智能嵌入式技术及应用开发

华南理工大学
计算机科学与工程学院
毕盛

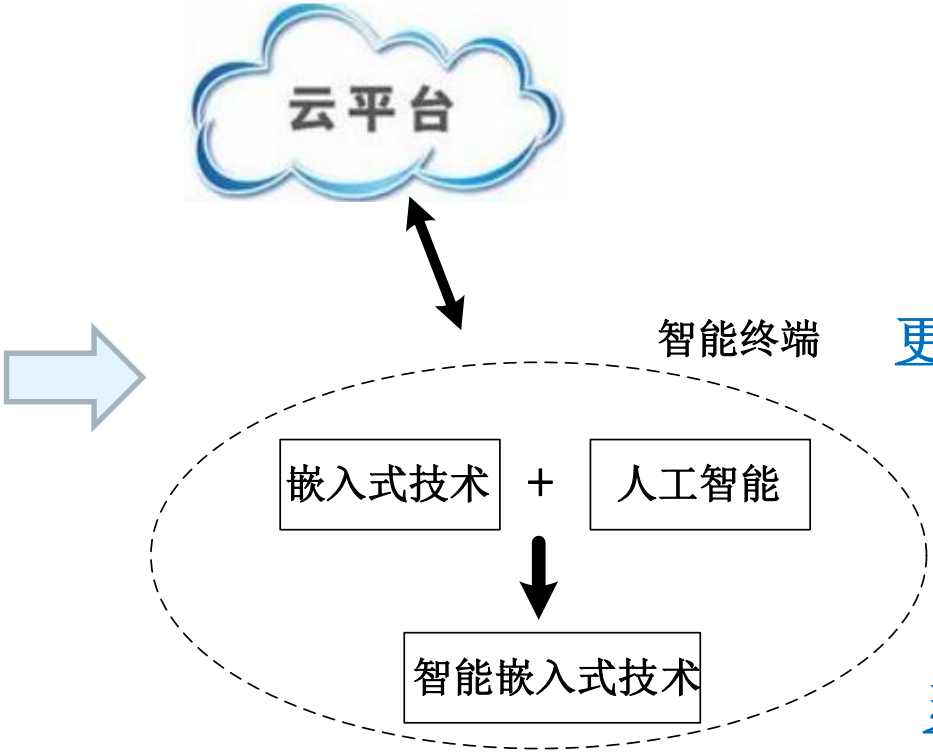
2019年12月19日

嵌入式系统 + 人工智能

云服务

物联网

终端设备



更低延时

可靠性高

安全合规

灵活部署

更加节能

缓解流量压力

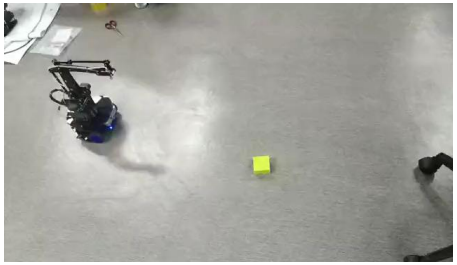
数据分析 → 智能终端控制

智能嵌入式应用

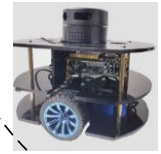
结合语义的目标识别

深度学习

智能抓取



视觉追踪



运动控制

全向运动

自主导航定位追踪系统

SLAM技术

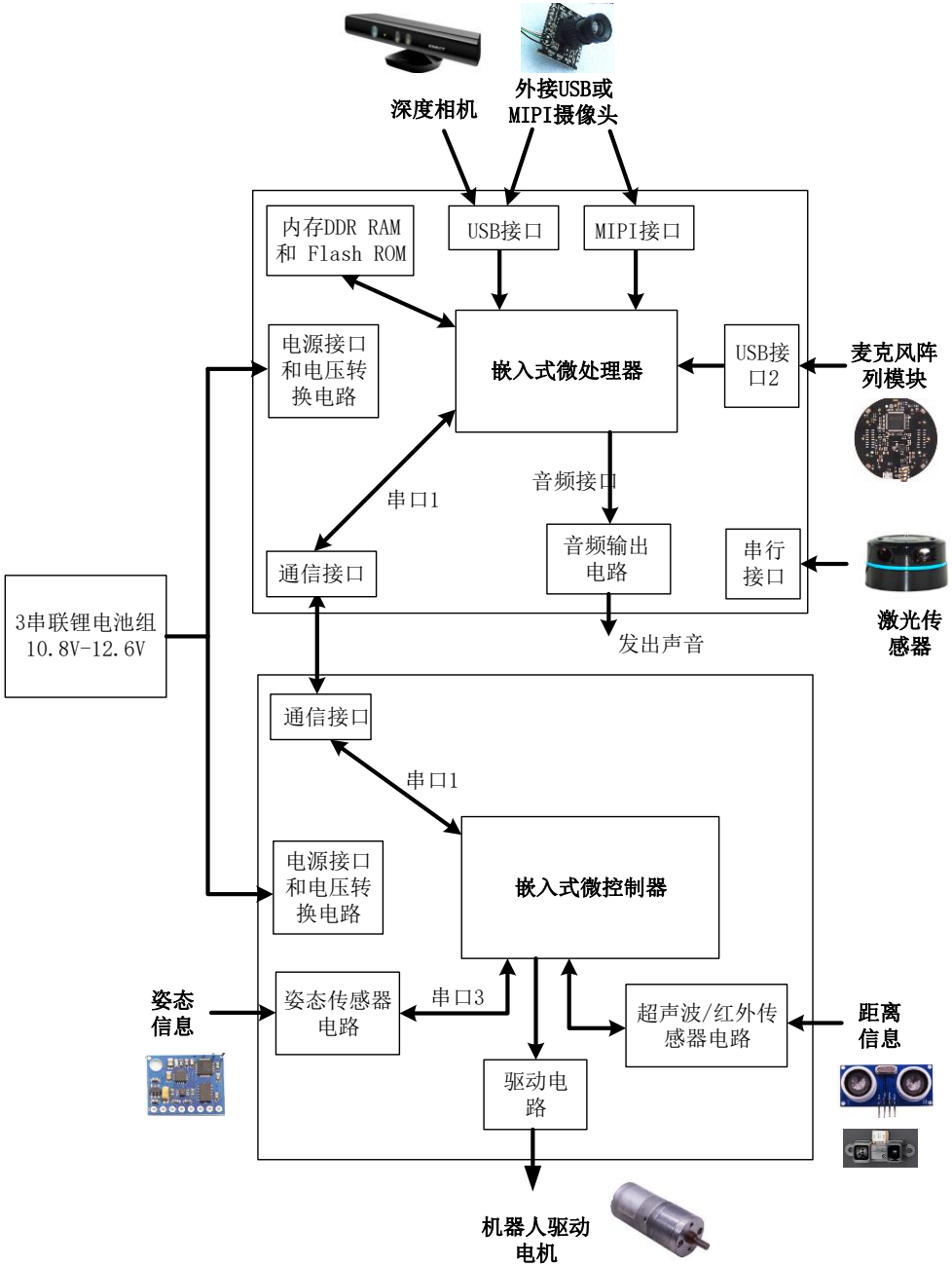
多机协作

全向移动机械臂

智能控制及强化学习

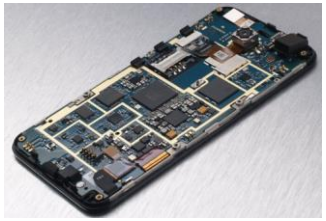
自主行走

智能嵌入式通用架构


深度相机
外接USB或
MIPI摄像头
内存DDR RAM
和 Flash ROM
USB接口
MIPI接口
电源接口
和电压转
换电路
嵌入式微处理器
USB接
口2
麦克风阵
列模块
串口1
音频接口
通信接口
音频输出
电路
串行
接口
激光传
感器
发出声音
3串联锂电池组
10.8V-12.6V
通信接口
串口1
电源接口
和电压转
换电路
嵌入式微控制器
姿态
信息
姿态传感器
电路
串口3
超声波/红外传
感器电路
距离
信息
驱动电
路
机器人驱动
电机

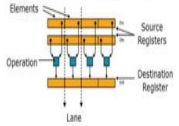
智能嵌入式硬件芯片

MPU-嵌入式微处理器



同构多核

NPU 神经网络处理器



异构多核

ARM

RISC-V

MCU-嵌入式微控制器

Mali GPU堆核

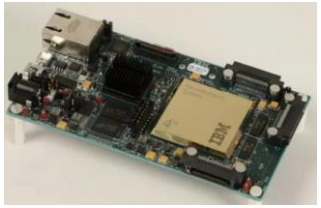
Adreno GPU大核

.....

DSP-数字信号处理器

SOC-片上系统

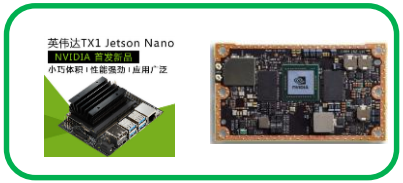
（神经网络芯片）NN



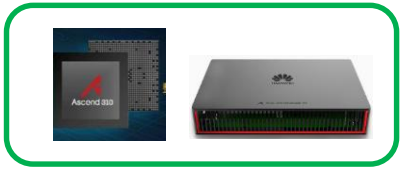
Open AI LAB EAIDK平台

中科创达 TurboX AI Kit

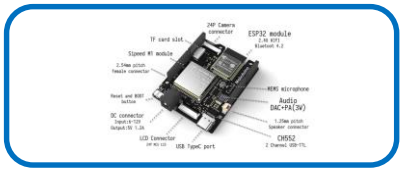
NVIDIA Jetson 平台



华为Atlas系列



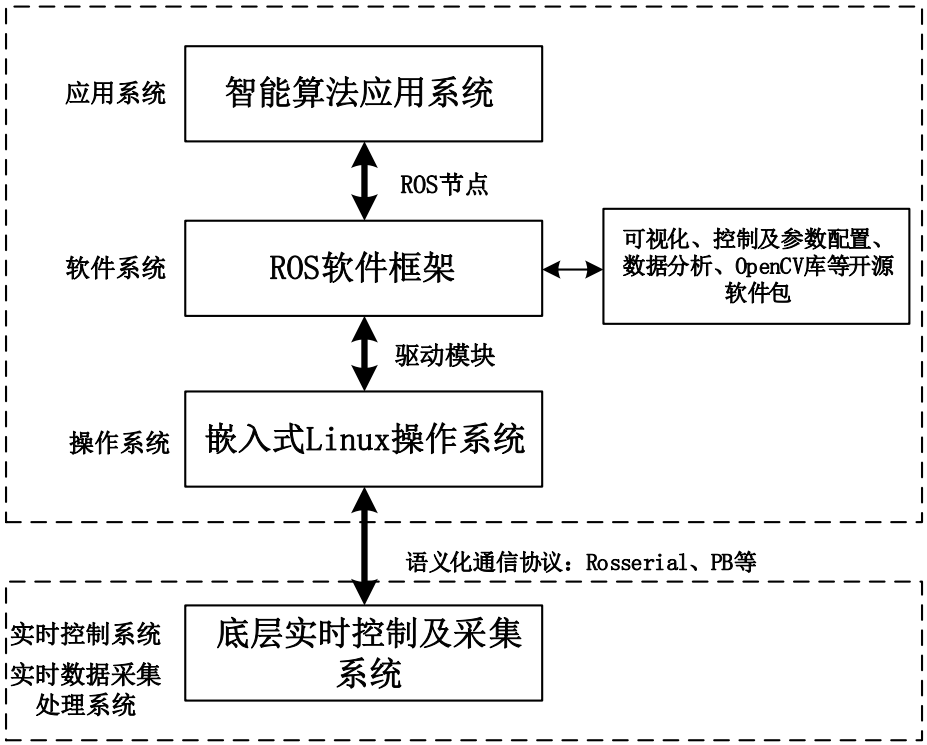
Kendryte K210



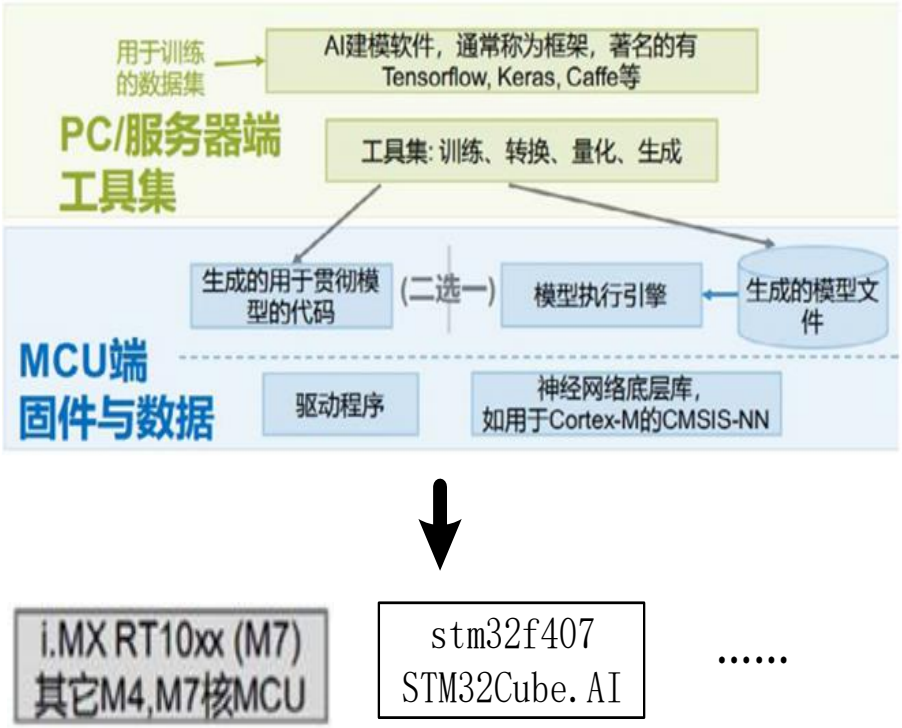
.....

智能嵌入式软件系统

MPU智能嵌入式软件系统

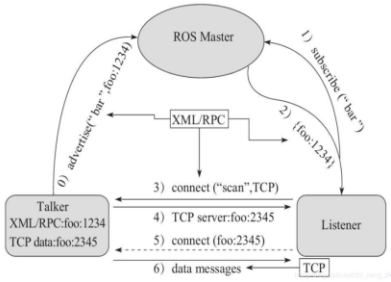


MCU智能嵌入式软件系统

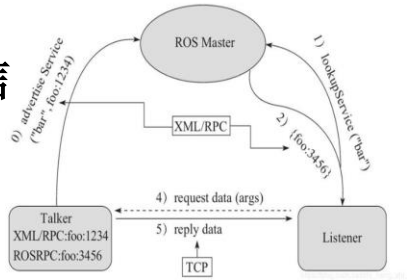


智能嵌入式软件系统--MPU

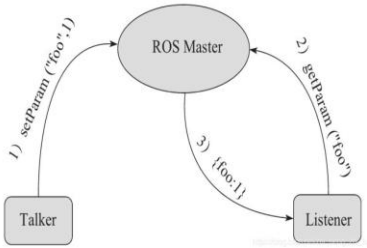
话题通信机制



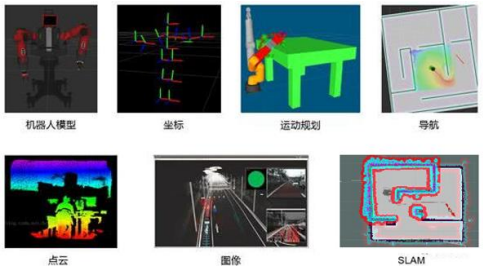
服务通信机制

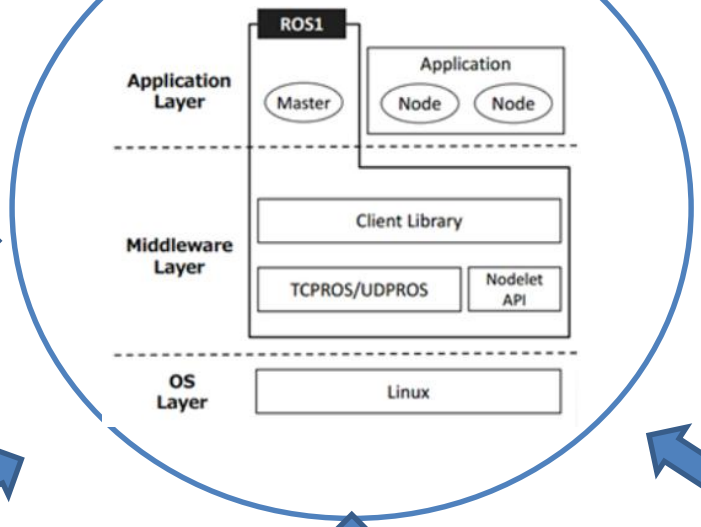


共享参数机制

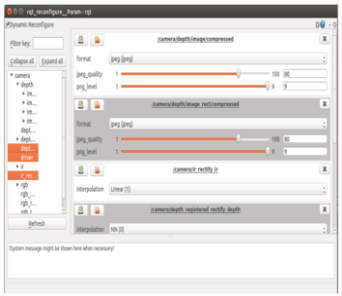


rviz可视化工具

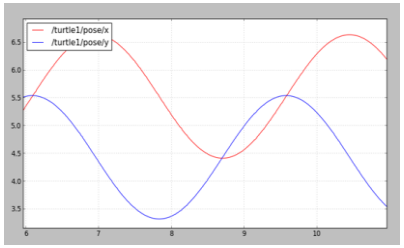




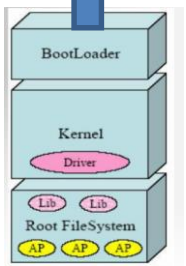
参数动态配置



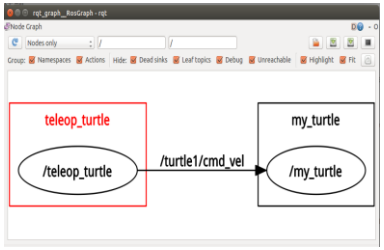
数据绘图



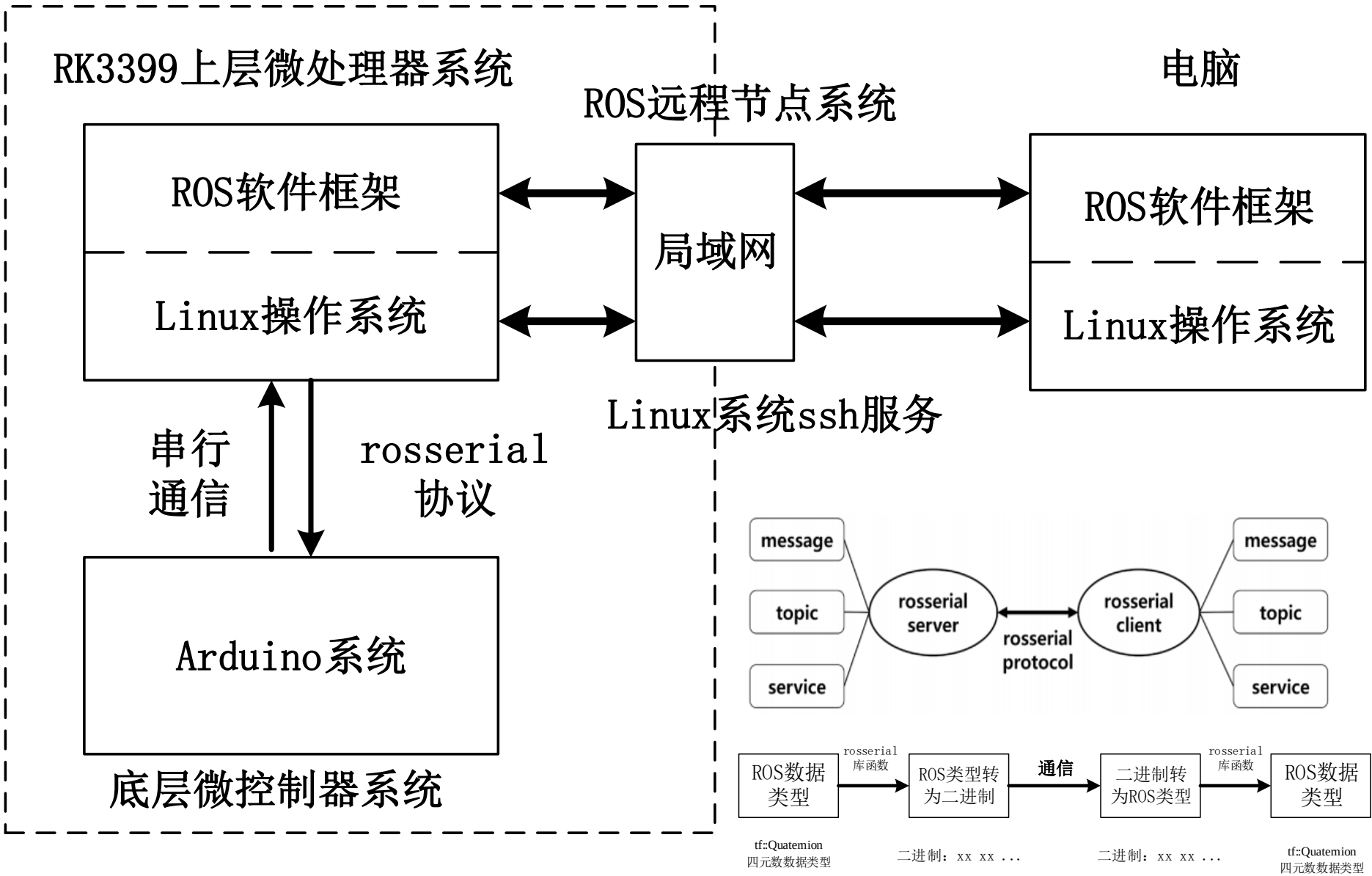
操作系统



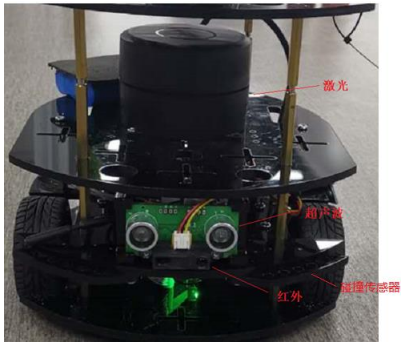
计算图可视化

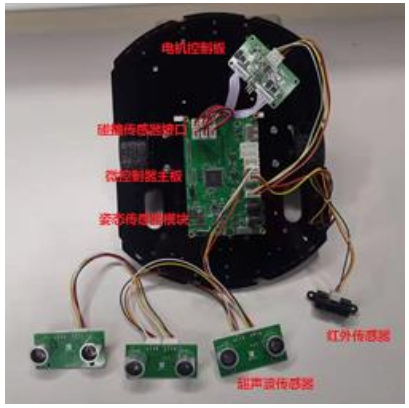


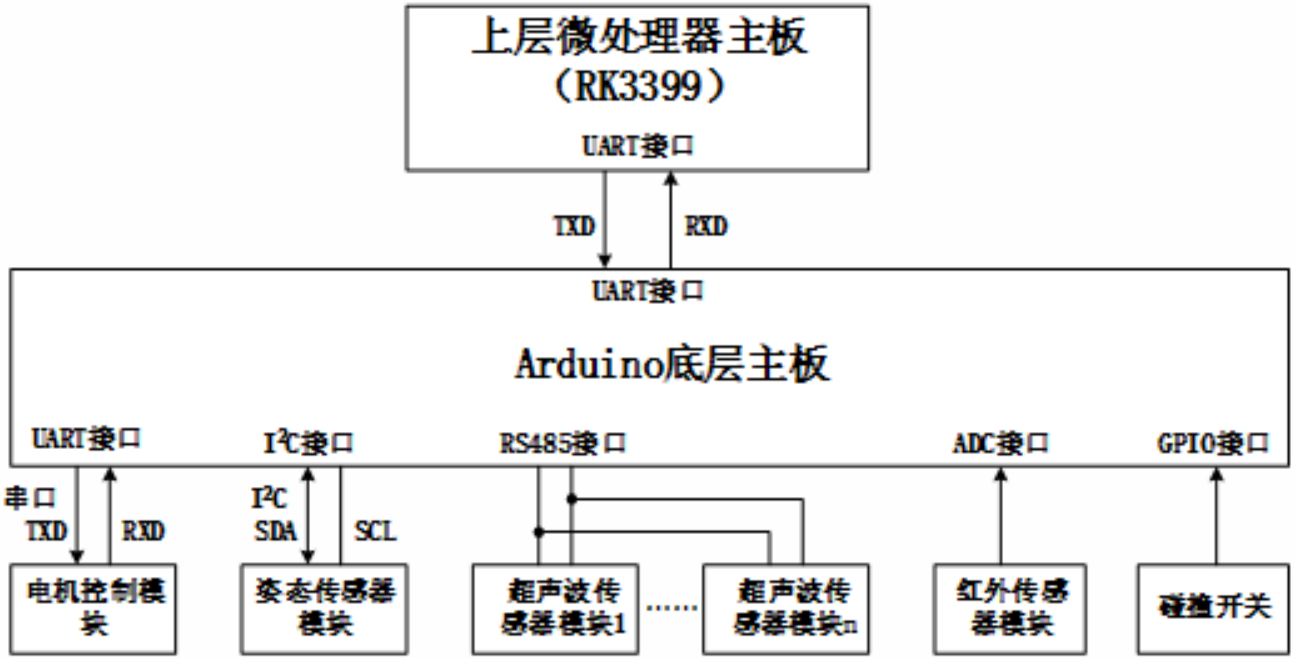
通信模式


RK3399上层微处理器系统
ROS远程节点系统
电脑
ROS软件框架
Linux操作系统
局域网
ROS软件框架
Linux操作系统
串行
通信
rosserial
协议
Linux系统ssh服务
Arduino系统
底层微控制器系统
message
topic
service
rosserial
server
rosserial
protocol
rosserial
client
message
topic
service
ROS数据
类型
rosserial
库函数
ROS类型转
为二进制
通信
二进制转
为ROS类型
rosserial
库函数
ROS数据
类型
tf::Quatemion
四元数数据类型
二进制：xx xx ...
二进制：xx xx ...
tf::Quatemion
四元数数据类型

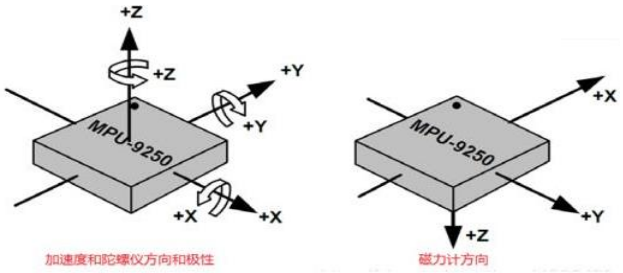
底层智能传感器系统



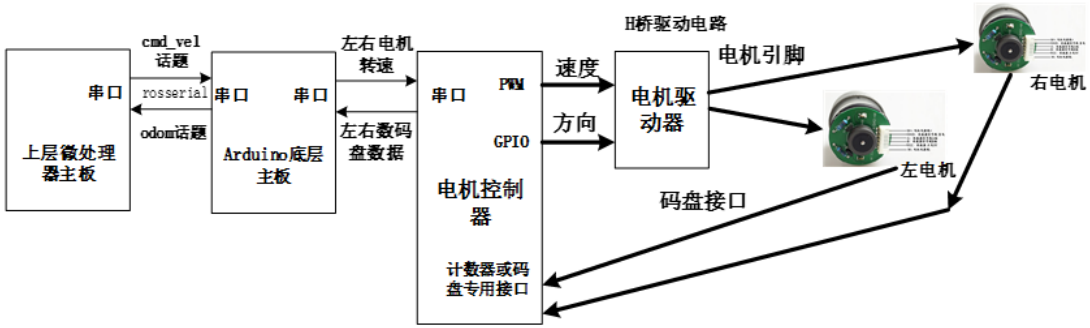




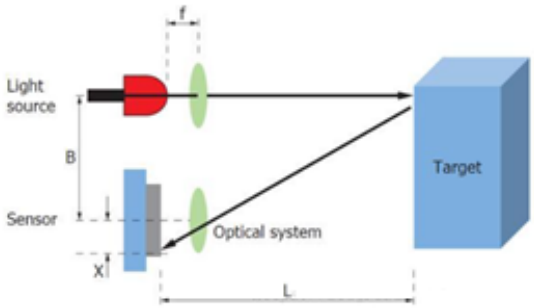
9轴融合:陀螺仪+加速度+电子罗盘四元数



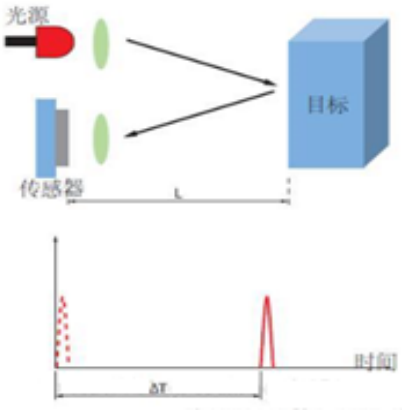
机器人小车控制



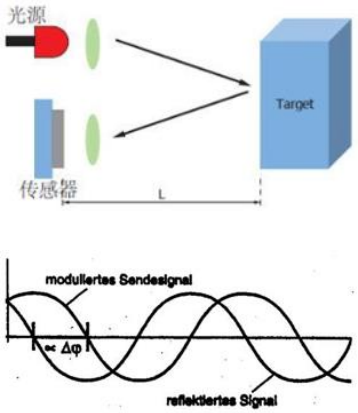
激光智能传感器系统



光学三角激光测距

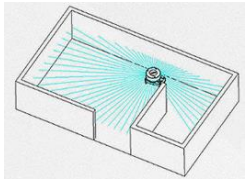


脉冲式TOF

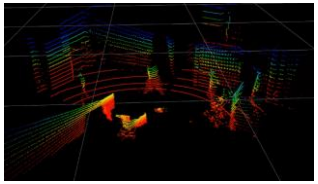


相位式TOF

单线激光



多线激光

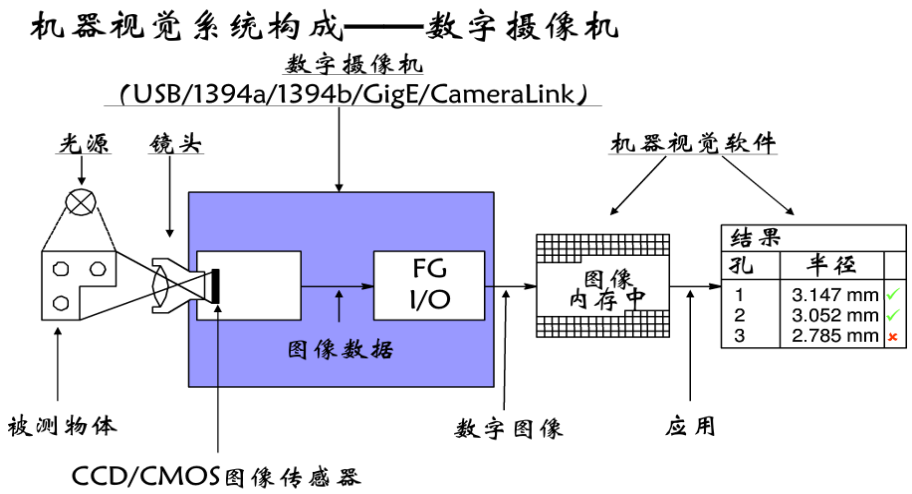


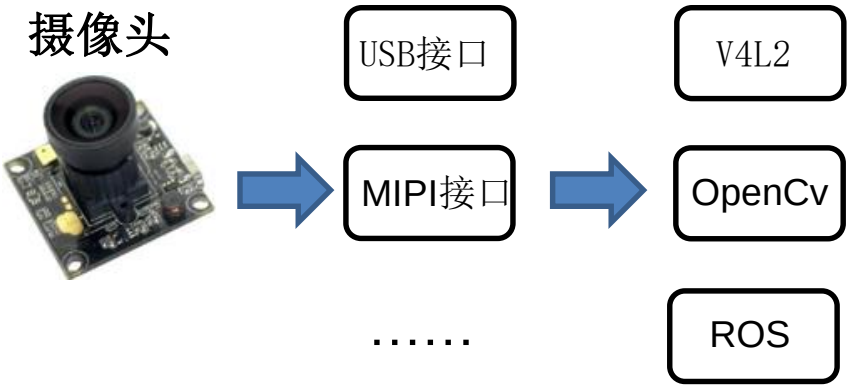
TOF传感器

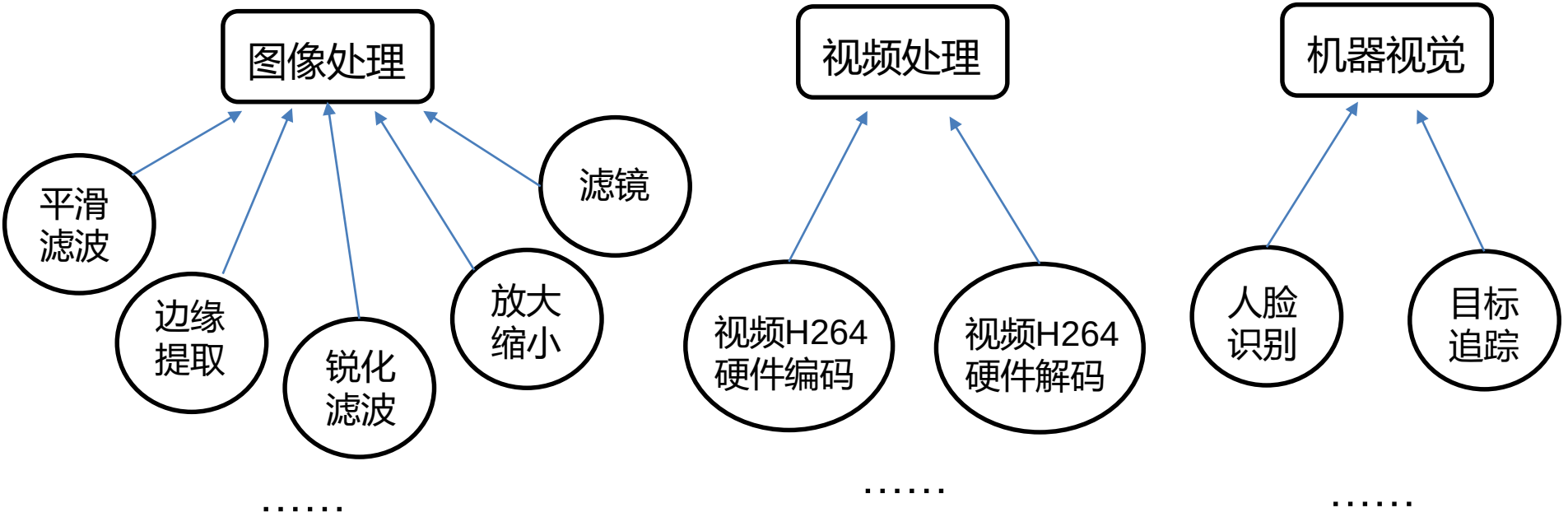
扫描匹配方法

- 卡尔曼滤波
- 迭代最近点（ICP）
- Hector方法
- 概率栅格方法
- 粒子滤波
- 正态分布变换（NDT）
- Cartographer方法

视觉智能传感器系统


机器视觉系统构成——数字摄像机
数字摄像机
(USB/1394a/1394b/GigE/CameraLink)
光源
镜头
机器视觉软件
FG
I/O
图像数据
图像
内存中
结果
孔
半径
1
3.147 mm
2
3.052 mm
3
2.785 mm
被测物体
CCD/CMOS图像传感器
数字图像
应用

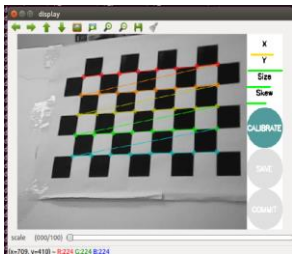

摄像头
USB接口
MIPI接口
.....
V4L2
OpenCv
ROS


图像处理
平滑
滤波
边缘
提取
锐化
滤波
放大
缩小
滤镜
.....
视频处理
视频H264
硬件编码
视频H264
硬件解码
.....
机器视觉
人脸
识别
目标
追踪
.....

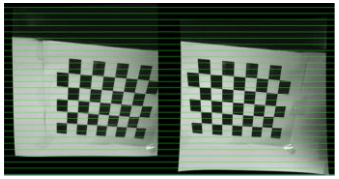
视觉定位系统

标定

单目相机

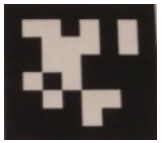


双目相机



RGBD深度相机

标签定位

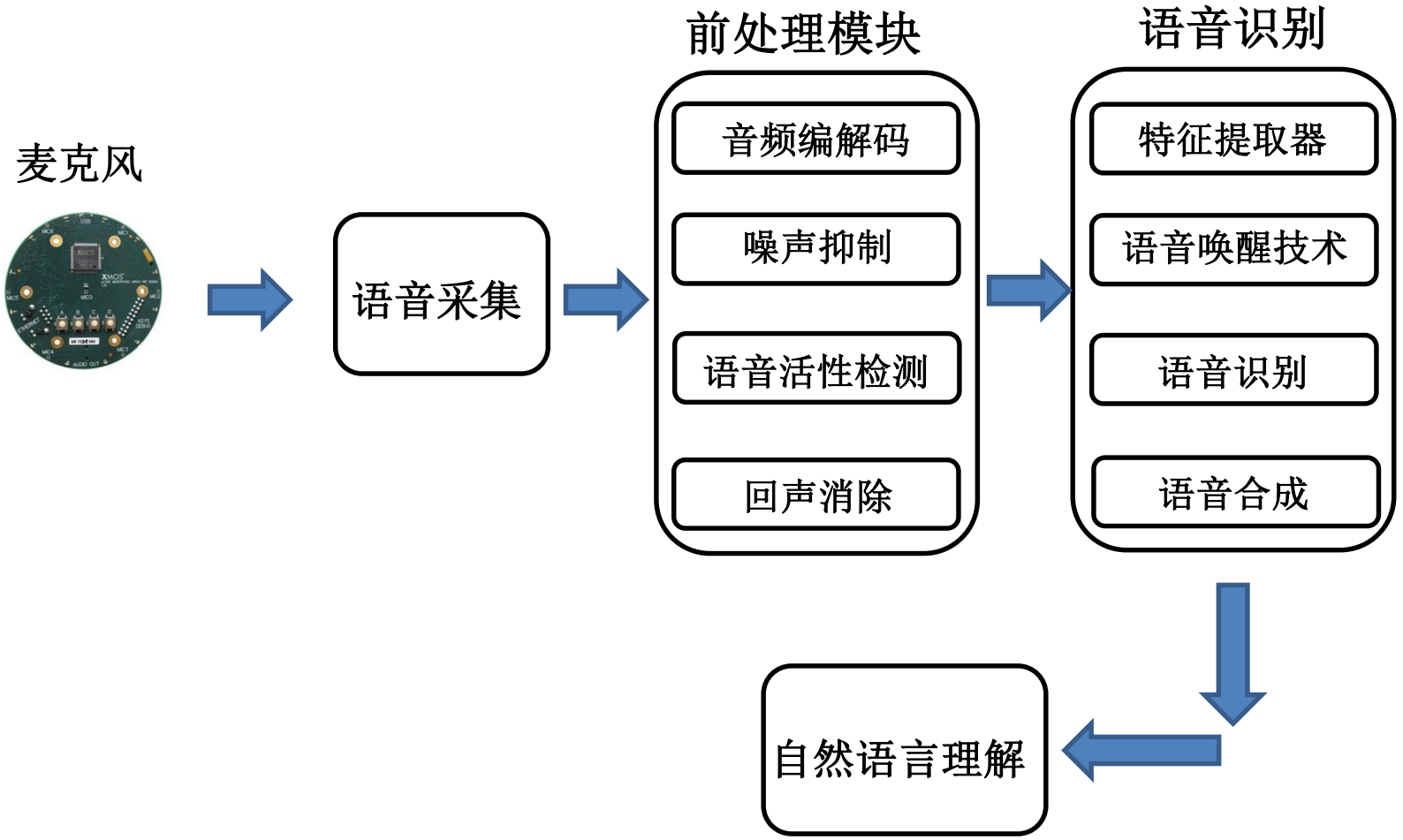


距离测量

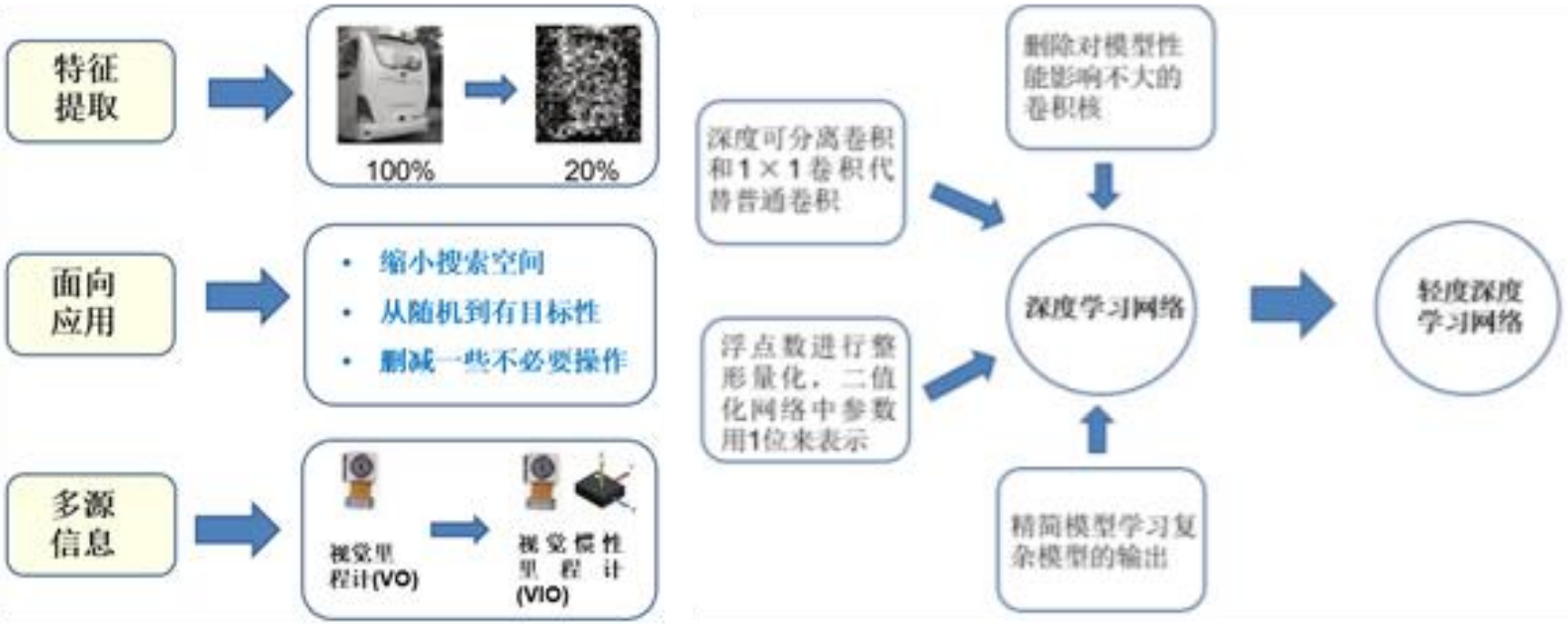
视觉SLAM

ORB和LSD方法等

自然语言系统


麦克风
语音采集
前处理模块
音频编解码
噪声抑制
语音活性检测
回声消除
语音识别
特征提取器
语音唤醒技术
语音识别
语音合成
自然语言理解

深度学习部署



在深度学习过程中，需要对网络进行简化，主要包括：1） 删除对模型性能影响不大的卷积核；2）深度可分离卷积和1×1卷积代替普通卷积；3）浮点数进行整形量化，二值化网络中参数用1位来表示；4）精简模型学习复杂模型的输出。

一些公司也开发了前端部署方案，例如ARM公司OPEN AI LAB的Tengine框架、Google公司的TensorFlow Lite、腾讯公司的NCNN框架 ， 小米公司的MACE框架、和百度公司的Mobile-deep-learning、亚马逊公司的TVM和美国高通公司的 SNPE等通过借助多核和加速单元实现卷积的快速计算,从而在移动设备上有效的实现深度学习算法。

深度学习--某个特定网络优化

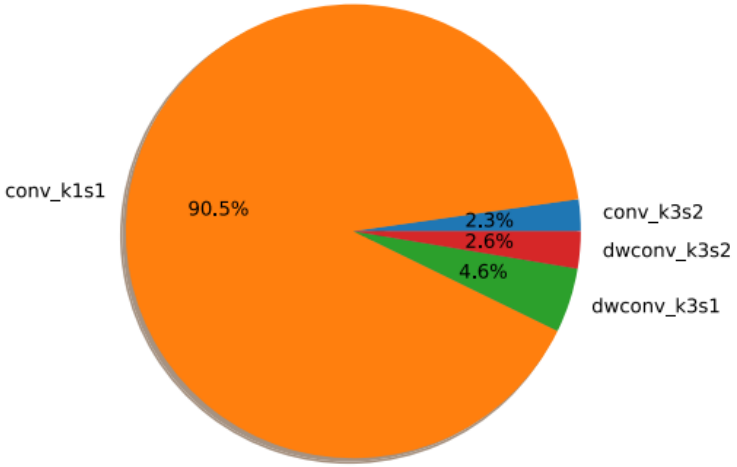
对某个特定网络的优化步骤：

1. 将一些简单的层与前一层合并，减少访存如BatchNorm层、激活层(ReLU)。
2. 测试网络中所有网络层的推理耗时。
3. 针对耗时非常严重的网络层进行专门优化（例如右图中的1x1卷积）。

3.1 对外层for循环，使用OpenMP将计算任务分配到多个CPU上。

3.2 对内存for循环，先进行循环展开，然后使用ARM NEON指令来优化单个CPU上的计算。

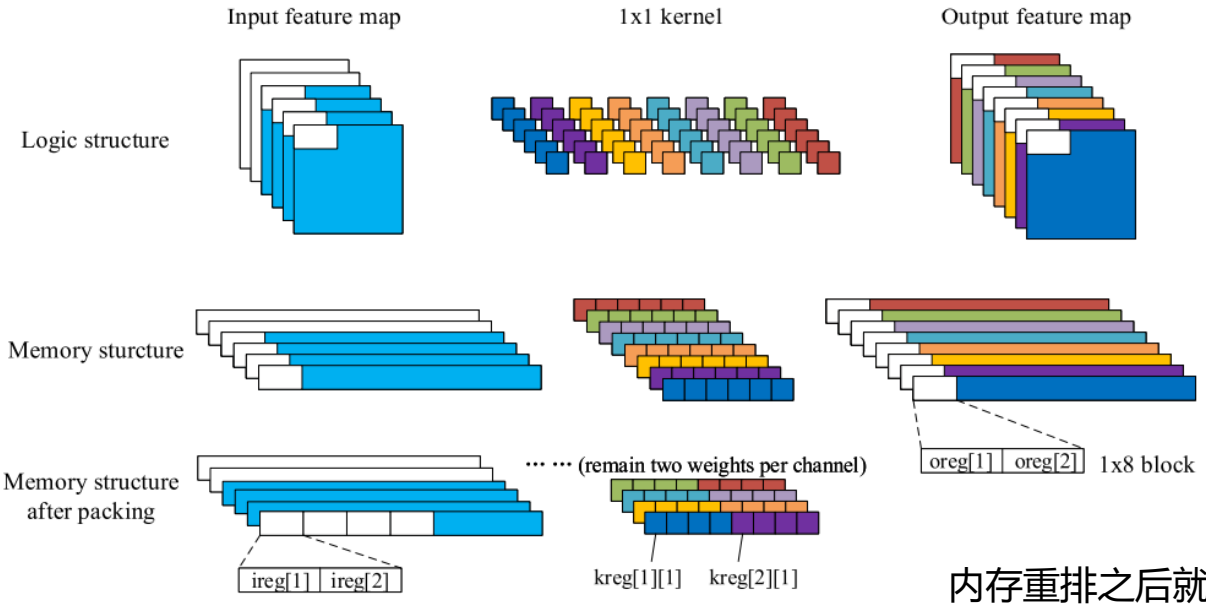
下图是MobileNet网络的耗时测试结果，图中并没有画出softmax和池化层，因为这些层的影响很小。



深度学习--某个特定网络优化

1x1卷积优化举例

对于输入数据量比较大的网络层，因为数据大，那么计算过程中所需要访问的内存就会大。而对于内存的访问模式我们是可以根据计算过程来提前了解到的。所以我们可以将计算的输入数据（输入特征图和卷积核）的内存根据计算过程中的内存访问模式来重排，增强访存的局部性，提高cache命中率，减少内存延迟对CPU计算的影响。



1x1卷积的优化方法：
最上面分别是：6个输入特征图、8个1x1卷积核、8个输出特征图的逻辑结构。
中间分别是：对应的在内存中的存储结构。
最下面是：经过内存重排后的内存布局。

内存重排之后就可以利用OpenMP和ARM NEON指令来优化1x1卷积的计算过程。针对特定的网络优化过的网络层，对于其他使用到该层的网路也会有一定的加速作用。

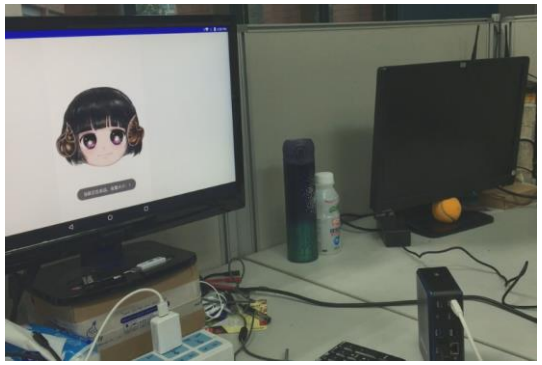
实例1-机器人运动控制定位及导航

SLAM方法 -- 创建地图、导航及追踪

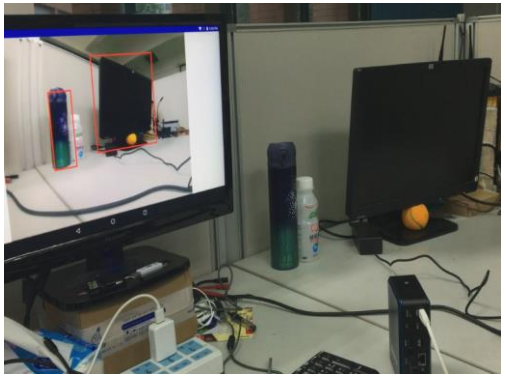
地图创建方法: 例如Gmapping,Hector, Cartographer方法等。
定位中匹配方法: 例如粒子滤波，迭代梯度下降(ICP，NDT)等。
导航方法: 全局导航A*，D*方法，局部导航DWA，TEB方法等。

实例2-结合自然语言理解和环境认知的智能服务系统

自然语言理解方法



基于深度学习物体认知方法



云端

语音输入 → 语音识别 → 语义分词

相机输入 → 基于深度学习物体认知 → 物体标签 概率 位置

语义和标签进行匹配 → 语义指称物体的位置 → 语音合成告知物体相对位置

喇叭发声

显示器显示

红框标记出识别到的物体

终端

- 谢谢！