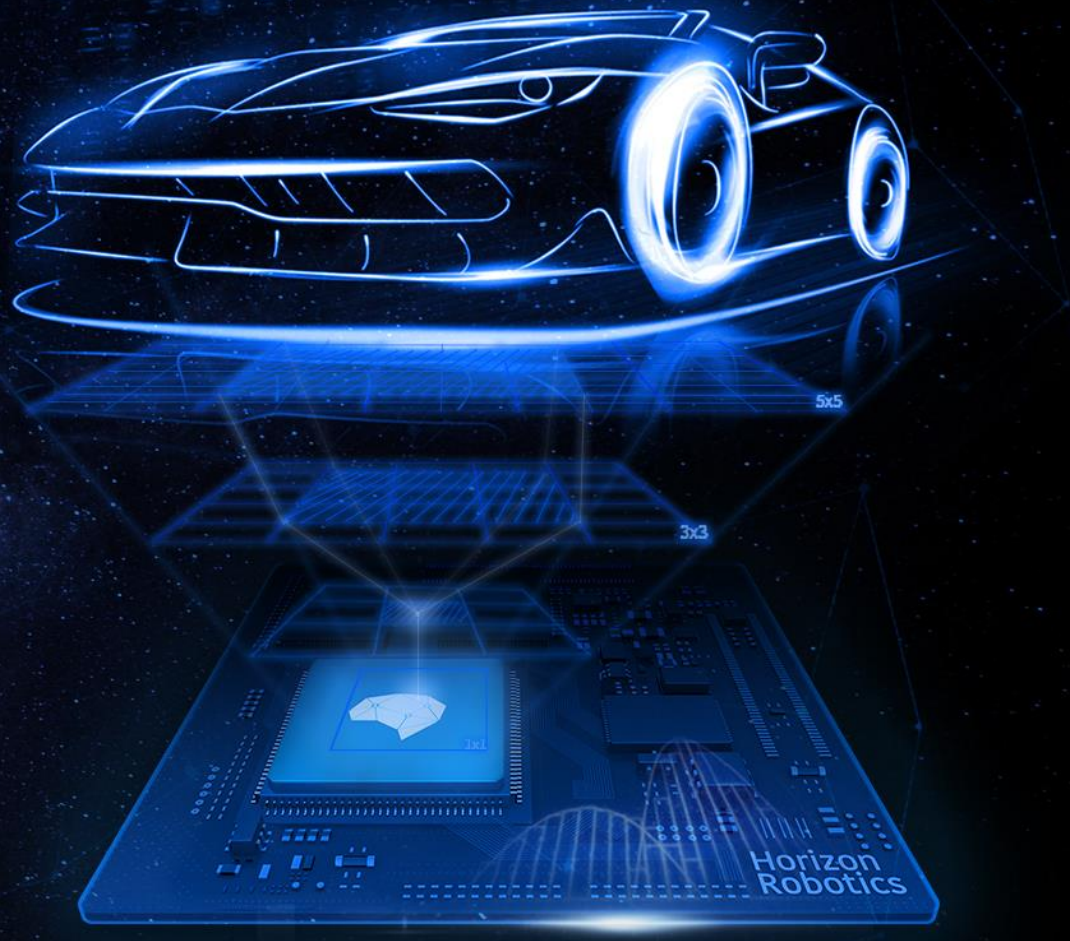
地 平 线
Horizon Robotics

BEV/Transformer 范式下的智驾量产实践



武 锐
rui.wu@horizon.ai

1. **经典单目感知算法方案**
2. BEV感知算法方案
3. Sparse4D感知算法方案
4. 感知量产方法论

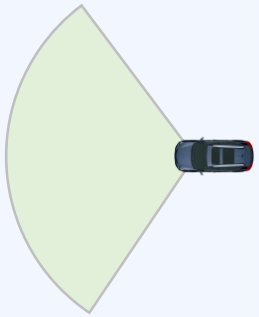
Horizon Mono 前视辅助驾驶解决方案

首个基于国产AI芯片的ADAS量产方案

高性价比方案

Matrix® Mono 2 (主打volume车型)

- Journey 2 AI Processor
- 2MP/3MP Camera@100° FOV



更安全的日常出行

C-NCAP 中国新车评价规程

面向ADAS标配基础功能

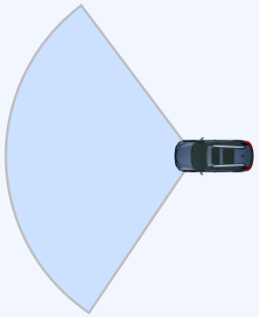
- 主动安全功能：FCW、LDW、AEB...
- 舒适性驾驶功能：ACC、LKA、TJA...

全球首个8MP前视ADAS量产方案

高性能方案

Matrix® Mono 3 (面向精英车型)

- Journey 3 AI Processor
- 8MP Camera@120° FOV



更舒适的驾驶体验

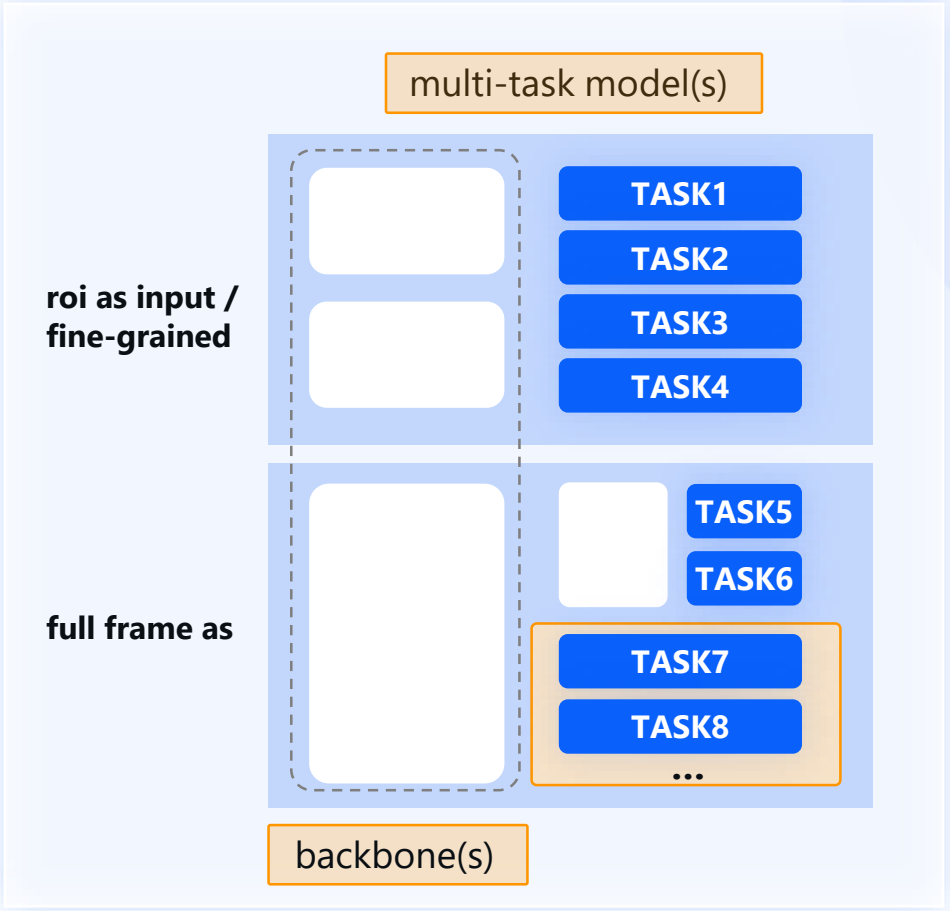
FOR SAFER CARS EURO NCAP

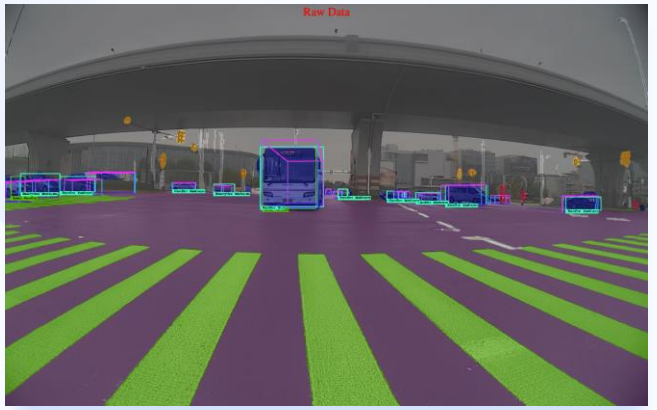
面向ADAS功能提升需求

- 匝道汇入/汇出
- 施工区域避让
- 复杂路口通行
- 智能调速
- 车道级导航
- 视觉定位

多任务模型组成感知系统

面向ADAS单目解决方案，我们采用多任务模型来获得更好的感知性能和更低的系统延迟
从backbone，任务设计和多任务模型改进三个方面展开





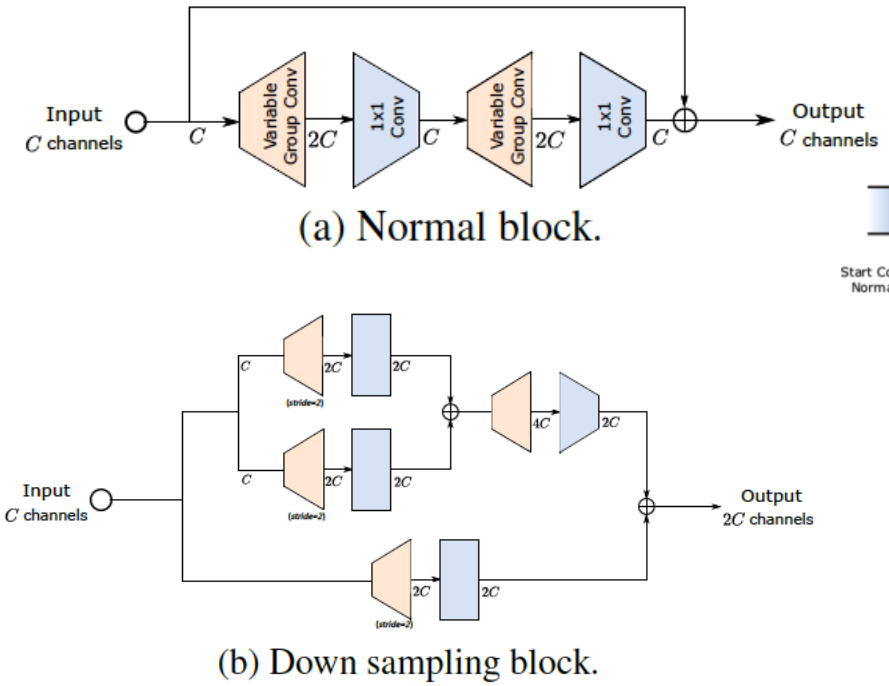
结合BPU特性的Backbone设计

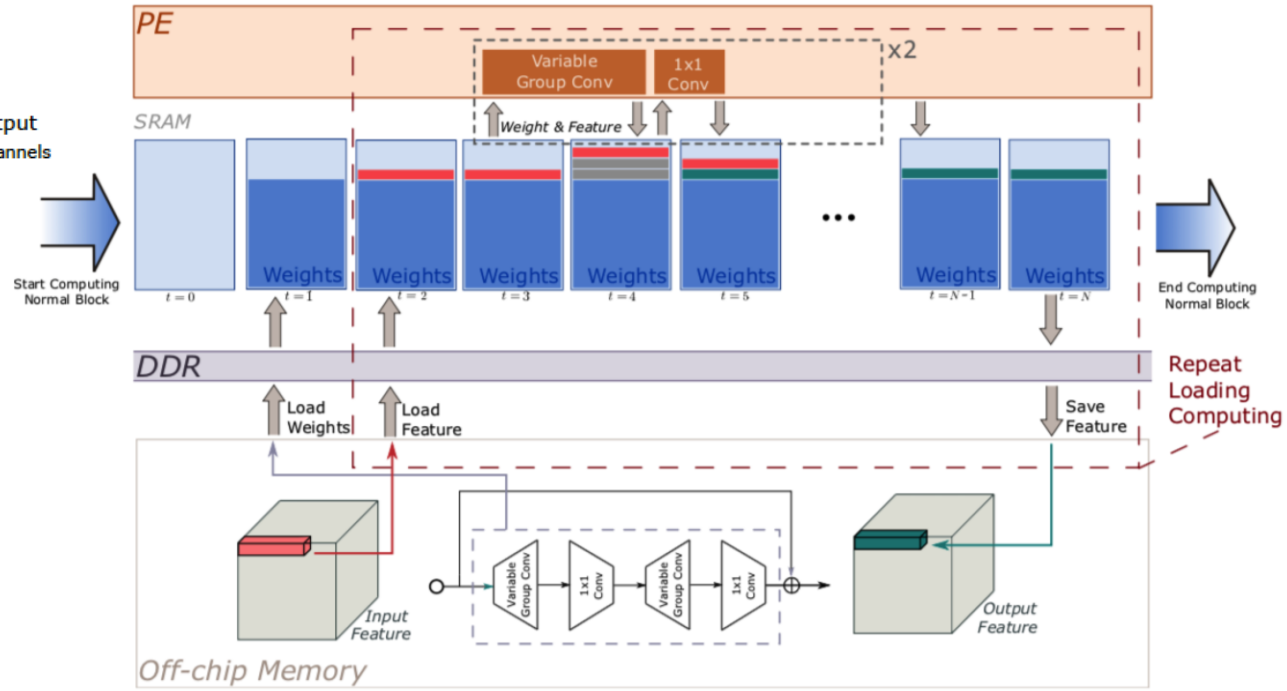
我们提出VarGNet

- 可变组(variable group)卷积升维
- Pointwise conv降维
- Group数可变，卷积中每个分组channel数一致

VarGNet的以下特性使其达到更好的计算和访存平衡

- Block内计算更均衡
- Block间采用更小的feature map

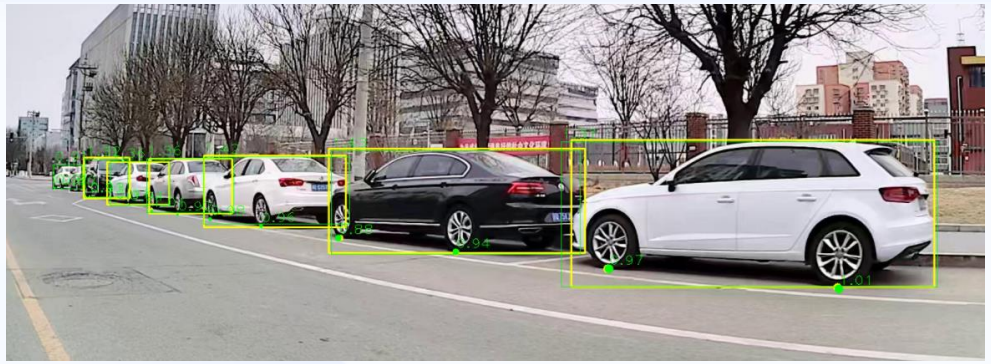




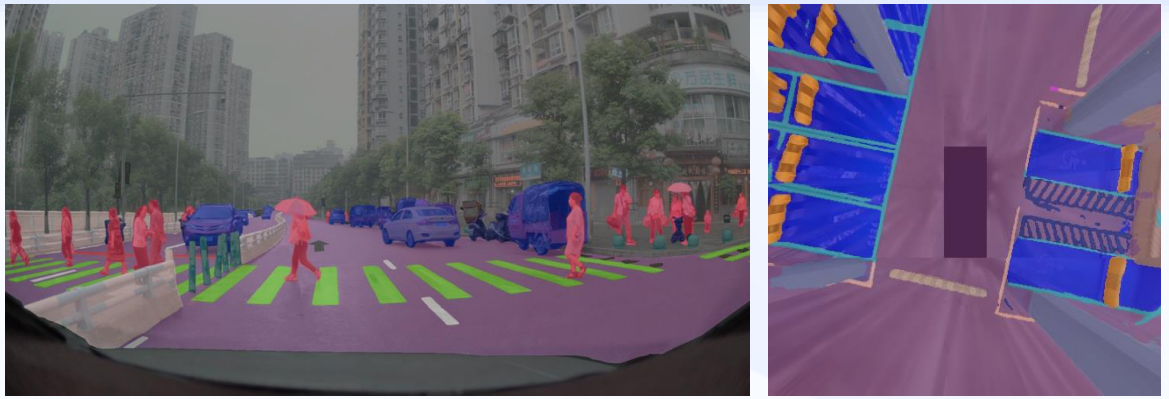
Zhang Q, Li J, Yao M, et al. VarGNet: Variable Group Convolutional Neural Network for Efficient Embedded Computing[J]. arXiv preprint arXiv:1907.05653, 2019.

感知任务设计 —— 2D

- 基于深度学习的计算机视觉经典任务构成了自动驾驶的2D感知
- 例如：检测、分割、分类、关键点检测等
- 2D感知的结构化输出需要结合先验进一步获取我们需要的信息



检测



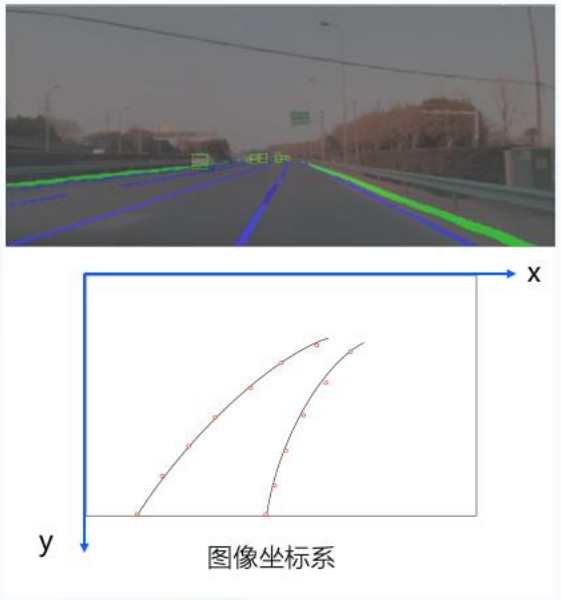
分割

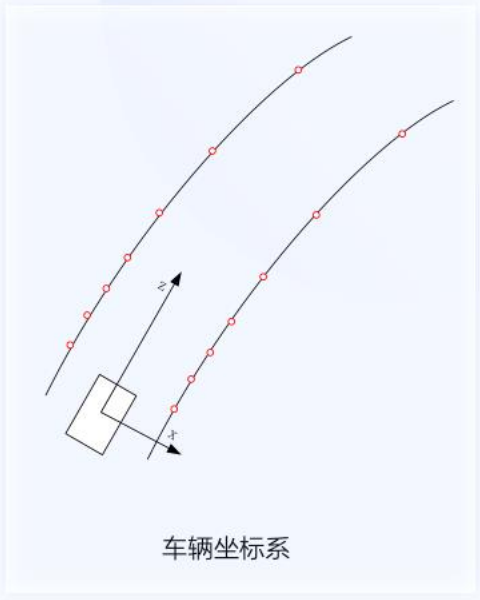
分类

关键点

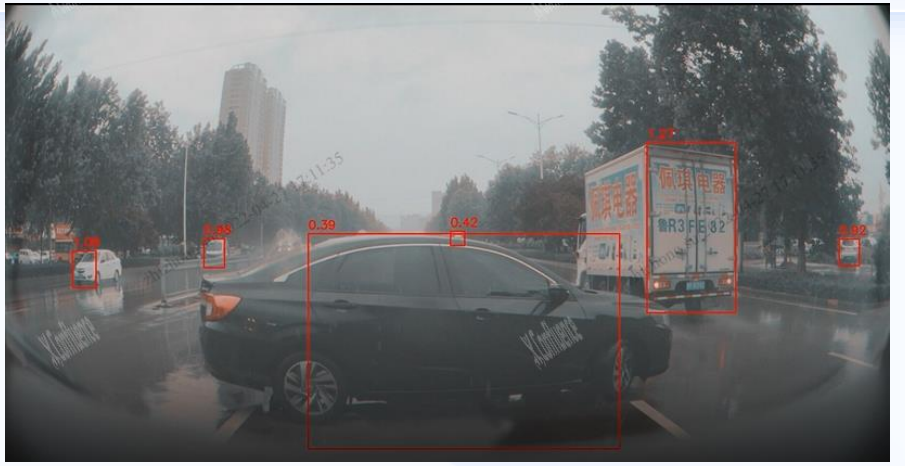
2D感知的局限性

- 2D感知依赖人工逻辑设计，难以规模化扩展
- 为了使用深度学习（数据和计算）端到端的优化，我们需要原生的3D环境感知





2D分割到车辆坐标系转换复杂



2D检测不能很好应对横穿车辆

2D感知不能很好的估计遮挡车辆的pose

感知任务设计 —— 单目深度估计

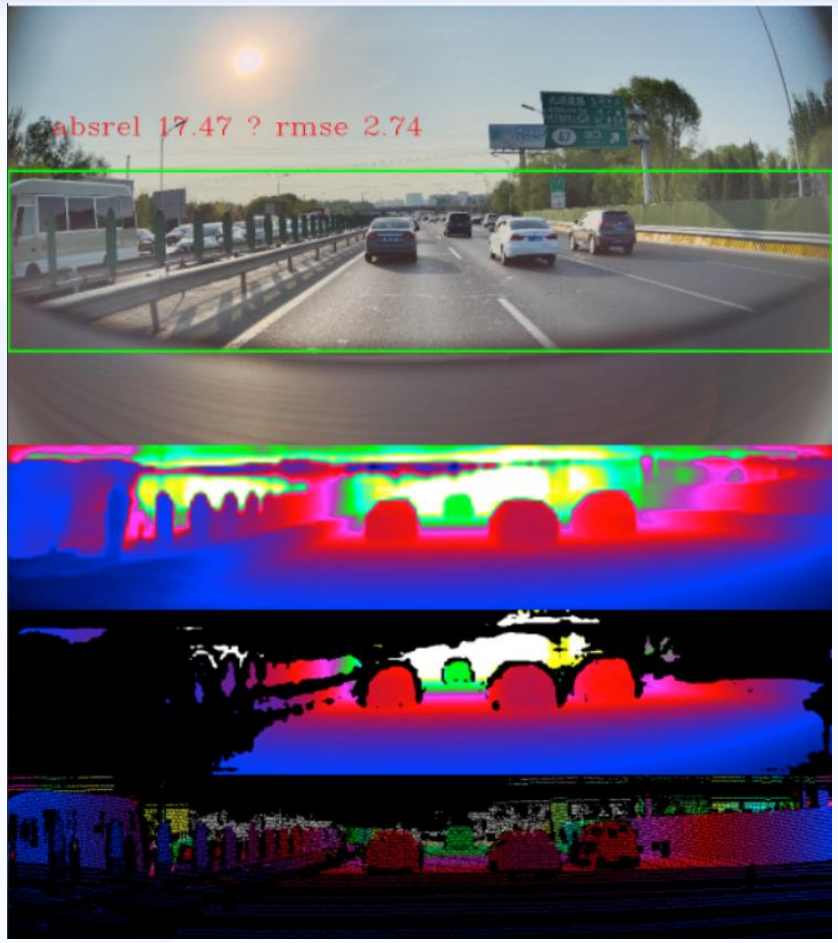
- 通过由激光雷达点云提供的稀疏监督，我们可以使用深度学习模型直接回归RGB图片中的深度
- 深度值可以帮助自动驾驶系统更好的理解场景和场景中目标的距离和速度
- 基于多摄像头和多帧的光度一致性，对深度预测模型进行无监督的优化

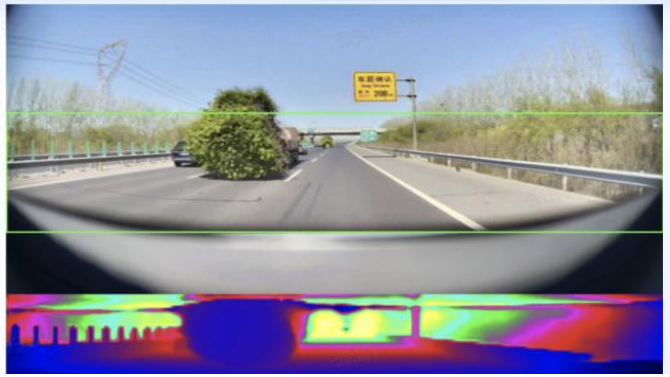
RGB图像

Depth预测

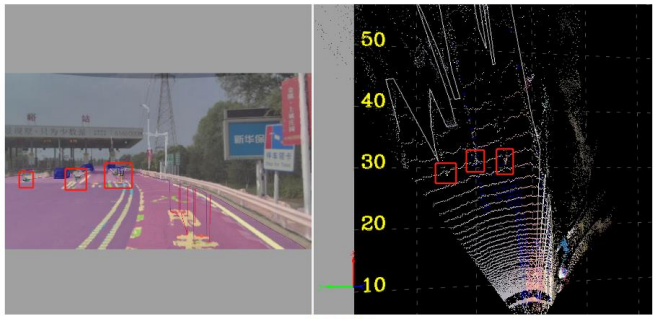
高置信度Depth

激光雷达提供的稀疏真值

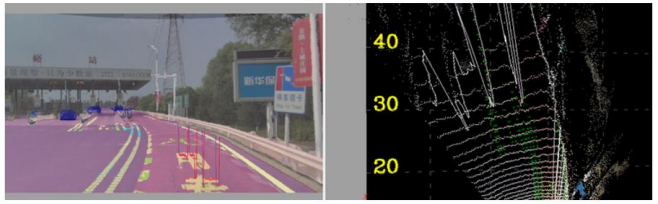




Depth对异形车的应对



外参测距

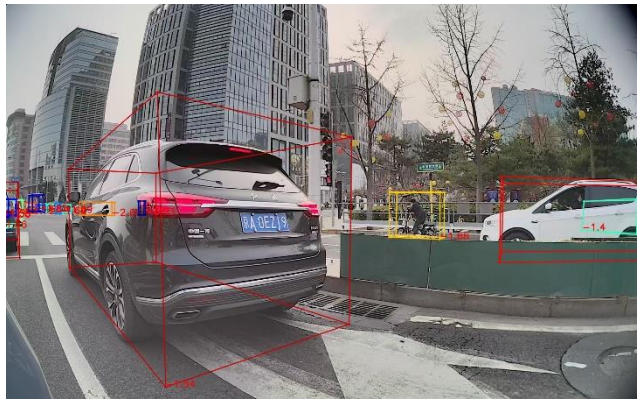


外参+depth

Depth对Freespace的改善

感知任务设计 —— 动态目标3D检测

- 车辆等动态目标的3D位姿对自动驾驶极其重要
- 通过Lidar检测得到的真实3D框作为监督，在对应的RGB图像中通过深度学习直接输出分类、位置、尺寸、朝向角



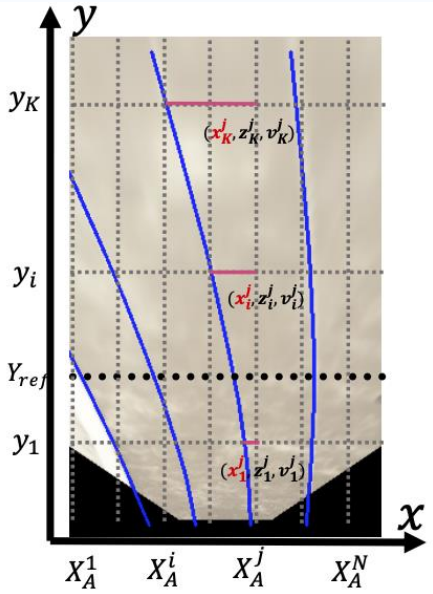
多类目标

横穿车辆

遮挡车辆

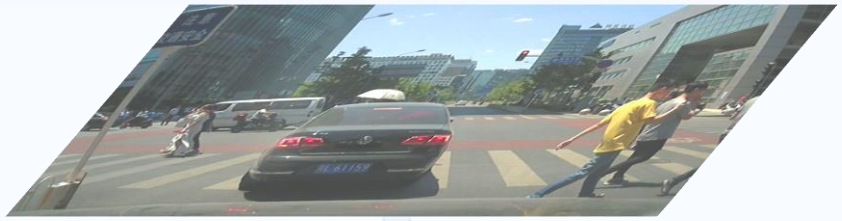
感知任务设计 —— 3D车道线

- 车道线是自动驾驶环境感知中最重要的静态目标
- 将竖直的列作为anchor进行anchor-based detection，用anchor出instance，再提取关键点
- 输入可以是单摄，也可以是跨摄感知融合的特征

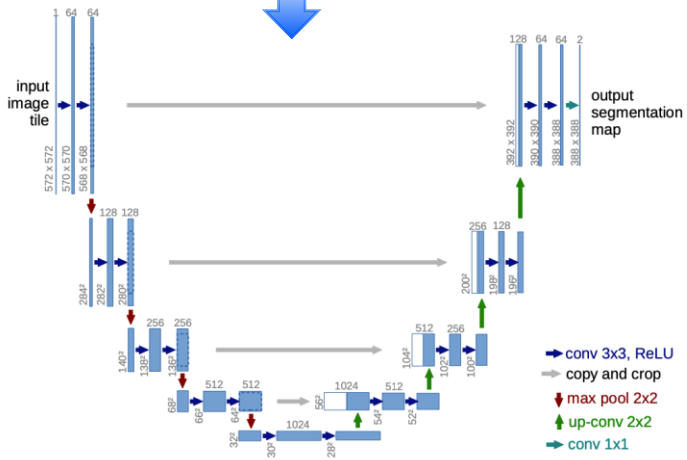


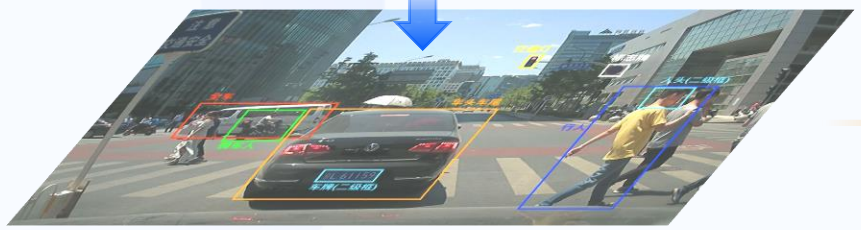
多任务模型的进化

- 全图输入的多任务模型得到图像中的感兴趣区域（ROIs）
- 将这些ROI送入针对不同类型目标的多任务模型中，以得到最终的感知输出



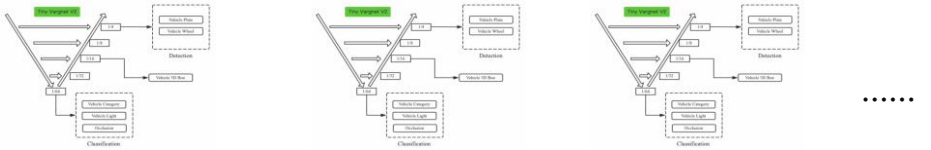
Multi-task model with full frame input





40+ tasks outputs

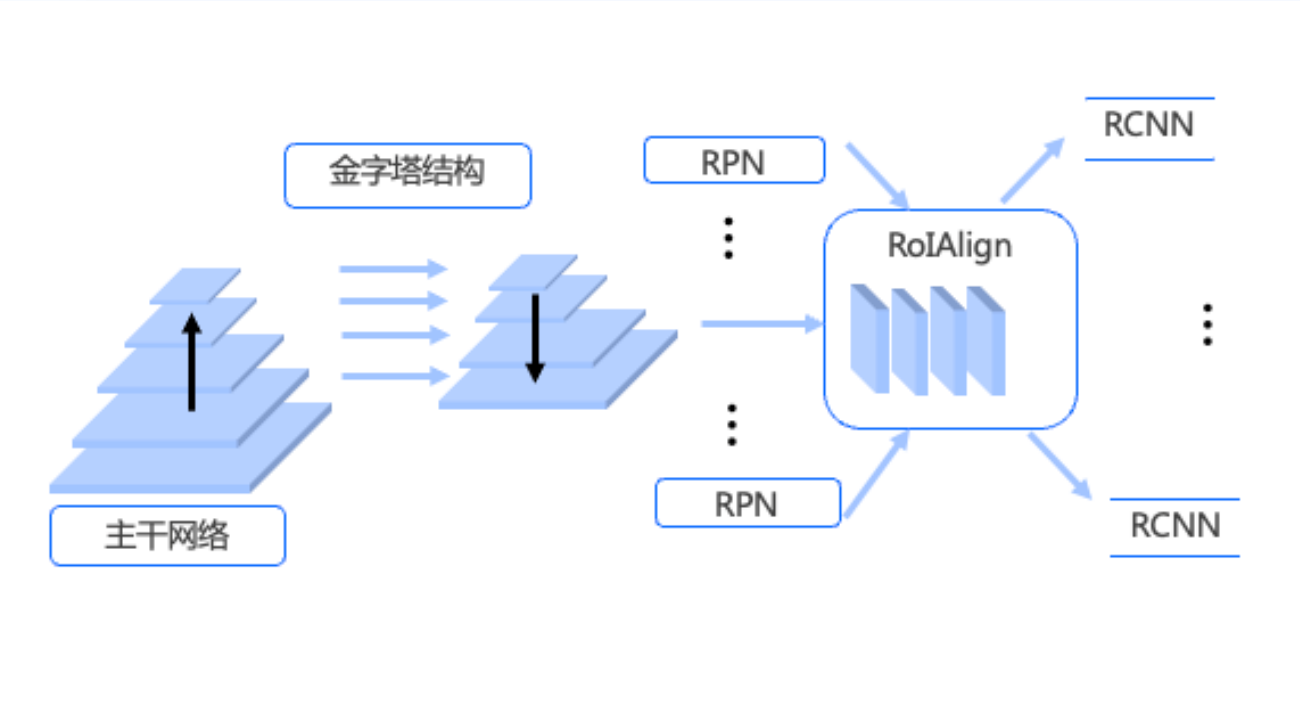
Multi-task models with cropped image ROIs as input



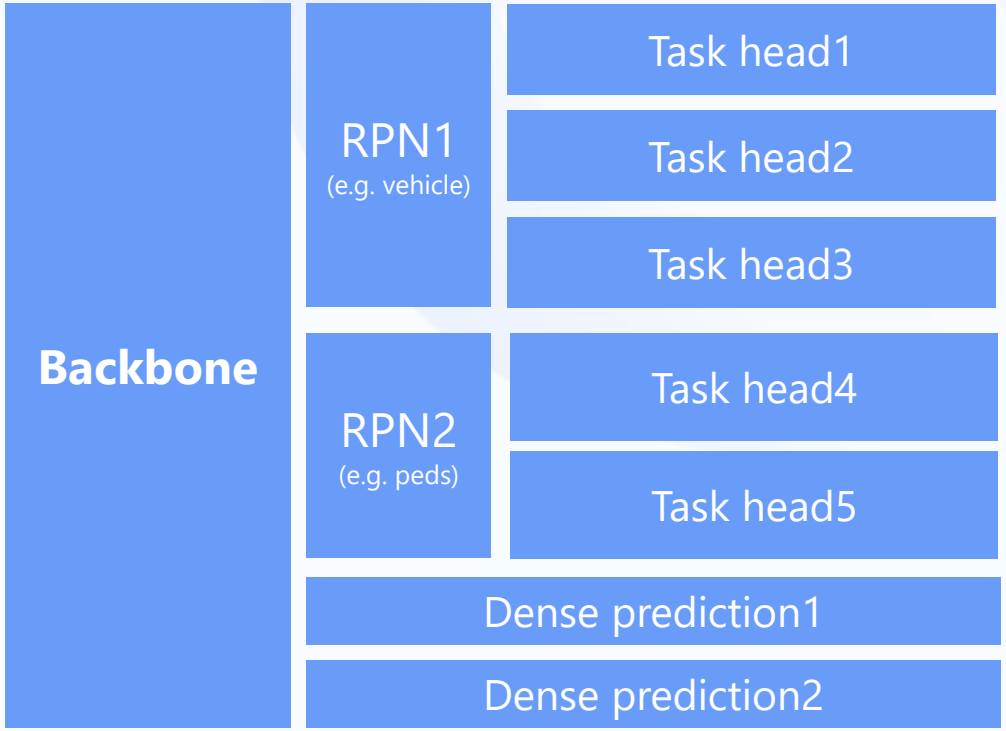
Cropped image ROIs

多任务模型的进化

- Journey3芯片支持高效的roi-align算子
- 与两阶段检测器类似，基于特征ROI的多任务模型抠取特征进行下游任务
- 用进一步共享了特征，降低了运算量



A typical 2-stage detector

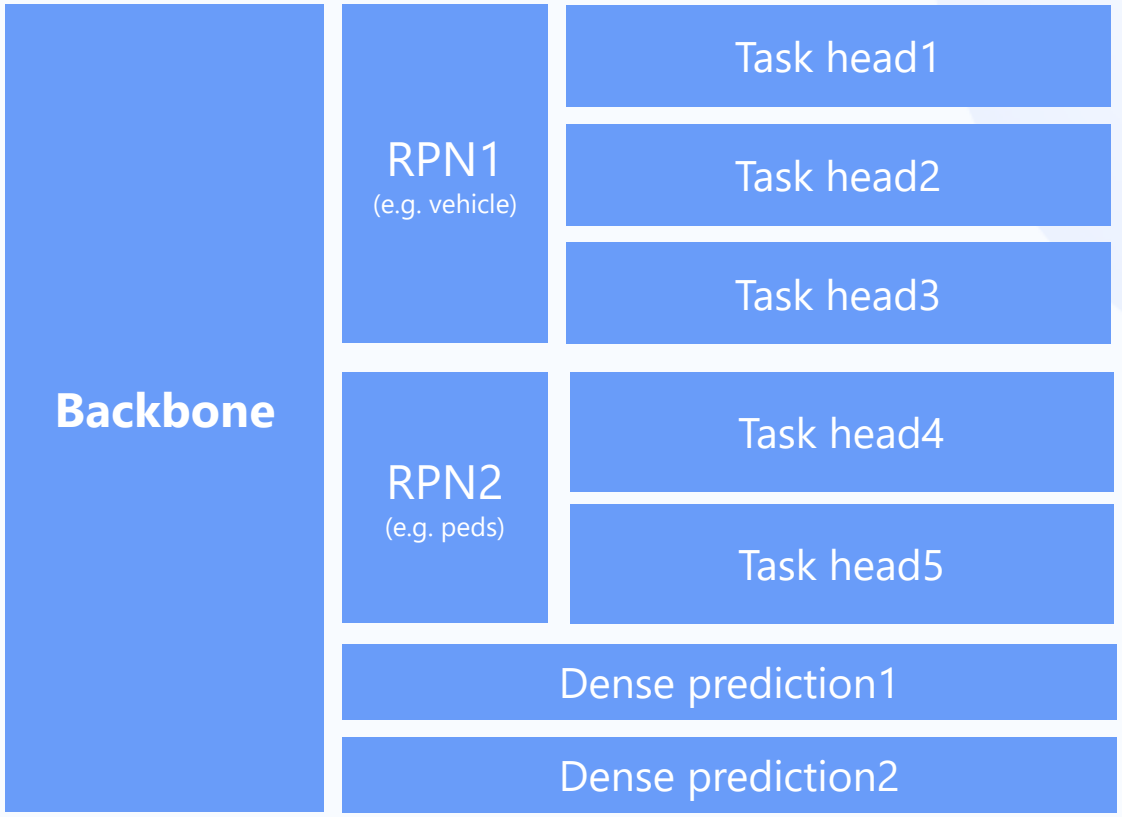


基于特征ROI的多任务模型示例

多任务模型的进化

A few techniques to optimize multi-task models

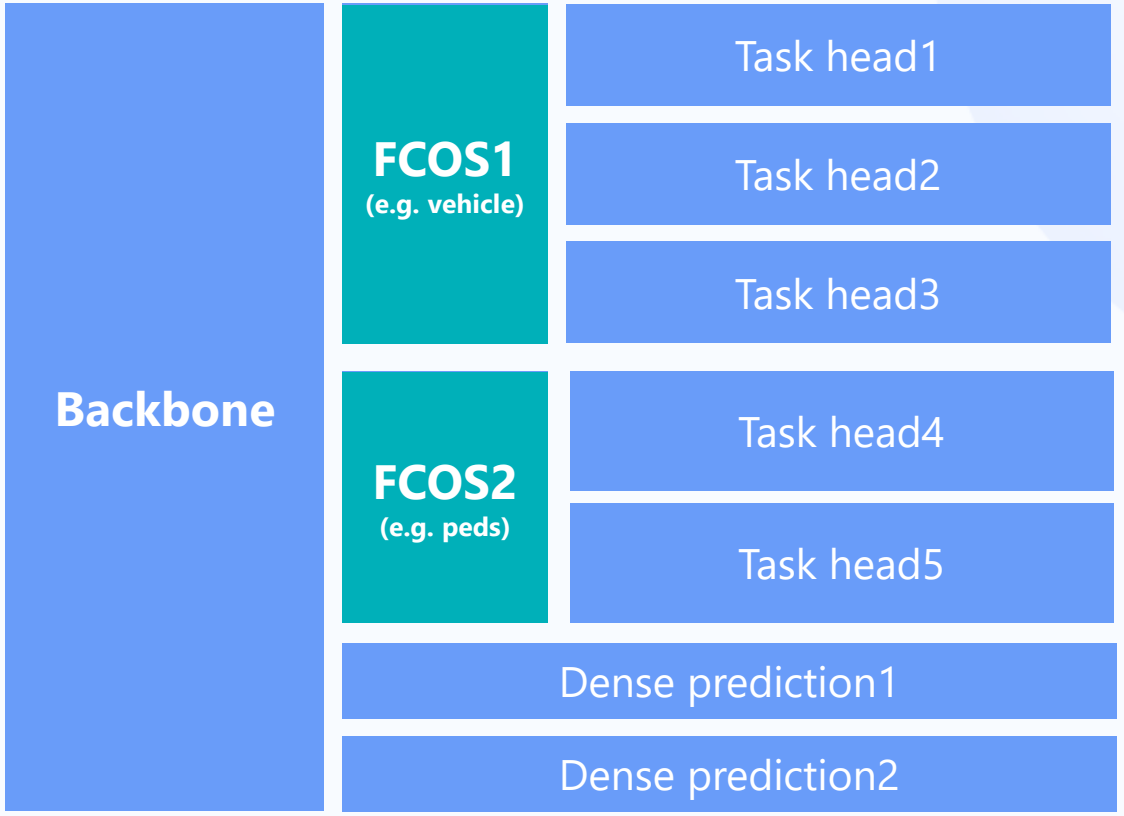
- Better proposal generator（supported by J3）
- Dense prediction -> sparse prediction
- Task head combination



多任务模型的进化

A few techniques to optimize multi-task models

- **Better proposal generator**（supported by J3）
- Dense prediction -> sparse prediction
- Task head combination

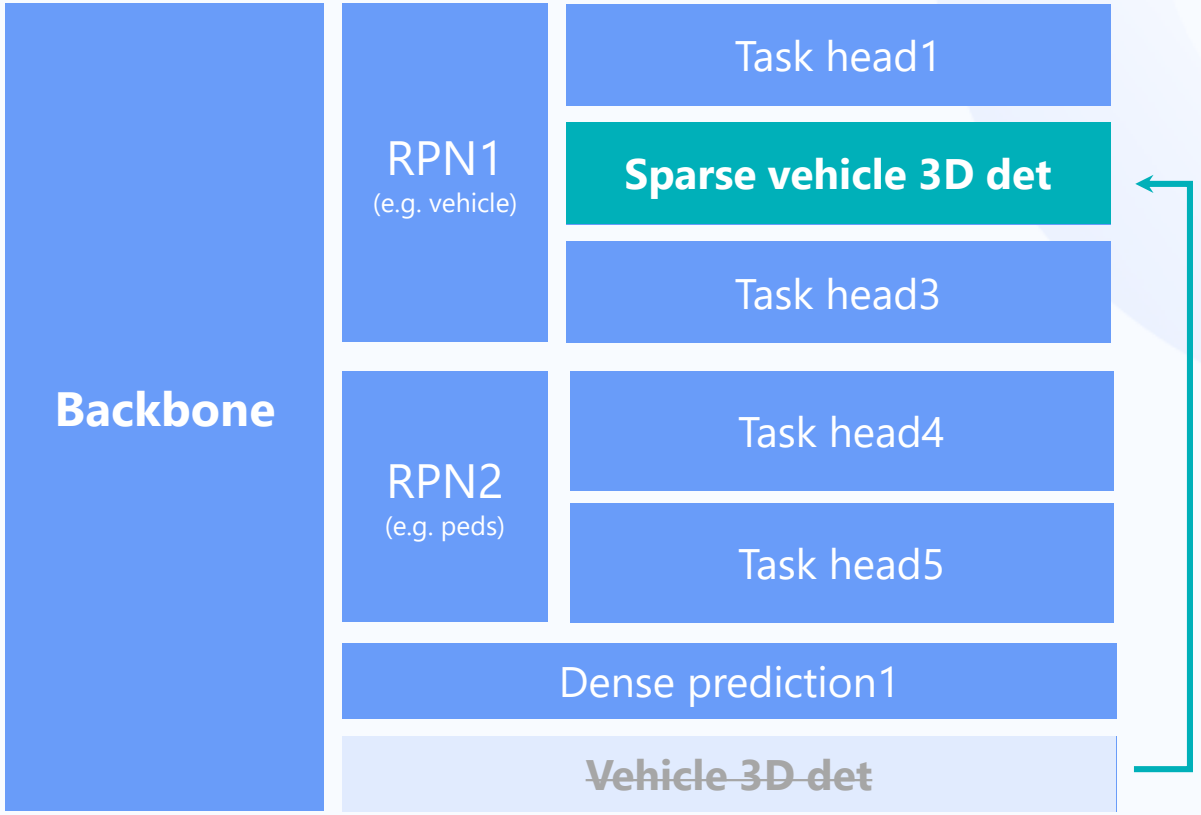


J3支持通过更复杂的检测器获取ROI，再进行roi-align和下游任务

多任务模型的进化

A few techniques to optimize multi-task models

- Better proposal generator（supported by J3）
- **Dense prediction -> sparse prediction**
- Task head combination



多任务模型的进化

A few techniques to optimize multi-task models

- Better proposal generator（supported by J3）
- **Dense prediction -> sparse prediction**
- Task head combination

Sparse 3D

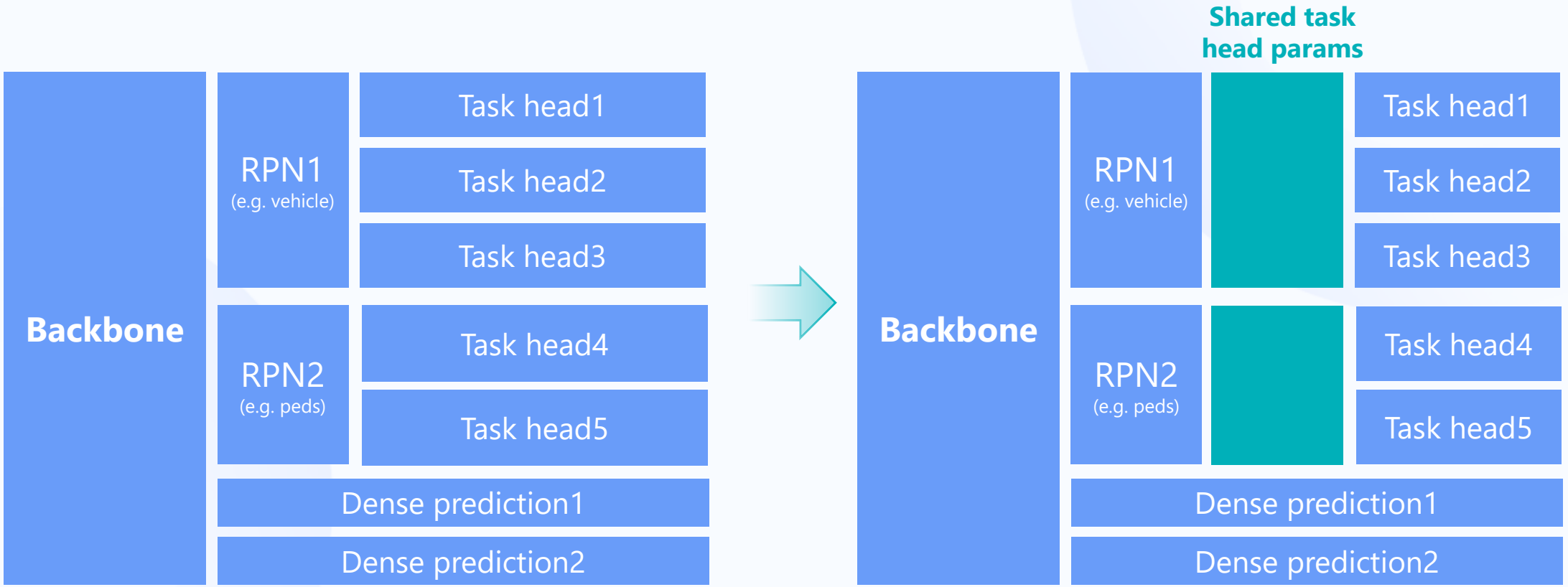
Dense 3D

Sparse3D在降低计算量的同时，提升了远距离的召回

多任务模型的进化

A few techniques to optimize multi-task models

- Better proposal generator（supported by J3）
- Dense prediction -> sparse prediction
- **Task head combination**



相关任务的task head特征共享进一步降低了模型的延迟

多任务模型的进化

基于特征ROI和图片ROI的模型并非简单的替代关系

	Pros	Cons
图片roi	• 研发工作解耦 • 针对单一任务的快速、低成本迭代 • 可以调度skip非关键目标的计算	• 更大的计算量 • 发版繁琐、维护复杂
特征roi	• 特征共享，节约算力 • 端到端的联合优化 • 特征ROI具有更大的感受野	• 训练难度大、耗时长 • Train from scratch代价高

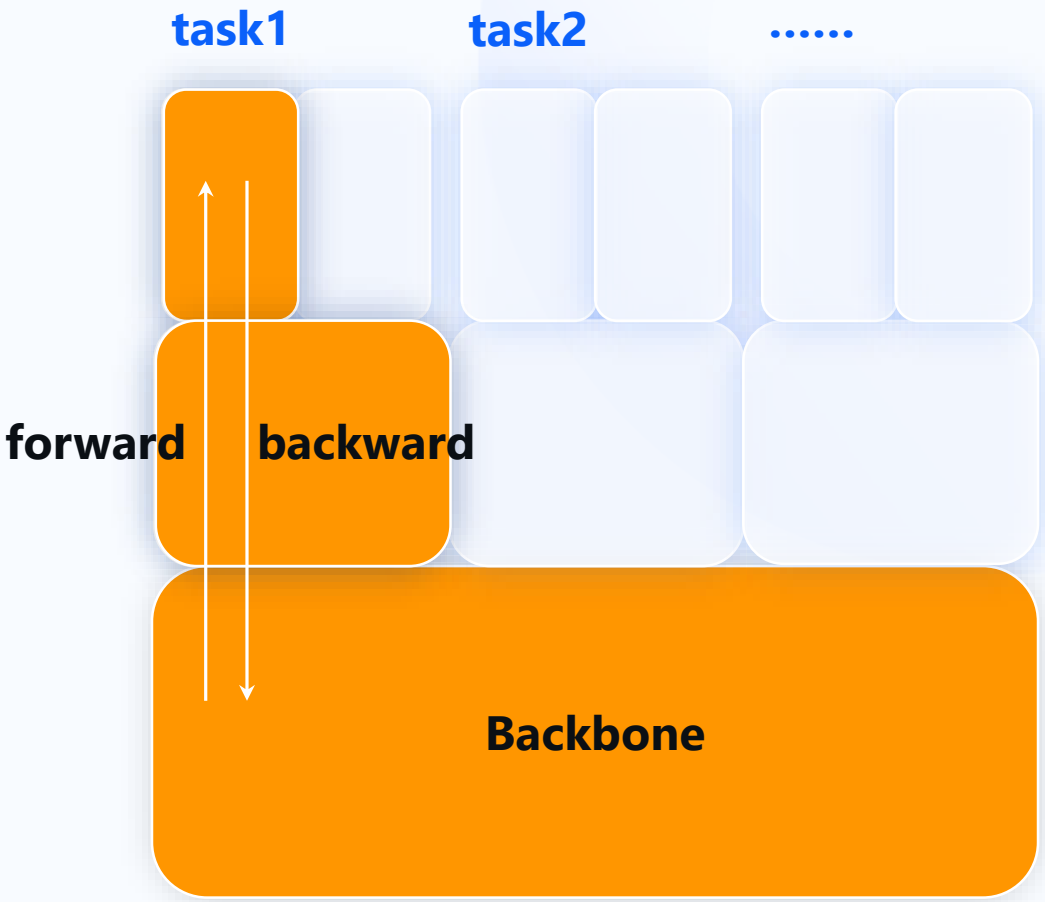
多任务模型的训练

训练

- 各任务单独构建计算图，独立的batch size
- 各任务backward次数便于控制
- 任务采样训练
- 支持图片仅做部分任务的标注

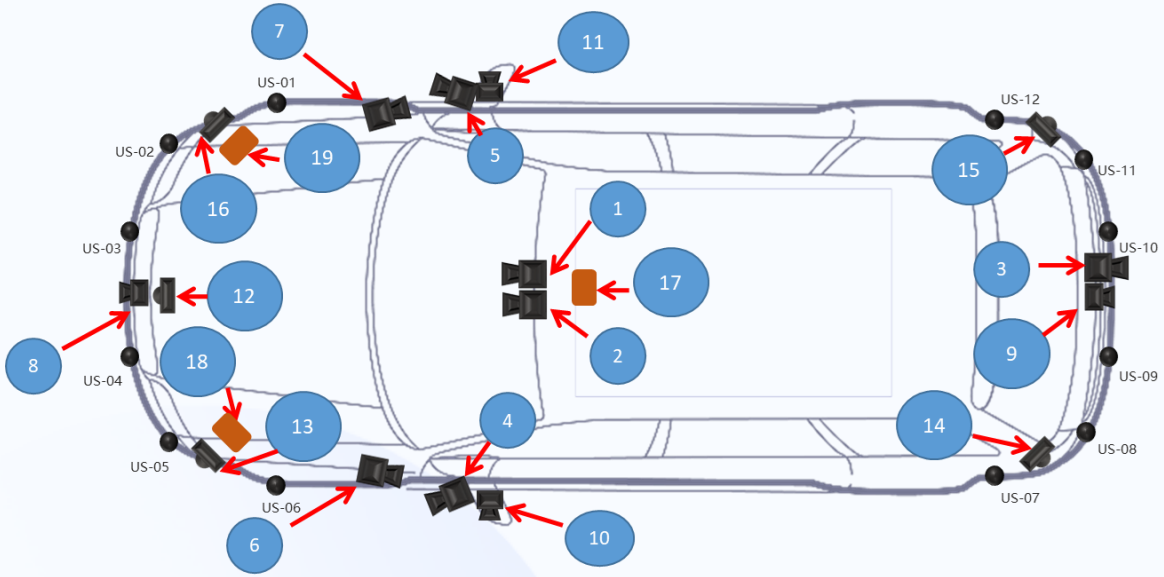
Repeat sample tasks：

- for task in tasks：
  - sample data of task
  - forward & backward



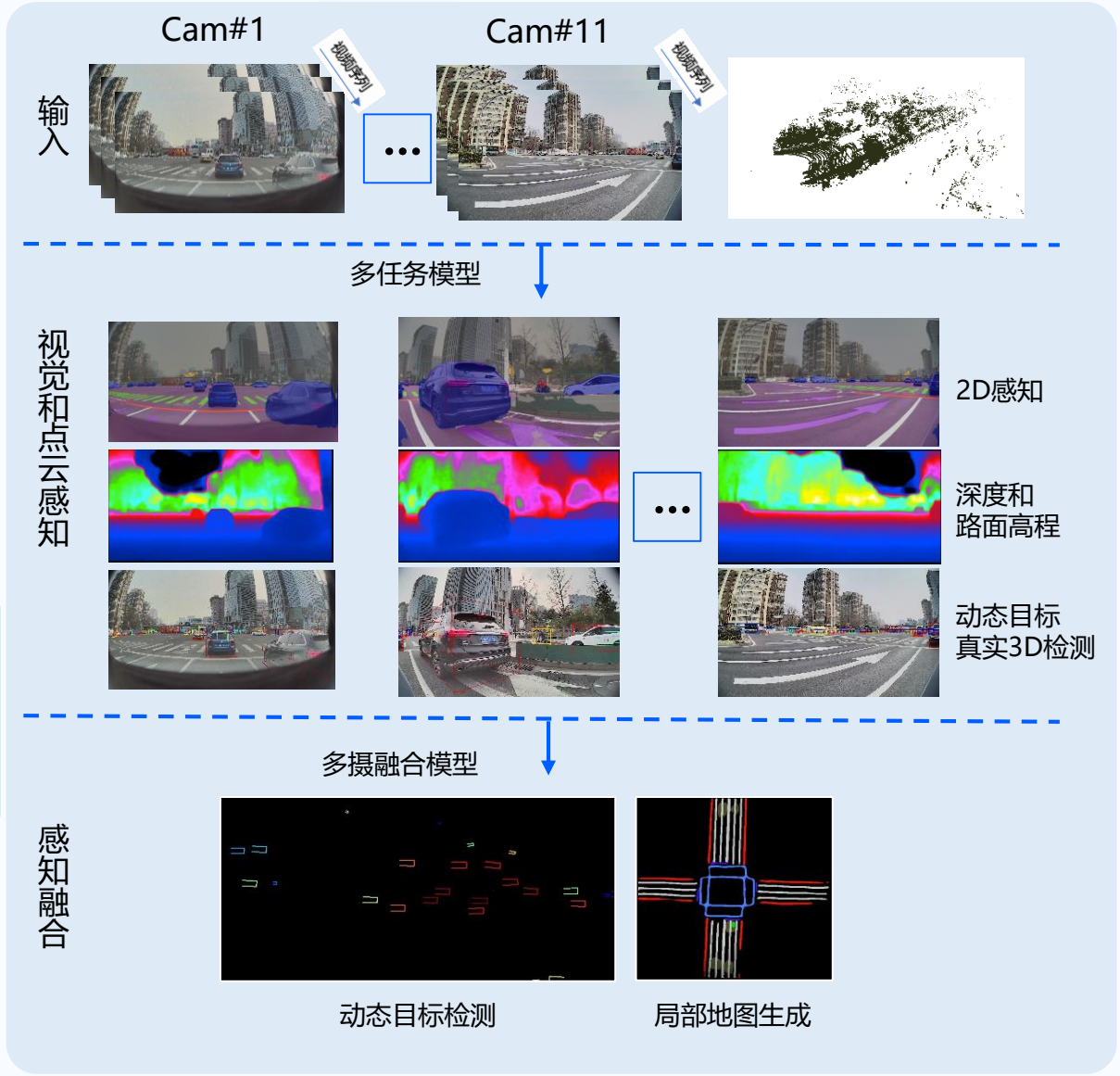
1. 经典单目感知算法方案
2. **BEV感知算法方案**
3. Sparse4D感知算法方案
4. 感知量产方法论

复杂环境的智驾感知：基于BEV的多帧多摄像多模态融合



摄像头编号

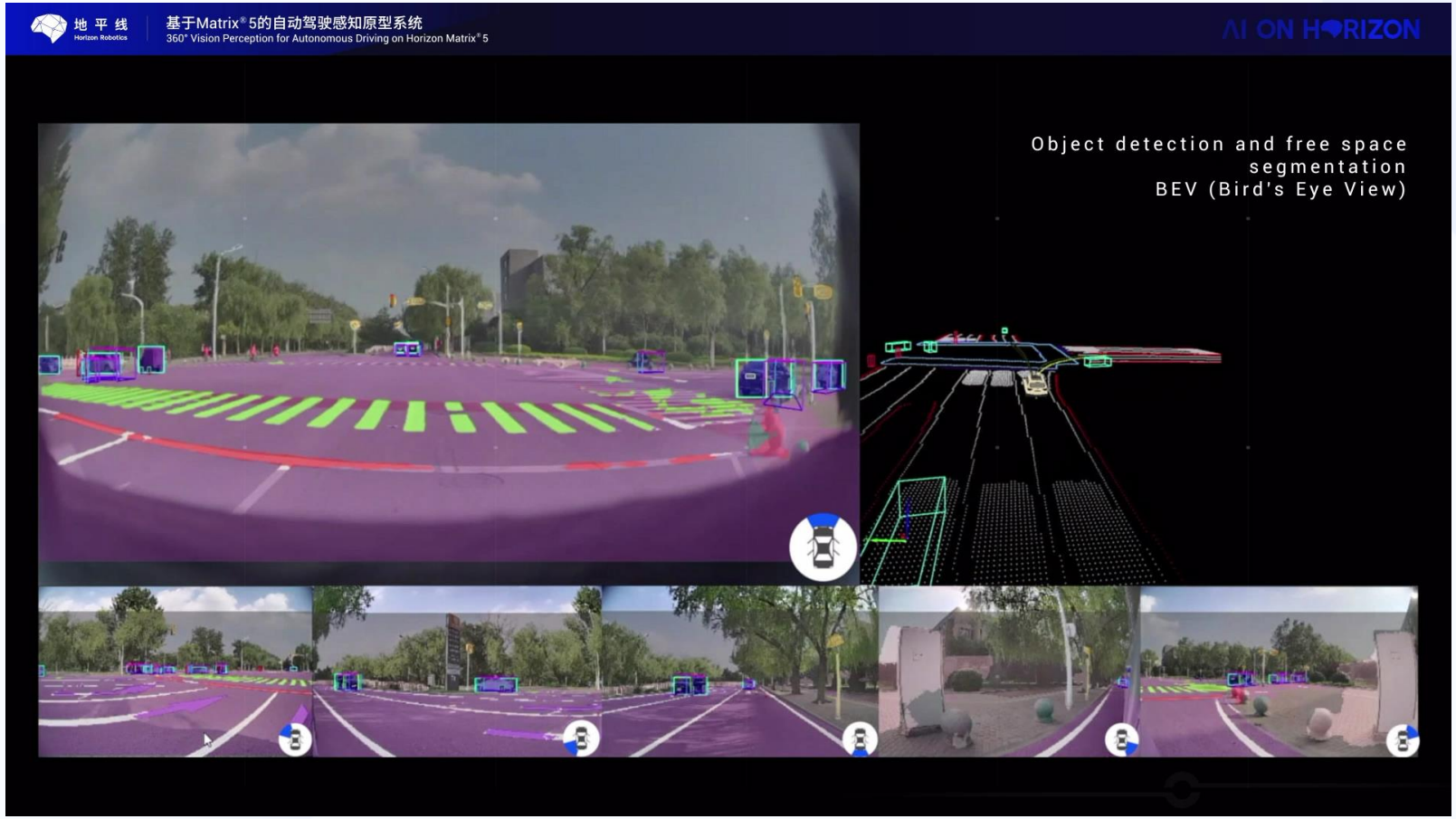
- 前向双焦距摄像头
  - **1. Main Camera**
  - **2. Narrow Camera**
- 后向摄像头
  - **3. Rear Camera**
- 侧前摄像头
  - **4. Left Corner Camera**
  - **5. Right Corner Camera**
- 侧后摄像头
  - **6. Left Wing Camera**
  - **7. Right Wing Camera**
- 泊车摄像头
  - **8. Front Fisheye**
  - **9. Rear Fisheye**
  - **10. Left Fisheye**
  - **11. Right Fisheye**
- 激光雷达(单Lidar方案时，安装位置为17；双Lidar方案时，安装位置为18/19;)
  - **17. Front Main Lidar**
  - **18. Front Left Lidar**
  - **19. Front Right Lidar**



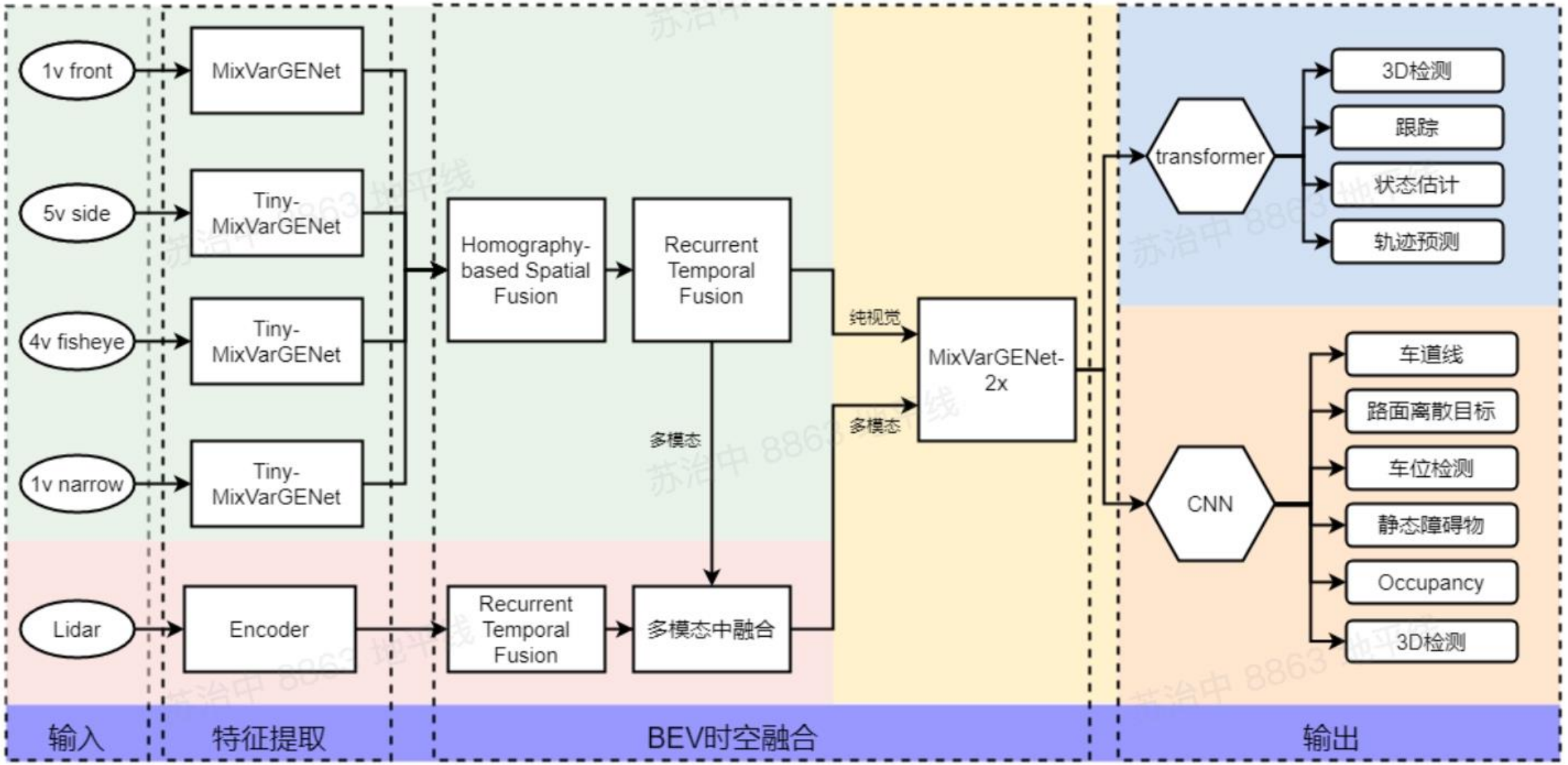
BEV感知：时空融合架构带来新一代感知“底座”

输入360度环视的图像，对自车周围的视觉信息进行融合，直接输出车辆坐标系的静/动态感知结果

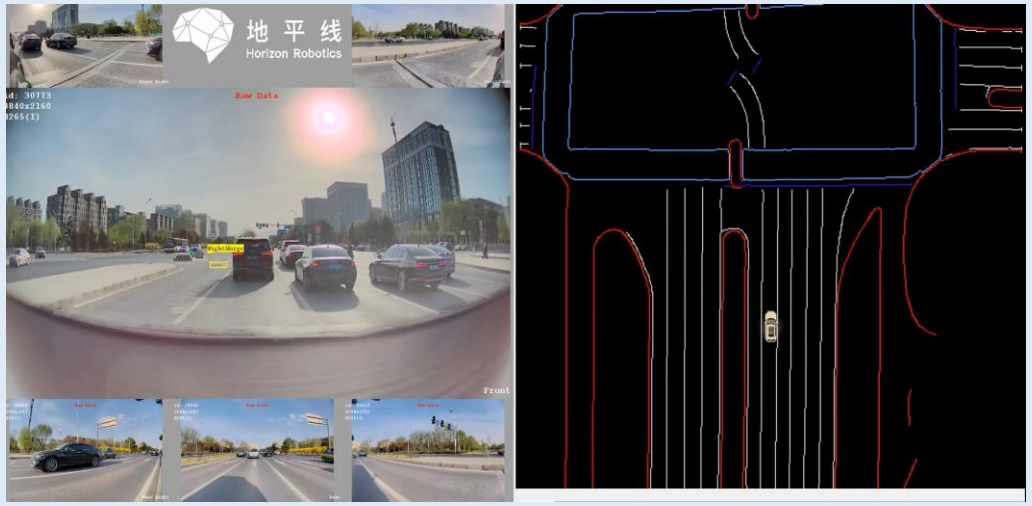
- 神经网络原生直接输出全要素360度环境感知结果
- 端上实时环境感知建图，无需依赖高精地图
- 多摄像头端到端感知、融合、环境建模、轨迹预测
- 复杂路况处理能力（十字路口、弯道、环岛等）



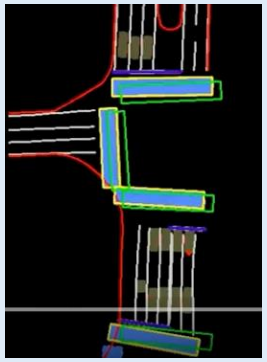
BEV感知：算法方案


1v front
5v side
4v fisheye
1v narrow
Lidar
MixVarGENet
Tiny-MixVarGENet
Tiny-MixVarGENet
Tiny-MixVarGENet
Encoder
Homography-based Spatial Fusion
Recurrent Temporal Fusion
Recurrent Temporal Fusion
多模态中融合
多模态
纯视觉
多模态
MixVarGENet-2x
transformer
3D检测
跟踪
状态估计
轨迹预测
CNN
车道线
路面离散目标
车位检测
静态障碍物
Occupancy
3D检测
输入
特征提取
BEV时空融合
输出

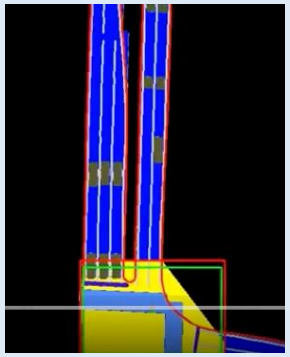
BEV下的动/静态要素



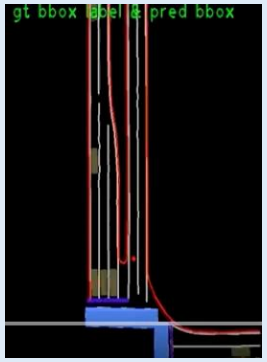
Lanes & Curbs Segmentation



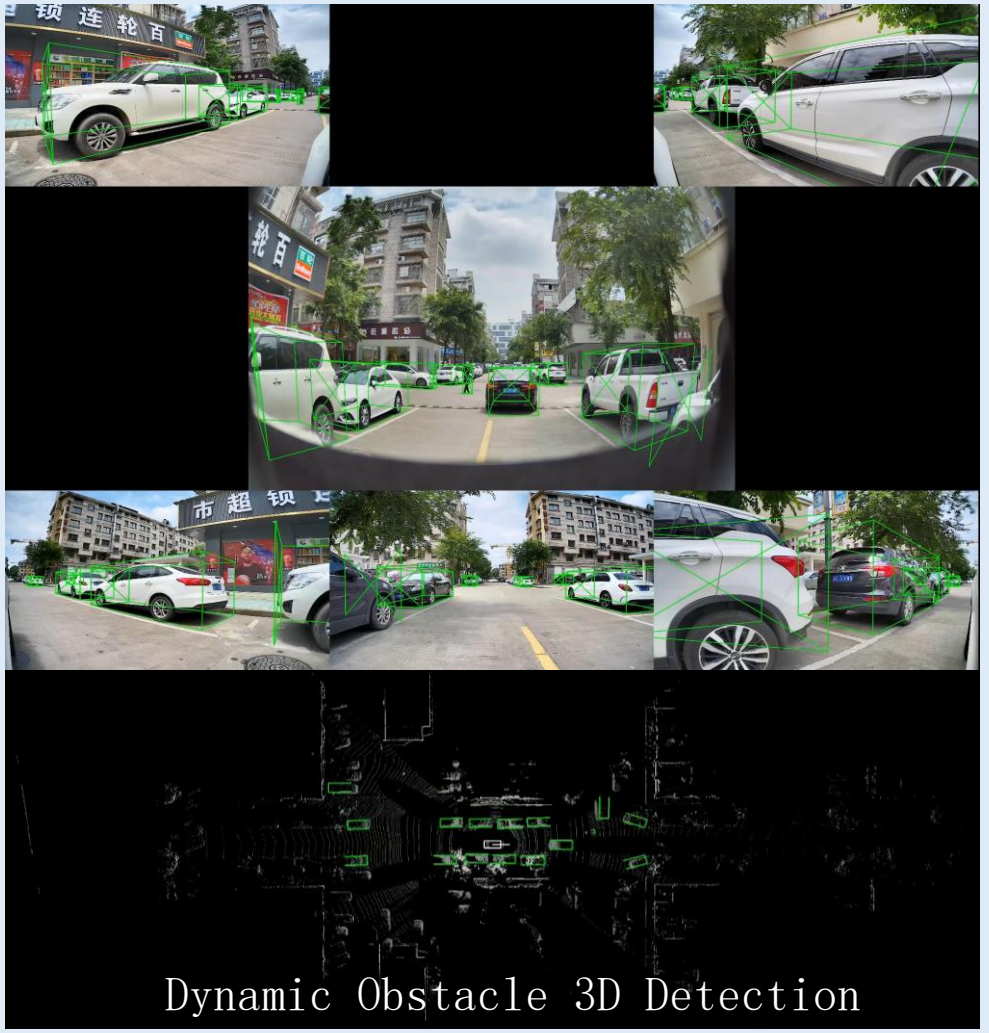
Crosswalks



Junctions



Road Signs

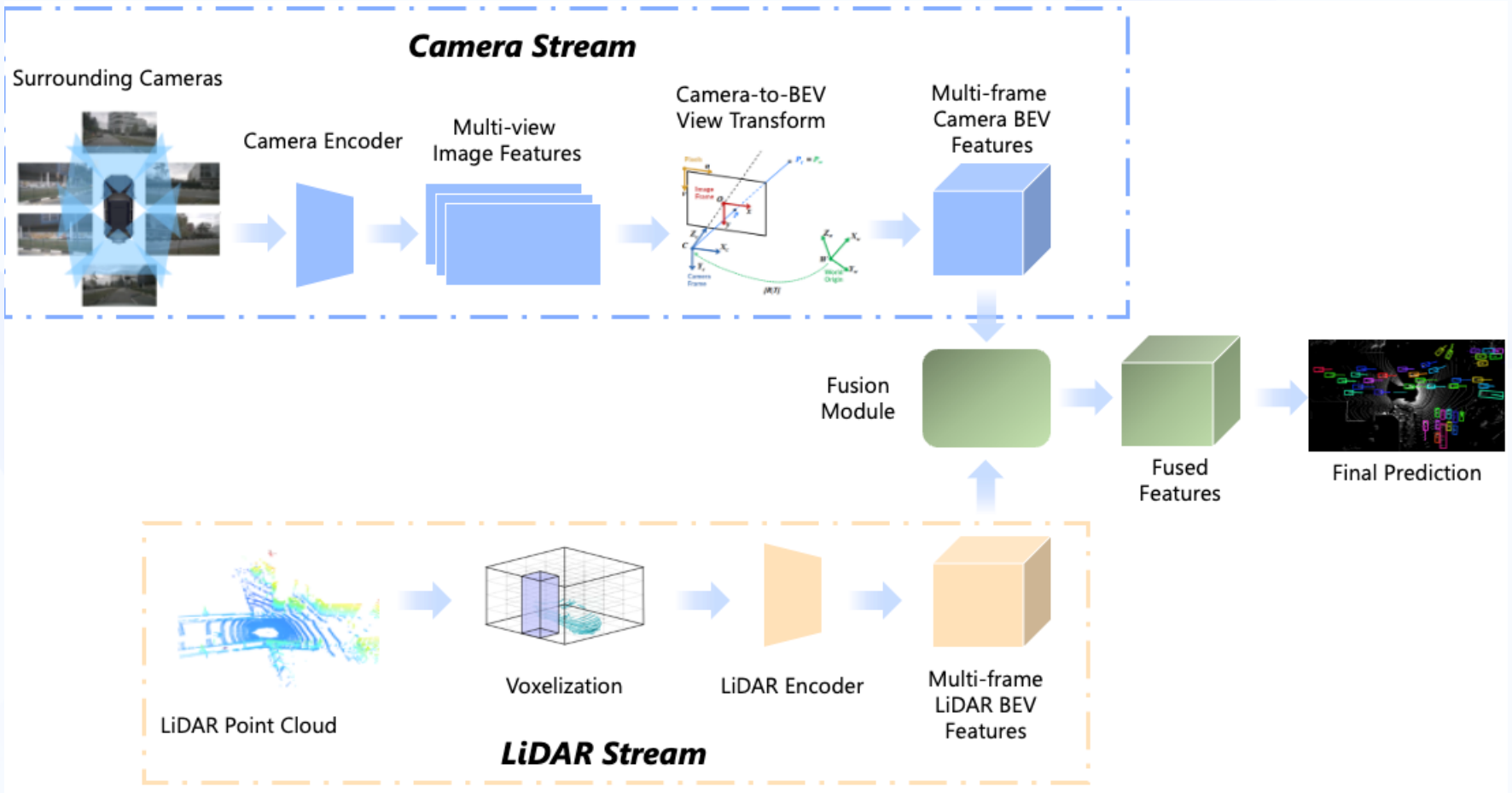


Error reduction up to 40%

BEV感知：更高效、简洁的多模态融合

将视觉BEV特征与Lidar点云特征融合，提升性能的同时减少逻辑计算代价

- 通过视觉BEV与Lidar特征融合，达到了有Lidar区域效果好于Lidar，无Lidar区域效果好于视觉的效果
- 在融合基础上，可以进行目标检测，freespace检测等多种任务



BEV感知性能

- BEV下的动静态感知相比单摄有显著提升
- 在征程5上，BEV多任务模型帧率可达100+FPS

	评测指标-分视角			
		dxp@recall0.9	dyp@recall0.9	drot@recall0.9
real3d	AP_3D:0.4077 AP:0.9477	0.0519	0.1477	5.4741
bev3d	AP_3D:0.6336 , AP:0.9496	0.0311	0.0974	3.8315

指标示例：BEV3D vs real3D

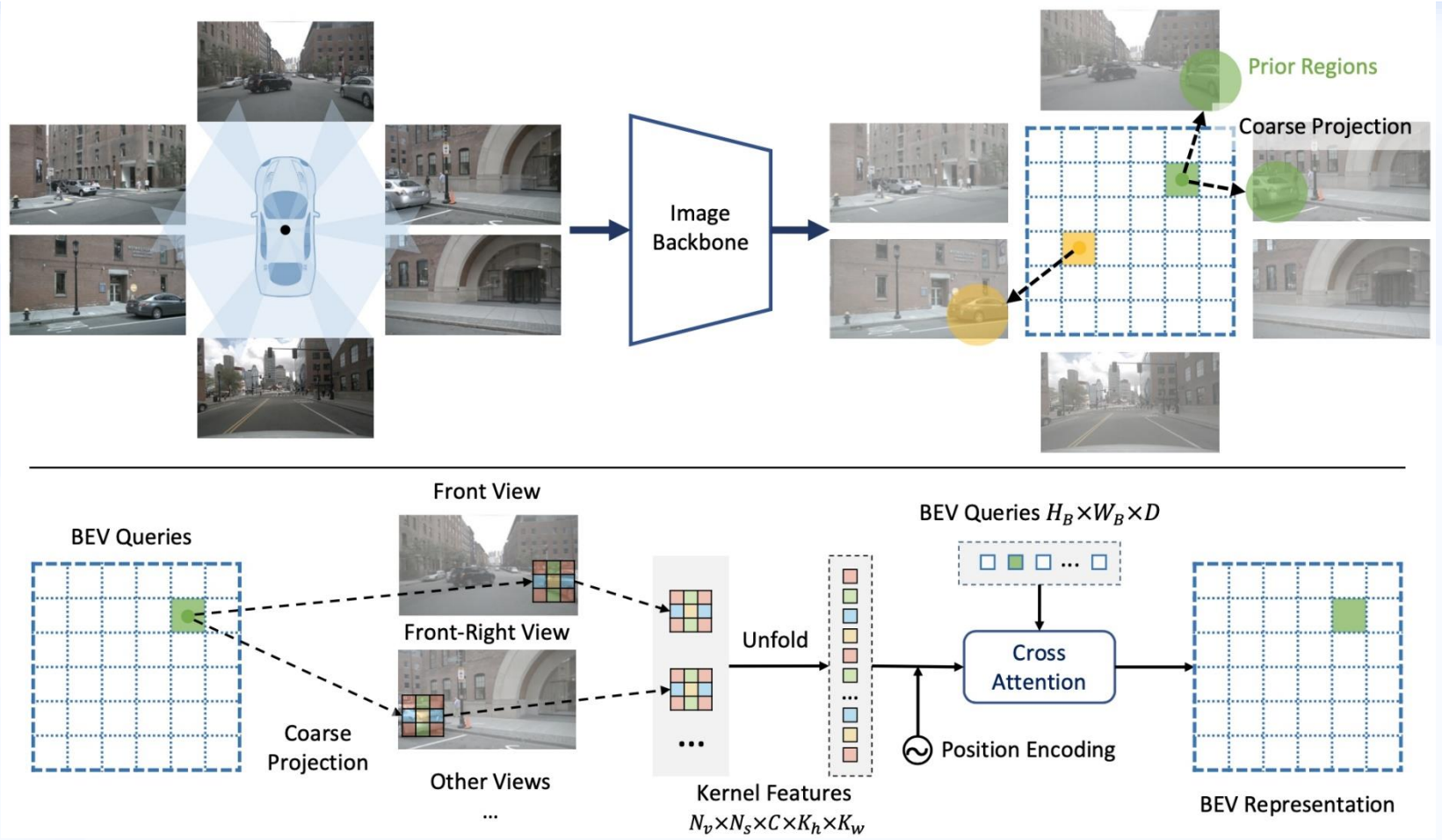
Bev车道线交叉点							
AP	Precision	Recall	Accuracy	dx	dy	dxy	dxyp
0.5645	0.6794	0.5654	0.9886	0.5178	0.2132	0.6116	0.0285

单摄车道线交叉点							
AP	Precision	Recall	Accuracy	dx	dy	dxy	dxyp
-	0.6285	0.1022	-	4.35	0.28	-	-

指标示例：车道线交叉点

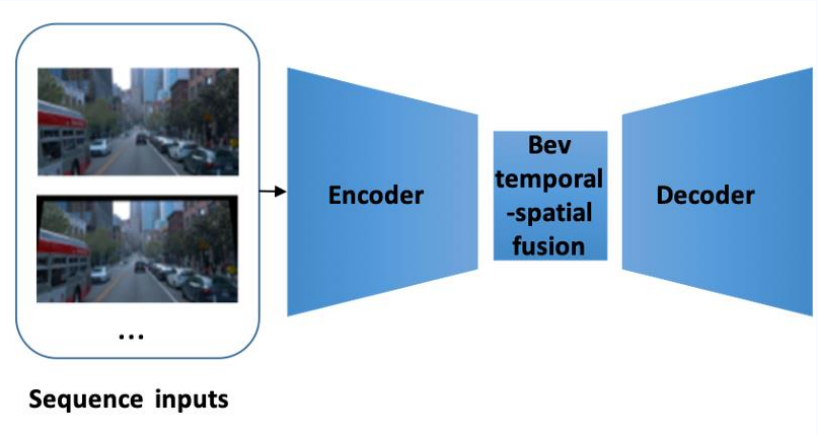
BEV感知的发展趋势

- 基于transformer的BEV感知相比warping有更好的泛化性，相比LSS对嵌入式平台更友好，且有更好的性能
- 我们提出：Efficient and Robust 2D-to-BEV Representation Learning via Geometry-guided Kernel Transformer
- https://arxiv.org/abs/2206.04584

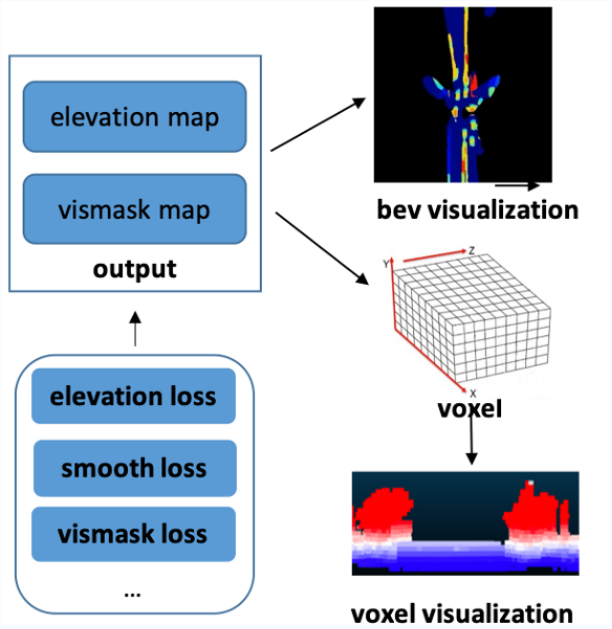


3D Occupancy强补充：利用几何解决通用场景问题

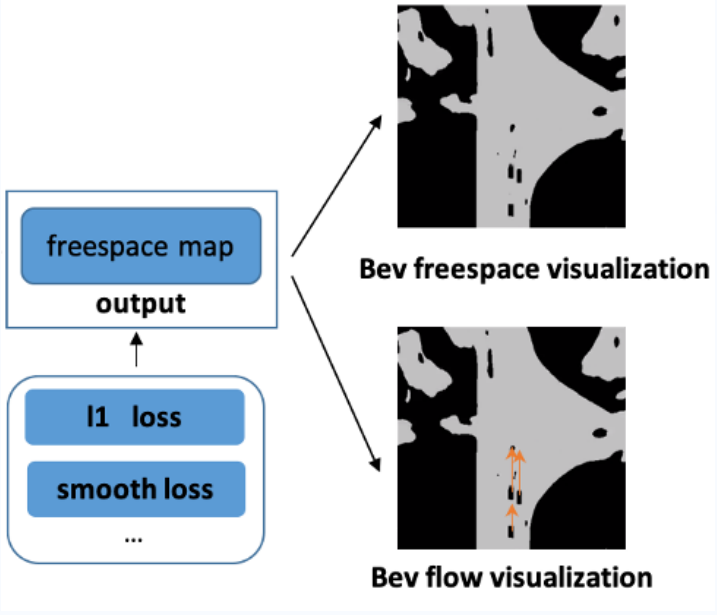
- 自动驾驶场景的障碍物类型可分为两类：主要的道路使用者和其它。对于广泛的一般障碍物，用semantic去解决是不现实的，而基于几何的occupancy会是更好的解决范式
- 基于前面提到的多模态融合的BEV架构，可进行 3D Occupancy的算法方案升级
- 与Tesla的方案不同，可以结构化分解为BEV 2D freespace + BEV elevation



BEV stage1 + stage2

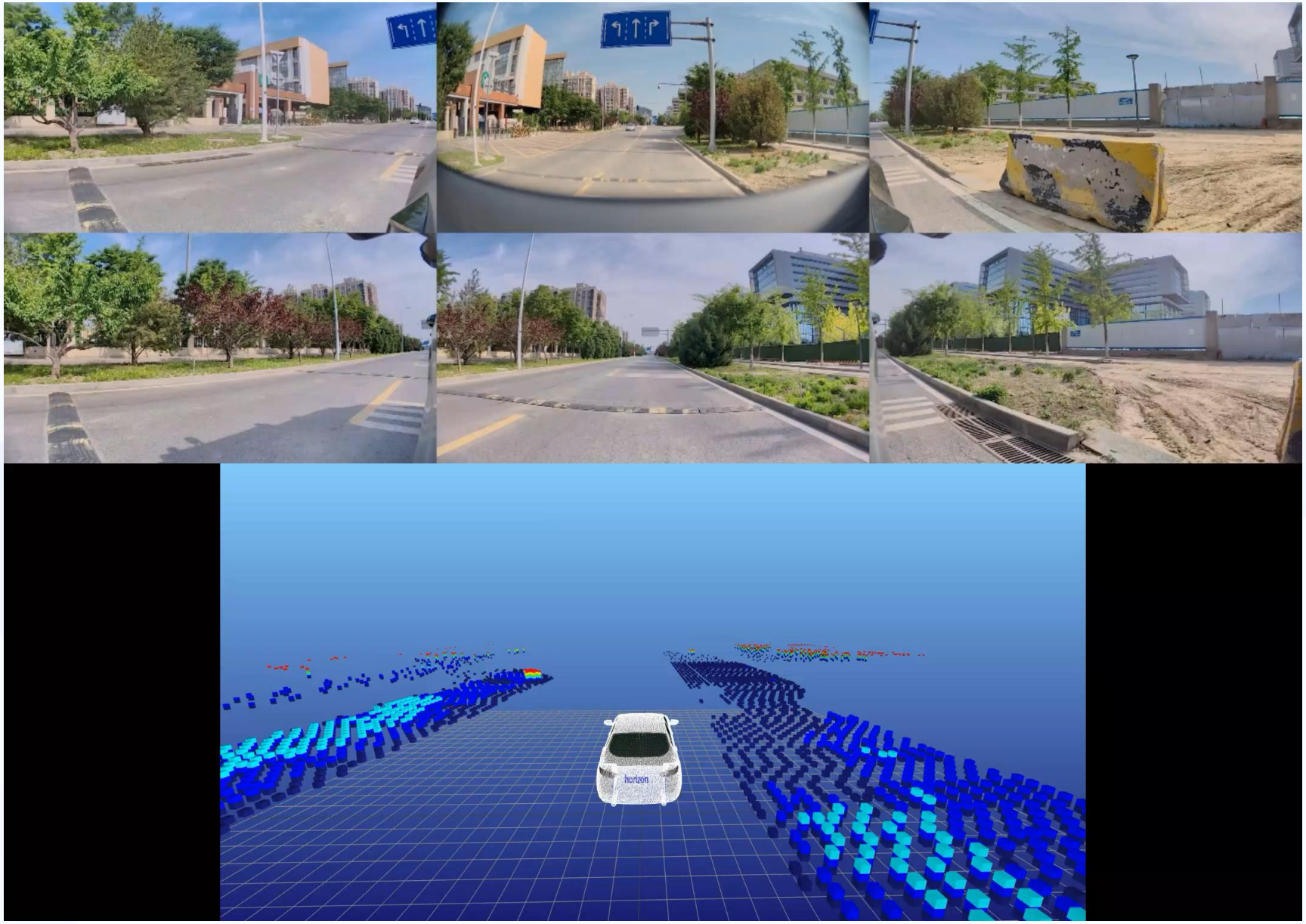


Elevation head



freespace head

3D Occupancy Demo



3D Occupancy的挑战

- 算力和带宽：Tesla的Occupancy方案输出一个4维tensor，XYZ和feature length，对计算平台挑战极大
- 稀疏性：以地面的一般障碍物为例，自动驾驶场景中地面出现一般障碍物的情况极少，模型需要大量数据才能学习
- 纯视觉真值生成：超大规模的occupancy真值不能依赖于lidar产生，而纯视觉生成occupancy真值的精度还有待提升

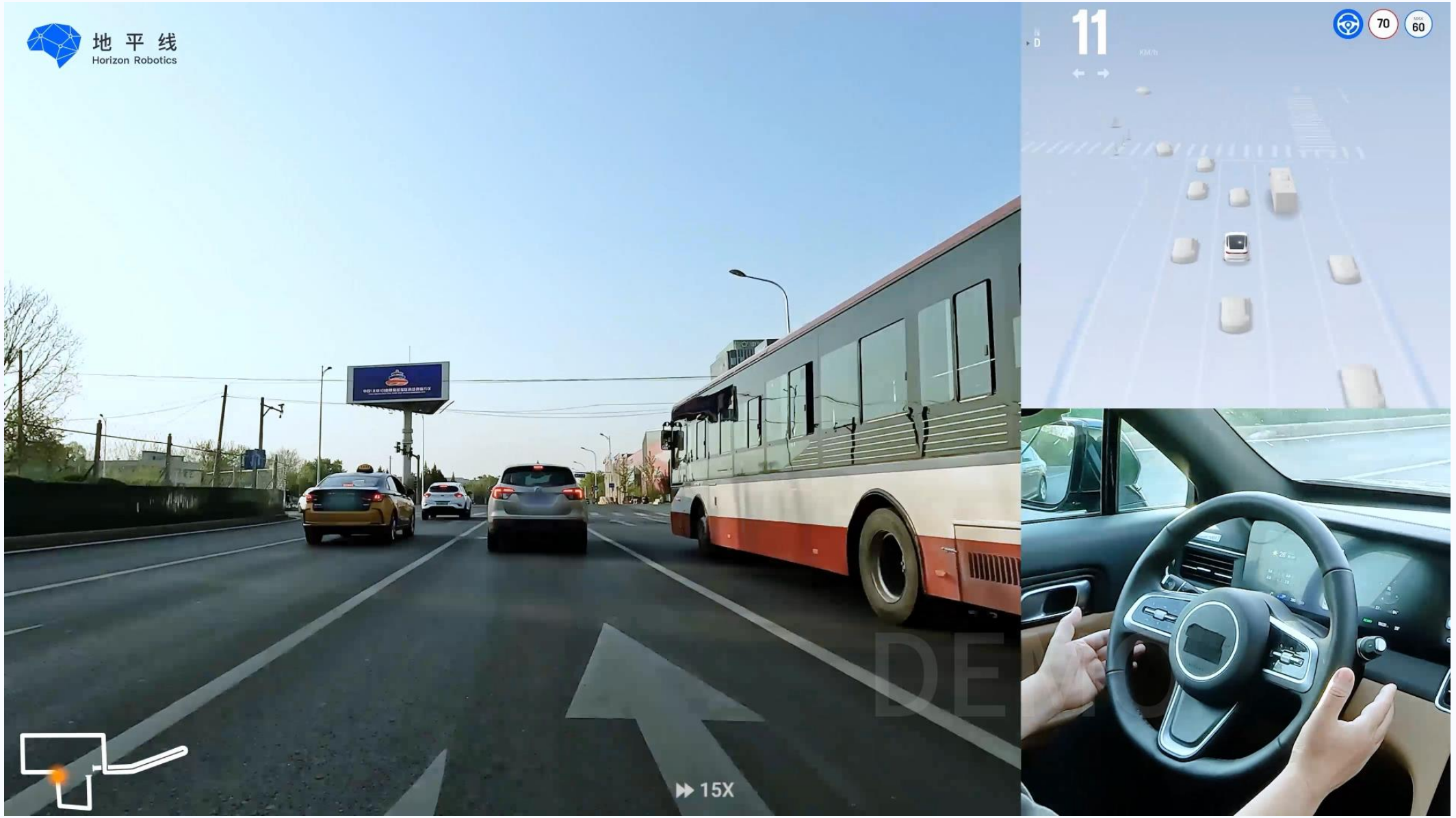
Tesla：14亿图片训练

Occupancy Networks
have some nice properties

- Volumetric occupancy
- Multi-camera & video context
- Dynamic occupancy
- Persistent through occlusions
- Resolution where it matters
- Efficient memory and compute
- Runs in ~10 ms

Tesla occupancy的公开信息，动态栅格，query able分辨率，10ms延迟

基于征程®5的城区复杂场景自动驾驶演示



1. 经典单目感知算法方案
2. BEV感知算法方案
3. **Sparse4D感知算法方案**
4. 感知量产方法论

Sparse4D：高性能高效率的长时序纯稀疏融合感知架构

接棒“BEV+Transformer”的下一代架构，带来感知新突破

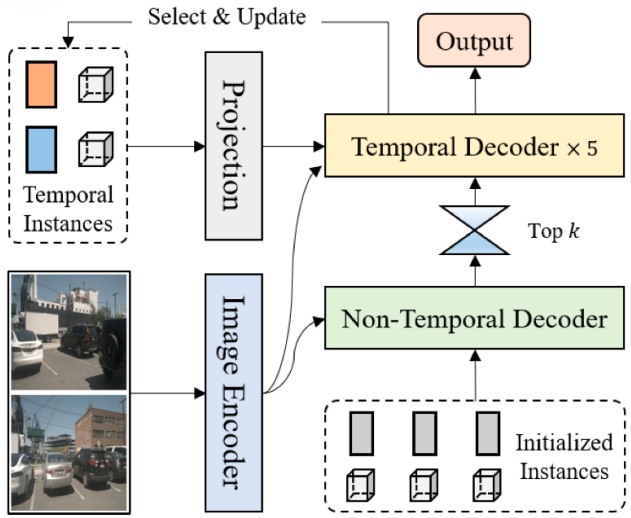
Dense BEV的不足

- BEV 特征空间十分稀疏，存在大量无用区域
- 感知范围、感知精度、计算效率难平衡
- 实用中通常还需要前视模型补充远距离感知，导致模型系统很难真正端到端
- BEV 特征由于压缩了高度方向，所以通常较难处理高度较高位置的一些任务，如红绿灯等

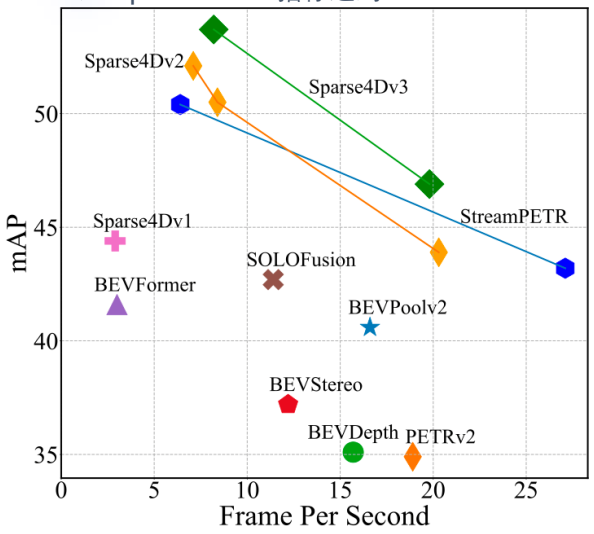
Sparse4D稀疏融合感知架构

- 基于Transformer，支持时序的多摄、多传感器融合，各视角传感器可插拔、拓展
- 无需显式的视角变换，稀疏方法消除了将图像空间转换为3D矢量空间的模块
- 极限稀疏实例化，更高效时序融合和存储
- 检测头中的恒定计算负载，与感知距离和图像分辨率无关

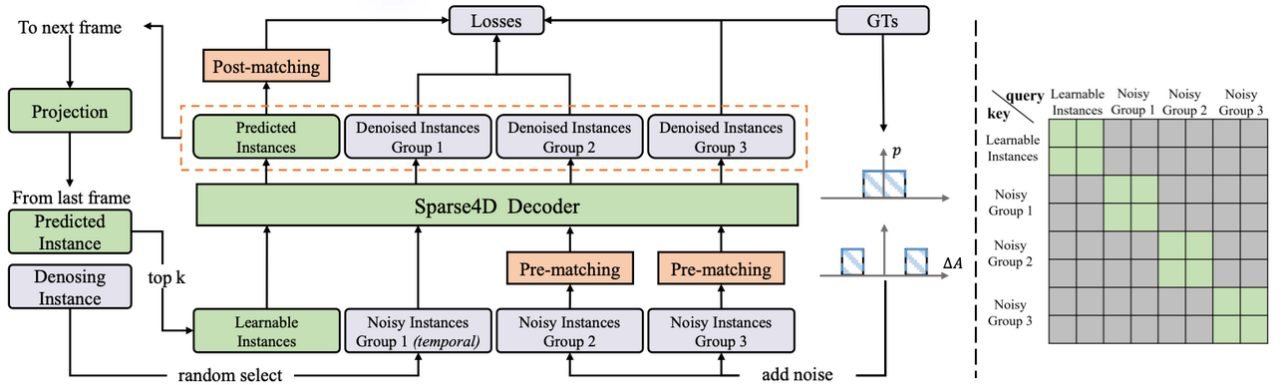
▼ Sparse4D系列架构概览



▼ Sparse4D V3指标达到SOTA



地平线Sparse4D V3详细架构 ▼



(a) Pipeline of training with temporal denosing

(b) Attention mask of self-attention

Sparse4D稀疏融合感知架构：指标优于主流Transformer BEV算法

公开数据集实验验证

nuScenes detection task

Leaderboard

在nuScenes detection 3D榜单达到SOTA效果，超越了BEVFormer v2、Stream PETR等一众新方法

优于Transformer BEV算法和其他稀疏融合算法

业务集验证

显著优于Mono 3D和BEV 3D

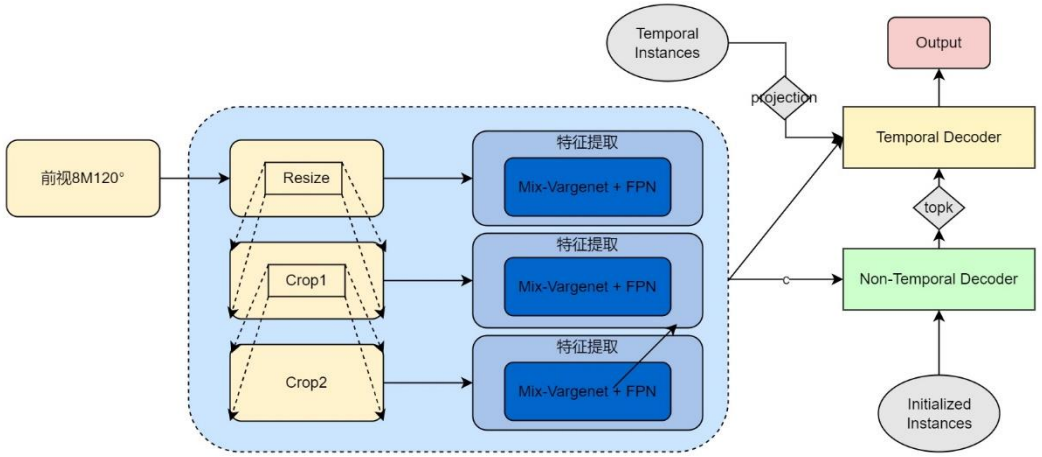
方案对比	前视范围内，距离：0-120m	
	AP_3D	AR_3D
Mono3D (1V)	0.843	0.4872
BEV3D (7V)	0.8824	0.5937
Sparse4D (1V)	**0.9133**	**0.6312**

Sparse4D单V前视感知平均精准度与平均召回率高于7V BEV 3D感知，也大幅领先于单V mono 3D感知

Sparse4D稀疏融合感知架构：具备更远、更精准、更稳定的感知能力

稀疏多尺度融合技术

- 通过Sparse4D提供的高效、稀疏的特征融合能力，在每个视角下，可以通过构建多个不同尺度的虚拟相机，并进行稀疏特征融合的方式，达成**远距离精准感知的能力**

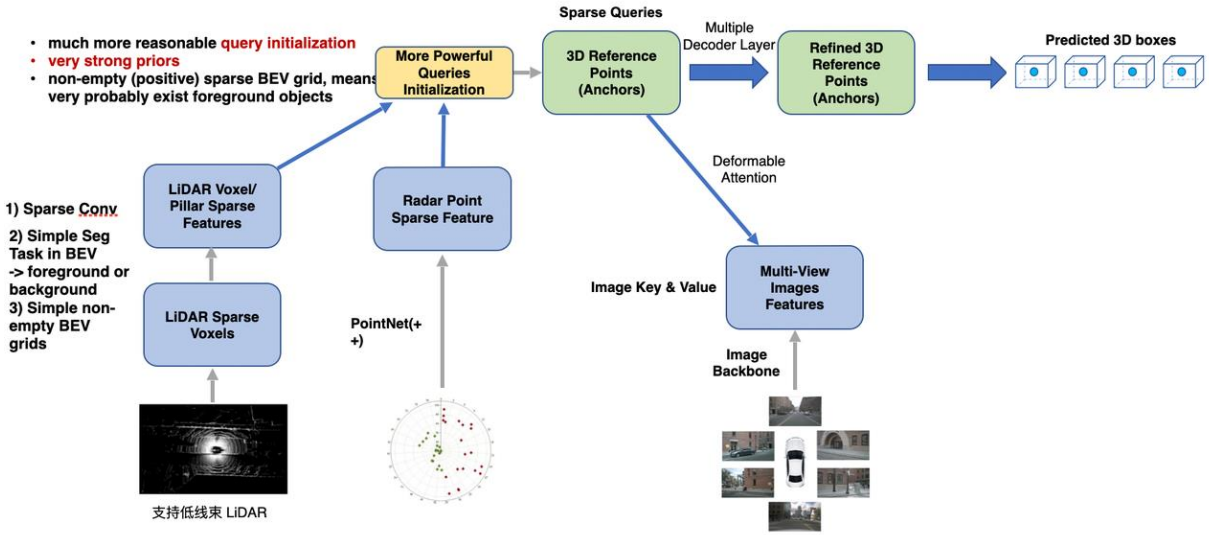


- 稀疏特征融合与传统Mono3D的全距离（0~300m）段指标对比，基于Sparse4D的特征融合指标提升显著

	AP_3D	AR_3D
Mono3D三尺度后融合	0.6558	0.3114
Sparse4D三尺度稀疏特征融合	**0.7894**	**0.3759**

稀疏多模态融合技术

- Sparse4D提供了非常自然的多模态（视觉、Lidar点云、Radar点云）融合框架
- 用(轻量化的)稀疏的LiDAR/Radar点云特征初始化3D Queries，通过强先验**提供更稳定的感知结果**



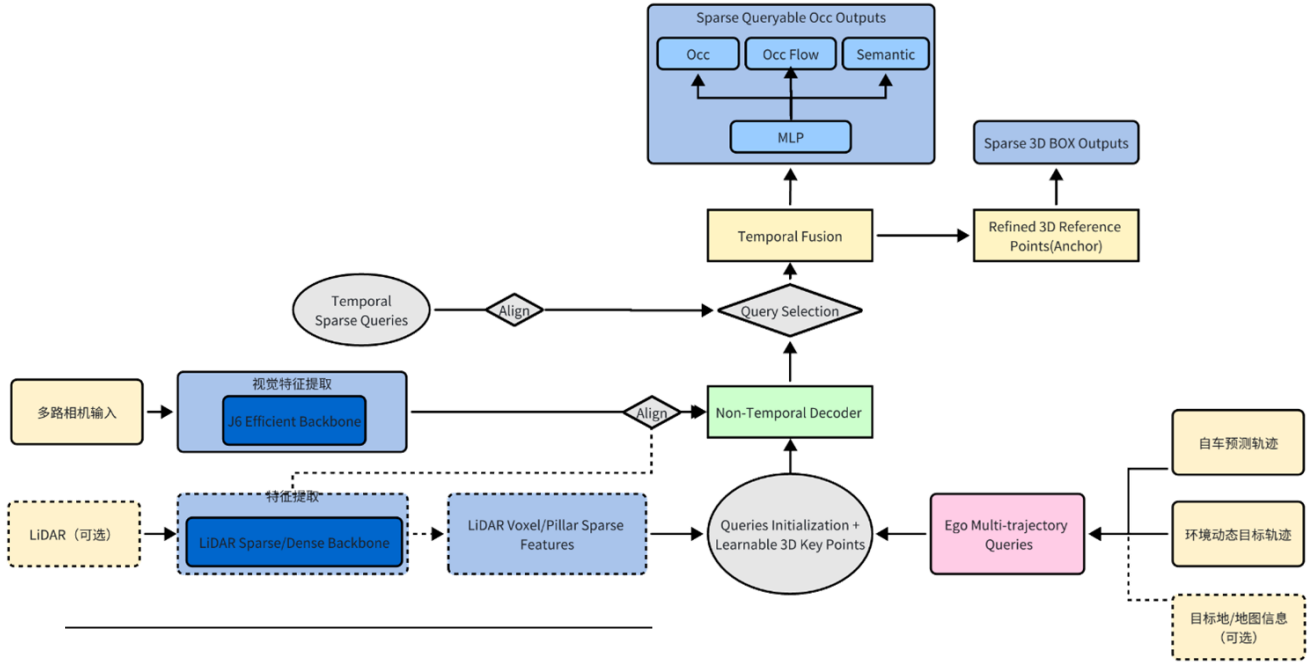
基于Sparse4D的稀疏3D Occupancy：解决3D Occupancy高算力高带宽难落地的行业难题

传统稠密3D Occupancy的不足

3D Occupancy是当前应对一般障碍物的最佳实践，但传统的稠密3D Occupancy输出全范围稠密的4维栅格（xyz，feature length），计算和带宽占用过大

稀疏3D Occupancy

- 结合多条轨迹确定感兴趣区域，保障高召回
- Sparse query，节约算力和带宽
- Queryable输出，精度和分辨率不打折扣



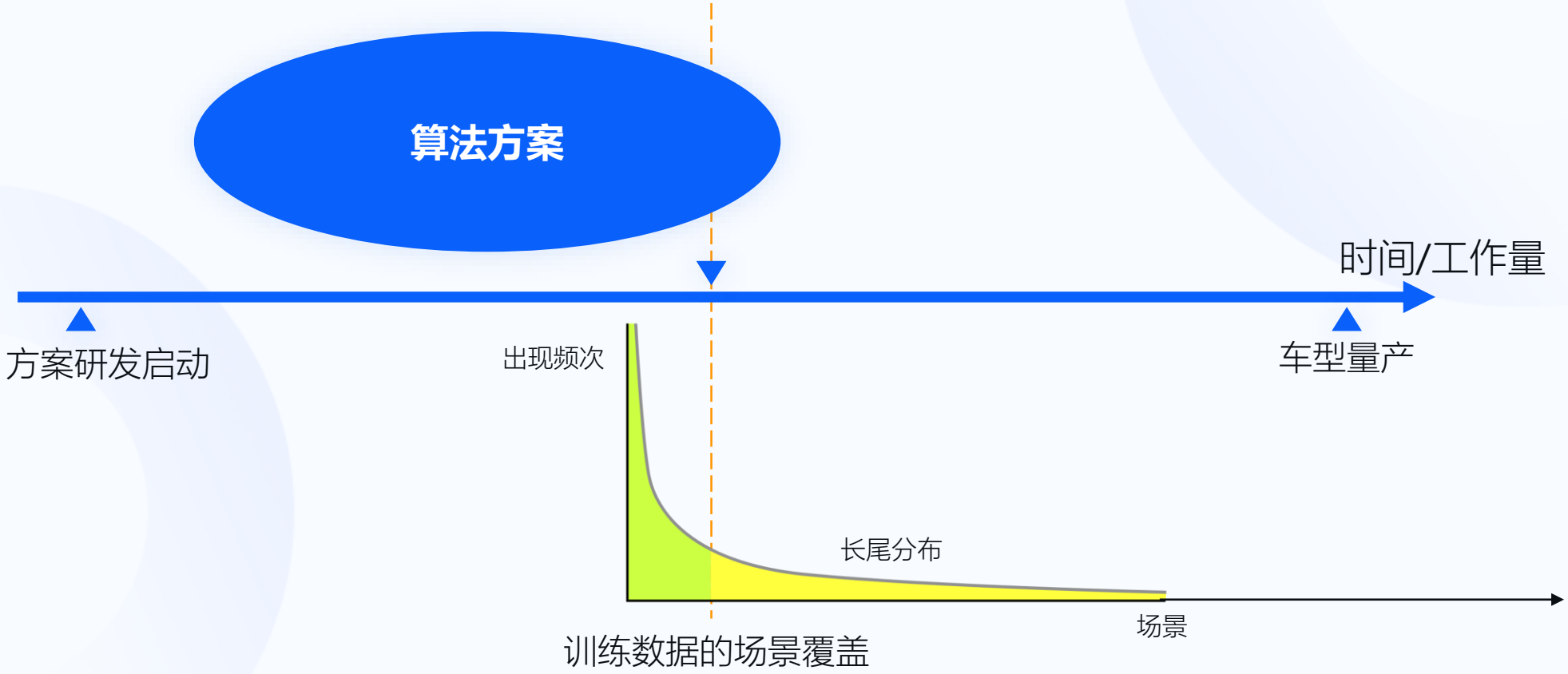
算法设计：

- 结合自车轨迹、他车轨迹和Lidar Voxel（可选）进行query初始化
- 经过与视觉、Lidar（可选）特征的交互进行空间融合，然后与历史帧sparse query完成时序融合
- 筛选有效query，并更新未来帧需要的时序sparse query
- 每个sparse query，输出可用于query占用、Flow和语义的特征向量

1. 软硬协同的感知算法方案
2. 加速量产的闭环平台基建
3. **BEV介绍和感知技术趋势**

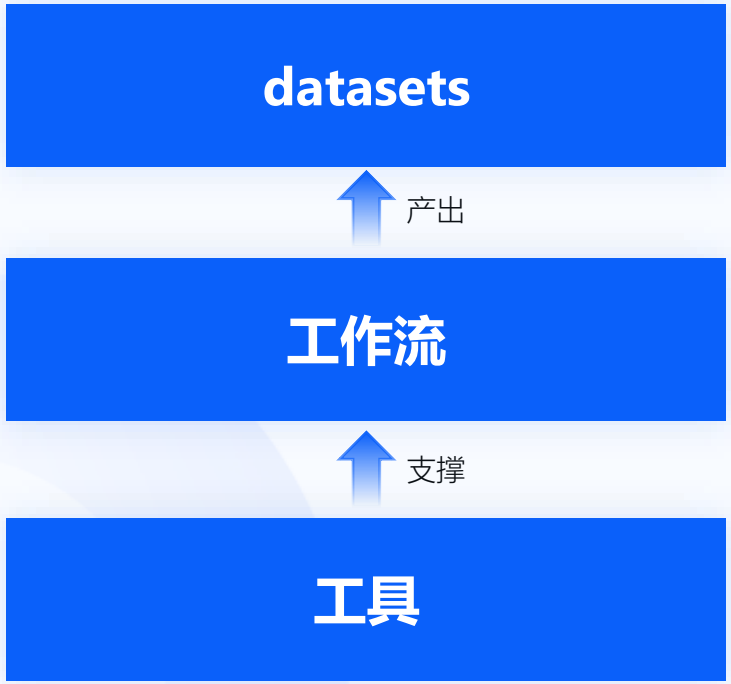
“行百里者半九十”

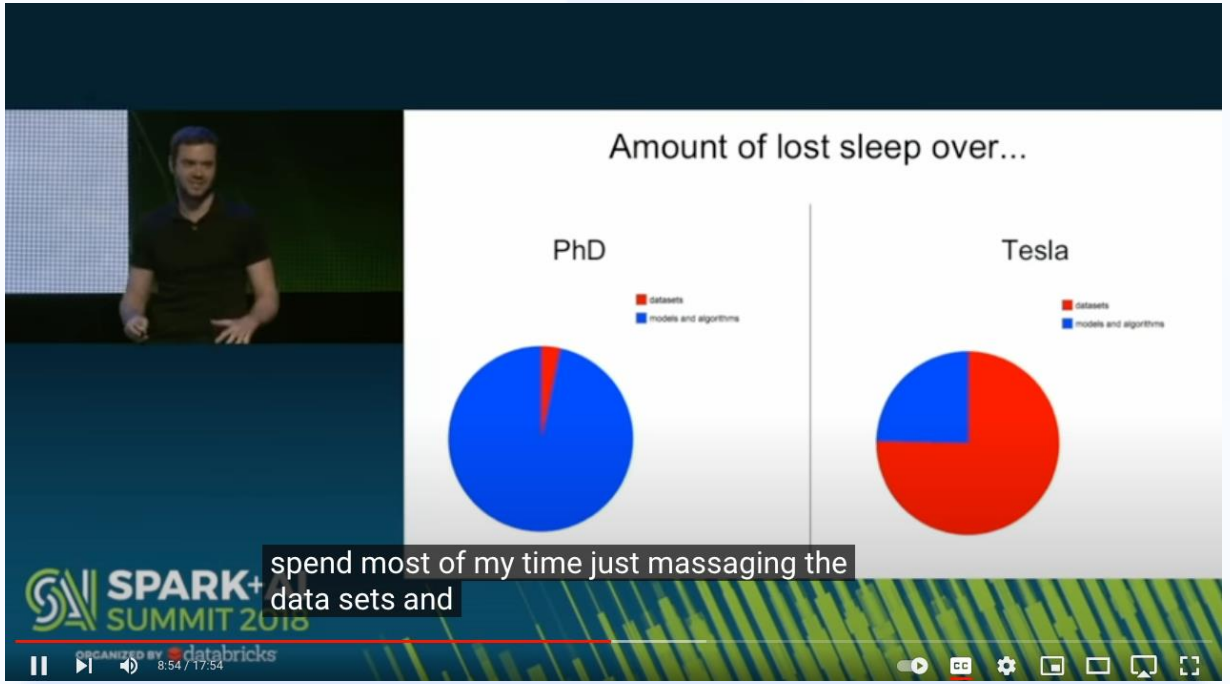
- **在算法方案大部分定型、数据覆盖大部分常见场景时，距离量产才完成大约50%的工作**
- **量产前仍有大量的细致打磨和解决长尾问题的工作**



解决方法

通过标准的工作流和好用的工具，构建更有效的datasets，去解决长尾问题

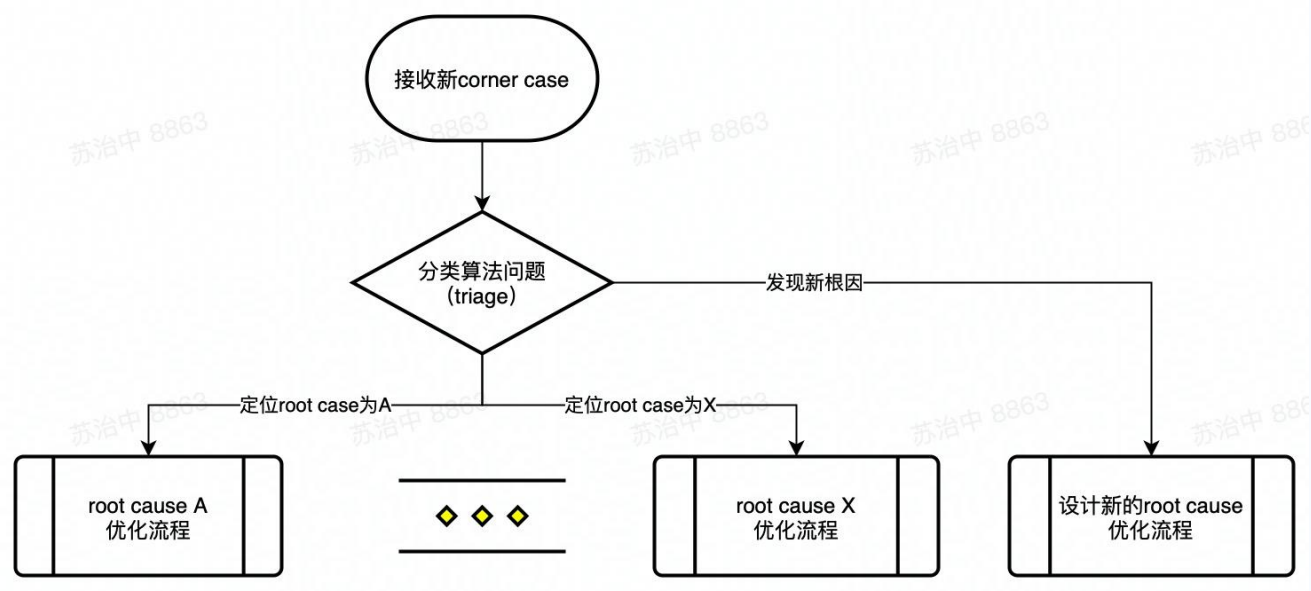


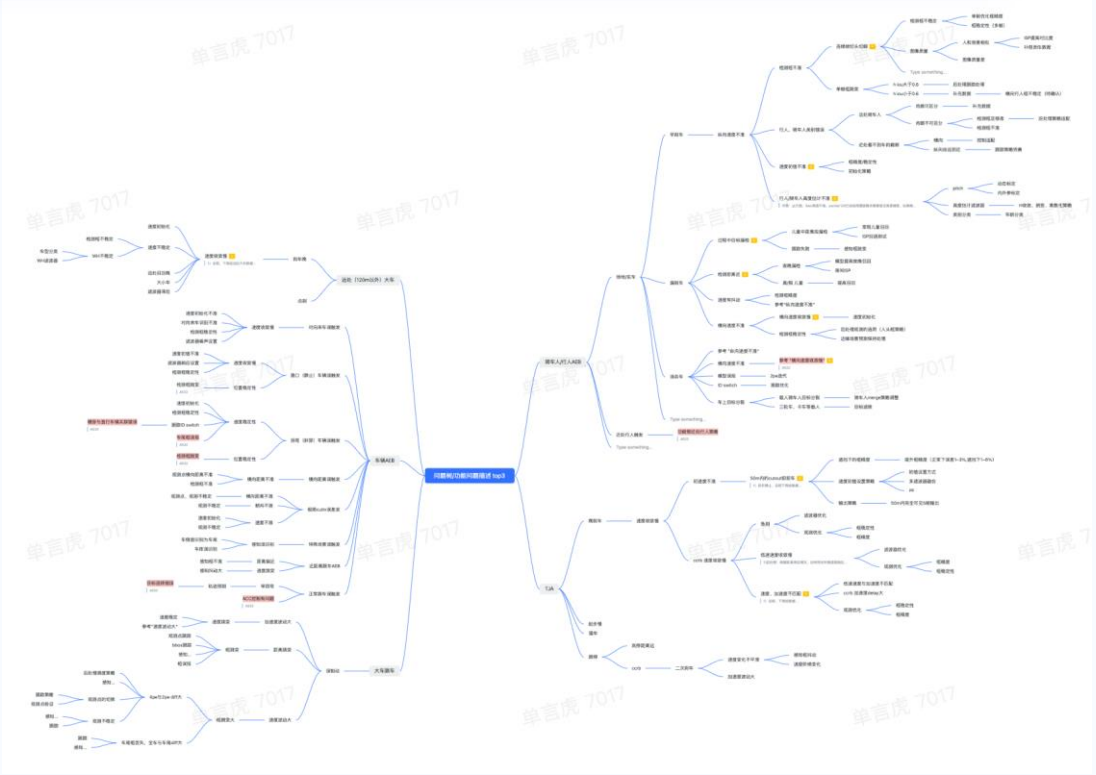


Andrej Karpathy at SPARK+AI SUMMIT 2018

标准工作流

- 通过root cause分析，我们将大量的corner case丢进相对收敛的一系列标准的优化工作流中
- Root cause分析过程部分为自动，部分为手动

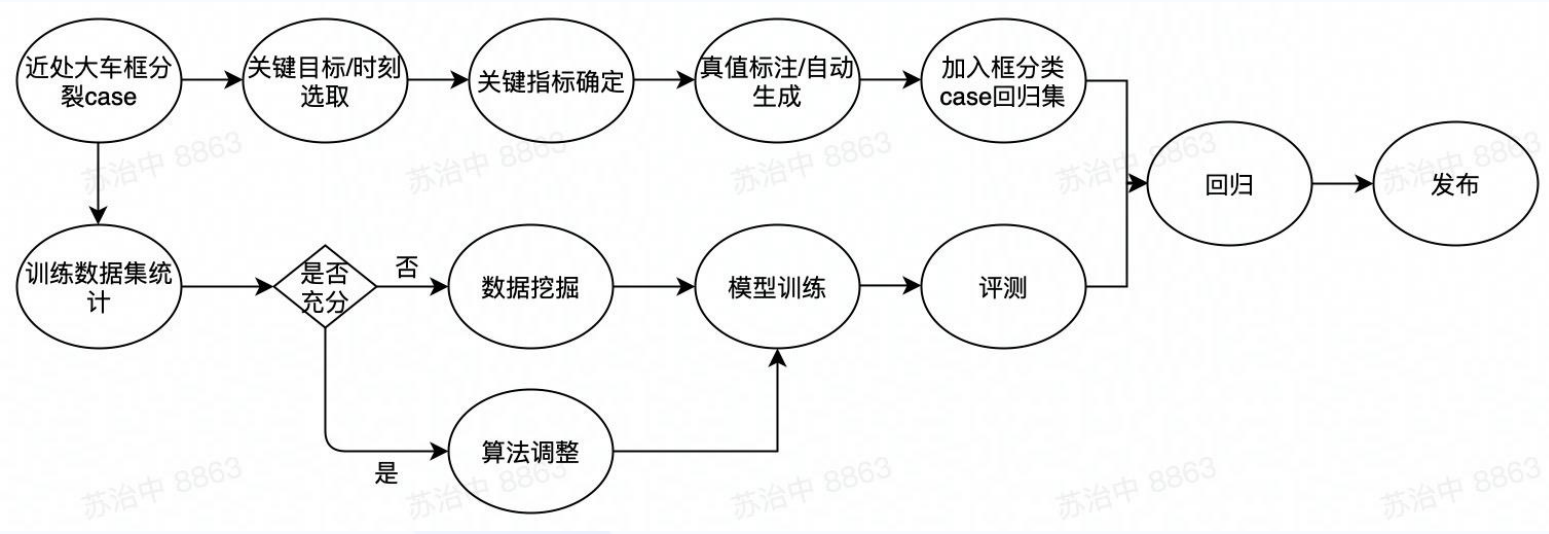




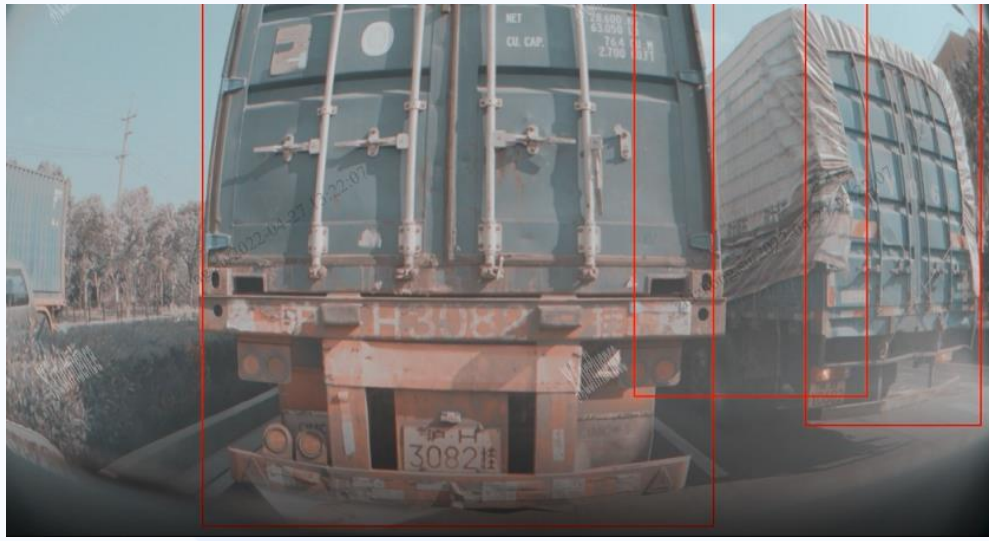
Overview of (part of) the triage tree

标准工作流

- **以近处大车造成的ACC点刹为例**



优化流程示例



改进前

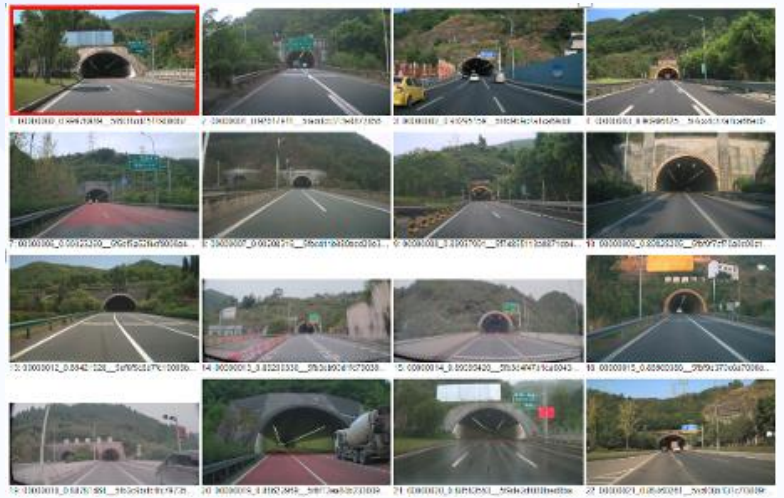
改进后

数据挖掘方法示例

- 这里展示了部分我们常用的数据挖掘方法和示例图

Model Committee
示例：近处大车

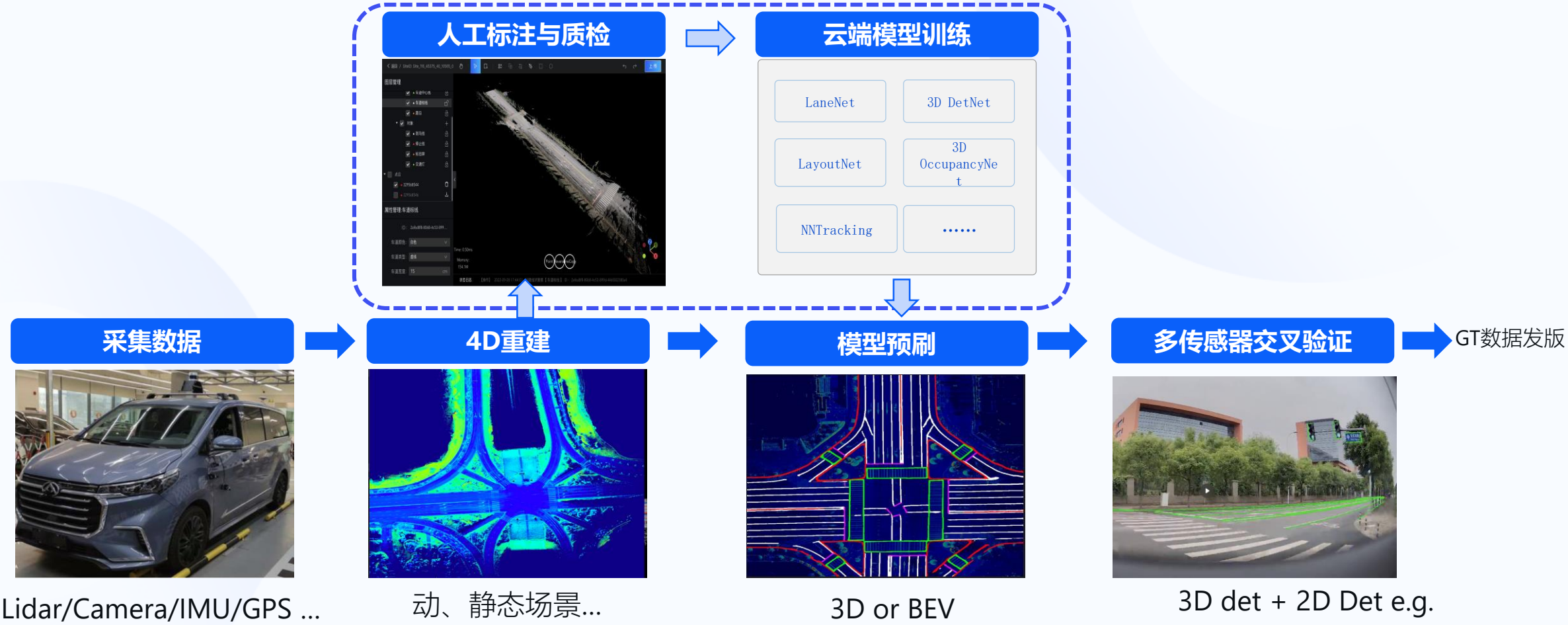
Scene tagging
示例：导流线



Similarity Search
示例：隧道口

BEV感知：真值系统-4D Label整体方案

- 通过4D重建实现点云级别或者object级别的重建，通过人工标注积累原始数据
- 随着数据积累到一定程度，可以训练云端大模型逐步替换人工标注，可提升80%+的标注效率



地 平 线
Horizon Robotics

征程与共，一路同行

拥抱价值共创、繁荣普惠的行业未来