CHAPTER 02

机器的"感官"

看见与听见

从像素到语义有多远？

本章学习目标

- 理解机器视觉原理与局限
- 掌握多传感器融合思想
- 认识感知五层阶梯



温故知新：PCA闭环中的"P"



回顾：感知是"世界的入口"——PCA闭环的第一个环节

本章定位：深入拆解"P"——感知层，从人眼到机器眼

核心对比：40亿年进化的人眼 vs 几十年工程的机器感官

思考：机器"看到"的世界，和人看到的世界一样吗？

人眼：进化40亿年的杰作



1.2亿 视杆细胞 + **600万** 视锥细胞

五大超能力

1. **双目视觉**：两眼间距6.5cm，大脑自动计算深度
2. **注意力机制**：仅5%视觉输入进入意识
3. **动态适应**：动态范围10000:1（摄像头仅1000:1）
4. **预测性追踪**：大脑预判物体轨迹，提前移动眼球
5. **颜色恒常性**：自动校正不同光源下的颜色

核心观点：人眼不是"摄像头"，是与大脑深度耦合的智能感知系统

有趣发现：卷积神经网络（CNN）"重新发现"了生物视觉的边缘检测原理

摄像头：机器的"像素眼"



工业摄像头家族

核心区别

摄像头看到"像素"，人眼看到"物体"和"意义"

从"像素矩阵"到"语义"——需要算法来"翻译"

摄像头家族

- 单目：二维图像，无法直接感知深度
- 双目：模仿人眼视差，计算深度
- 环视：4-8个摄像头360度覆盖
- 红外：黑暗中"看见"，但分辨率低

事件相机（Event Camera）

微秒级时间分辨率 · 30毫秒内捕获"鬼探头" · 极高动态范围

从像素到语义：目标检测

传统方法

手工设计特征：边缘、角点、纹理

在图像中搜索特征组合

像在没有索引的书里找内容

效率低、准确率低

深度学习革命

让机器自己学习"什么是重要的"

就像教小孩认动物

不需要解释特征，反复指认即可

效率高、准确率高

YOLO

You Only Look Once

从"两阶段检测"到"一步到位"

YOLO v1速度：**每秒45帧**——实时处理视频流

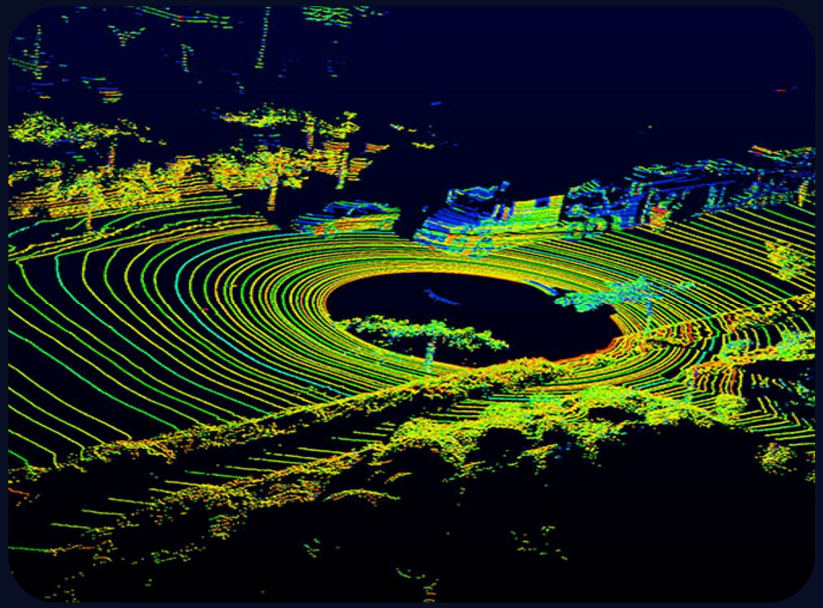
进化史：v1 → v2 → v3 → v4 → v5 → v8 → v10

实战案例

2025中国智能车未来挑战赛：改进YOLOv5模型

基于BDD100K数据集训练，同时检测13类目标——行人、车辆、交通信号灯、可行驶区域、车道线

激光雷达：机器界的"回声定位"



激光雷达三维点云图

原理

发射激光 → 接收回波 → 计算时间差

精确测距，误差仅几厘米

输出

三维"点云"——环境的"数字骨架"

演进

机械旋转式 → 固态（MEMS/OPA/Flash）

三大优势

- 测距精度极高（厘米级）
- 不受光照影响
- 直接输出三维信息

三大局限

- 怕雨雾天气
- 对低反射率物体弱
- 成本较高

毫米波雷达：穿透雨雾的"千里眼"

原理

发射毫米波电磁波 → 接收回波

多普勒效应测速

频段

77GHz（远距200m+）vs 79GHz（高分辨率）

最大优势

穿透雨雾 · 直接测量相对速度



短板

· 分辨率低 · "ghost object"虚警问题

· 2016年特斯拉事故：电磁波穿过白色拖车底盘

新方向：4D成像雷达

低成本替代激光雷达的潜力

8发16收 → 角度分辨率大幅提升

纯视觉 vs 多传感器融合：特斯拉的"豪赌"

特斯拉 · 纯视觉路线

8个摄像头 360度覆盖

2021年取消毫米波雷达

"人类用两只眼睛开车，机器也可以"

优势：成本低（摄像头$50）

劣势：光照敏感、夜间性能下降

华为/小鹏 · 多传感器融合

摄像头 + 激光雷达 + 毫米波雷达

各司其职、取长补短

优势：全天候鲁棒性强

劣势：成本高、系统复杂

2025小米SU7事故教训：夜间纯视觉未能识别施工路障

核心思想：不是"谁对谁错"，而是"场景决定方案"





INTERMISSION

课间休息

播放传感器工作原理科普短视频

同学间交流思考题

多传感器融合：破解"盲人摸象"

"盲人摸象"——摸到腿的人说像柱子，摸到耳朵的人说像扇子，摸到尾巴的人说像绳子。
每个传感器都是"盲人"，只摸到了"象"的一部分。

摄像头

像摸到耳朵的盲人

能看清颜色和纹理

能识别交通标志

但测不准距离

激光雷达

像摸到腿的盲人

能精确测距

能构建三维模型

但分不清红绿灯颜色

毫米波雷达

像摸到尾巴的盲人

能穿透雨雾测速

但"看不清"物体形状

不知道是小轿车还是纸板

融合的核心挑战：不同数据格式如何"对齐"？

摄像头=二维图像 · 激光雷达=三维点云 · 毫米波雷达=距离-速度-角度



三级融合策略：前融合、深度融合、后融合

01 数据级融合（前融合）

原始数据直接拼接

信息保留最完整

优势：信息最多

劣势：难度最大

数据格式差异巨大

02 特征级融合（深度融合）

BEV Transformer

将多传感器数据映射到

统一"鸟瞰图"空间

优势：主流方案

"上帝视角"

03 决策级融合（后融合）

各传感器独立判断

→ 投票加权

优势：最简单

劣势：信息损失最大

早期自动驾驶常用

BEV视角："上帝视角"

把世界"展开"成二维平面

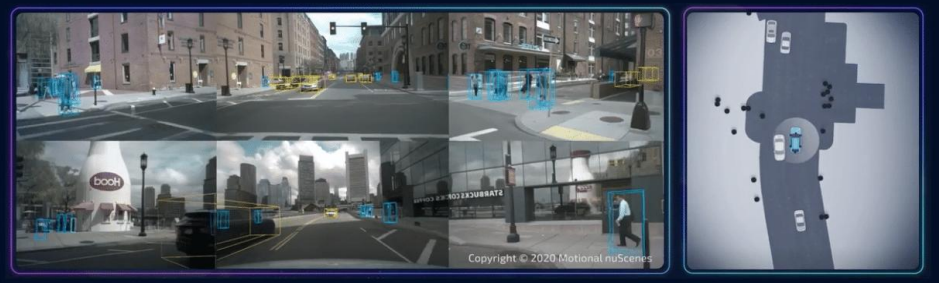
距离变成"位置"

核心思想：

不是"传感器A看到什么"

而是"在车的左前方30米处

有什么物体"



时空对齐：融合成功的关键

时间同步

摄像头 30fps vs 激光雷达 10Hz

几十毫秒误差 = 1米错位

解决方案：

GPS授时 / PTP协议

精度达微秒级

空间对齐

摄像头坐标系 vs 激光雷达坐标系

vs 车辆坐标系

解决方案：

外参标定（厘米级精度）

统一转换到车辆坐标系

"工程细节决定系统成败"

时空对齐做不好，融合就是"把错误的数据拼在一起"

类比：交响乐团

每个传感器是不同乐器，时空对齐就是指挥家

没有指挥 → 各乐器自顾自演奏 → 混乱的噪音

实战案例：自动驾驶如何破解"鬼探头"

"鬼探头"：行人从盲区突然冲出——自动驾驶的终极考验

方案A技术路线

YOLOv5
目标检测
→
筛选
关注目标
→
构建
安全区域
→
鸟瞰图
转换
→
侵入判断
+ 距离估算

核心创新

用边界框聚合替代像素级分割——"以工程效用为导向"

时序融合：对比前n帧安全区域，检测"异常侵入"

成果：2025中国智能车未来挑战赛 全国第五名

40多支高校队伍中，唯一使用纯视觉方案进入前五

感知的五层阶梯：从"检测"到"理解"

L1	检测	"那里有东西"——感知系统的起点
L2	分类	"那是一个人"——从形状到语义
L3	跟踪	"那个人在移动"——从静态到动态
L4	预测	"那个人可能要过马路"——从观察到预判
L5	语义理解	"那个人闯红灯了，需要紧急制动"——开放难题

当前水平：**L3-L4之间** | L5是开放难题

感知中的"不确定性"：机器也会"没把握"

不确定性是 feature，不是 bug

—— 它反映了真实世界的本质

置信度意识

85% confidence this is a car → 15% 的概率是别的

自动驾驶系统必须知道"自己不知道什么"

保守策略

宁可误报，不可漏报

安全优先于舒适

灵敏度 vs 特异度

永恒的平衡

太敏感→误报多

太保守→漏检多

类比：学生考试——有的题100%确定，有的题"好像是A"

好的感知系统 = 知道"哪些题有把握，哪些题没把握"

开放世界：感知的终极挑战

训练数据不可能覆盖所有物体：兔子、床垫、恐龙服装、掉落的轮胎.....

"长尾分布"：99%常见场景 + 1%罕见场景（最致命）

应对方向

开放词汇检测

用自然语言描述任意物体

"检测所有红色的圆柱形物体"

异常检测

识别"不属于任何已知类别"的物体

"这是异常"

持续学习

部署后不断学习新物体

"终身学习"

世界模型

让机器拥有"常识"

"理解世界如何运作"

核心洞察：感知不是终点，是起点

从"看见"到"看懂"到"理解"——感知只是智能的第一步

第2章核心总结

人眼：40亿年进化的奇迹

机器"人造感官"的工程替代

六大传感器：各有特长、各有短板

没有"完美传感器"

多传感器融合：1+1+1>3

时空对齐是基础

感知五层阶梯：从"看见"到"看懂"

L1检测 → L5语义理解

开放世界：训练数据永远不够

机器需要"想象力"——世界模型、持续学习

感知不是"看见"，而是"理解"

从像素到语义，从数据到意义

课堂讨论 & 实验二预告

课堂讨论

问题：如果你要设计一辆自动驾驶汽车，你会怎么配置传感器？

预算：1000美元

· 摄像头 $50 · 激光雷达 $500 · 毫米波雷达 $200

· 超声波 $10 · IMU $30 · GPS $20

实验二预告

"从车的视角看世界"

· 虚拟小车第一视角体验

· 观察"像素→语义"的感知过程

· 对比不同传感器的"世界观"

· 亲手配置传感器组合，感受融合效果

谢谢聆听

THANK YOU

感知不是"看见"

而是"理解"

下章预告

机器的"大脑"——从强化学习的"试错进化"

到世界模型的"主动想象"