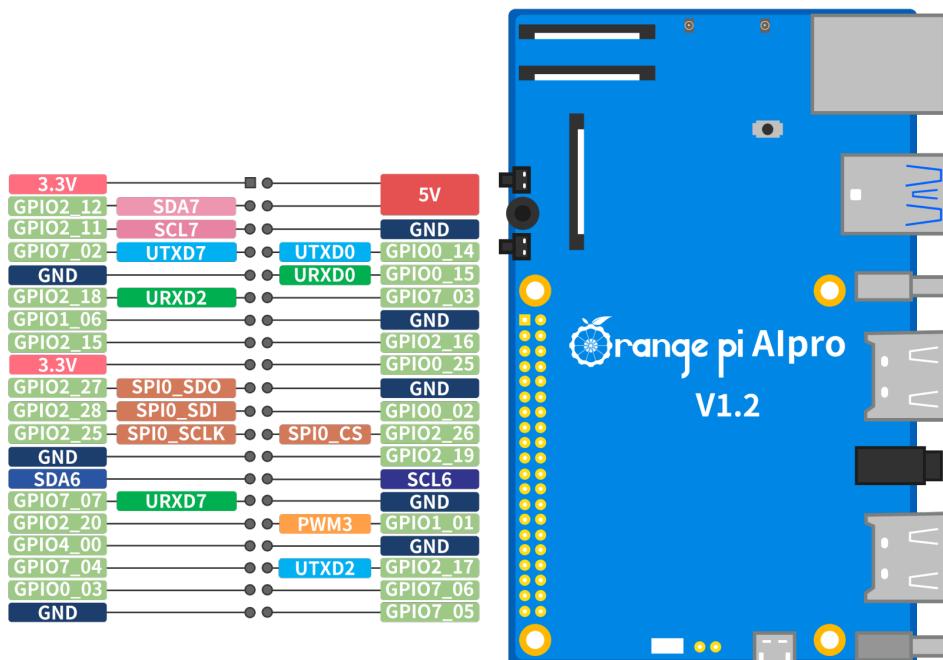
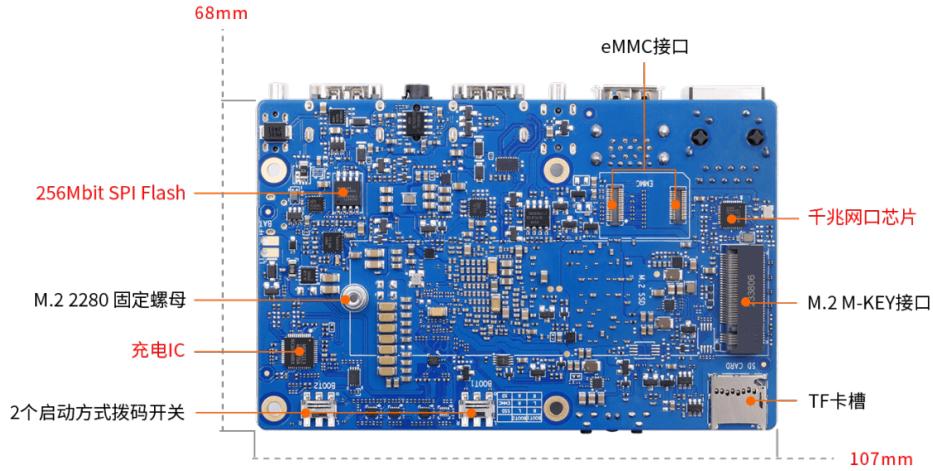
OrangePi AIpro(8T) 开发板是香橙派联合华为精心打造的高性能 AI 开发板，采用昇腾 AI 技术路线，搭载的昇腾 310B 为 4 核 64 位处理器 +AI 处理器，集成图形处理器，支持 8TOPS INT8 的 AI 算力，拥有 8GB/16GB LPDDR4X 内存，可以外接 32GB/64GB/128GB/256GB eMMC 模块，支持双 4K 高清输出。OrangePi AIpro(8T) 引用了相当丰富的接口，包括两个 HDMI 输出、GPIO 接口、Type-C 电源接口、支持 SATA/NVMe SSD 2280 的 M.2 插槽、TF 插槽、千兆网口、两个 USB3.0、一个 USB Type-C 3.0、一个 Micro USB（串口打印调试功能）、两个 MIPI 摄像头、一个 MIPI 屏等，预留电池接口，可广泛适用于 AI 边缘计算、深度视觉学习及视频流 AI 分析、视频图像分析、自然语言处理、智能小车、机械臂、人工智能、无人机、云计算、AR/VR、智能安防、智能家居等领域，覆盖 AIoT 各个行业。OrangePi AIpro(8T) 支持 Ubuntu、openEuler 操作系统，满足大多数 AI 算法原型验证、推理应用开发的需求。



11.1.1 开发板详细视图







11.1.2 开发板硬件规格

11.1.3 所需配件

1. TF 卡容量最小为 32GB，速率率为 Class10 级以上的闪迪品牌的 TF 卡，如下图所示。建议使用 64G 及以上的 TF 卡，以避免在开发过程中出现磁盘空间不足的问题。



2. TF 卡读卡器用于读写 TF 卡，刷写系统，建议选择速率为 USB3.0 以上的，减少系统刷写的等待时间。



3. HDMI 线或 HDMI 转 mini-HDMI 线主要取决于显示器的接口类型该开发板的视频输出接口为标准 HDMI 接口。





- 电源该开发板的电源输入为 PD 20V，需要搭配支持 PD 协议 20V 挡位的 65W 电源适配器。



5. USB 接口的鼠标以及键盘在无远程访问的条件下对开发板进行本地调试。

香橙派

手感舒适

让你爱不释手

无线传输

灵敏精准



6. 金属配套外壳用于保护开发板硬件。



7. 12V 散热风扇以及散热鳍块开发板的风扇接口为 2pin，输出电压为 12v，支持 PWM 调速。
由于该开发板的 CPU 发热较大，强烈建议安装主动扇热设备。



8. Type-C 转 USB 3.0 转接线（可选）OrangePi AIPro 开发板具有一个 Type-C 接口，协议为 USB3.0（不支持 USB 2.0），可外接支持 USB3.0 以上协议的外置设备。



9. M.2 接口 2280 规格的 PCIe Nvme SSD（可选）开发板的背部设计有 M.2 接口，可外接一个 M.2 的 SSD 作为开发板的系统盘或者存储。

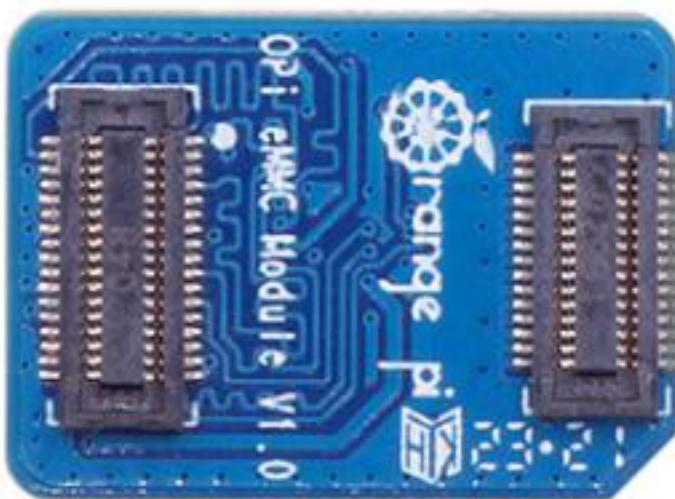


10. M.2 接口 2280 规格的 Sata Ngff SSD (可选) 同样，开发板的 M.2 接口不仅支持 PCIe 协议，也支持 Sata 协议，因此也可以使用 Sata 协议的 SSD。



11. 香橙派的 eMMC 模块 (可选) eMMC (嵌入式多媒体卡) 是一种集成了闪存和控制器的低成本存储解决方案，主要用于智能手机、平板电脑和低端笔记本电脑等消费电子产品。其读写速度适中 (100-400MB/s)，比传统机械硬盘快但不及固态硬盘 (SSD)，具有体积小、功耗低和易于集成的特点。开发板支持使用 eMMC 模块作为存储，但需要额外购置 eMMC 模块。





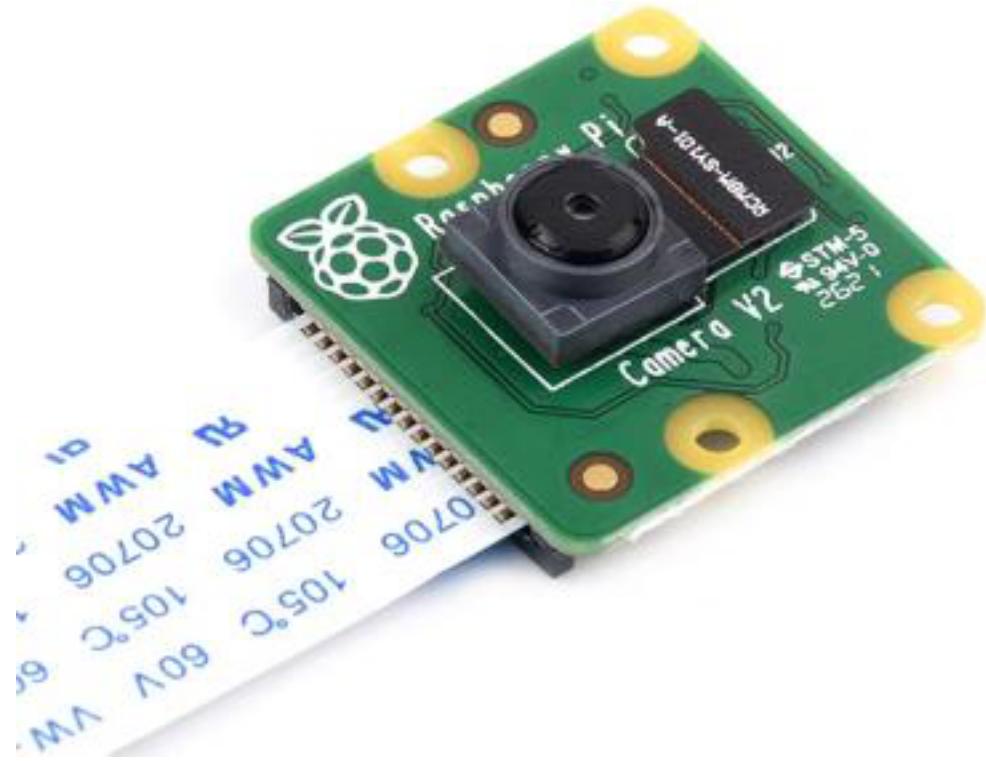
12. USB 摄像头模块（可选）可用于图像识别、视频通话等多方面用途。



13. 网线（可选）开发板自带 wifi 模块可用于连接 wifi，若需要更稳定的网络连接，建议使用网线连接。



14. 树莓派 IMX219 型号摄像头 (MIPI-CSI) (可选) 开发板带有两个 MIPI-CS1 接口，可以兼容树莓派的 MIPI 摄像头，无需占用 USB 接口。



15. 树莓派 5 寸 MIPI LCD 显示屏 (可选) 开发板带有一个 MIPI-DSI 显示输出接口，可以直接驱动 MIPI 的显示屏，而无需外接显示器。



16. Micro USB 数据线（可选）开发板自带了 CH343P 芯片，将 UART 转发为 Micro USB 接口，若需要使用串口对开发板进行调试，则需要使用 Micro USB 数据线。



11.1.4 下载开发板的系统镜像

作为华为生态中重要的一员，开发板不仅支持 Ubuntu 系统，也支持 openEuler 系统，但由于开发板自身并无存储，我们在使用开发板的过程中需要使用电脑对 TF 卡进行系统的刷写，建议使用安装有 Windows11 或 Ubuntu22.04 以上版本的 PC。

首先，打开香橙派官网的[技术支持界面](#)。

The screenshot shows the official website for Orange Pi. At the top, there's a navigation bar with links for 'Orange pi' (logo), '开源硬件' (Open Source Hardware), '开源软件' (Open Source Software), '定制化服务' (Customized Services), '论坛' (Forum), '资讯' (Information), '服务与下载' (Services and Downloads), '关于我们' (About Us), and 'EN/中文' (EN/Chinese). Below the navigation bar, there are tabs for '概述' (Overview), '下载' (Download), and '参数' (Parameters). The main content area features a large image of the OrangePi Alpro(8T) board. Below the board, the text 'OrangePi Alpro(8T)' is displayed. Further down, there's a section titled '官方资料' (Official Documentation) with links to various resources: '外壳及散热器安装资料' (Case and Heat Sink Installation Instructions), '官方工具' (Official Tools), '用户手册' (User Manual), '原理图' (Circuit Diagram), '机械图' (Mechanical Drawing), and 'linux源码' (Linux Source Code). Each link has a small '下载' (Download) button.

向下滑动网页，找到官方镜像部分，分为 Ubuntu 和 openEuler 两个部分，两个系统都是官方为我们编译完成的，且预装了部分昇腾 NPU 的应用环境以及软件，非常方便新手用户上手使用。

官方镜像



Ubuntu

1. 点击下载



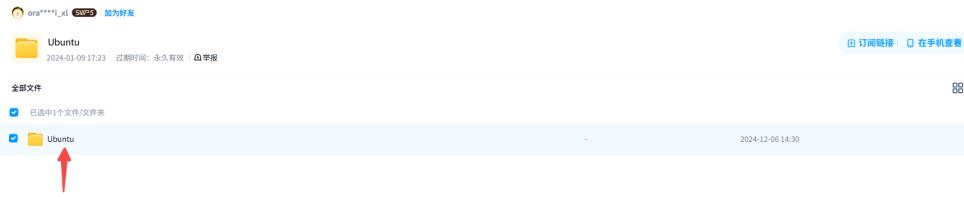
ubuntu镜像



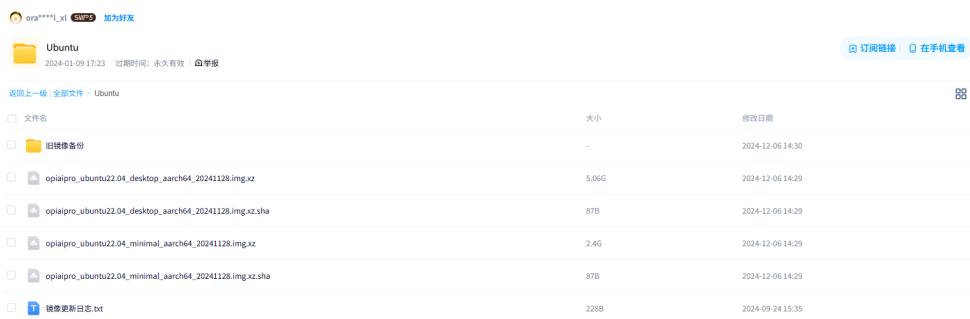
2. 复制提取码并跳转



3. 打开百度网盘的链接后有一个命名为 Ubuntu 的文件夹，点开该文件夹



4. 文件夹中，后缀为.xz 的文件是镜像压缩包文件，.sha 文件是压缩包的 md5 校验码文件，用于校验镜像包文件是否完整。
5. 文件夹中的镜像有两种，一种文件名带有 Desktop 的，是带有 GUI 图形化界面的，另一种文件名带有 minimal 的，是不具有图形化界面的，只有命令行界面。建议新学习的用户使用带有 desktop 的镜像。



6. 下载后先校验压缩包是否完整，后解压压缩包

openEuler

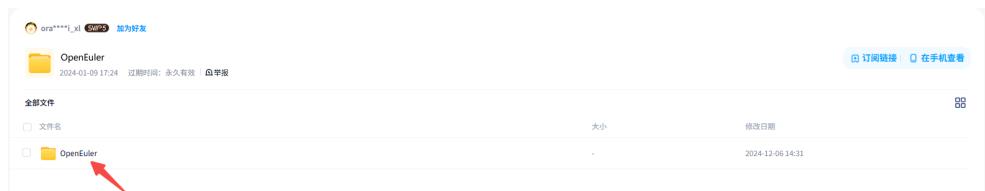
1. 点击下载



2. 复制提取码并跳转



3. 打开百度网盘的链接后有一个命名为 OpenEuler 的文件夹，点开该文件夹



4. 文件夹中，后缀为.xz 的文件是镜像压缩包文件，.sha 文件是压缩包的 md5 校验码文件，用于校验镜像包文件是否完整。
5. 文件夹中的镜像只有一种，即具有 GUI 图形化界面的 openEuler 系统。



6. 下载后先校验压缩包是否完整，后解压压缩包

使用 md5 校验下载的文件

在 Windows 系统下，可以使用 certutil –hashfile <filename> md5；在 Ubuntu 系统下，可以使用md5sum <filename>；在 MacOS 系统下，可以使用md5 <filename>进行计算，此处以 Windows 系统为例：在文件夹按住 Shift 键并单击鼠标右键，选择“在终端（Powershell/命令提示符）中打开”



,然后在打开的窗口中输入 certutil -hashfile opiaipro_ubuntu22.04_desktop_aarch64_20241128.img.xz md5

```
Microsoft Windows [版本 10.0.26100.4652]
(c) Microsoft Corporation。保留所有权利。

D:\BaiduNetdiskDownload>certutil -hashfile opiaipro_ubuntu22.04_desktop_aarch64_20241128.img.xz md5
MD5 的 opiaipro_ubuntu22.04_desktop_aarch64_20241128.img.xz 哈希:
c2504fd63b2cc222106c30a29ac06386
CertUtil: -hashfile 命令成功完成。

D:\BaiduNetdiskDownload>
```

, 将得到的 md5 值与 opiaipro_ubuntu22.04_desktop_aarch64_20241128.img.xz.sha 文件进行对比, 若一致可进行下一步操作, 否则需要重新下载。

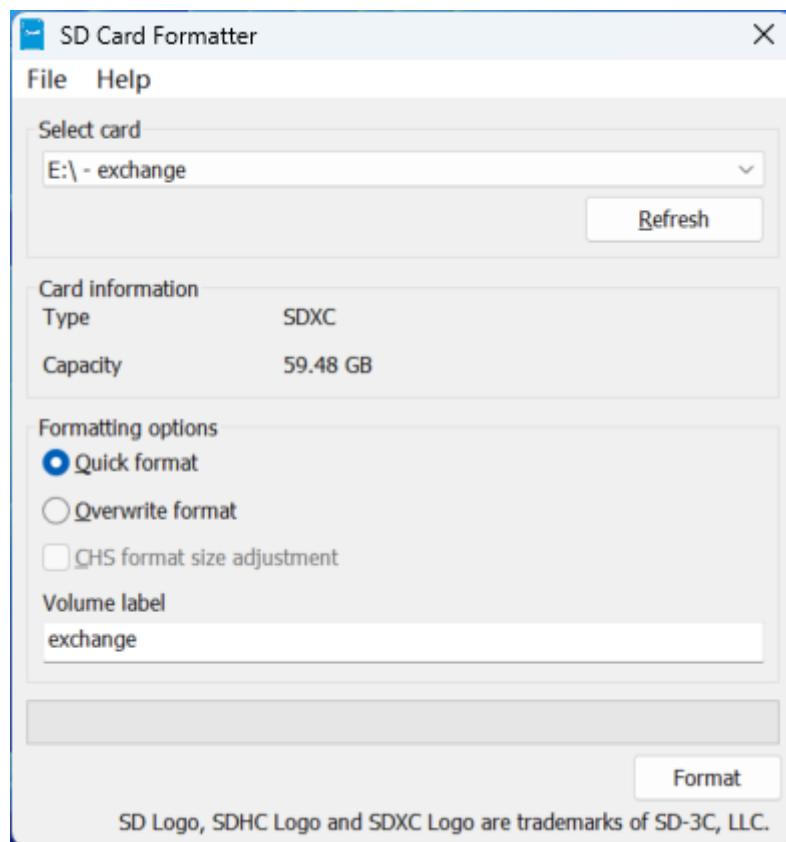
11.1.5 刷写系统到 TF 卡

下载并安装必要的工具

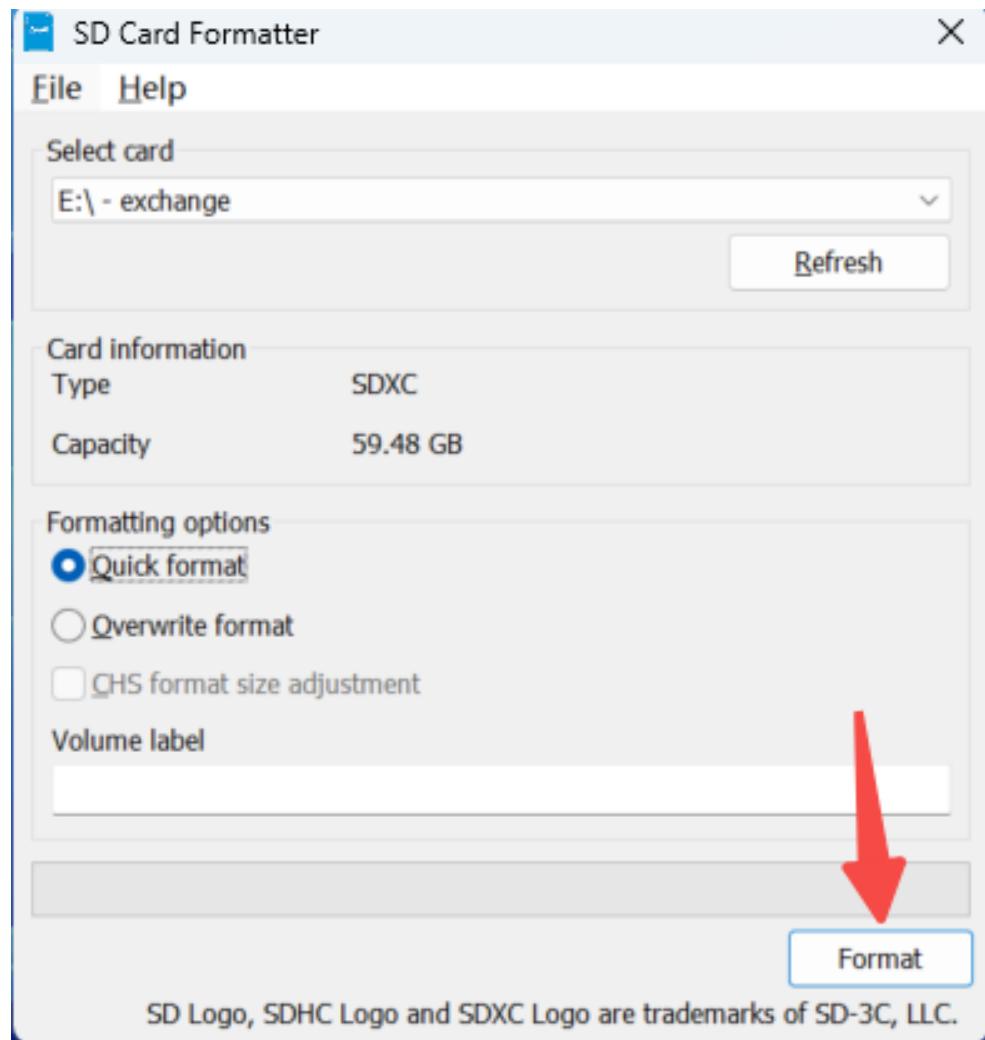
下载链接：[官网 百度网盘](#) 1. SD Card Formatter 这个是 TF 卡的快速格式化工具，在每次需要刷写系统之前，都必须先对 TF 卡进行格式化操作，若不格式化在后续的刷写系统过程中有较大概率出错。2. balenaEther 这个是系统镜像的刷写工具，用于刷写 img 镜像文件进入 TF 卡。

格式化 TF 卡

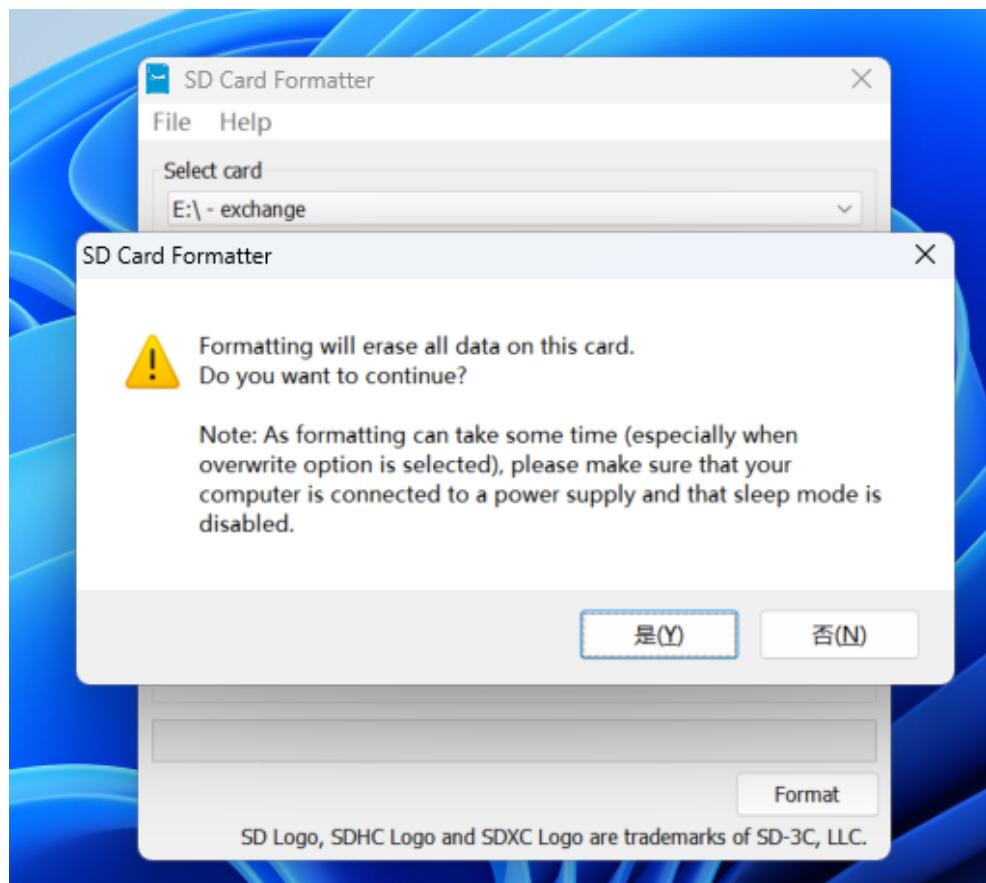
1. 将 TF 卡插入读卡器中，并将读卡器插入电脑
2. 打开 SD Card Formatter 软件



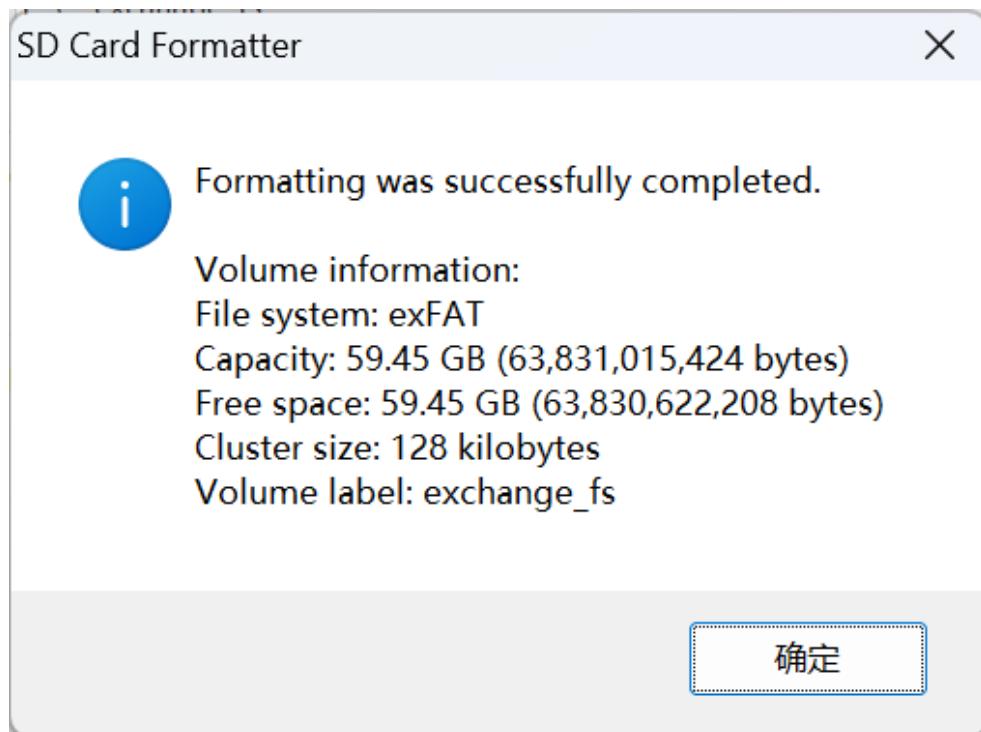
3. 点击右下角 Format 按键，格式化 TF 卡



> 警告内容是关于格式化操作会清除 TF 卡上原有的所有数据，此处选是



4. 等待软件格式化完成，并点击确定



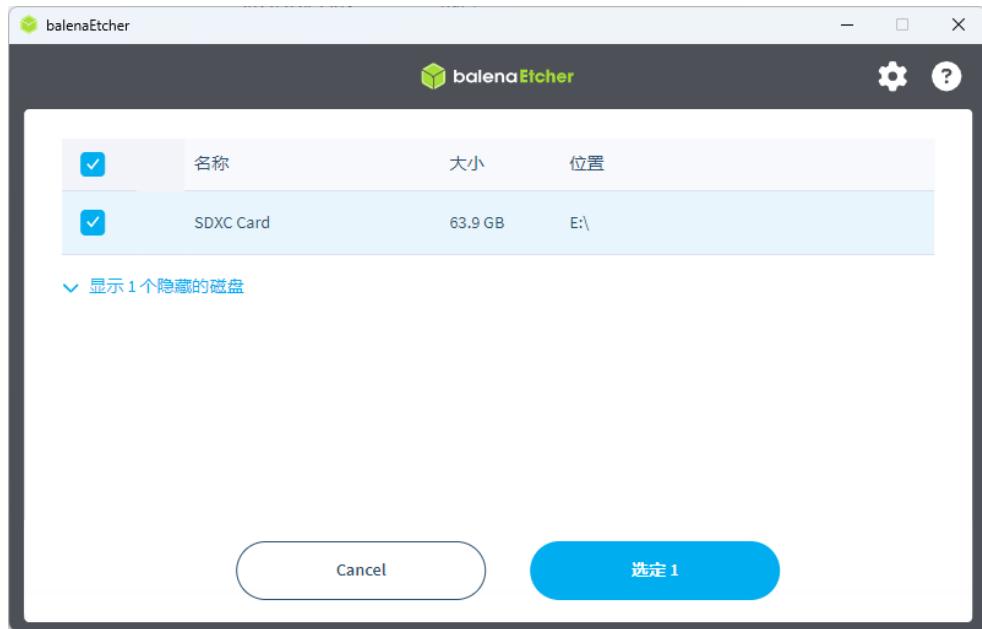
刷写系统到 TF 卡 (以 Ubuntu 为例)

此处以刷写 Ubuntu 为例

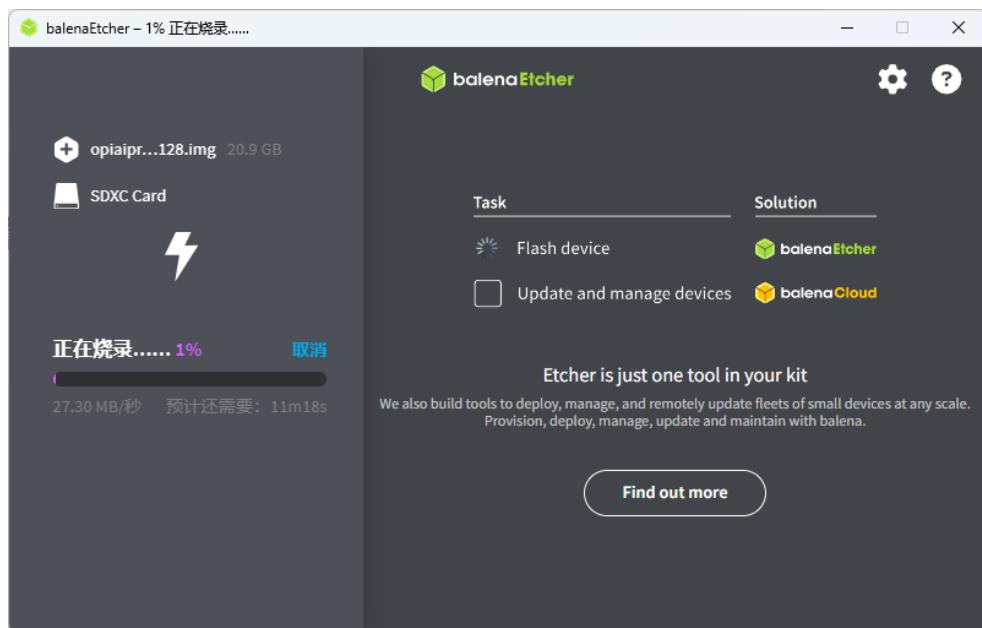
1. 打开 balenaEtcher, 选择“从文件烧录”



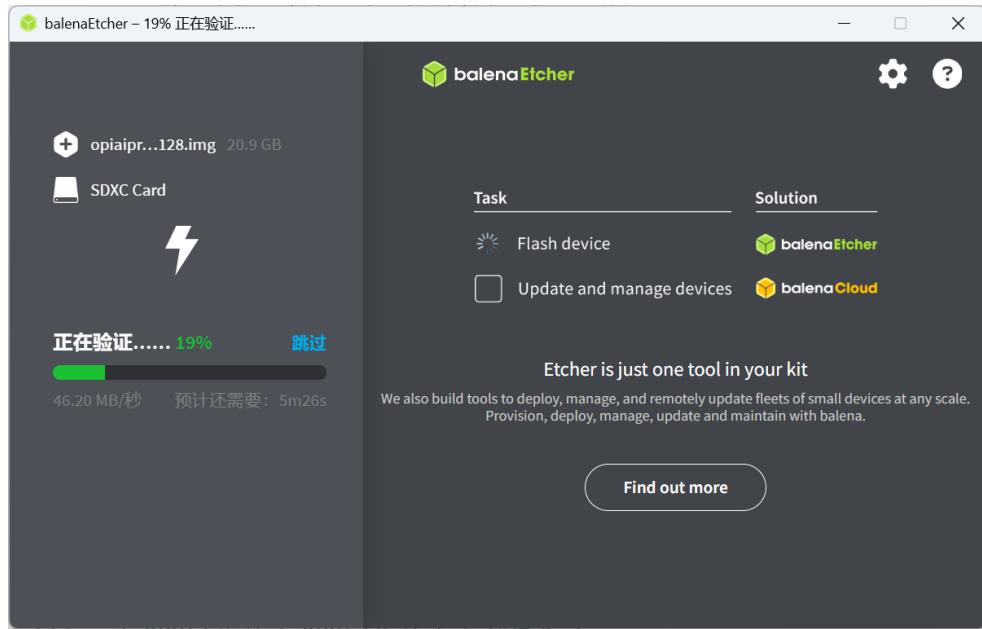
2. 选择好要烧录的镜像文件 (.img 格式), 再选择目标磁盘为 TF 卡对应的位置, 如图中名称为“SDXC Card”的位置, 选中并选择“选定 1”。



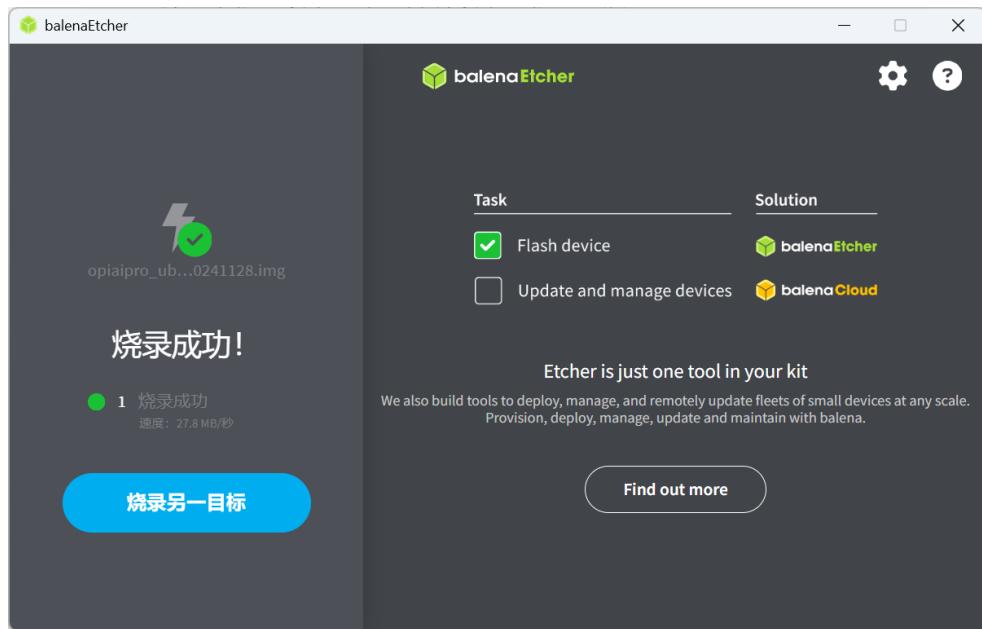
3. 点击“现在烧录！”，耐心等待烧录完成。



4. 烧录完成后进入校验过程, 也请耐心等待。



5. 烧录完成后即可关闭程序，并安全弹出 TF 卡



刷写系统到 eMMC

由于板上并不自带 eMMC 模块，若要想使用需要额外购买香橙派的 eMMC 模块，此处暂时不列入参考，若需使用，请查阅香橙派的用户手册。

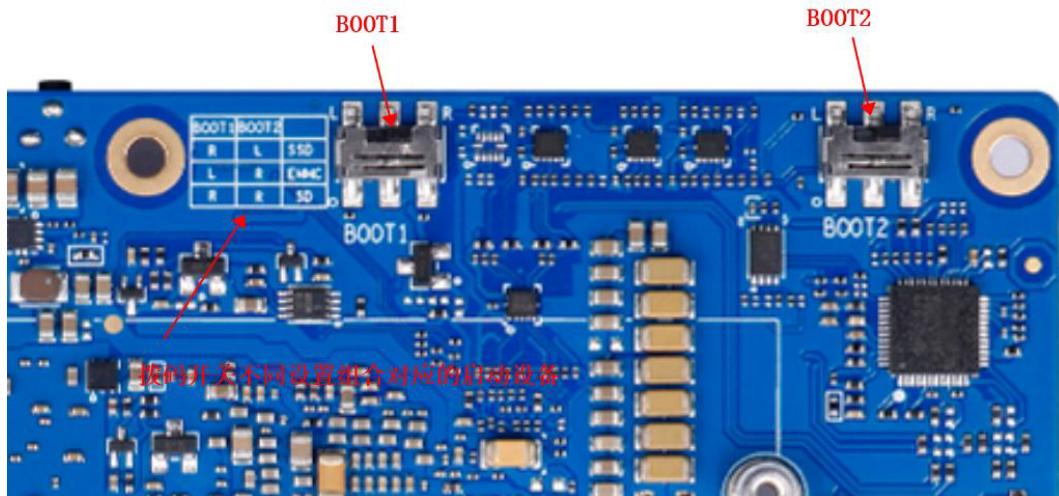


Figure 11.1: boot 开关

刷写系统到 SSD

开发板带有 M.2 接口，可以使用 SSD 作为启动设备。但 SSD 需要自行准备，且根据香橙派的兼容性说明，该开发版仅支持少数品牌的 SSD，因此不推荐使用 SSD 作为系统安装位置。

调整设备启动方式的拨码开关

开发板支持多种启动方式，包括 TF 卡、eMMC 以及 M.2 SSD，当这些存储设备都同时存在时，需要让开发板选定一个存储设备作为启动来源。

两个开关都有左、右两种状态，因此共有 4 种状态，但是目前开发板仅使用 3 种模式，对应的参数表如下：

| Boot1 开关 | Boot2 开关 | 启动设备 |
|----------|----------|-----------------------|
| 左 | 左 | 未使用 |
| 右 | 右 | TF 卡 |
| 左 | 右 | eMMC |
| 右 | 左 | M.2 SSD (Nvme 或 Ngff) |

切换拨码开关后，必须要将开发板完全断电再重新上电才能使新的启动配置生效，使用 RESET 按键重启则不会使新的启动配置生效。

11.1.6 启动开发板 (Ubuntu)

- 图形化界面

- 将系统刷写完成的 TF 卡从读卡器中取出，插入开发板的 TF 卡插槽中，并确保两个启动开关的位置均在右边，接入 HDMI 数据线到靠近 USB3.0 接口的 HDMI0 接口，然后将 Type-C 电源线插入开发板最边缘的 TYPE-C 供电口，等待风扇的声音变小以及屏幕出现系统登录界面。



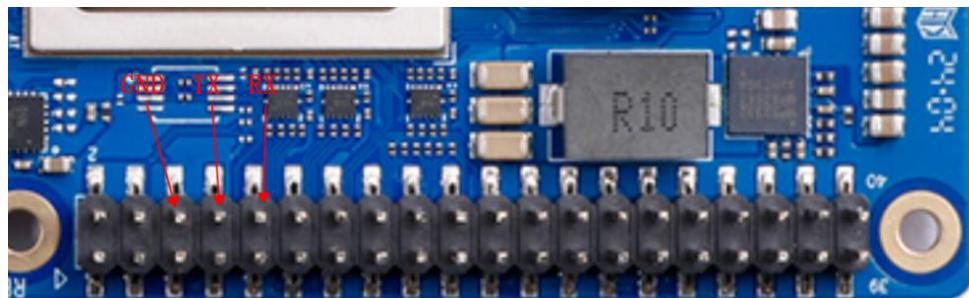
2. 进入登录界面后，将键盘接入开发板的 USB 接口中，默认的登录用户名是HwHiAiUser，输入该账户的密码Mind@123，登录进入系统。



> 若无法登陆请检查输入的密码是否正确，大小写以及符号是否正确默认账户表格：| 用户名 |
密码 | | :—: | :—: | | root | Mind@123 | | HwHiAiUser | Mind@123 |

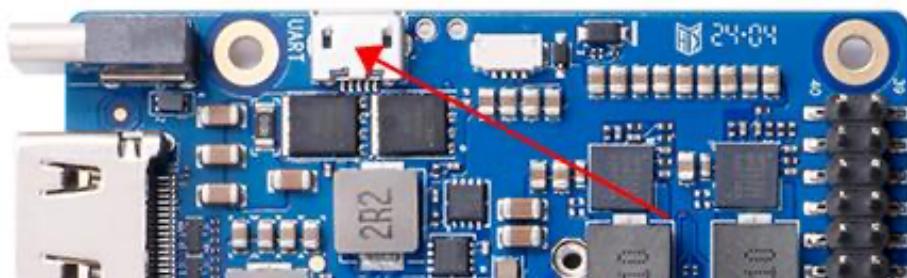
- 串口界面

1. 使用 USB2TTL 模块，与开发板的 GPIO 口进行连线



，开发板的 TX (GPIO8) 接入 USB2TTL 模块的 RX 接口，开发板的 RX (GPIO10) 则接入模块的 TX 接口，并连接好 GND 接地，在 Windows 电脑下可以使用 PUTTY 连接串口。

2. 使用开发板自带的 Micro USB 接口进行串口调试，该方法更为方便，只需要一根 Micro USB 数据线，接入电脑后打开设备管理器查询对应的串口，然后使用 PUTTY 进行链接即可。

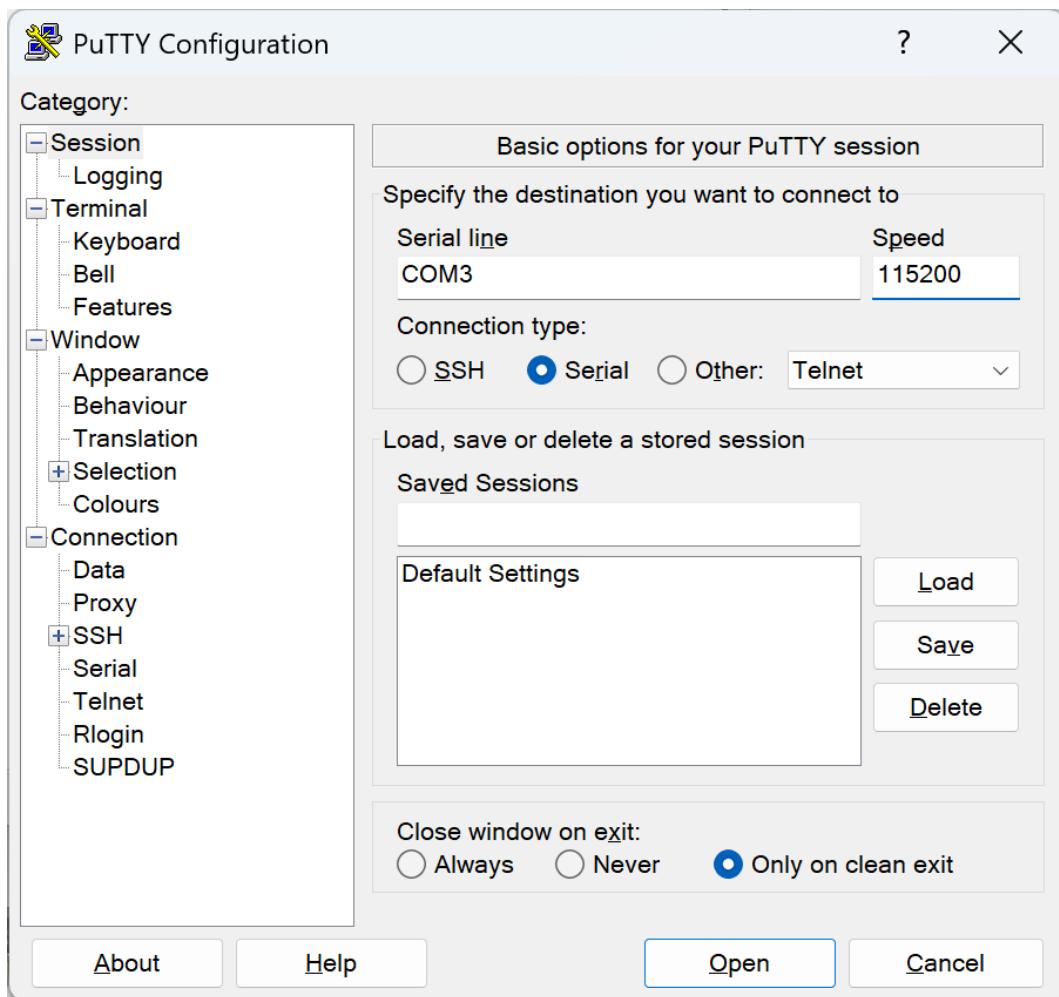


以 Micro USB 接口为例：1. 使用 Micro USB 数据线连接开发板和电脑 2. 打开电脑的设备管理器，选择端口，寻找开发板对应的串口端口号

▼  端口 (COM 和 LPT)

 USB-Enhanced-SERIAL CH343 (COM3)

3. 打开串口调试软件 (PUTTY)



, 将 Connection Type 选择为 Serial, 然后在 Serial Line 处将端口号修改为设备管理器中查到的端口号, 如作者此处端口号为COM3, 此外, 还需要将 Speed 从 9600 修改为 115200, 最后点击 Open 打开串口。4. 等待出现Ubuntu 22.04.3 LTS orangepiaipro ttyAM0字样, 输入登录的用户名 HwHiAiUser 并回车, 然后输入密码 Mind@123 并回车, 注意在输入密码的时候屏幕并不会显示任何东西, 登陆后的界面如图所示。

```

NOTICE: SubSysID:0xff, DeviceID:0x0, SubSysNum:0x0
NOTICE: RECOVERABLE!
NOTICE: HostNotifiedOS
NOTICE: [RasCbbCommonHandler]:[89] Handler end
NOTICE: base = 0xc1260000
NOTICE: ERR_FRL = 0x142aa2
NOTICE: ERR_FRH = 0x0
NOTICE: ERR_CTRLL = 0x515
NOTICE: ERR_CTRLH = 0x0
NOTICE: ERR_STATUSL = 0xfc30050e
NOTICE: ERR_STATUSH = 0x0
NOTICE: ERR_ADDRL = 0x10080010
NOTICE: ERR_ADDRH = 0xe0000001
NOTICE: ERR_MISCOL = 0x0
NOTICE: ERR_MISCOH = 0x0
NOTICE: ERR_MISC1L = 0xe718005
NOTICE: ERR_MISC1H = 0x800122
NOTICE: el3_int exit!
cpu 0 entering scheduler
>>>>>>>>>LiteOS start succeed!<<<<<<<
Ubuntu 22.04.3 LTS orangepiapro ttyAMA0
orangepiapro login: 

```

```

cpu 0 entering scheduler
>>>>>>>>>LiteOS start succeed!<<<<<<<
Ubuntu 22.04.3 LTS orangepiapro ttyAMA0
orangepiapro login: HwHiAiUser
Password:
Welcome to Orange Pi Ai Pro
This system is based on Ubuntu 22.04.3 LTS (GNU/Linux 5.10.0+ aarch64)

This system is only applicable to individual developers and cannot be used for commercial purposes.

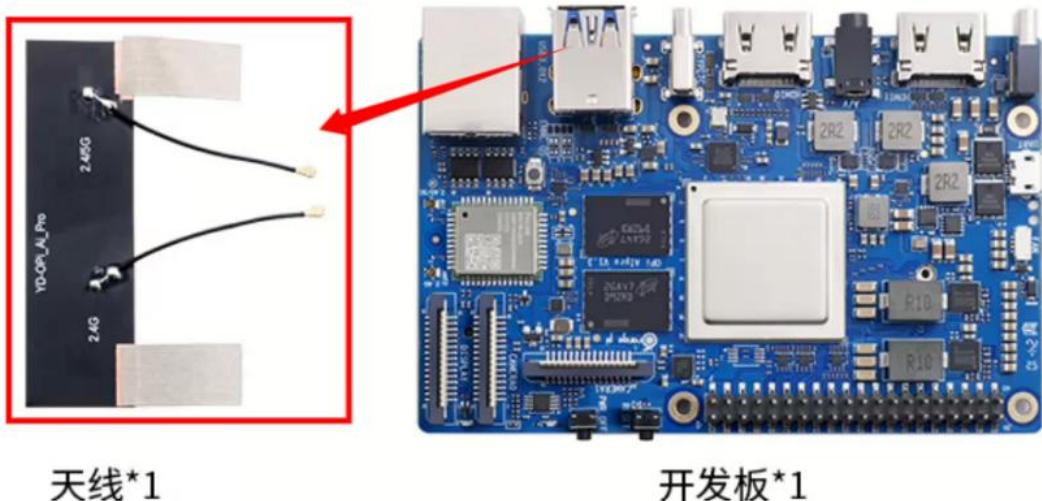
By using this system, you have agreed to the Huawei Software License Agreement.
Please refer to the agreement for details on https://www.hiiscom/software/protocol

(base) HwHiAiUser@orangepiapro:~$ 

```

11.1.7 WIFI 天线安装指南

开发版的 wifi 天线如左侧红色矩形框内所示



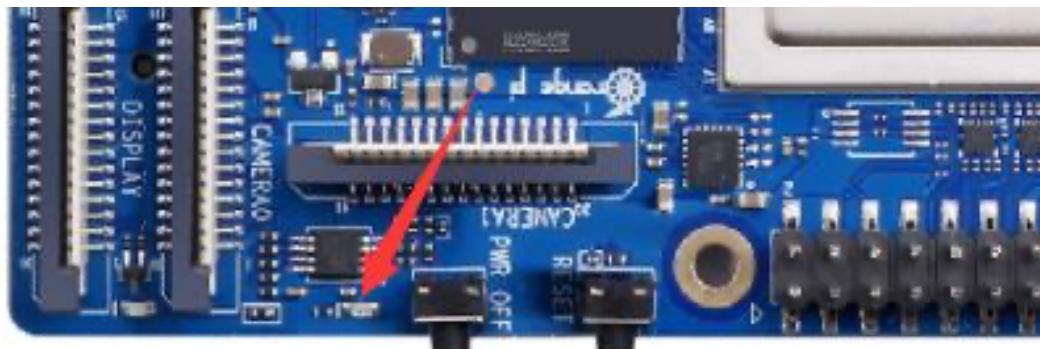
将其对准开发版的天线接口安装牢固即可，注意不要将天线贴到开发版的背面，也需要注意天线下方的导电胶布也不要接触开发版，否则有可能导致 PCB 短路烧坏开发版。

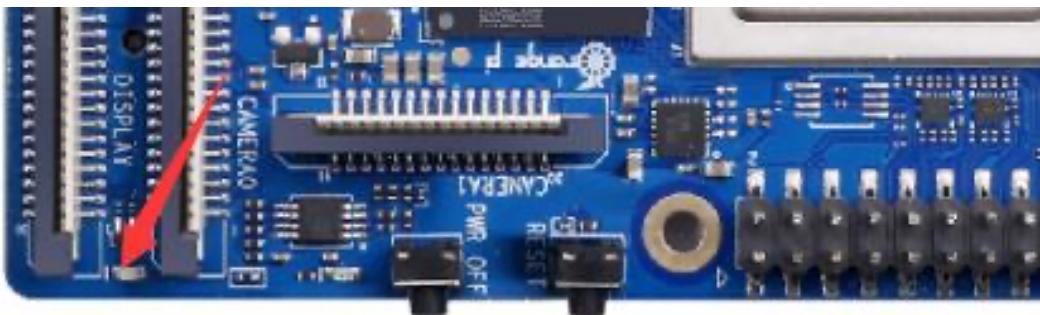
11.1.8 Ubuntu Xfce 桌面使用说明

目前系统仅支持 Ubuntu 22.04 - Jammy 系统，内核版本为 Linux 5.10 ##### 当前版本适配情况请详见香橙派官方的用户手册，有部分功能仅支持使用官方程序进行测试，无法直接从系统中调用，在使用过程中需注意这些限制。

板载 LED 灯

开发版上有两个绿色的 LED 灯，一个为电源指示灯，另一个为 Linux 内核指示灯。





Linux 内核指示灯由 GPIO4_14 控制，默认情况下则在 Linux 启动后该灯就会点亮，如需要修改该灯的点亮条件，需要修改内核 DTS 文件并重新编译 Linux 系统。

有线网络连接

1. 将网线一端连接开发版的网口，另一端连接交换机/路由器
2. 在 Ubuntu 系统中打开终端 (terminal)，输入 ip a s eth0，ip 地址显示在输出的 inet 一列

使用无线网络连接

- 使用 nmcli 连接,首先nmcli dev wifi 扫描 WIFI,然后使用sudo nmcli dev wifi connect wifi_name password (将 wifi_name 和 wifi_passwd 替换为实际的 SSID 和密码, 不支持中文), 连接成功后使用ip a s wlan0查看 wifi 地址。
- 使用 nmtui 连接, 在终端输入sudo nmtui后即可使用键盘对图形界面进行操作, 包括连接网络、断开网络、设置静态 IP 地址等。

11.1.9 HDMI 口使用

开发板有两个 HDMI2.0 接口，目前只有 HDMI0 支持显示 Linux 系统的桌面，当 Linux 系统的桌面系统关闭时，HDMI0 和 HDMI1 还可以用于 NVR 二次开发场景输出图片。

使用 HDMI VDP 模式

1. 将显示器连接至 HDMI0 接口，登陆进入系统
2. 打开终端，输入如下命令：

```
sudo -i
cd /opt/opi_test/hdmi0_pic
./update_dt.sh
```

3. 等待系统重启后，注意此时 HDMI 接口不会再有输出，使用远程 ssh 或者串口登陆系统，输入如下命令：

```
sudo -i
cd /opt/opt_test/hdmi0_pic
./test.sh
```

可以发现显示屏会输出一张图片，若需要使用 hdmi1 输出，则只需要将上文的 hdmi0 修改为 hdmi1。

恢复 HDMI DRM 模式

进入终端，输入如下命令：

```
sudo -i
cd /opt/opi_test/hdmi_desktop
./update_dt.sh
```

等系统重启后即可。

11.1.10 USB 摄像头使用

将 USB 摄像头插入开发版的 USB3.0 接口中，然后输入如下命令查询摄像头：

```
sudo apt-get update
sudo apt-get install -y v4l-utils
sudo v4l2-ctl —list-devices
```

接着安装 fswebcamsudo apt-get install -y fswebcam，就可以使用 fswebcam 进行拍照。

或者使用内置的 USBCamera 测试代码，运行如下命令，获得一张 yuv 格式的图片：

```
sudo -i
cd /opt/opi_test/USBCamera
./main /dev/video0
```

使用 ffplay 查看ffplay -pix_fmt yuyv422 -video_size 1280*720 out.yuv

11.1.11 音频使用

Linux 内核没有适配耳机和 HDMI 等的 ALSA 音频驱动，此部分驱动还在开发中，目前只能通过音频样例代码来测试耳机、HDMI 的音频播放和板载 MIC 的录音功能。或者自行购买 Linux 系统免

驱的 USB 外置声卡，经测试可以正常使用。若想要使用 USB 音频，需要将自行准备的 USB 声卡或者 USB 接口的耳机连接至 USB3.0 接口，使用arecord -l命令查看录音设备的编号，得到编号后，即可开始测试。

```
sudo -i
cd /opt/opi_test/USBAudio
./main plughw:0 # 录制音频
over # 结束录制
```

若需要播放音频，则使用ffplay -ar 44100 -ac 2 -f s16le audio.pcm

开发版具有 3.5MM 的接口，但是如前文所述，目前 Linux 系统内核并无驱动，使用 3.5MM 接口播放与录制需要使用指定的测试程序。播放：

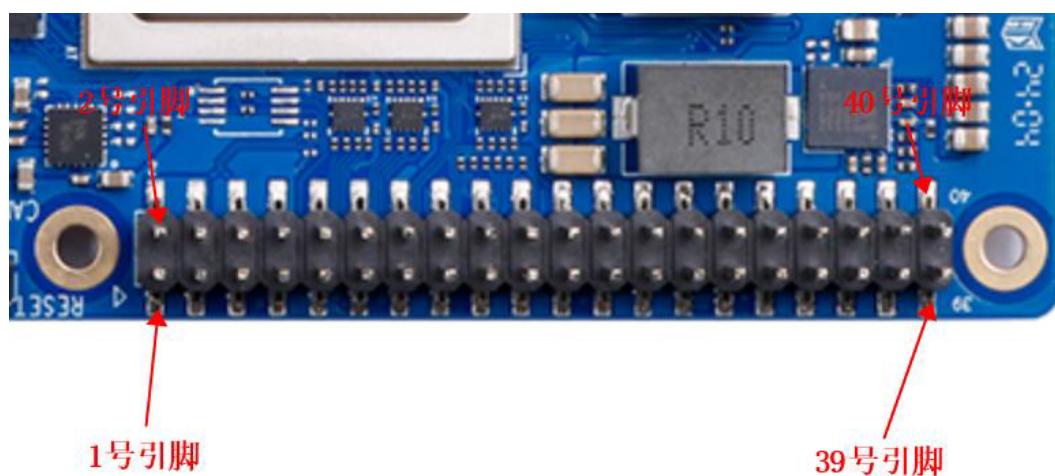
```
sudo -i
cd /opt/opi_test/audio
./sample_audio play 2 qzgy_48k_16_mono_30s.pcm
```

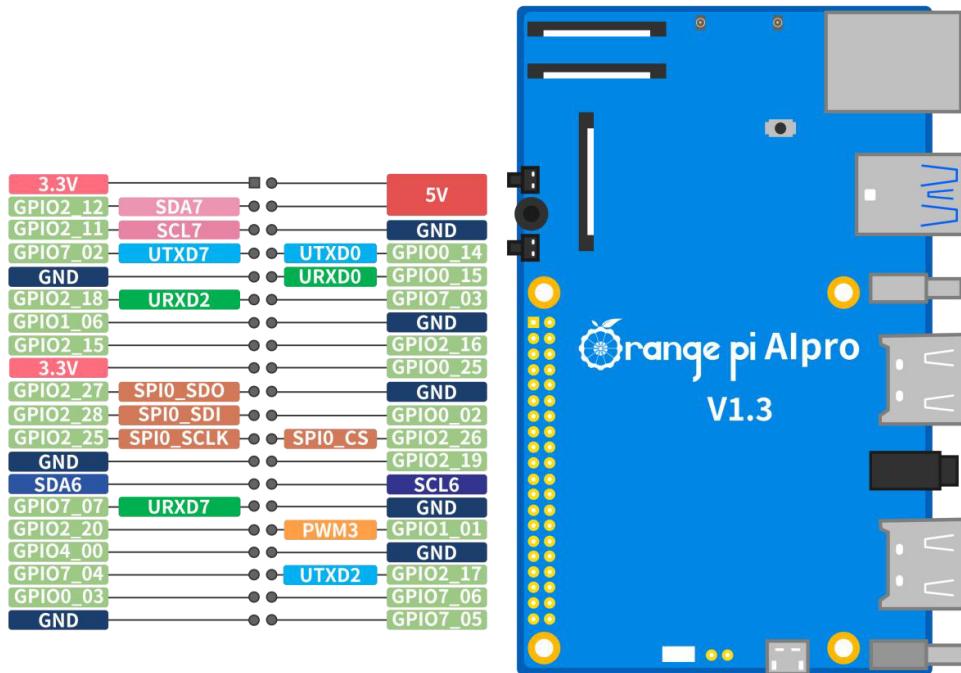
录音：

```
sudo -i
cd /opt/opi_test/audio
./sample_audio capture test.pcm # 录音
./sample_audio play 2 test.pcm # 播放
```

11.1.12 GPIO 口的引脚顺序

如图，单号引脚和双号引脚分别在一排。





注意事项：1. 40pin 接口中总共有 26 个 GPIO 口，但 8 号和 10 号引脚默认是用于调试串口功能的，并且这两个引脚和 Micro USB 调试串口是连接在一起的，所以这两个引脚请不要设置为 GPIO 等功能。2. 所有的 GPIO 口的电压都是 3.3v。3. 40pin 接口中 27 号和 28 号引脚只有 I2C 的功能，没有 GPIO 等其他复用功能，另外这两个引脚的电压默认都为 1.8v。

GPIO 测试工具

目前开发板的系统镜像已经预装了 gpio_operate 工具，该工具可用于设置 GPIO 管脚的输入与输出方向，也可将每个 GPIO 管脚独立的设为 0 或 1。查阅该工具的帮助：

```
sudo -i
gpio_operate -h
```

查询 GPIO 管脚方向使用 gpio_operate get_direction gpio_group gpio_pin, 其中 gpio_group 是 GPIO 口对应的分组，取值在 [0,8] 之间，gpio_pin 则是 GPIO 口的管脚号，取值 [0,31] 之间，例如 GPIO1_01，获取其方向的代码为 gpio_operate get_direction 1 01，得到的结果如图所示，value 为 0 是输入方向，value 为 1 则为输出方向，如此处 GPIO1_01 的方向为输入方向。

```

/---\ /---\ /---\ /---\ /---\ /---\ /---\ /---\ /---\
| | | | | | | | | | | | | | | | | | | | | | | | | |
\---/ \---/ \---/ \---/ \---/ \---/ \---/ \---/ \---/
Welcome to Orange Pi Ai Pro
This system is based on Ubuntu 22.04.3 LTS (GNU/Linux 5.10.0+ aarch64)

This system is only applicable to individual developers and cannot be used for commercial purposes.

By using this system, you have agreed to the Huawei Software License Agreement.
Please refer to the agreement for details on https://www.hiascend.com/software/protocol

Web console: https://orangepiapro:9090/

(base) HwHiAiUser@orangepiapro:~$ sudo -i
[sudo] password for HwHiAiUser:
(base) root@orangepiapro:~# gpio_operate -h
Usage: gpio_operate <Command|-h> [Options...]
gpio_operate Command:
  -h                               : This command's help information.
  set_value                         : Set gpio pin value.
  get_value                          : Get gpio pin value.
  set_direction                     : Set gpio pin direction value.
  get_direction                     : Get gpio pin direction value.
(base) root@orangepiapro:~# gpio_operate get_direction 1 01
Get gpio pin direction value successed, value is 0.
(base) root@orangepiapro:~# █

```

若需要修改 GPIO 的管脚方向,则使用另一条命令gpio_operate set_direction gpio_group gpio_pin direction, gpio_group、gpio_pin 和 direction 的定义与上文一致,只需要根据需要将 gpio 口修改为需要的方向即可。另外还能通过这个工具查询和设置 GPIO 管脚的电平信号, gpio_operate get_value gpio_group gpio_pin 用于查询管脚的状态为高电平(1)亦或是低电平(0),如gpio_operate get_value 1 01,得到的 value 为 1,说明是高电平,若 value 值是 0,则说明是低电平。

```

(base) root@orangepiapro:~# gpio_operate -h
Usage: gpio_operate <Command|-h> [Options...]
gpio_operate Command:
  -h                               : This command's help information.
  set_value                         : Set gpio pin value.
  get_value                          : Get gpio pin value.
  set_direction                     : Set gpio pin direction value.
  get_direction                     : Get gpio pin direction value.
(base) root@orangepiapro:~# gpio_operate get_value 1 01
Get gpio pin value successed, value is 1.
(base) root@orangepiapro:~# █

```

同时,也可以使用gpio_operate set_value gpio_group gpio_pin value来设置默认的管脚电平,注意设置管脚值前,请确保已将 GPIO 管脚的方向设置为输出!

SPI 测试

开发板具有 SPI 功能，且 Ubuntu 系统默认配置了 SPI 的 Master 功能，SPI 总线为 SPI0，SDO 对应的 GPIO 为 19 (GPIO2_27)，SDI 对应的 GPIO 为 21 (GPIO2_28)，SCLK 对应的 GPIO 为 23 (GPIO2_25)，CS (片选) 对应的 GPIO 为 24 (GPIO2_26)。查看 ubuntu 系统中存在的 spi 设备 ls /dev/spidev*

```

Welcome to Orange Pi Ai Pro
This system is based on Ubuntu 22.04.3 LTS (GNU/Linux 5.10.0+ aarch64)

This system is only applicable to individual developers and cannot be used for commercial purposes.

By using this system, you have agreed to the Huawei Software License Agreement.
Please refer to the agreement for details on https://www.hiascend.com/software/protocol

Web console: https://orangepiapro:9090/

Last login: Fri Sep 19 16:42:55 2025 from 172.16.1.185
(base) HwHiAiUser@orangepiapro:~$ ls /dev/spidev*
/dev/spidev0.0  /dev/spidev1.0  /dev/spidev3.0  /dev/spidev4.0  /dev/spidev5.0
(base) HwHiAiUser@orangepiapro:~$ 
```

首先测试一下 SPI 在未连接 MISO (SDI) 和 MOSI (SDO) 两个管脚情况下的输出，在终端输入 sudo spidev_test -v -D /dev/spidev0.0，得到如下结果

```

(base) HwHiAiUser@orangepiipro:~$ ls /dev/spi*
/dev/spidev0.0  /dev/spidev1.0  /dev/spidev3.0  /dev/spidev4.0  /dev/spidev5.0
(base) HwHiAiUser@orangepiipro:~$ sudo spidev_test -v -D /dev/spidev0.0
[sudo] password for HwHiAiUser:
spi mode: 0x0
bits per word: 8
max speed: 500000 Hz (500 KHz)
TX | FF FF FF FF FF FF 40 00 00 00 95 FF F0 0D |.....@.....|.....|
RX | FF |.....|.....|
```

可以发现 TX 和 RX 的结果不尽相同，说明此时开发板已经调用 SPI 接口的驱动在向外发送数据，但是没收到数据，接下来我们使用杜邦线将 SDI 和 SDO 连接，构成一个回环，再次运行上述命令，可以发现 TX 和 RX 的数据一致，说明 SPI 的发送和接收功能正常，可以通过 spidev 命令调用 SPI 接口了。

wiringOP

这是一个高性能的 GPIO