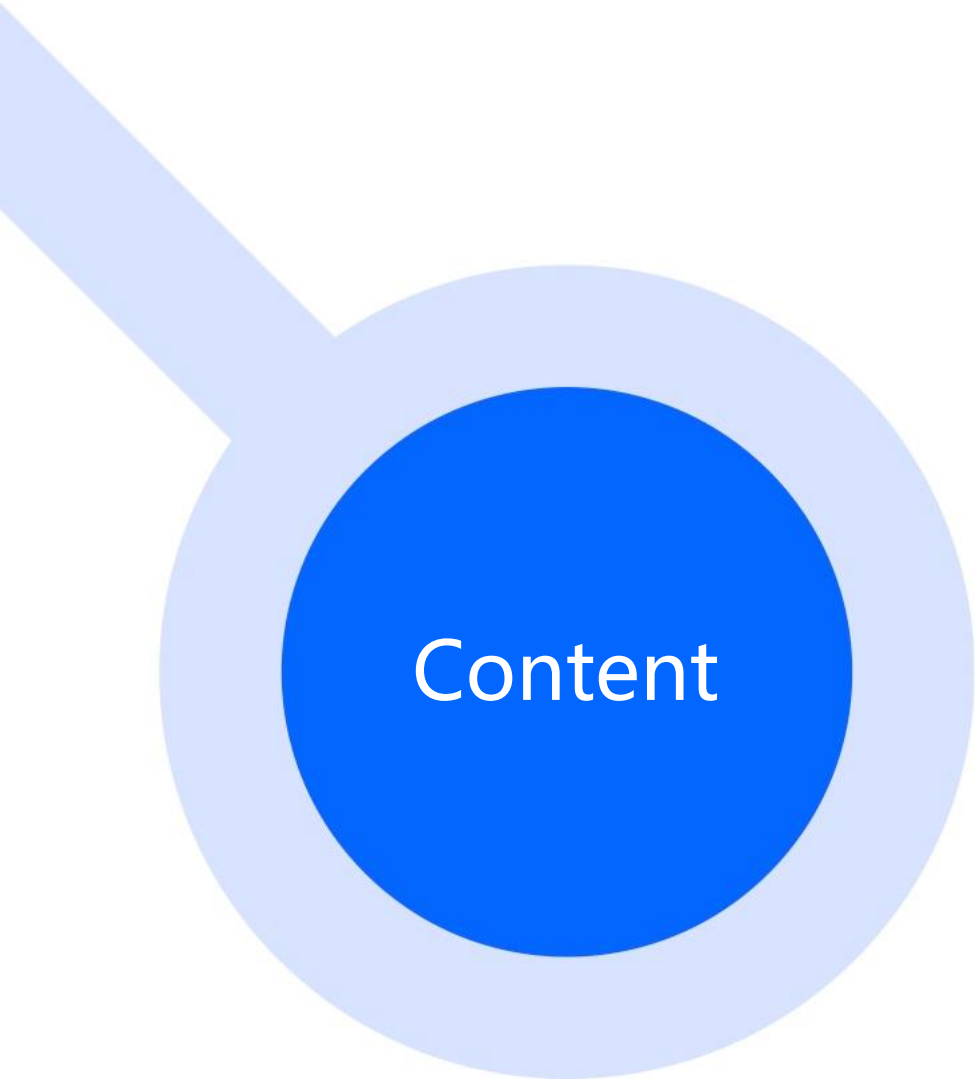


自动驾驶感知新范式

张家馨



Content

上半节

- 01. 自动驾驶感知技术发展
- 02. BEV感知框架
- 03. BEV感知应用

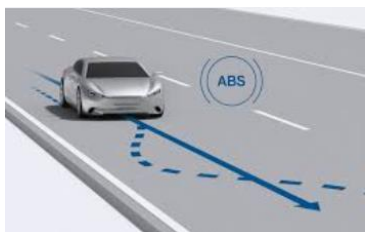
下半节

- 04. 云端自动标注
- 05. 仿真系统
- 06. 端到端自动驾驶

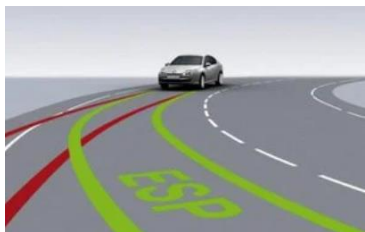
01. 自动驾驶感知技术发展

- 从L2主要面向高速场景的辅助驾驶场景采用前视单目传感器即能满足需求
- L2+ 增加了自主超车、无保护左转等高级辅助驾驶功能，这些功能的基础都是自动变道辅助系统
- L2+ 面临的场景更加复杂，从高速场景延伸到更加复杂的城区场景

L0~L1辅助驾驶功能



防抱死刹车系统 (ABS)



电子稳定程序 (ESP)

L1~L2辅助驾驶系统



自动紧急刹车系统 (AEB)



自适应巡航控制系统 (ACC)



车道保持辅助系统系统 (LKA)

L2+ 高级辅助驾驶系统



自动变道辅助系统 (ALC)



自动汇入汇出

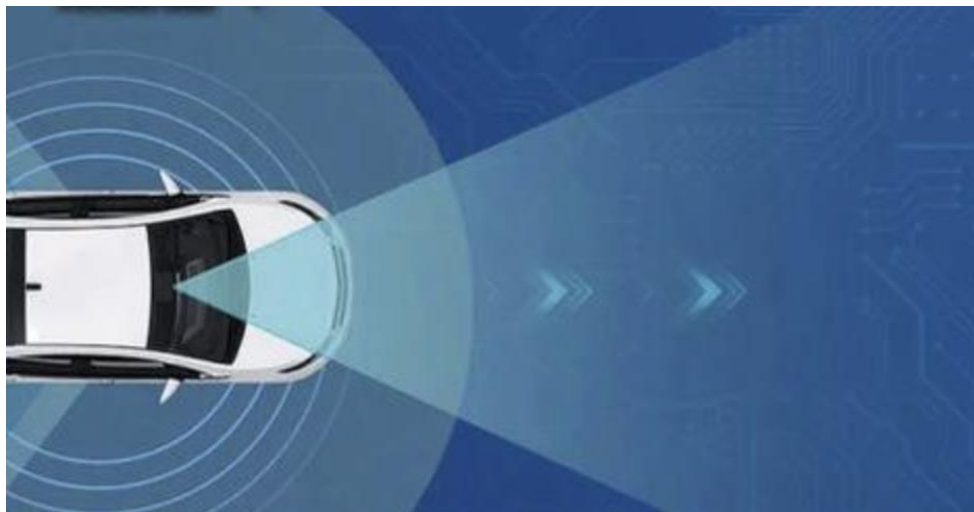


城区路口左转

01. 自动驾驶感知技术发展

- L2辅助驾驶系统主要是前方的车辆感知，单目前视相机即能够满足需求
- L2+功能的增加和场景的复杂化，要求车身360度感知，业界一般采用surround-camera周视布局的方式，同时为了远距离感知，前视一般会保留窄角相机

L2辅助驾驶系统的视觉传感器布局



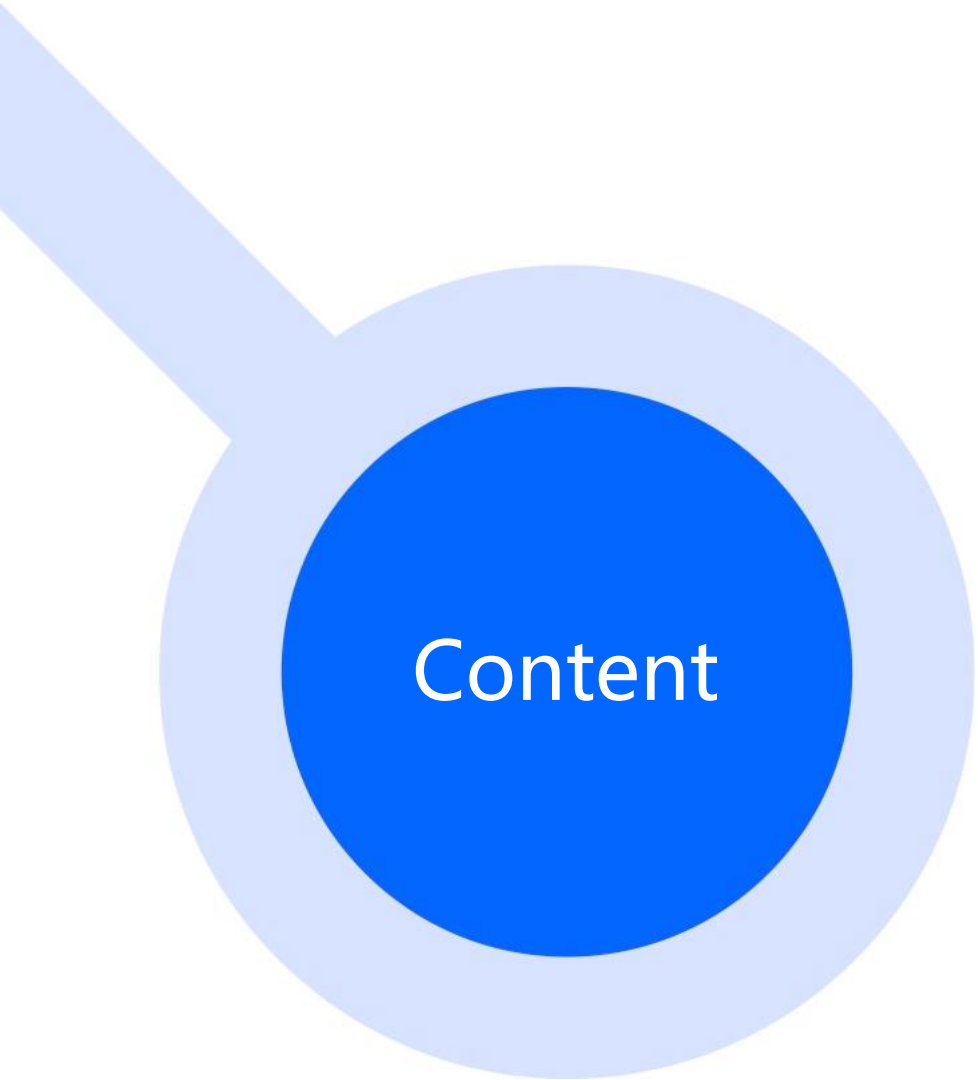
L2+ 高级辅助驾驶系统的视觉传感器布局



01. 自动驾驶感知技术发展



BEV感知下的城区辅助驾驶



Content

02. BEV感知框架

(a) Transformer

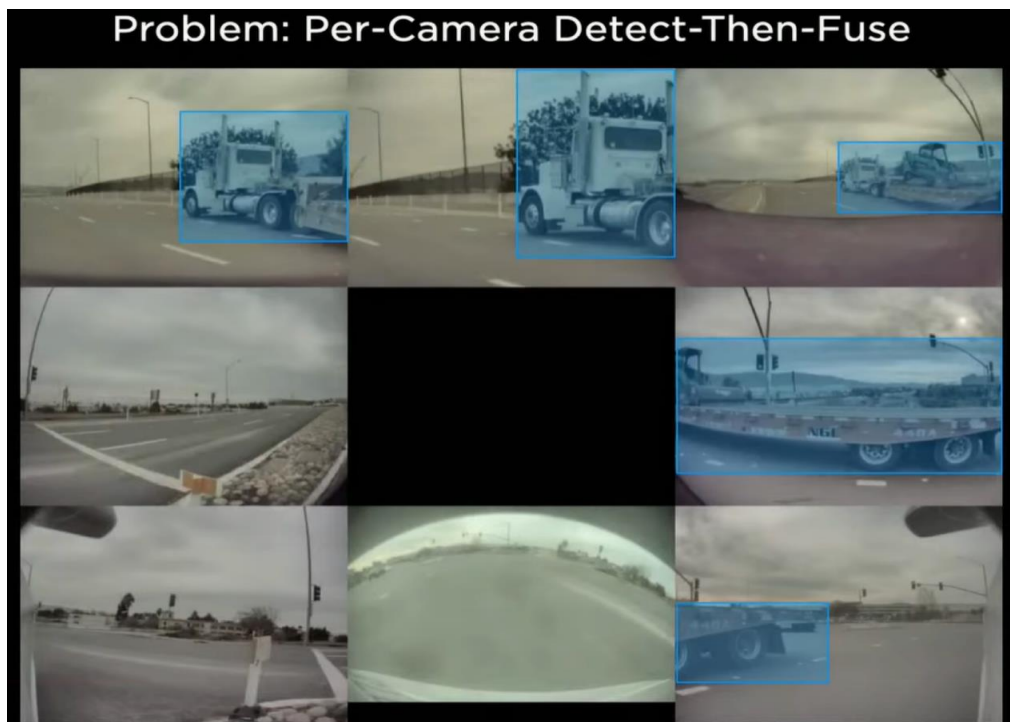
(b) 空间融合

(c) 时序融合

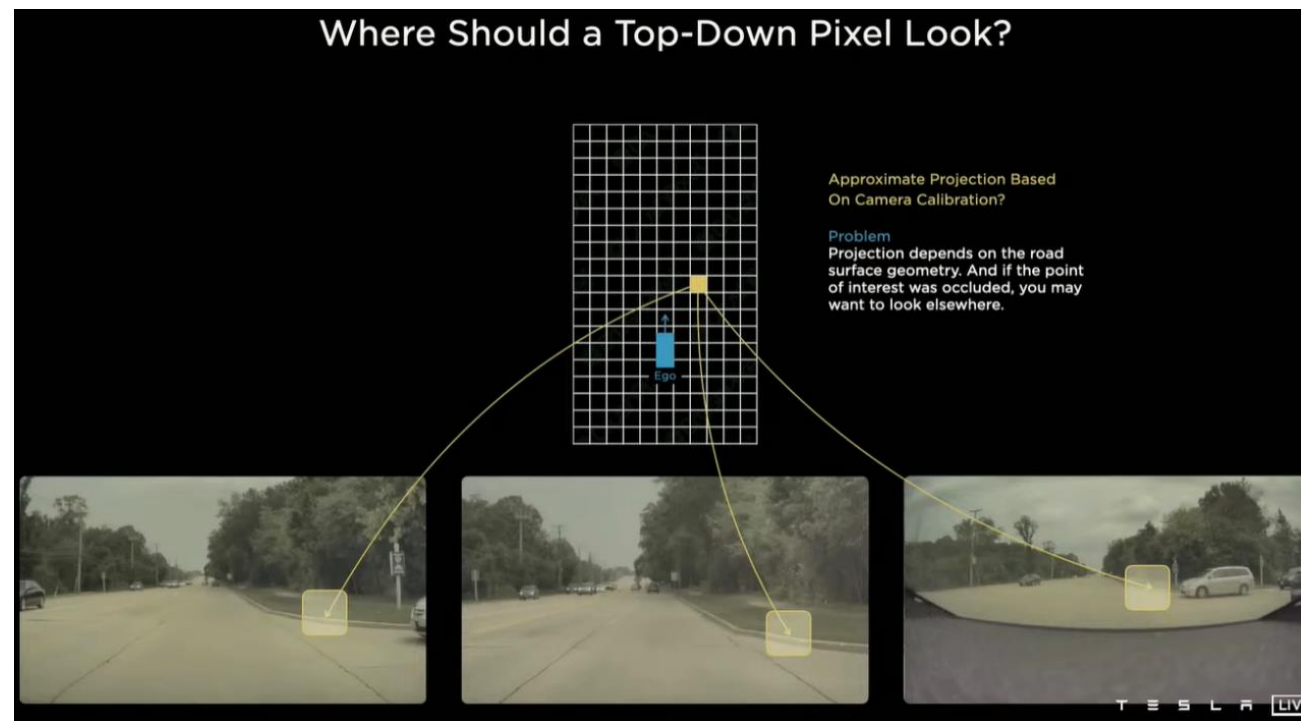
02. BEV感知框架-什么是BEV感知

BEV感知和传统感知的区别

传统感知



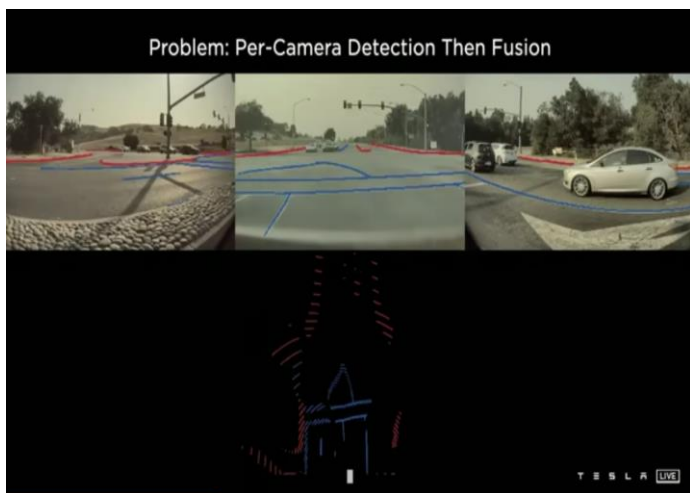
BEV感知



02. BEV感知框架-什么是BEV感知

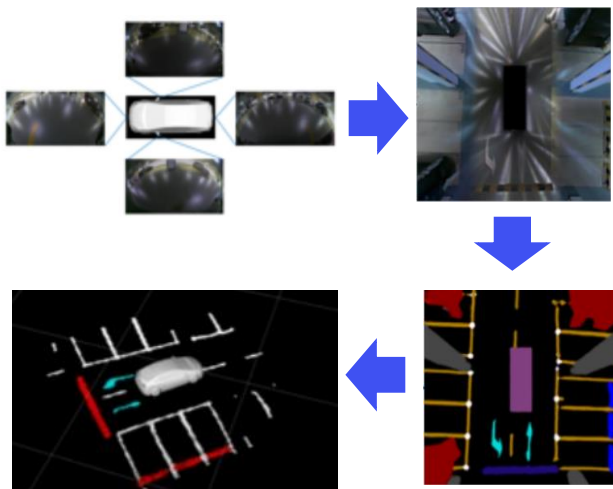
- 神经网络输入空间为图像空间(Perspective View), 输出空间为3D空间 (Bird Eye View)
- 模型将图像空间的特征变换到3D空间进行融合, 即中融合

单视角感知+rule-based后处理重建



图片来自于Tesla AI Day 2021

IPM 前融合方案



图片来自于华为AVP-SLAM

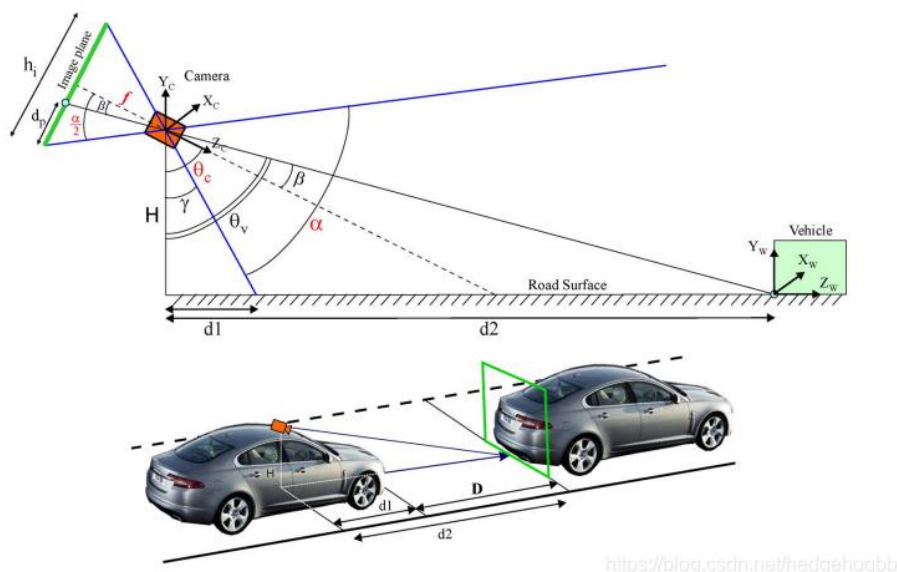
BEV 中融合方案



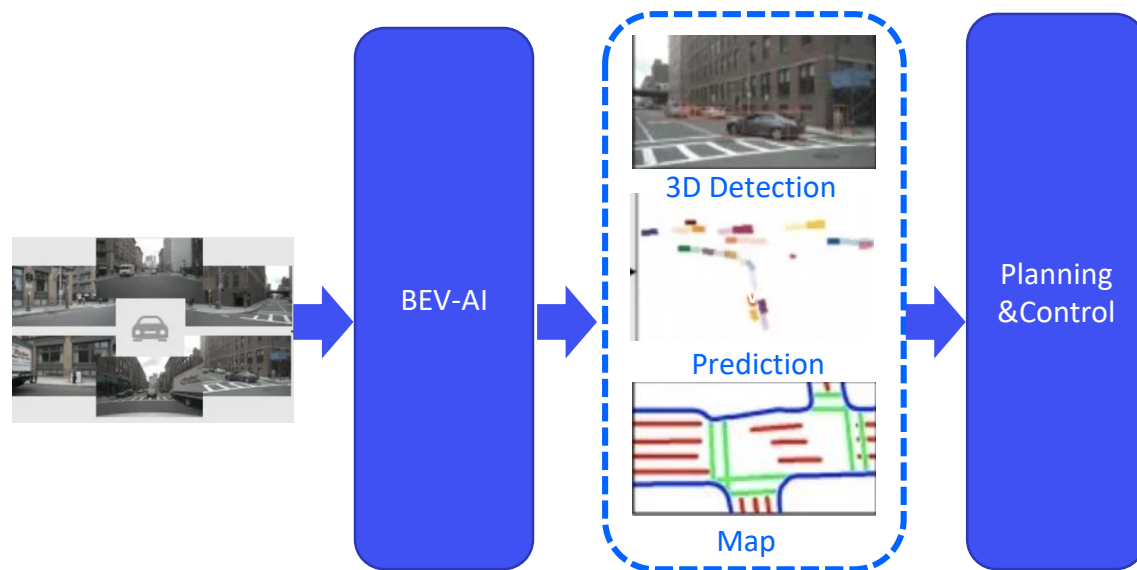
图片来自于地平线SuperDrive

02. BEV感知框架-BEV 感知的优势

- 可更加直观的展示车辆周围的3D环境,有利于下游任务游任务（规划、控制）的开发
- 测距更加准确、尺度稳定，避免了图像视角下的尺度和遮挡问题，可以有效地识别有遮挡或者交叉的交通标志
- 中融合感知，输出结果一致性高，避免了Per-camera Detect-Then-Fuse 的ruled based操作
- 方便做多传感器、时序融合以及端到端规控感知



单目测距法依赖相机内外参、地平面假设、以及物体的真实尺寸，测距性能不稳定，无法适用复杂场景

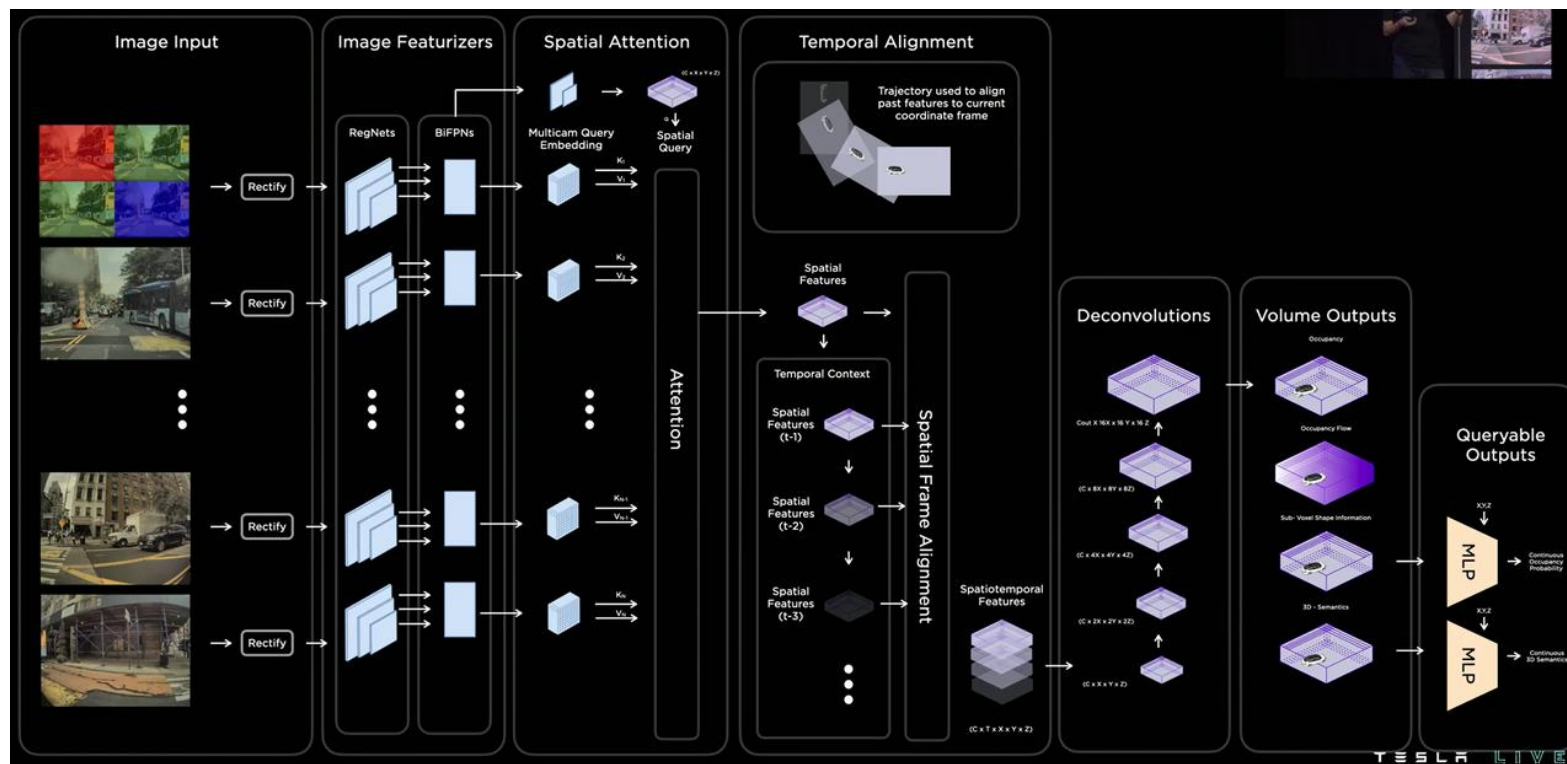


BEV输出空间为3D 空间，让端到端的感知成为可能

02. BEV感知框架-BEV 感知的优势

- 感知网络由backbone、空间融合模块、时序融合模块、任务head构成, 结构类似于bevformer
- 空间融合模块采用Transformer、时序融合模块采用Spatial RNN (参考AI Day 2021)
- 采用Rectify层对相机的外参进行归一化, 显著提升感知性能
- 特征融合之后BEV feature 分辨率上采样了32倍, 说明用于融合BEV query可以很小, 我们目前的实验结果能够佐证 (0.4m/pixel 和1.6m进行融合 之后上采样到同样分辨率的精度基本一致), **小物体需要更大尺度的图像特征, 而不是**

BEV融合分辨率

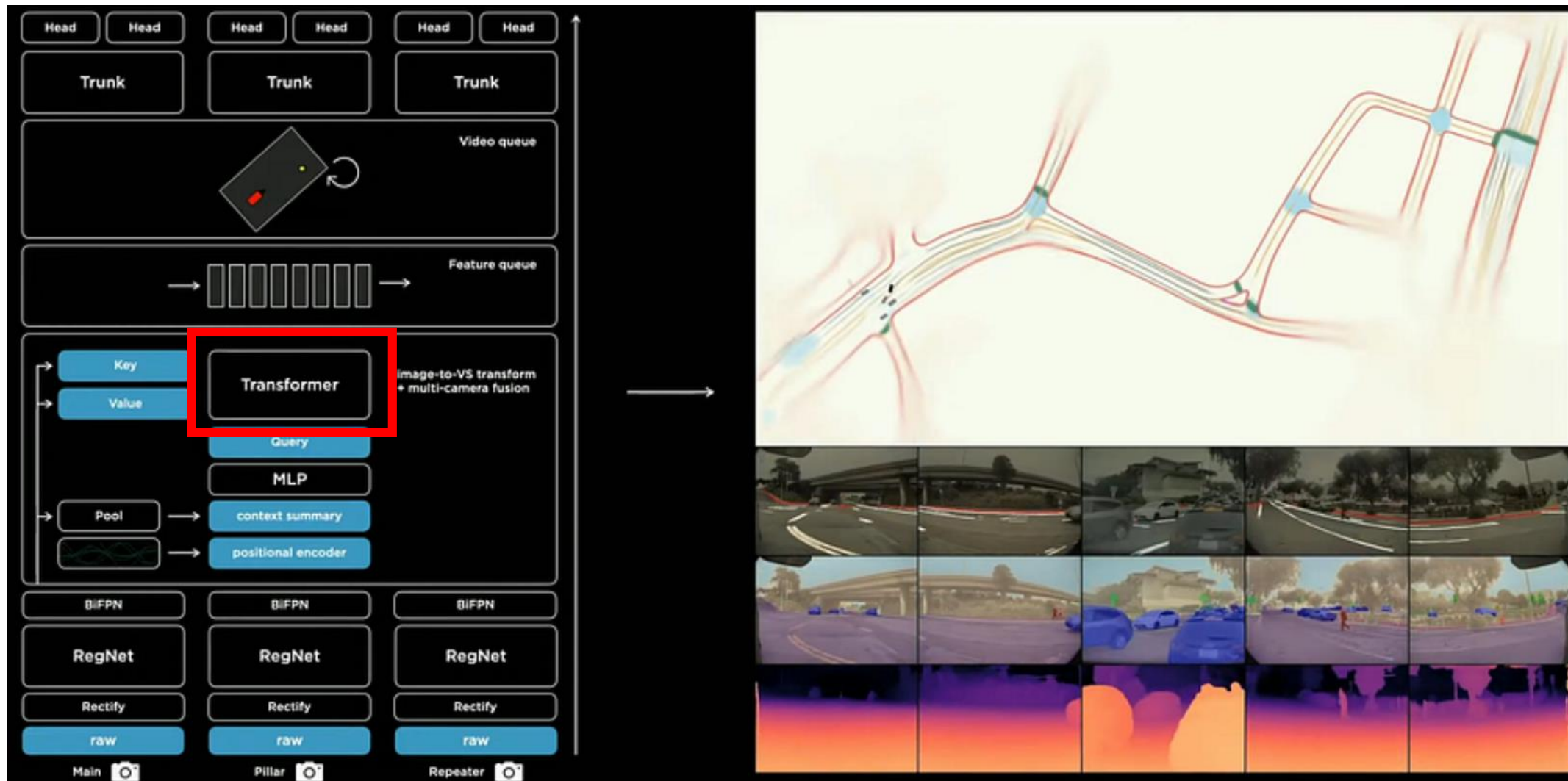


02. BEV感知框架 - 一个基础+两大模块

Transformer

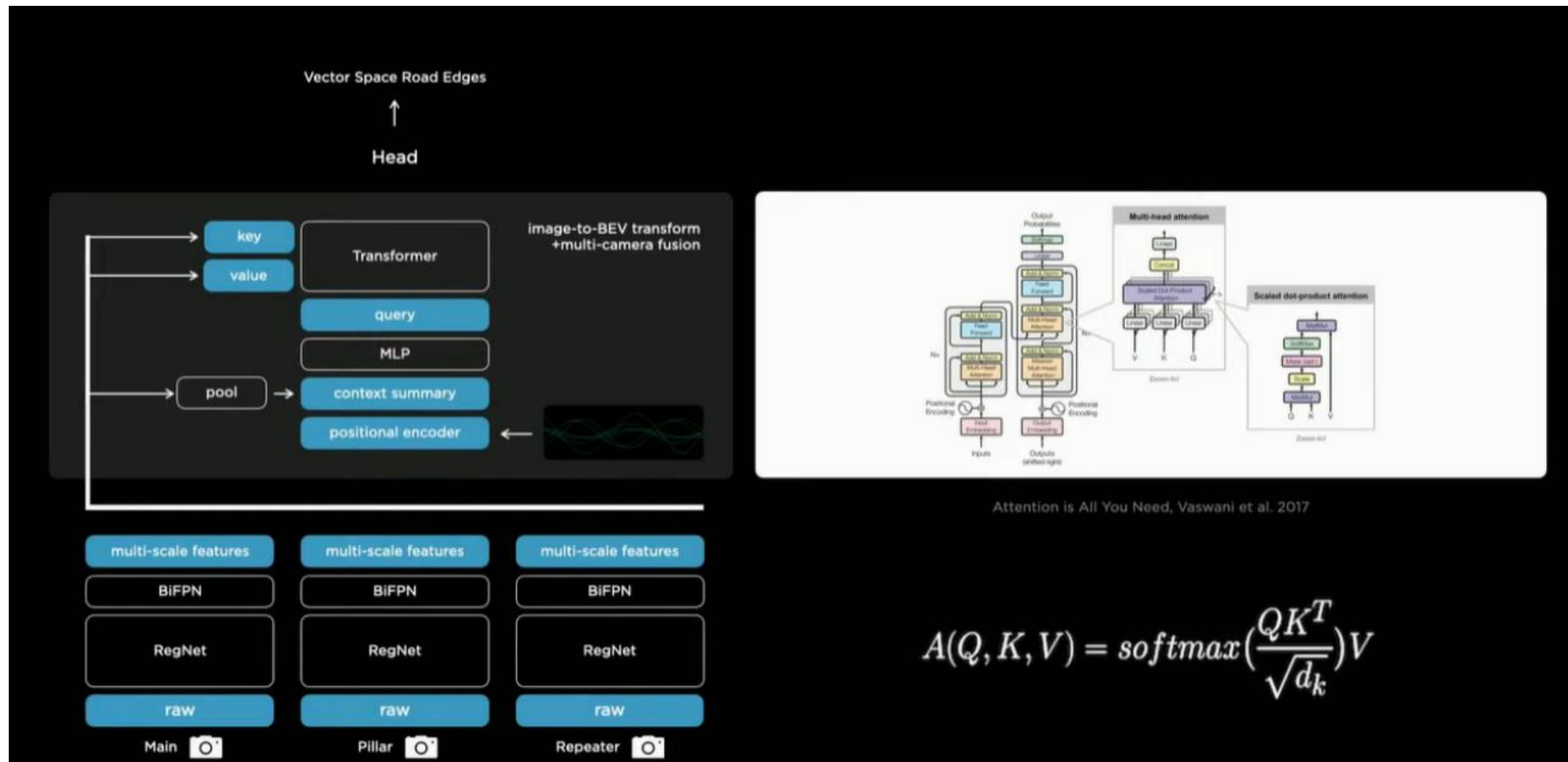
+

1. 空间融合 (View Transformation)
2. 时序融合 (Temporal Fusion)



02. BEV感知框架 - Transformer

Transformer作为BEV感知backbone



02. BEV感知框架 - Transformer

- Transformer最初被提出主要用来解决语言翻译问题
- 替代RNN/LSTM， 得益于良好的并行训练能力

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

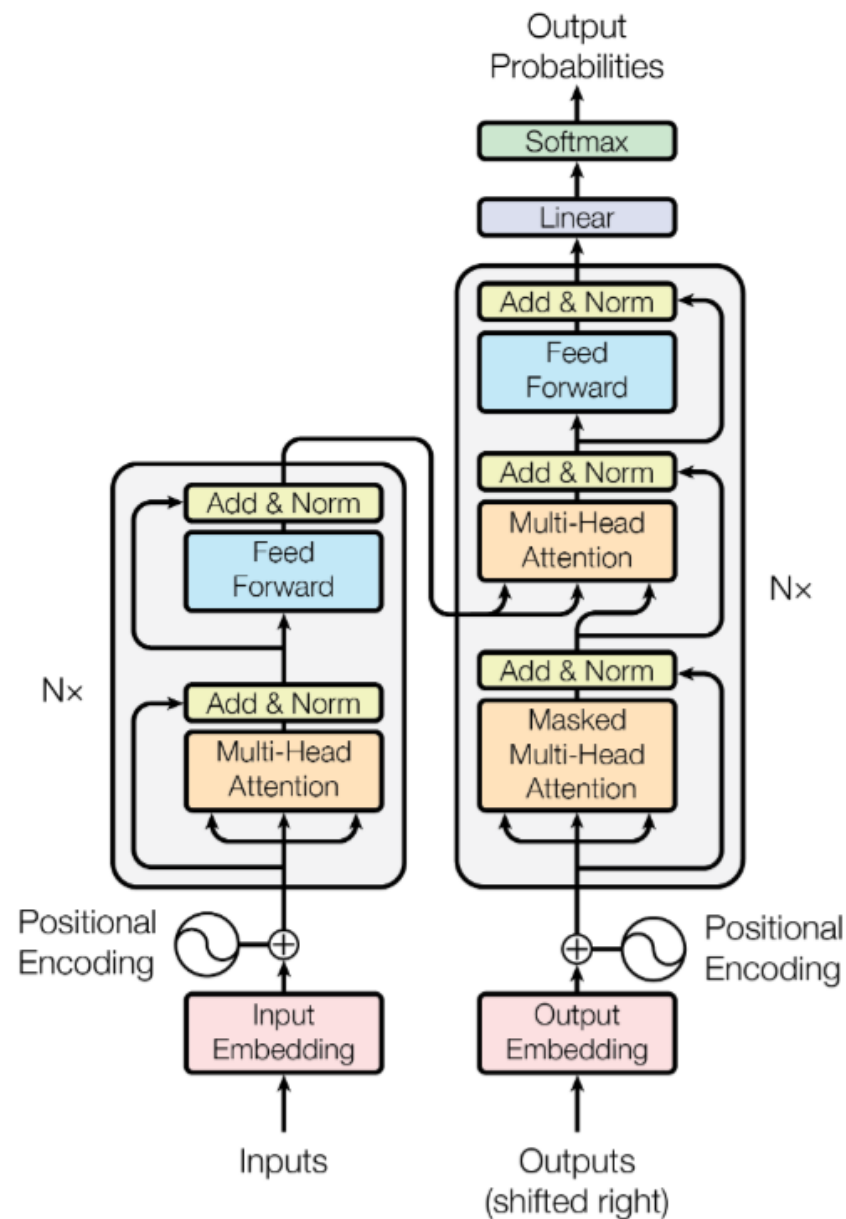
Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

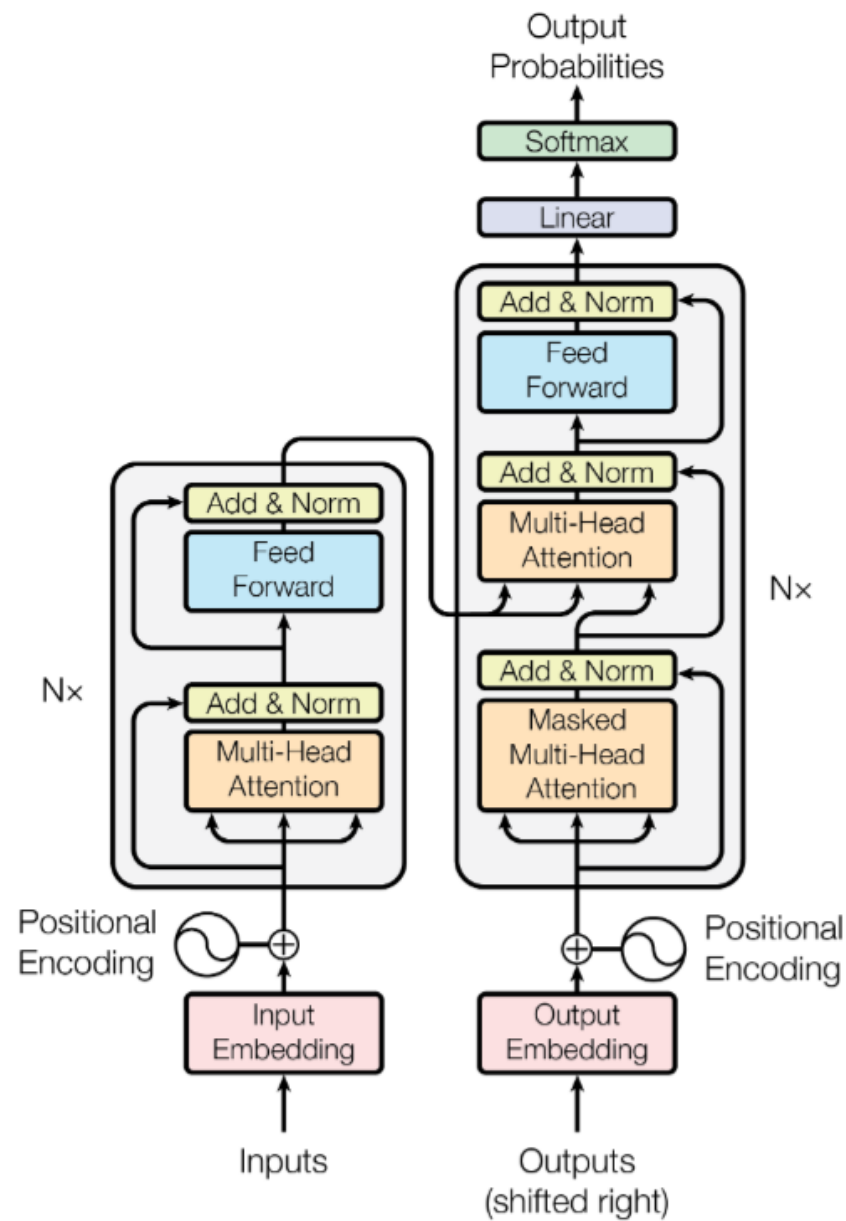
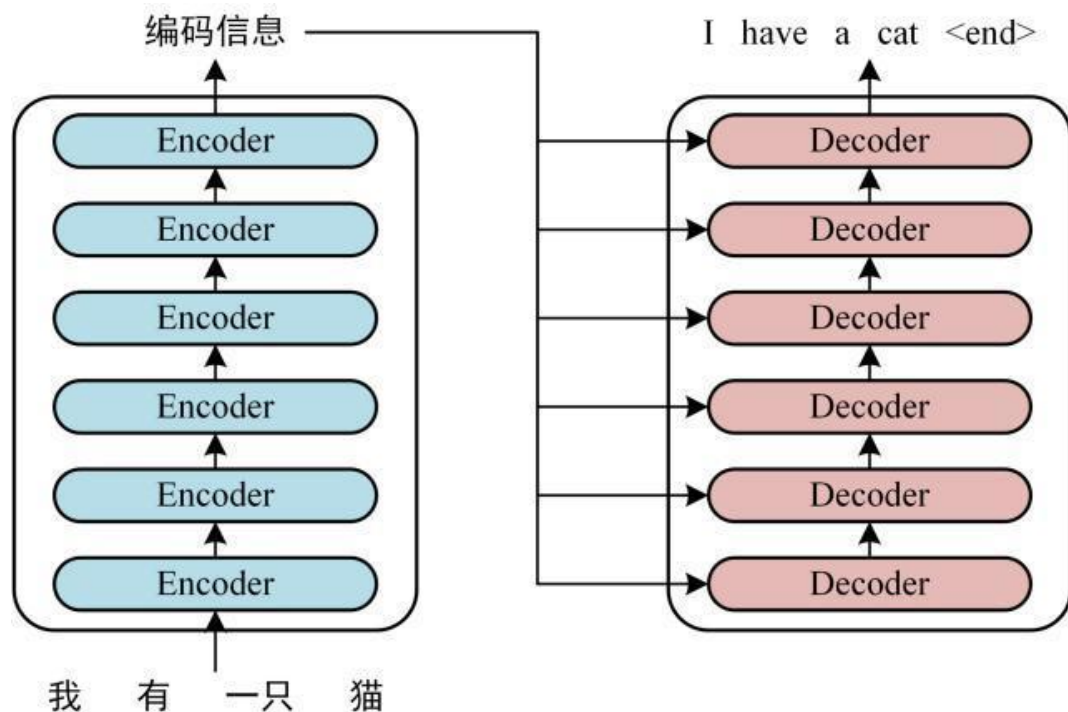
Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com



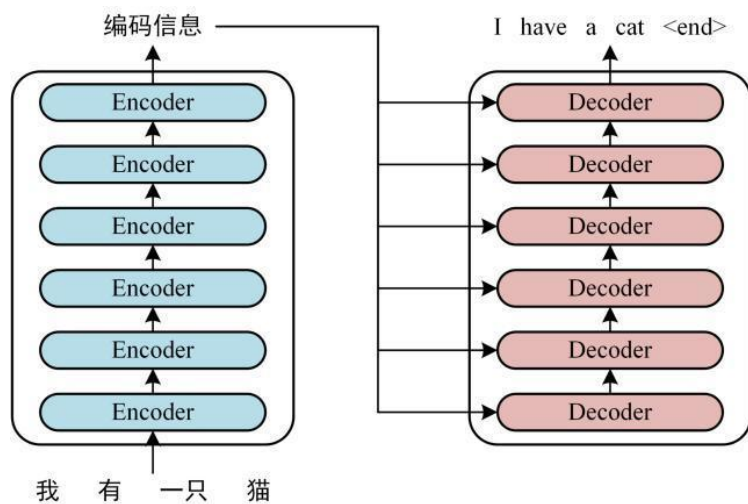
02. BEV感知框架 - Transformer

原始Transformer用于翻译任务

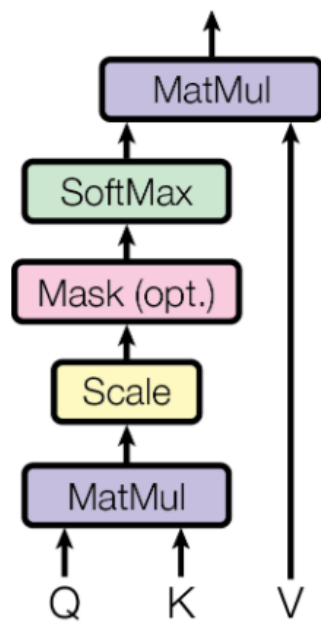


02. BEV感知框架 - Transformer

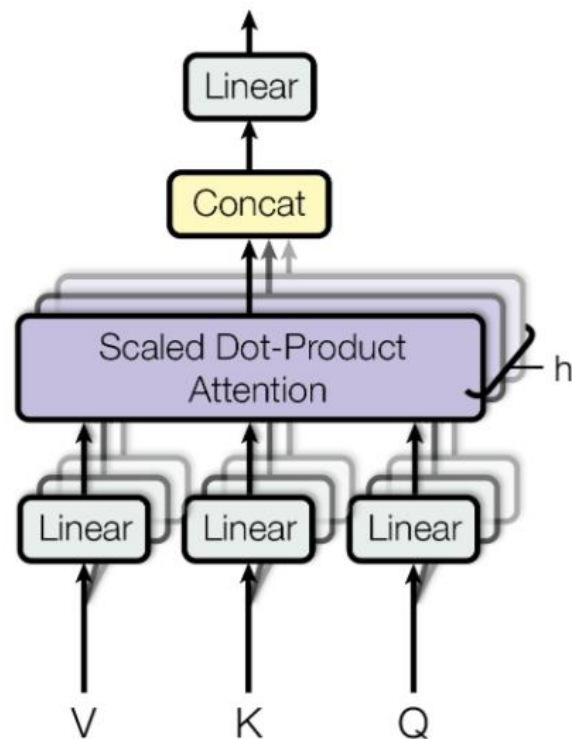
Transformer核心思想



可并行的序列编码



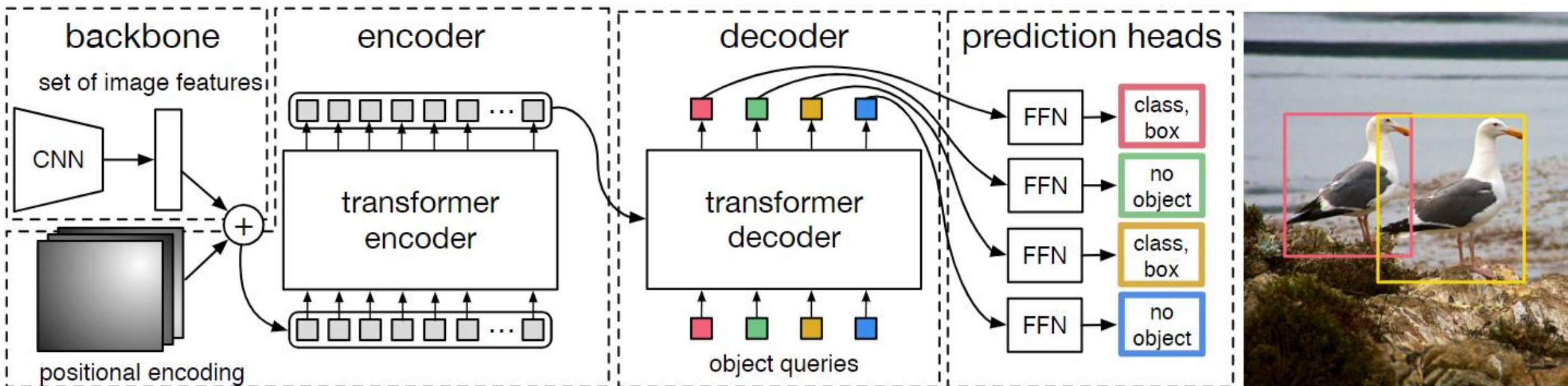
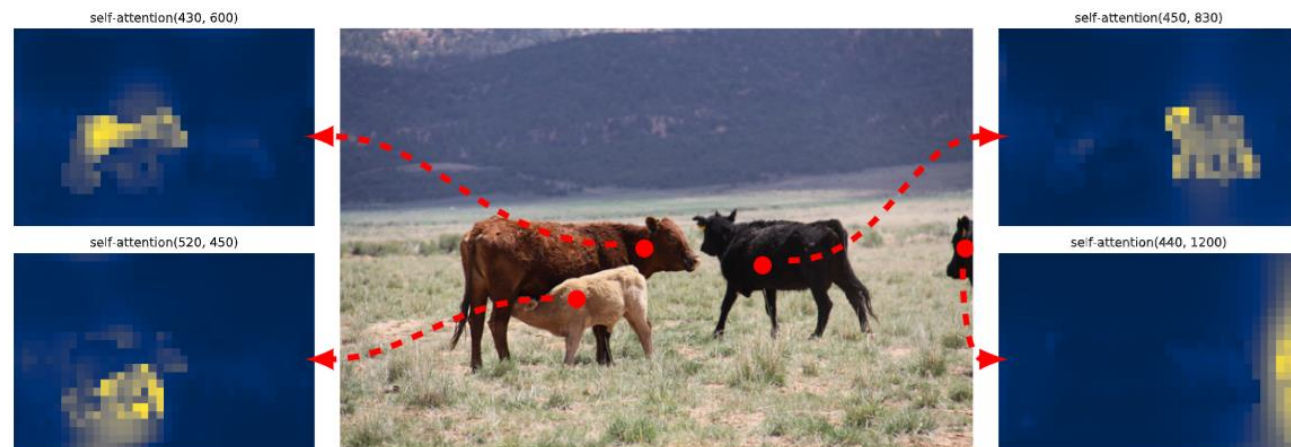
QKV注意力机制



多头注意力机制

02. BEV感知框架 - Transformer

- Transformer框架在视觉领域大展身手
- DETR 2D图像检测 超越CNN



02. BEV感知框架-BEV空间融合



输入：2D图像 (PV)

如何转换



输入：3D空间 (BEV)

IPM-Based

- 计算效率高
- 地面平坦 + 目标接地
- 远距离感知效果差 (<50m)

Depth-Based

- 2D -> 3D 自底向上
- 稠密BEV特征表示
- 每个像素都预测一系列深度概率
- 远距离感知效果差 (<50m)
- 需要NMS后处理, 超参数多, 难优化
- 无法端到端

Transformer-Based

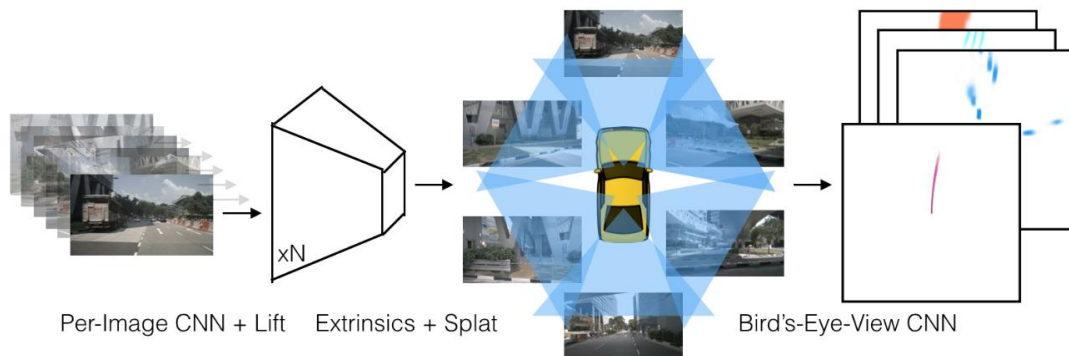
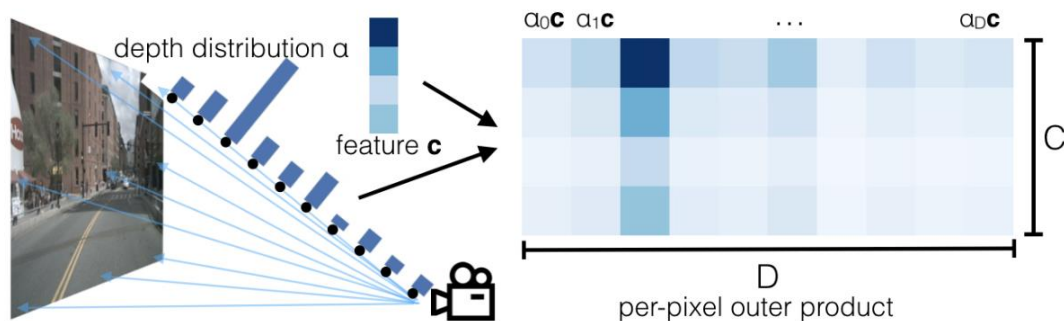
- 3D -> 2D 自顶向下
- 端到端
- 便于多传感器融合
- 利于多种感知任务统一框架

02. BEV感知框架-BEV空间融合

Depth Based

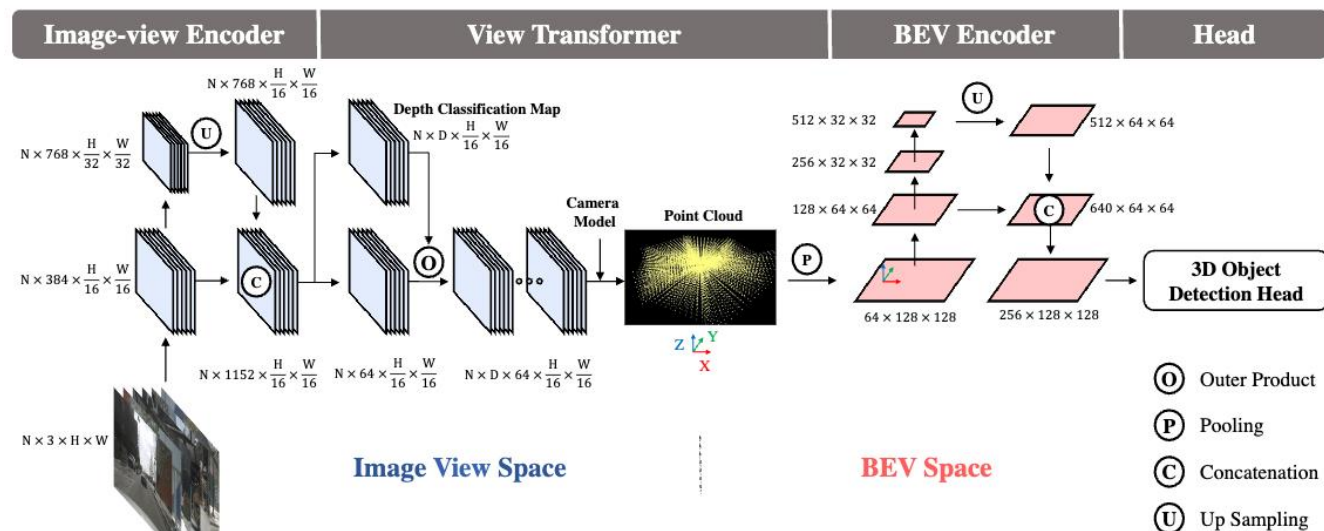
Lift, Splat, Shoot (LSS)

- 远距离效果不好
- 计算量大, 比较依赖depth 真值, 需要depth监督
- 泛化性不好



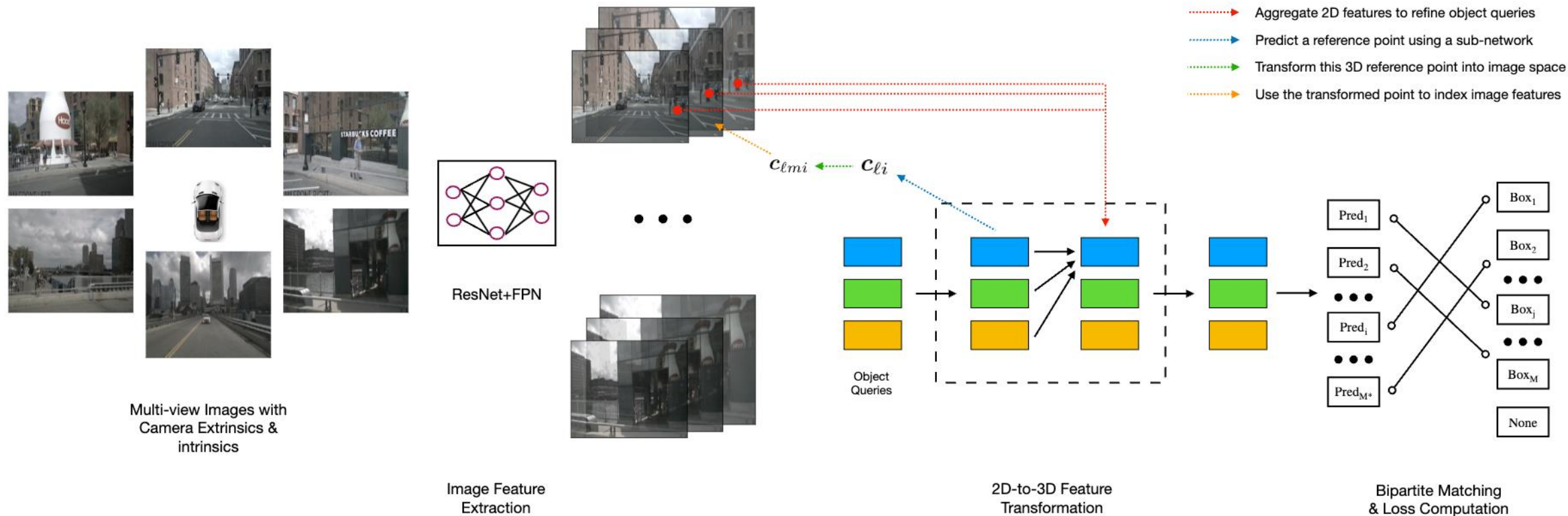
BEVDet

- 基于LSS做了检测+heatmap的检测
- 对于原始图像分辨率要求高, (行人4x或8x, 车辆16x)
- BEV feature 分辨率和IPM类似, 不同点是IPM有平面假设



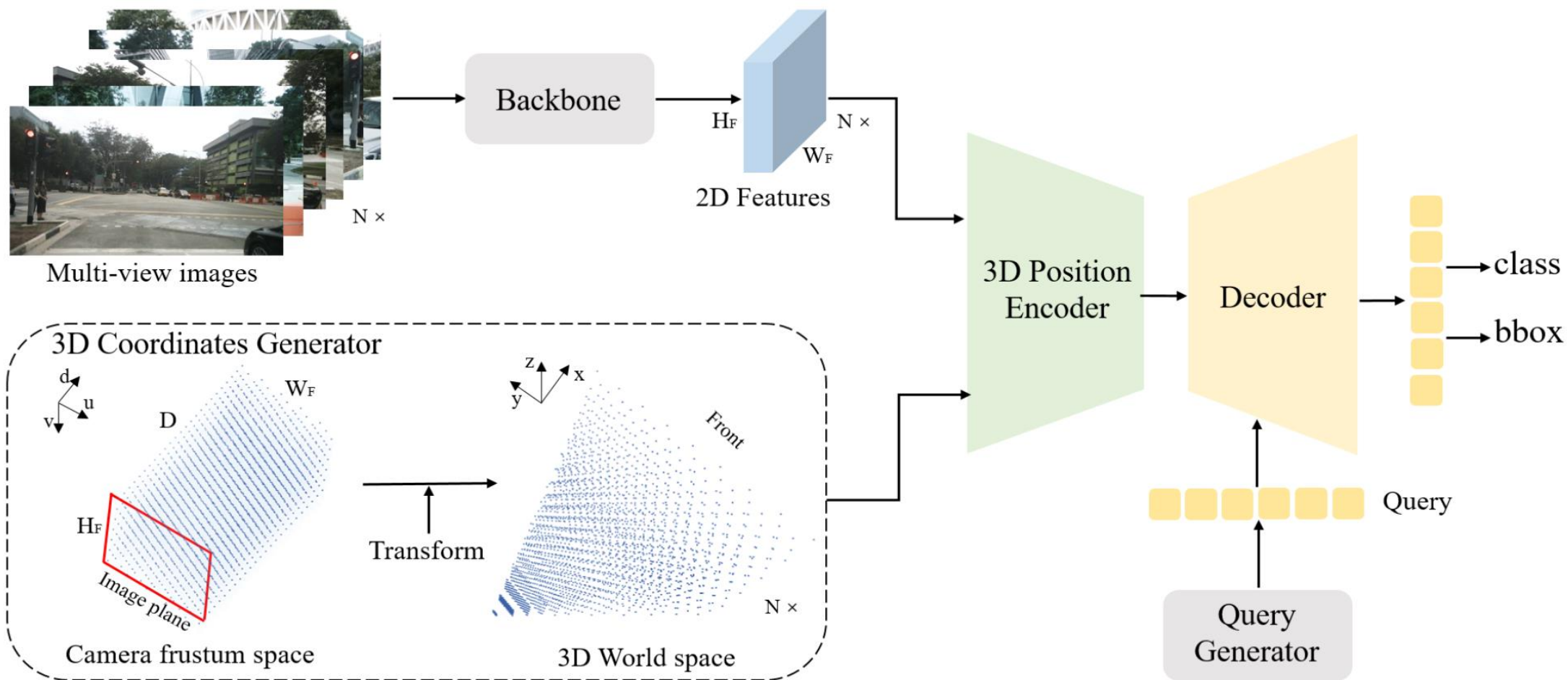
02. BEV感知框架-BEV空间融合

Transformer based (Sparse): DETR3D



02. BEV感知框架-BEV空间融合

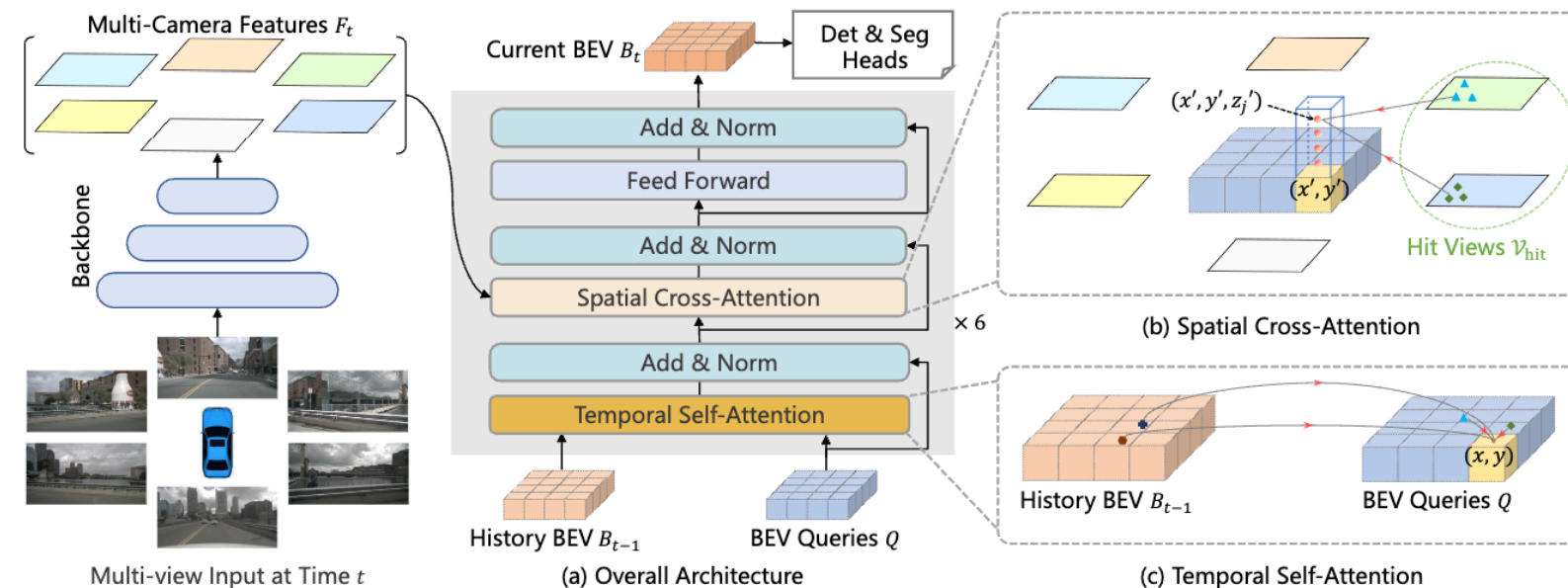
Transformer based (Sparse):
PETR



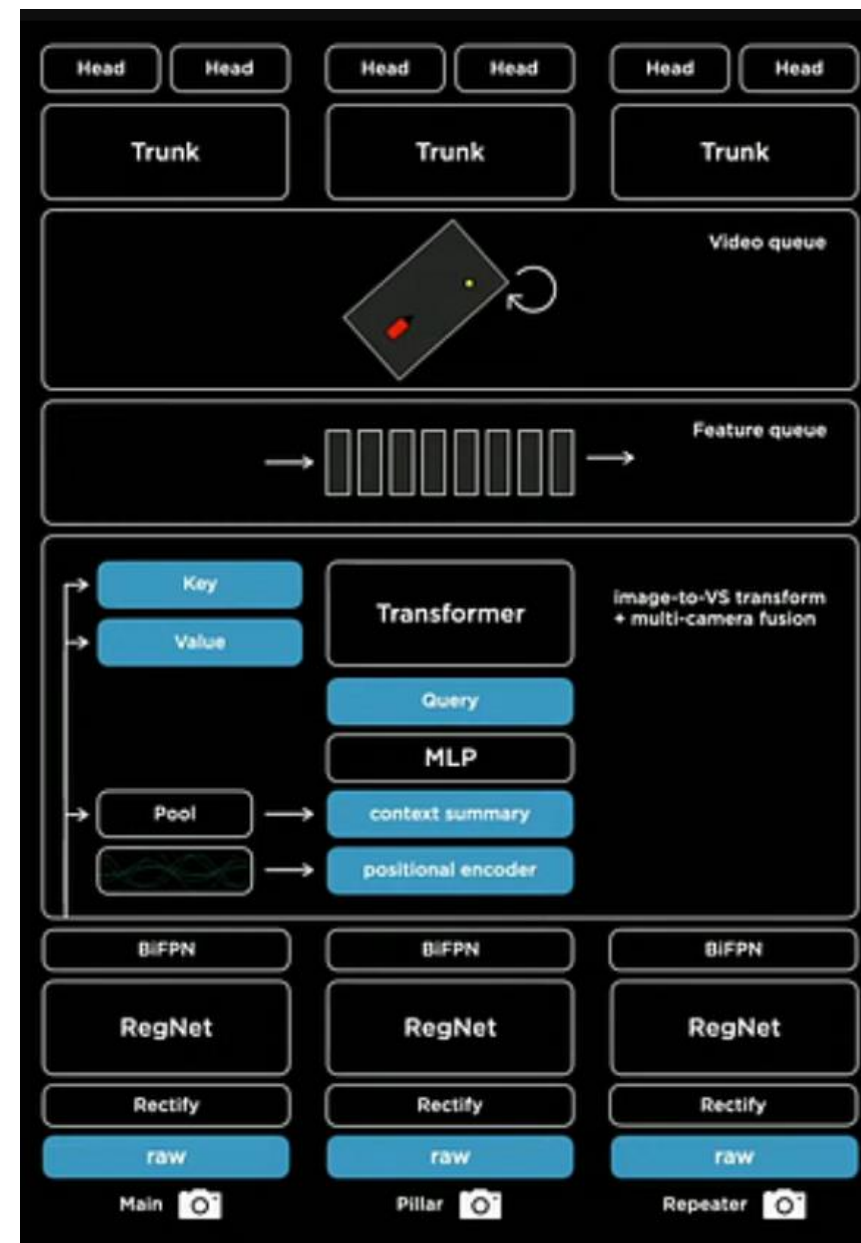
02. BEV感知框架-BEV空间融合

Transformer based (Dense)

BEVFormer



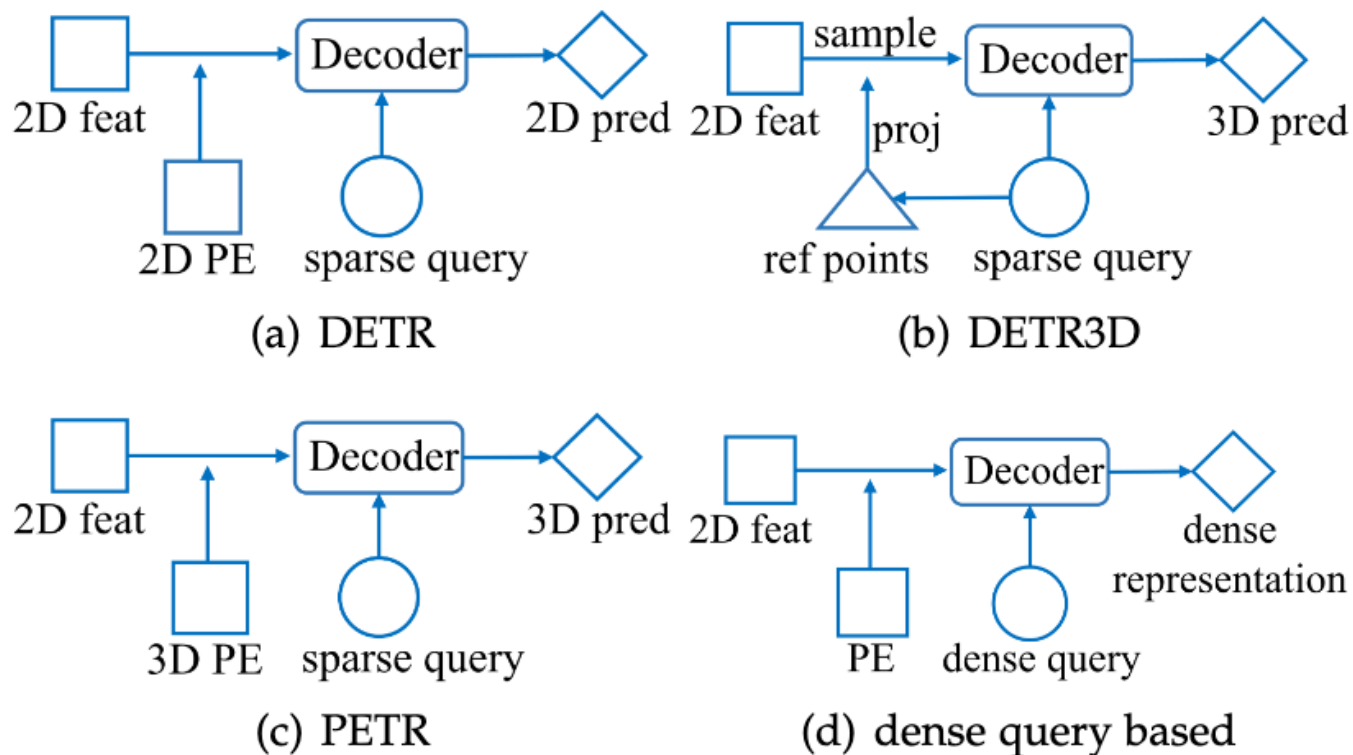
Tesla AI Day



02. BEV感知框架-BEV空间融合

Transformer based

- 分为sparse 和 dense query两种方法



- 性能: $c) = d) > b)$, b)类方法的ref points 不好学
- 对外参的依赖: $c) = d) < b)$ PETR 和 dense query based 方法都用了全局的cross attention, 对外参依赖变弱
- Dense query based 方法如果投影就是BEVFormer, 不投影就是无参数的方法

02. BEV感知框架-BEV空间融合

Transformer based

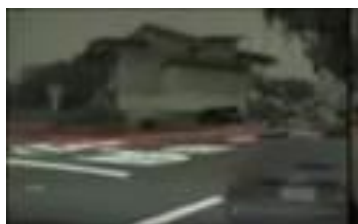
- Sparse query检测效果很好，效率更高，可针对不同感知任务，单独设置不同的queries，如目标检测和语义分割，单分割的话也是强行等价成dense query,比如 $625=25 \times 25 \rightarrow \text{cross attention} \rightarrow 200 \times 200$, 强行做的，没有实际物理意义，
- Dense query 有稠密的中间态表示，可以进一步进行特征提取和数据增强，而且更容易支持多种检测、分割、预测头，但效率较低且受感知范围的限制。
- 实际应用中，dense 的方法更通用，但是受到带宽、距离和分辨率限制，感知距离和精度难以平衡
- 车道线可以用sparse query做成实例级别，但是occupancy只能用dense query的方法

02. BEV感知框架-BEV空间融合

总结：BEV空间融合的主要方法

模型名称	空间融合	Decoder	稀疏/稠密	特征编码	参考点	使用NMS
LSS	自底向上	CNN	稠密	无	无	不涉及
BEVDet	自底向上	CNN	稠密	无	无	有
DETR3D	自顶向下	Transformer	稀疏	2D	Query映射	无
PETR	自顶向下	Transformer	稀疏	3D	无	无
BEVFormer	自顶向下	Transformer	稠密	2D	固定	无

02. BEV感知框架-BEV时序融合



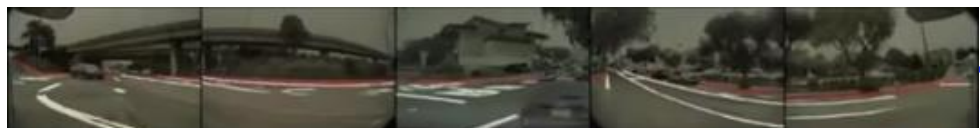
2D感知

后处理
(空间+时序)

左后相机

前视相机

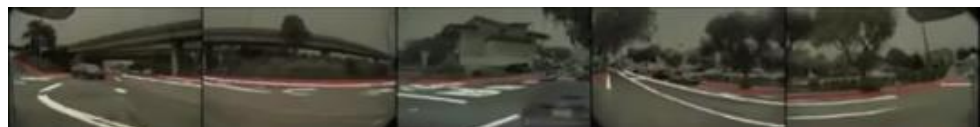
右后相机



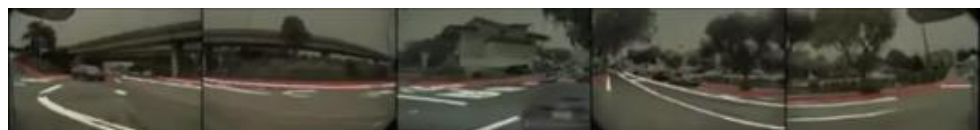
空间融合

后处理
(时序)

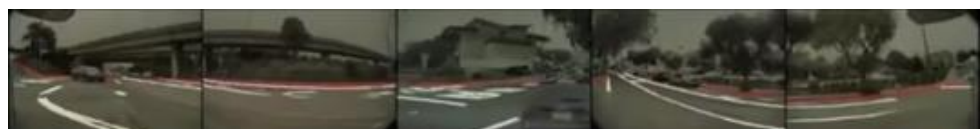
t=1



t=2



t=3

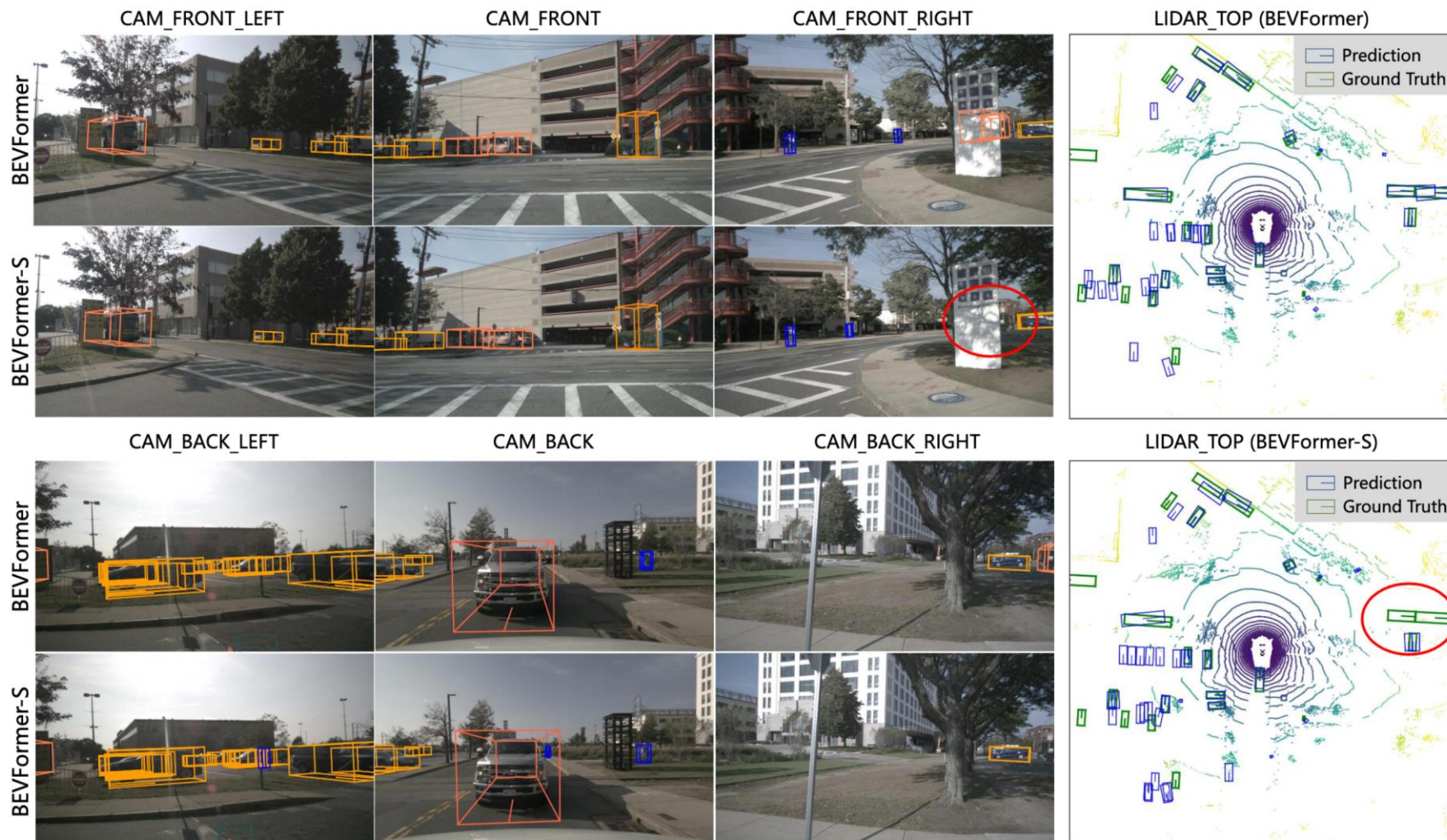


时序融合

接规控

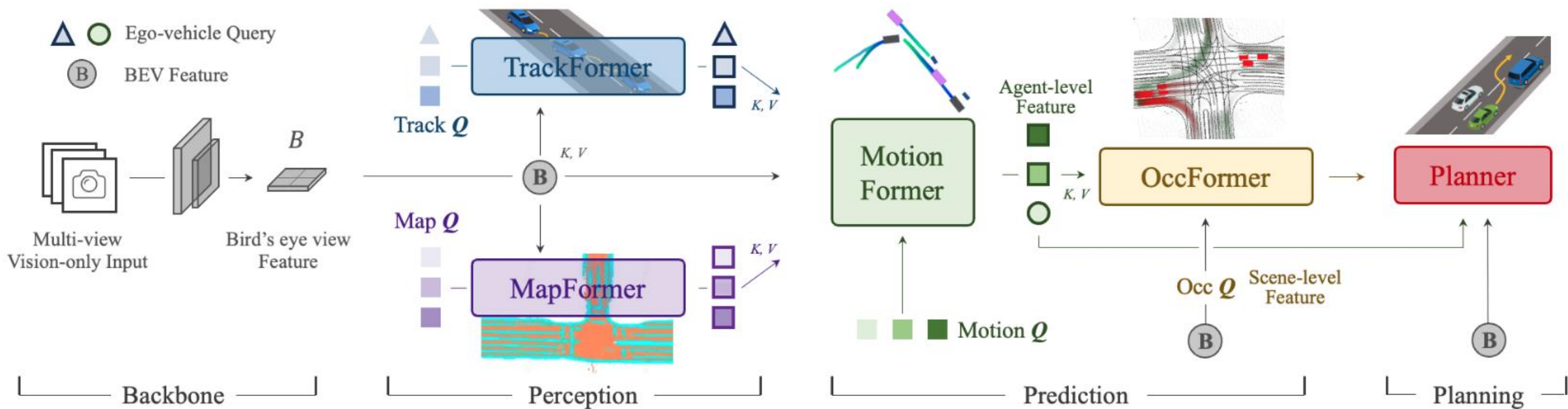
02. BEV感知框架-BEV时序融合

作用一：解决遮挡问题，利用时序信息提升感知的稳定性



02. BEV感知框架-BEV时序融合

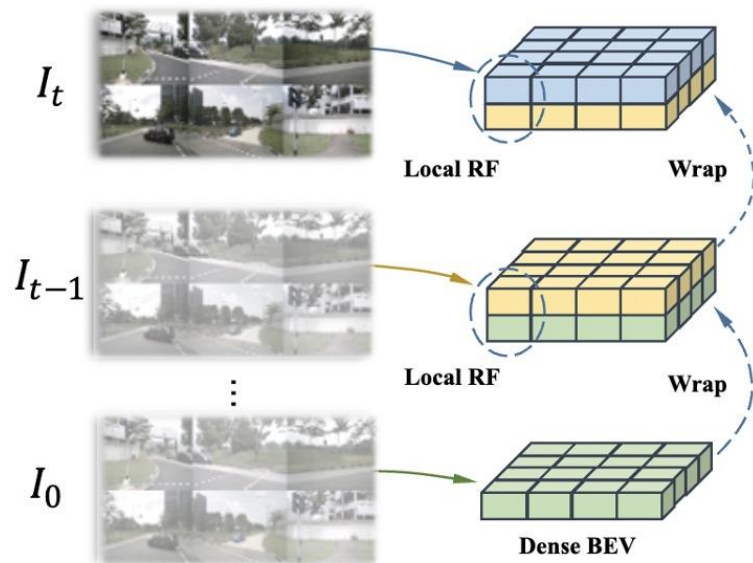
作用二：给tracking以及prediction相关的任务提供时序信息，以实现BEV下的端到端感知预测



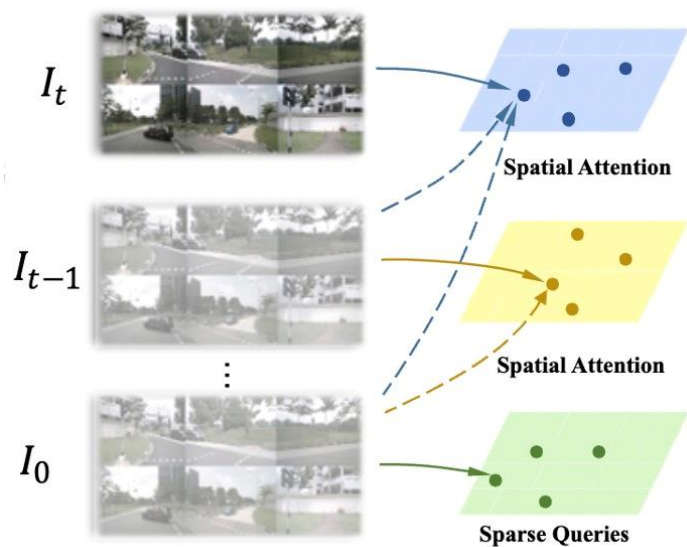
02. BEV感知框架-BEV时序融合

三种主流方法

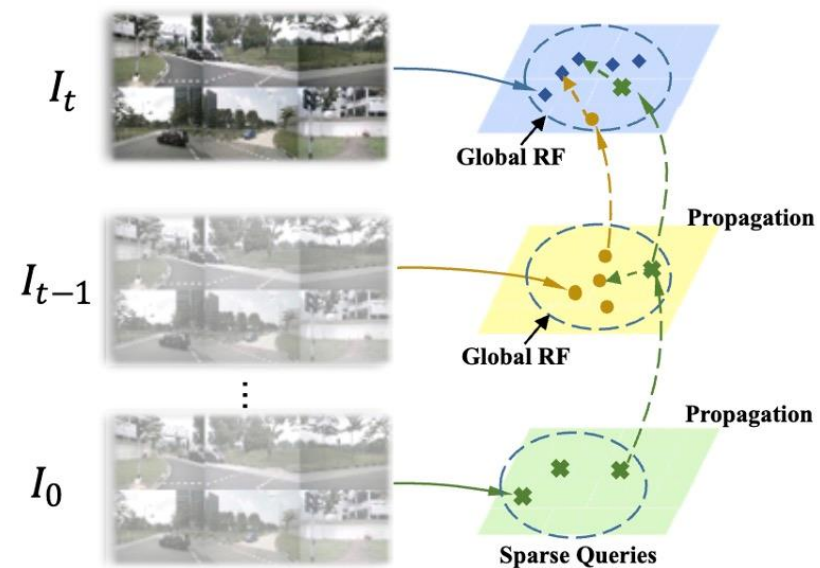
Dense BEV Method



Perspective Method

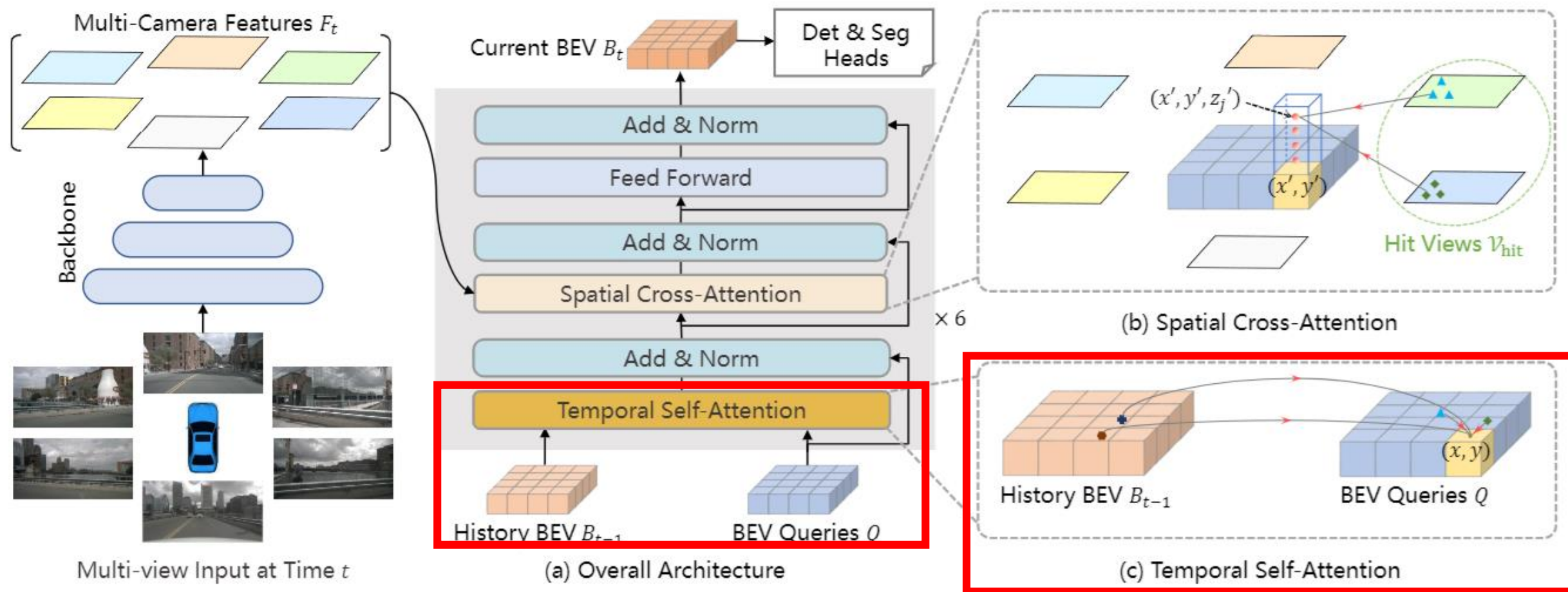


Object-centric Method



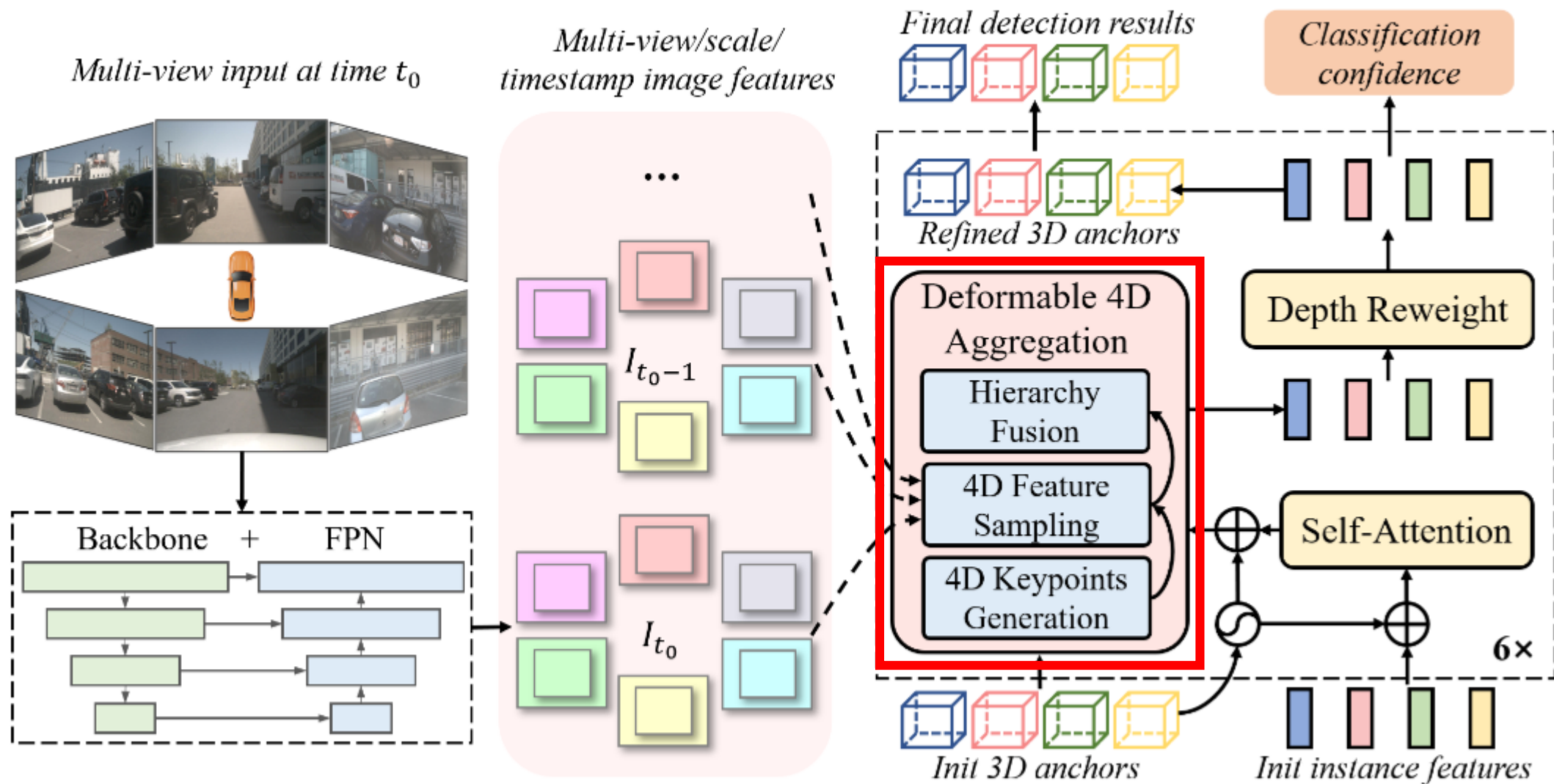
02. BEV感知框架-BEV时序融合

DenseBEV Method: BEVFormer



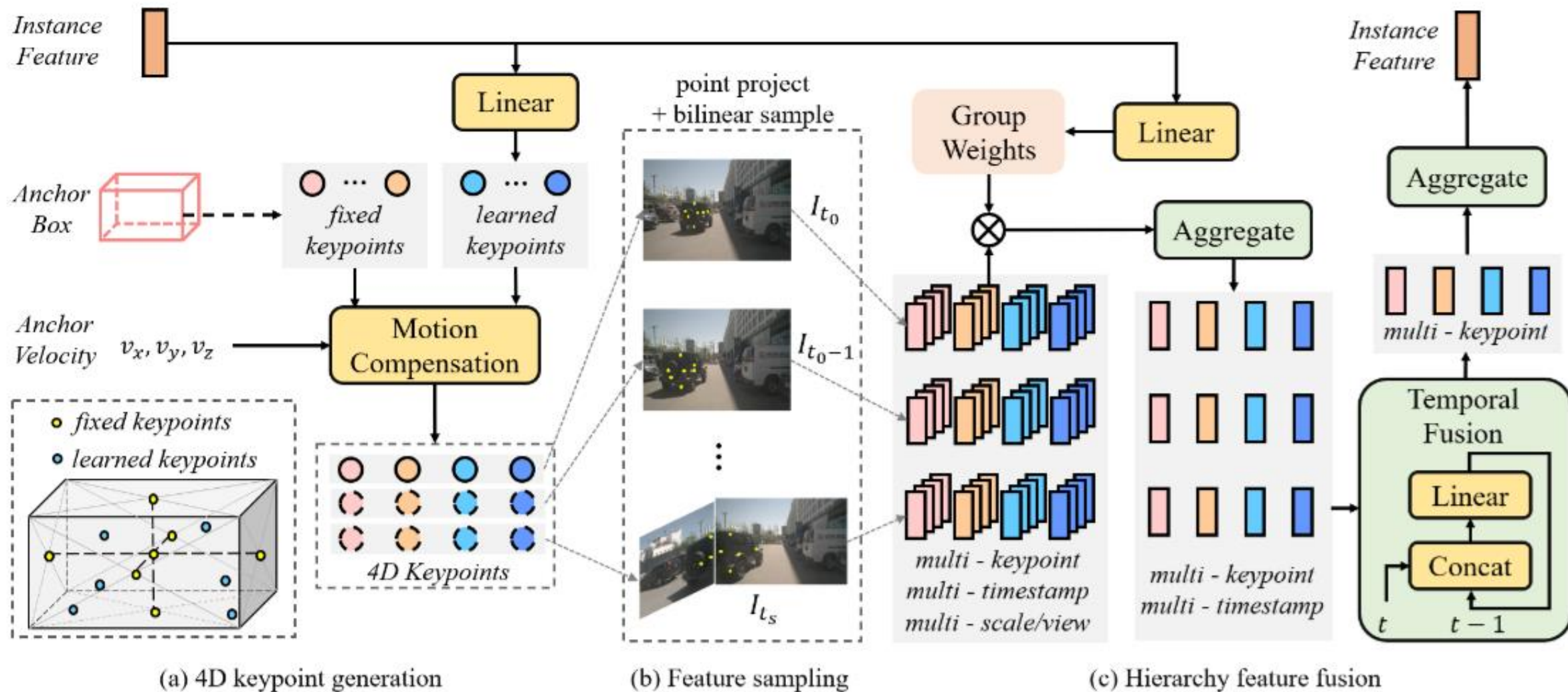
02. BEV感知框架-BEV时序融合

Perspective Method: Sparse4D



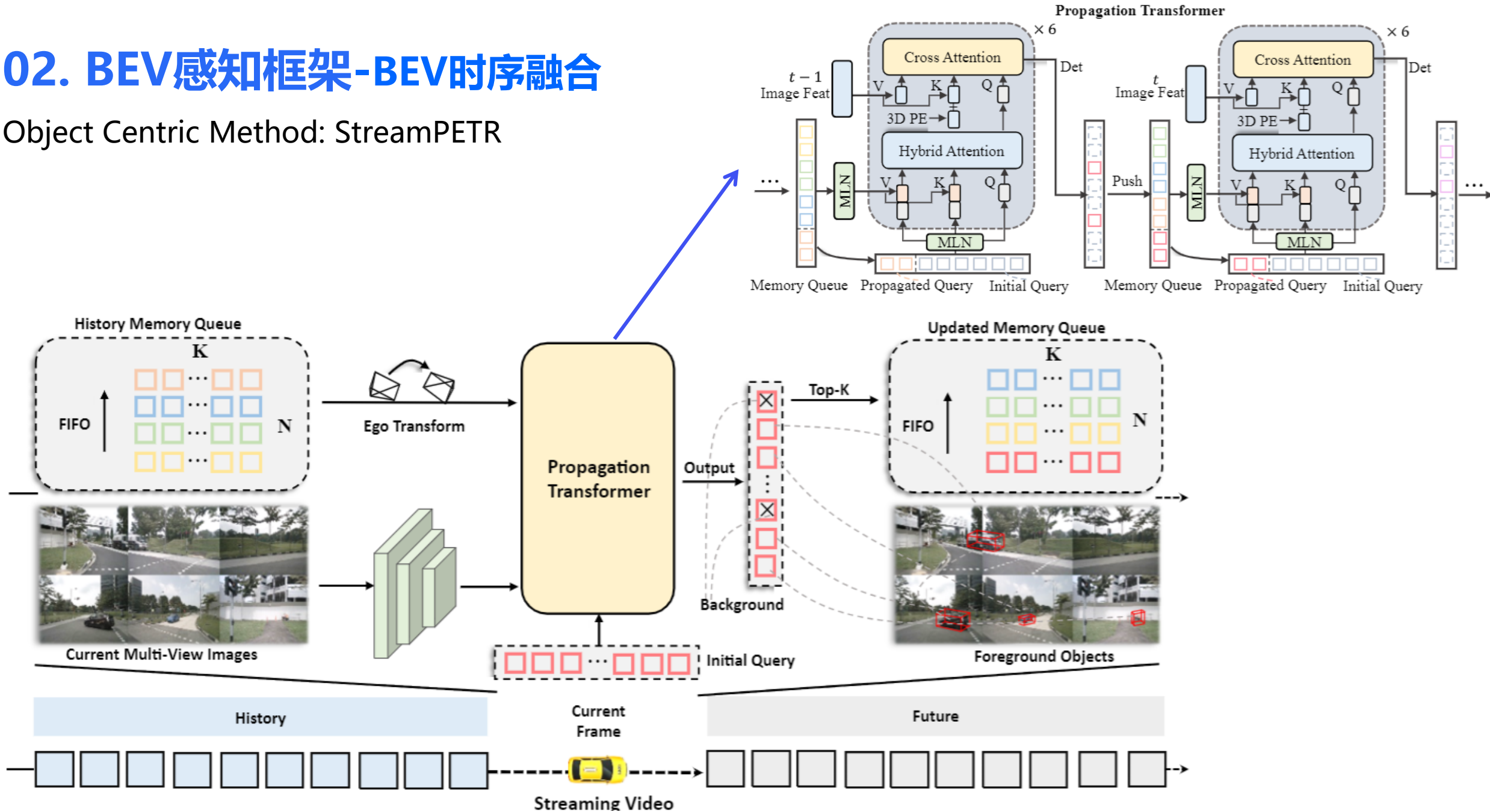
02. BEV感知框架-BEV时序融合

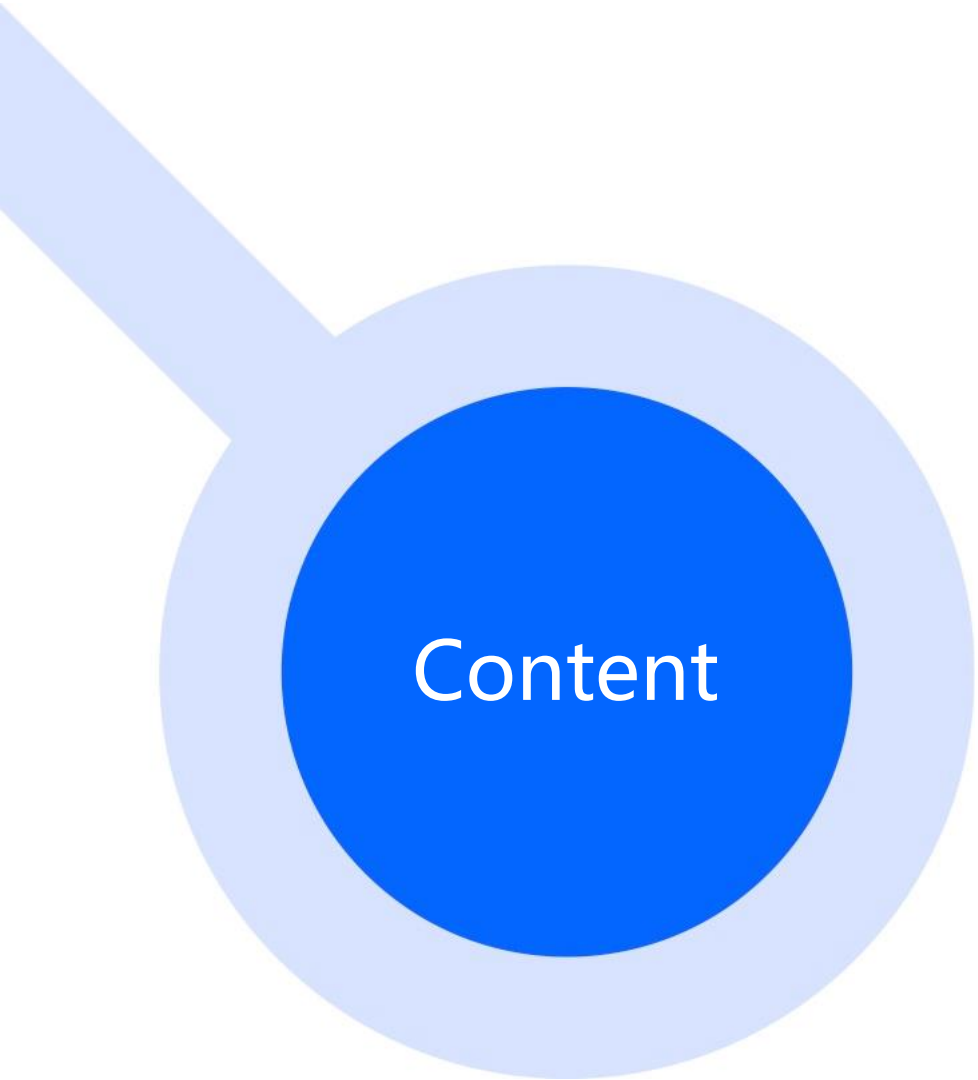
Perspective Method: Sparse4D



02. BEV感知框架-BEV时序融合

Object Centric Method: StreamPETR





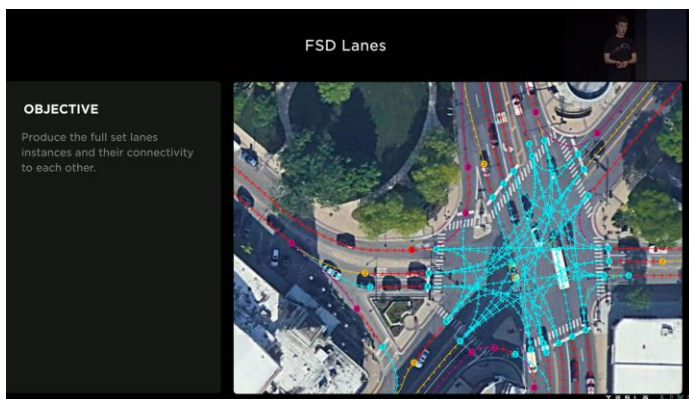
Content

03. BEV感知应用

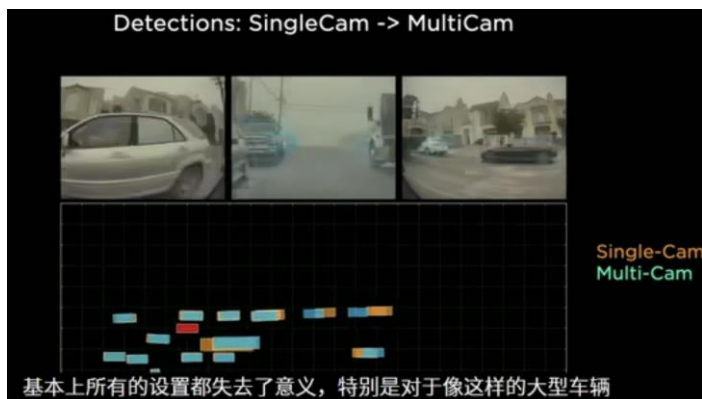
- (a) 静态结构化感知
- (b) 动态结构化感知
- (c) Occupancy

03. BEV感知应用

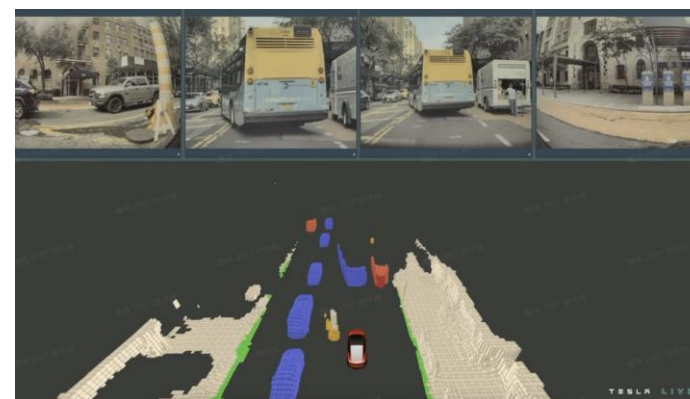
■ 常用的三种不同任务的感知网络



静态感知（路面要素、锥桶）

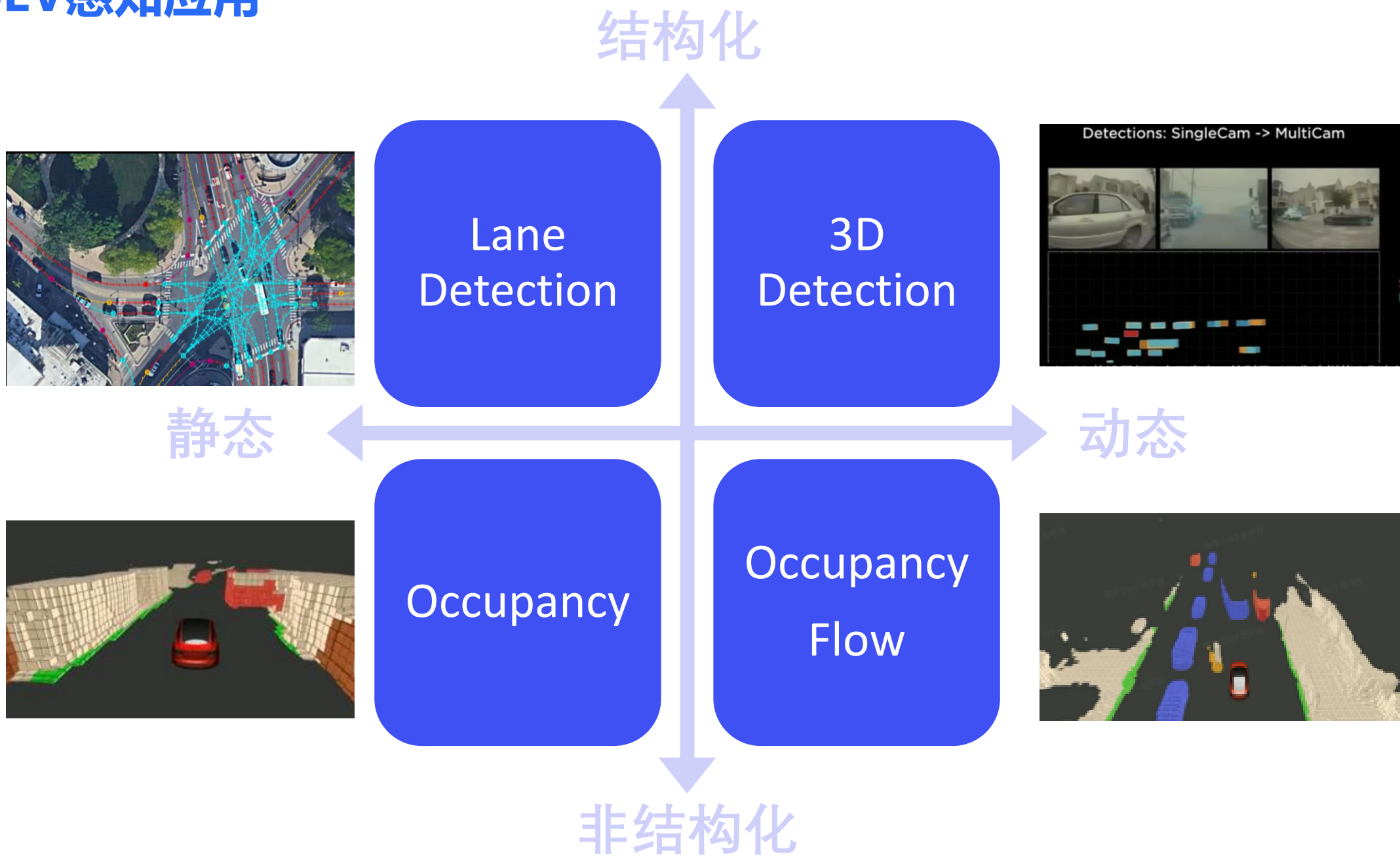


动态感知（车辆、行人）



通用障碍物感知（Occupancy）

03. BEV感知应用

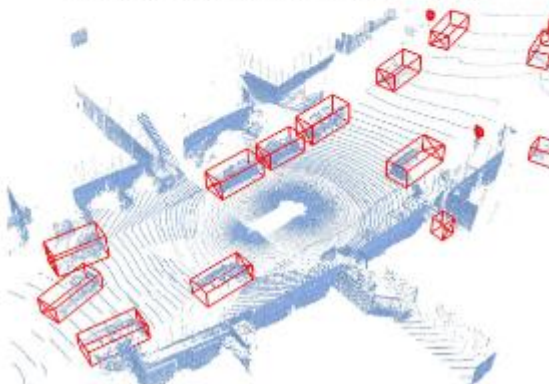


03. BEV感知应用-动态结构化感知

3D Object Detection



3D Object Detection from Image



3D Object Detection from Point Cloud

位置, 大小, 朝向

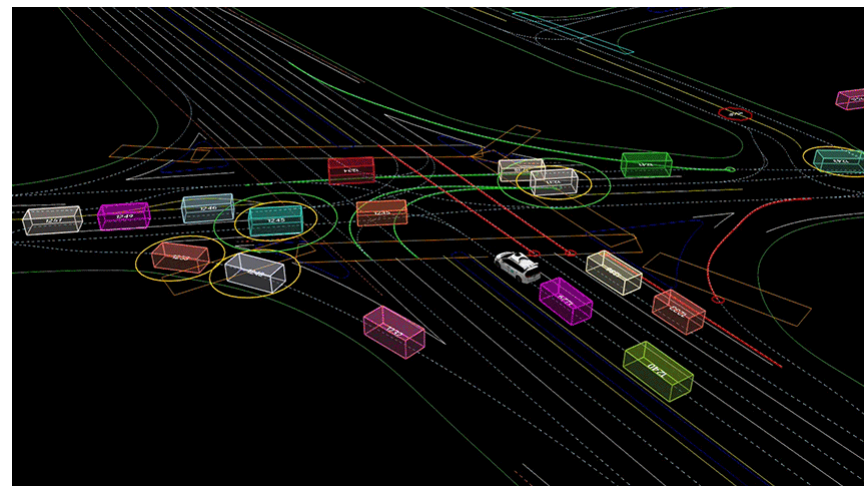
$\mathbf{r}$ Position

$\frac{d\mathbf{r}}{dt} = \dot{\mathbf{r}}$ Velocity (speed)

$\frac{d^2\mathbf{r}}{dt^2} = \ddot{\mathbf{r}}$ Acceleration

$\frac{d^3\mathbf{r}}{dt^3} = \dddot{\mathbf{r}}$ Jerk

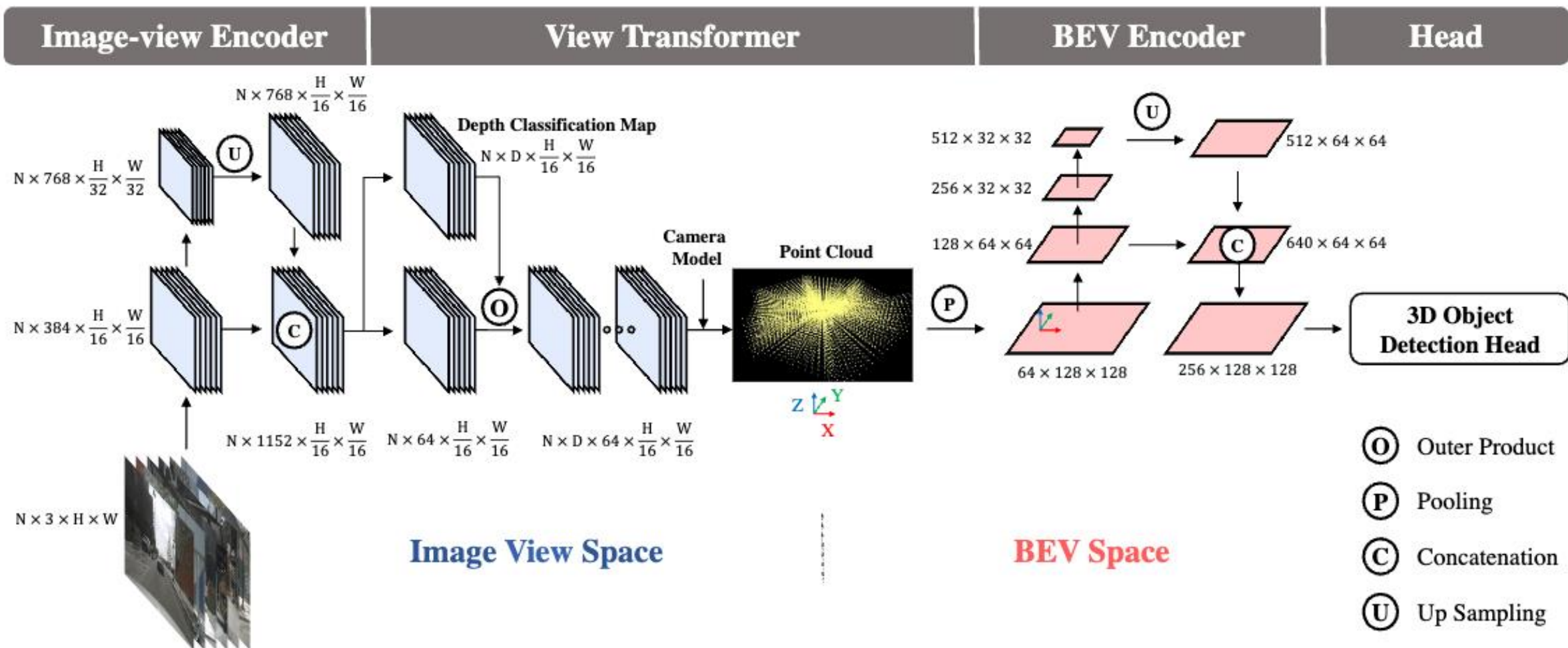
速度, 加速度, 加加速度



轨迹预测

03. BEV感知应用-动态结构化感知

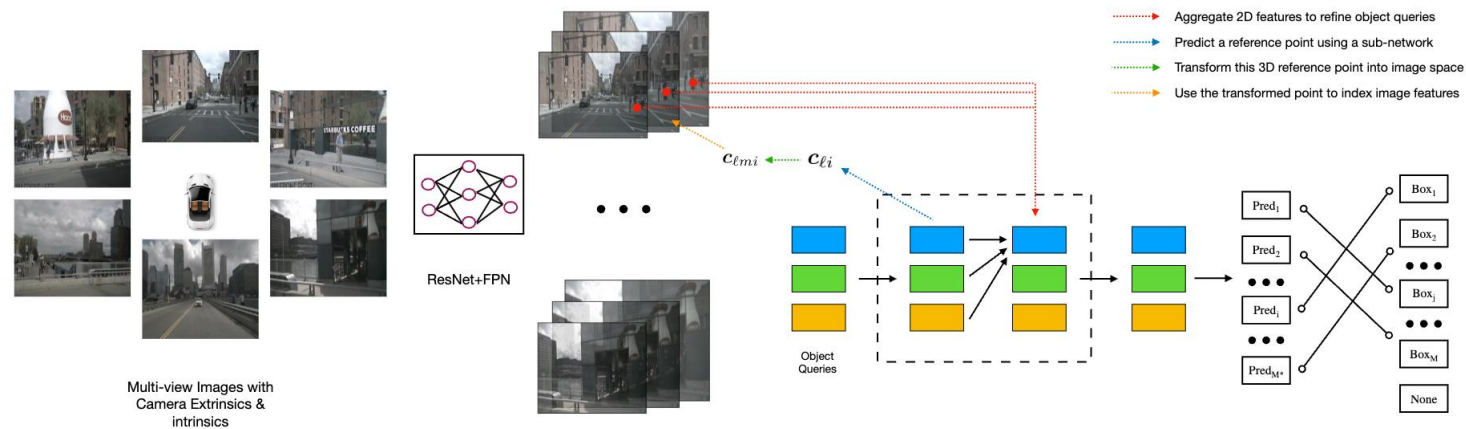
3D Detection 的代表方法-BEVDet: Heatmap + NMS



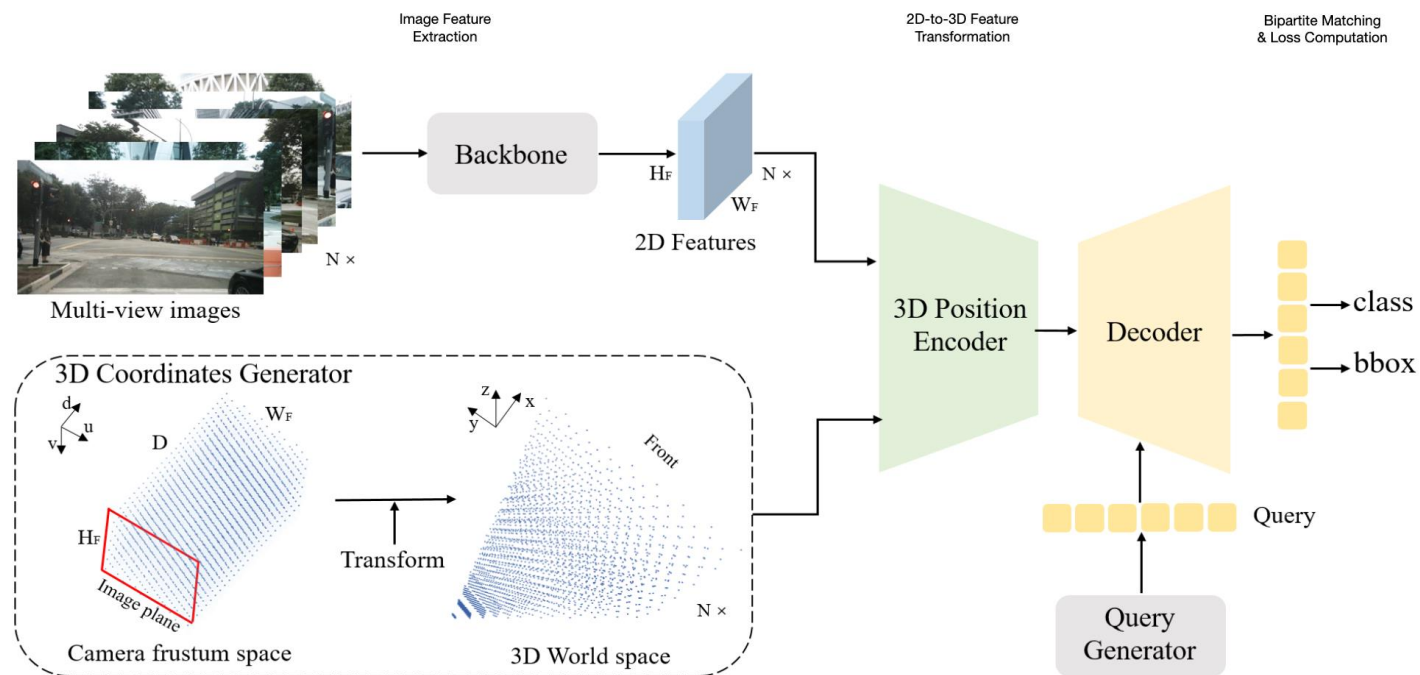
03. BEV感知应用-动态结构化感知

3D Detection 的代表方法-DETR3D、PETR

DETR3D

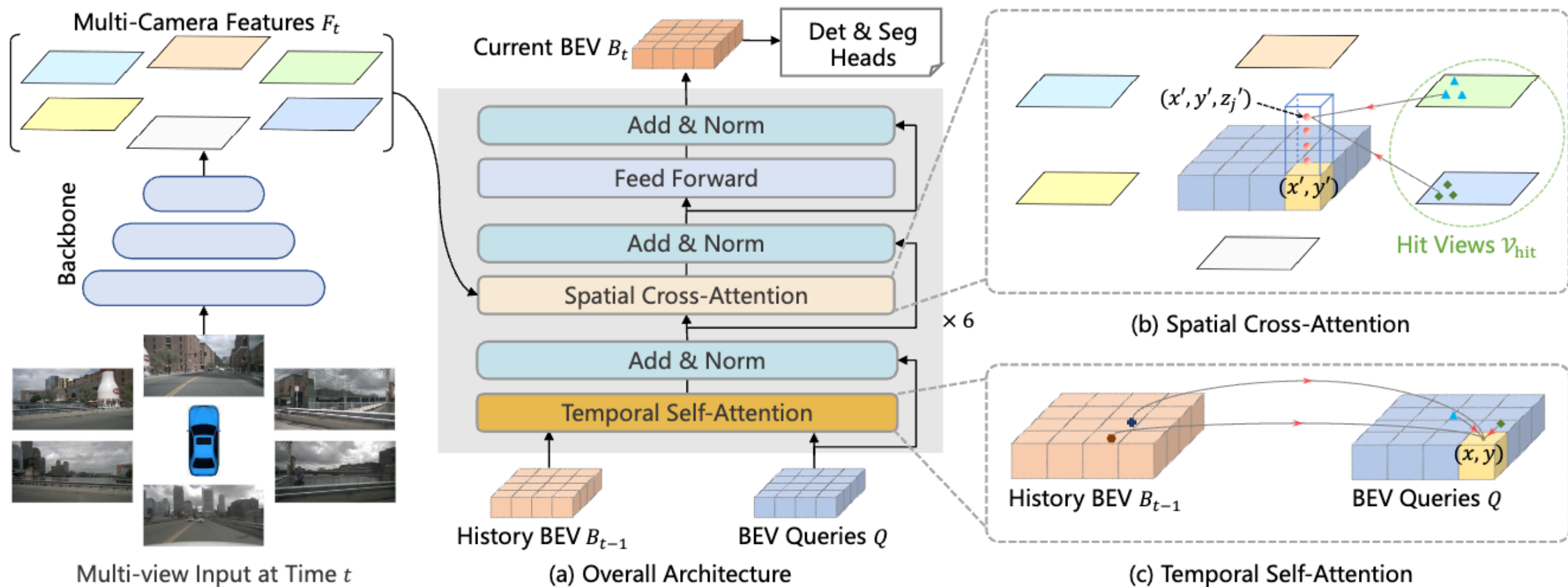


PETR

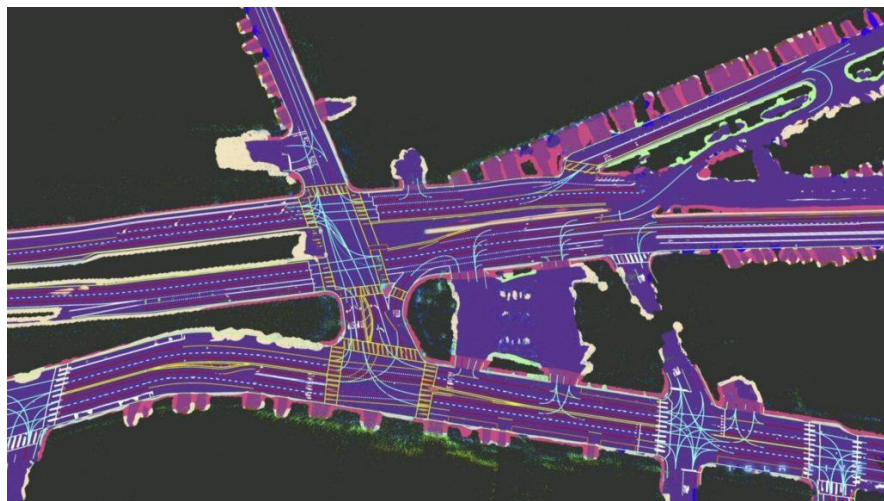


03. BEV感知应用-动态结构化感知

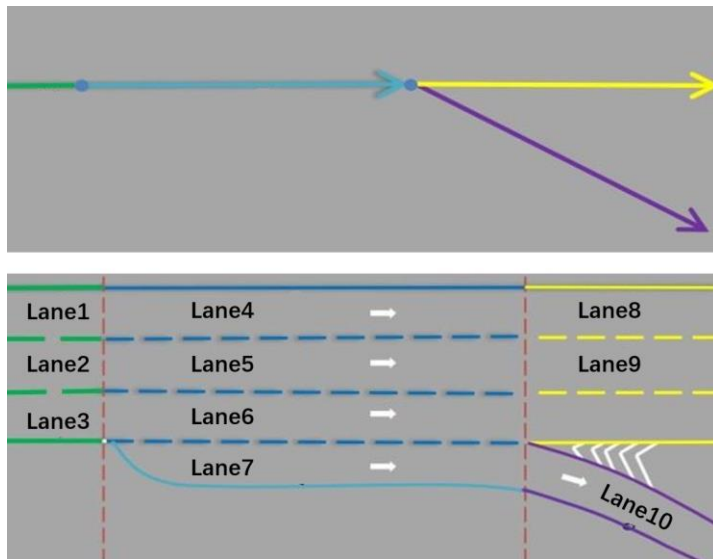
3D Detection 的代表方法-BEVFormer



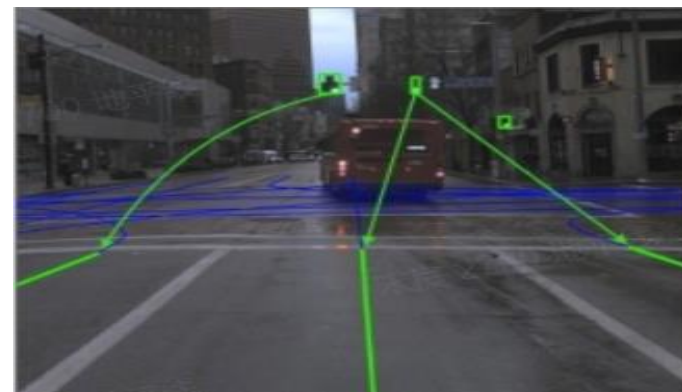
03. BEV感知应用-静态结构化感知



道路环境 几何层



道路链接关系 逻辑层

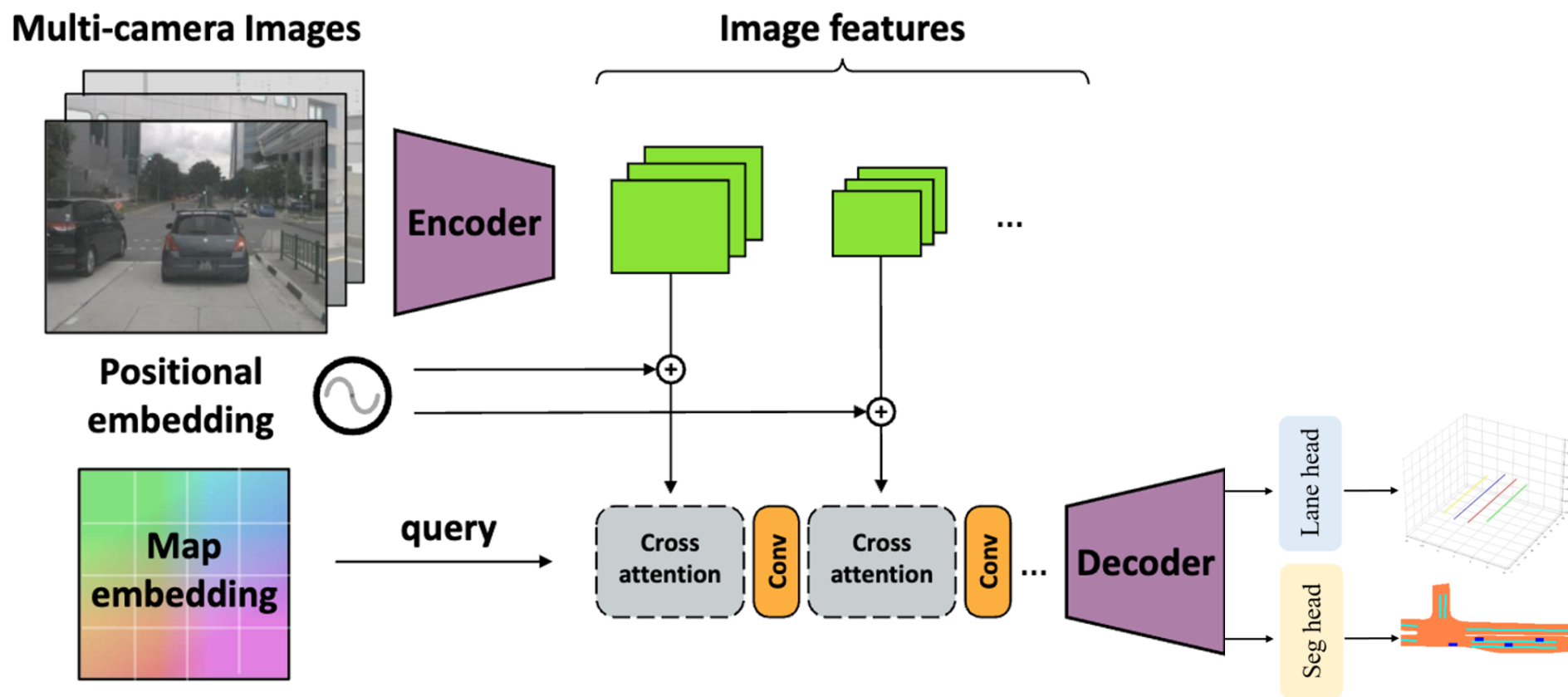


红绿灯理解 逻辑层

03. BEV感知应用-静态结构化感知

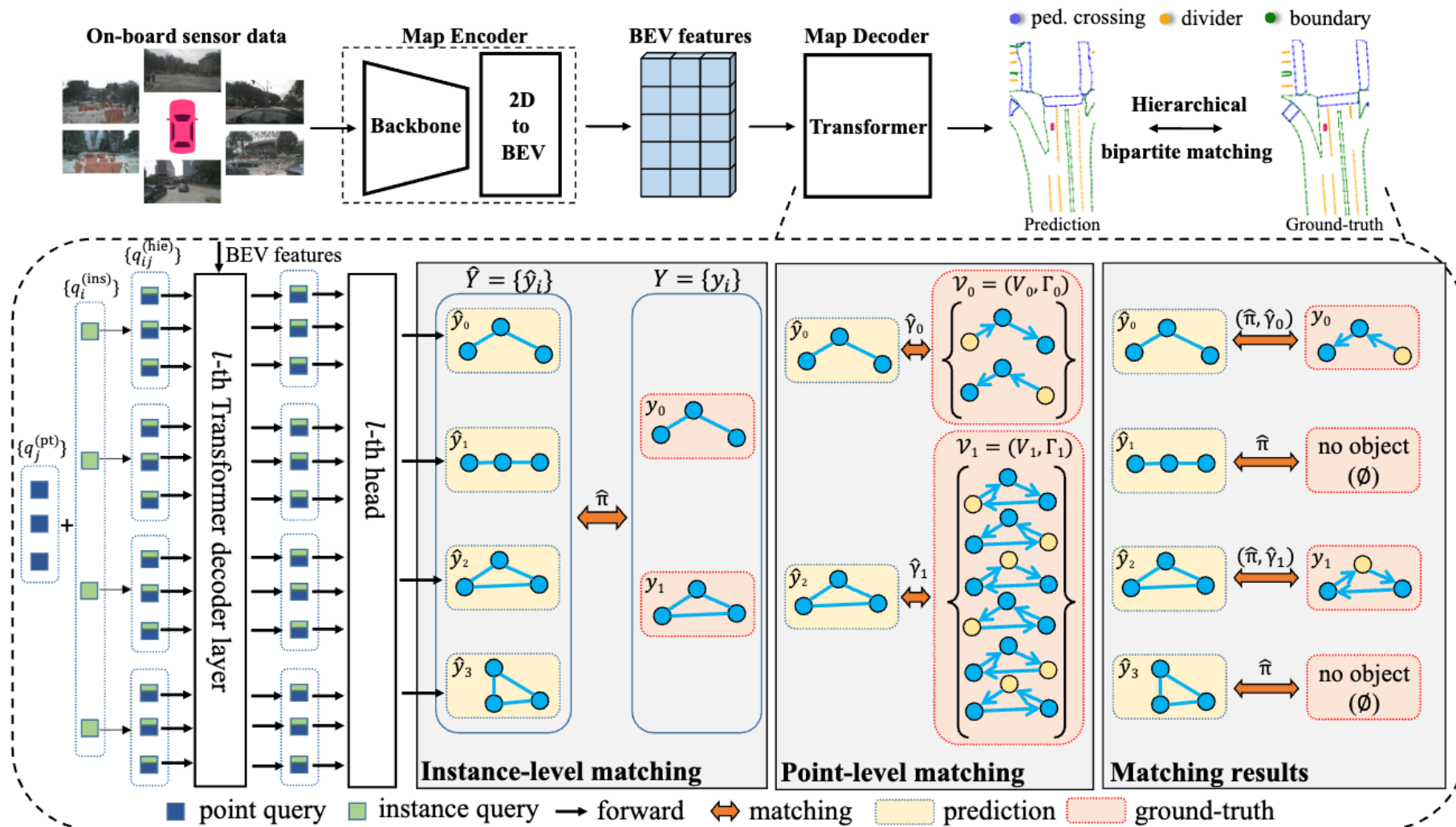
BEV-SEG

- 输出为BEV平面的2D像素分割



03. BEV感知应用-静态结构化感知

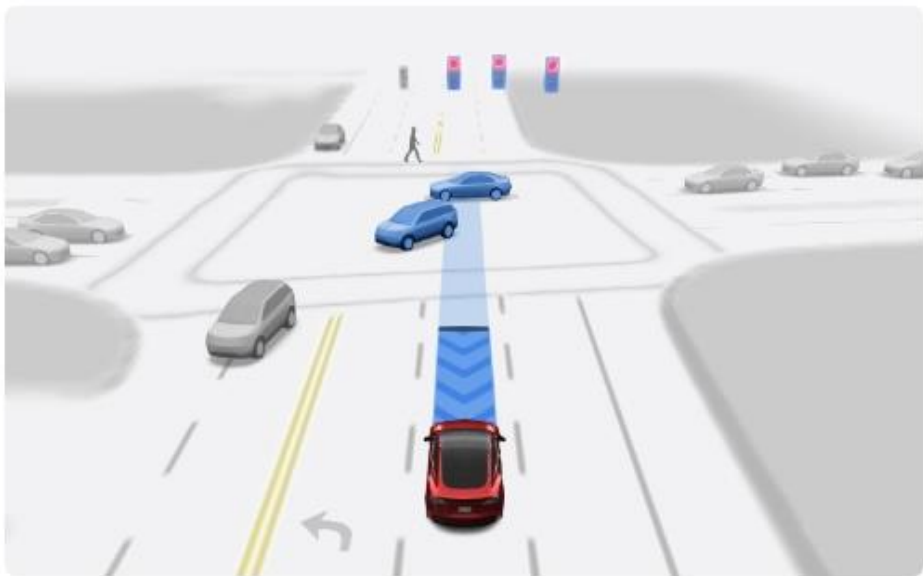
MapTR



03. BEV感知应用-静态结构化感知

引入long-range roadway信息

FSD Beta v11.4.1



Available in the United States

Models:

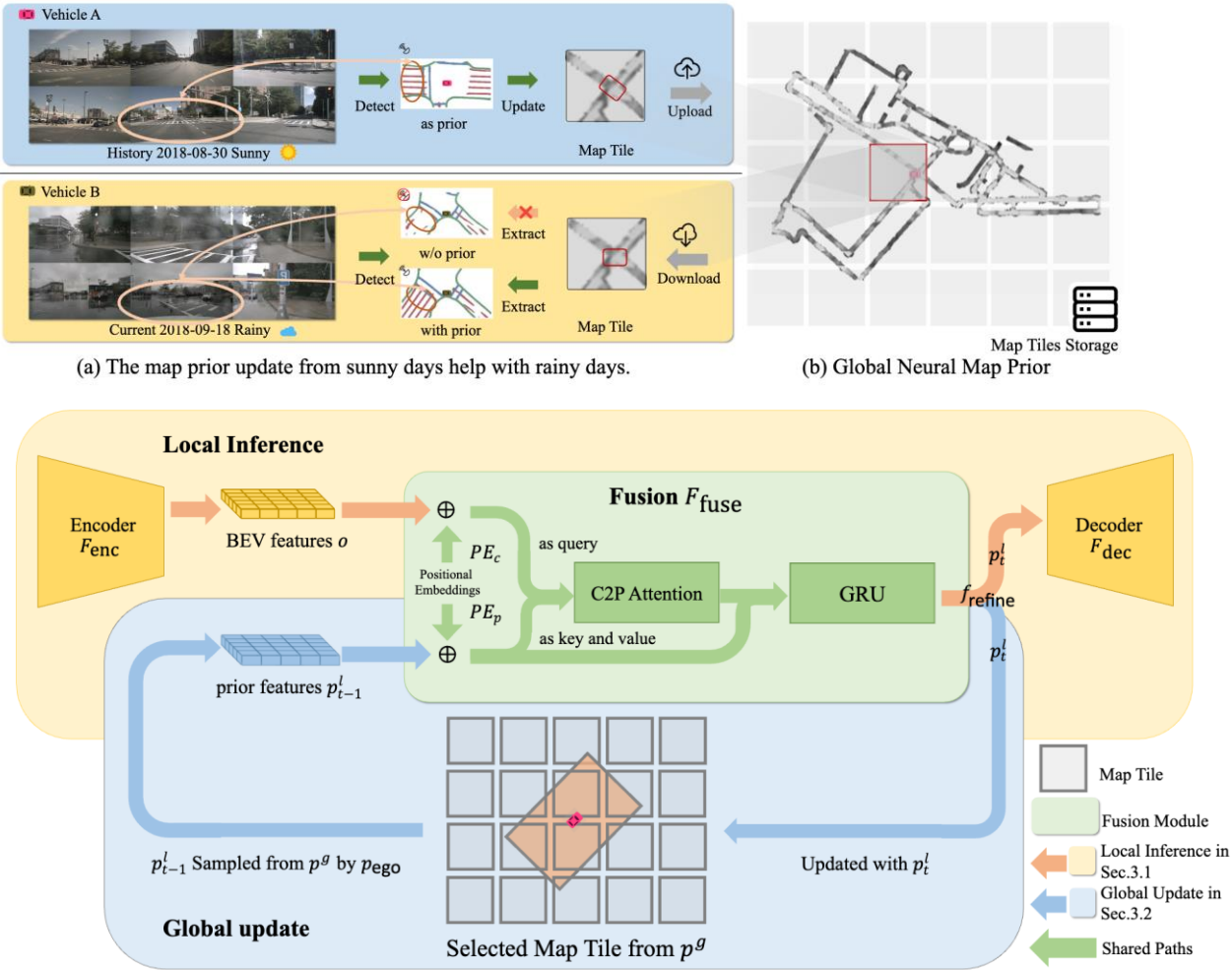
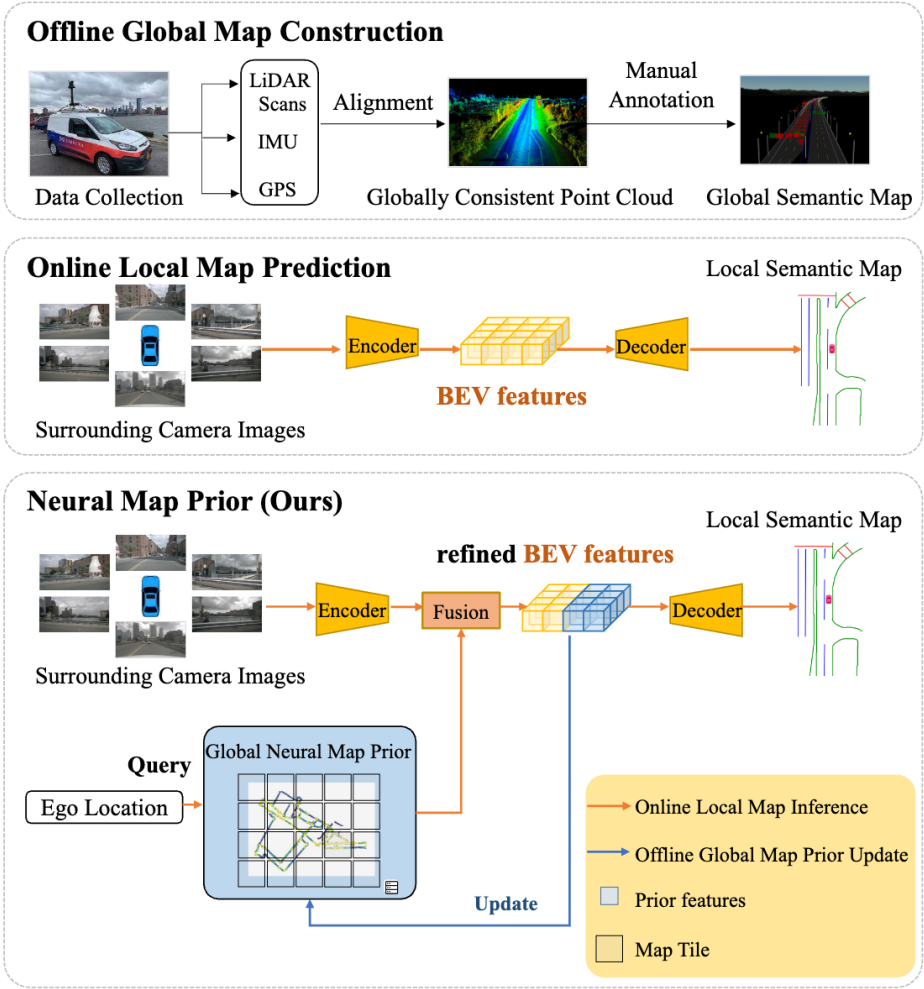
S 3 X Y

- Improved control through turns, and smoothness in general, by improving geometry, curvature, position, type and topology of lanes, lines, road edges, and restricted space. Among other improvements, the perception of lanes in city streets improved by 36%, forks improved by 44%, merges improved by 27% and turns improved by 16%, due to a bigger and cleaner training set and updated lane-guidance module.

- Added lane-guidance inputs to the Occupancy Network to improve detections of long-range roadway features, resulting in a 16% reduction in false negative median detections.

03. BEV感知应用-静态结构化感知

引入long-range roadway信息

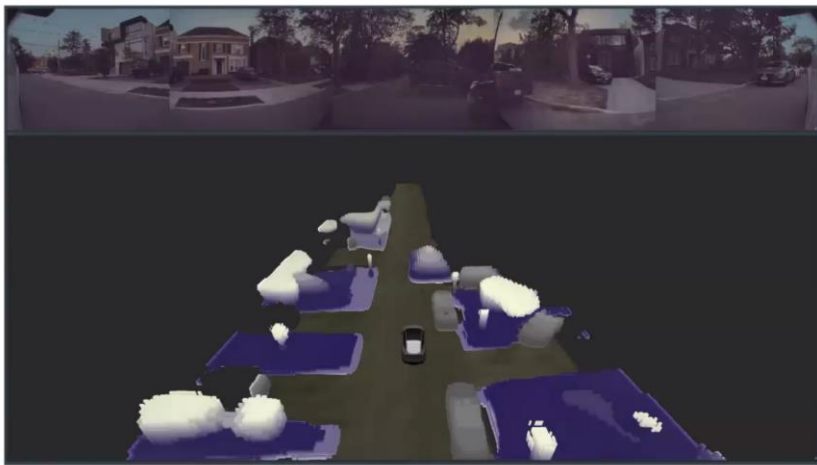


03. BEV感知应用-Occupancy

Tesla CVPR2022 WorkShop

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 AIDAY 2022

Properties

- Volumetric Occupancy
- Multi-Camera & Video Context
- Persistent Through Occlusions
- Occupancy Semantics
- Occupancy Flow
- Resolution Where It Matters
- Efficient Memory and Compute
- Runs in ~10 Milliseconds

TESLA LIVE

03. BEV感知应用-Occupancy

学术界Occupancy

Occupancy: 三维语义占据栅格最早来自MonoScene: Monocular 3D Semantic Scene, 语义补全任务

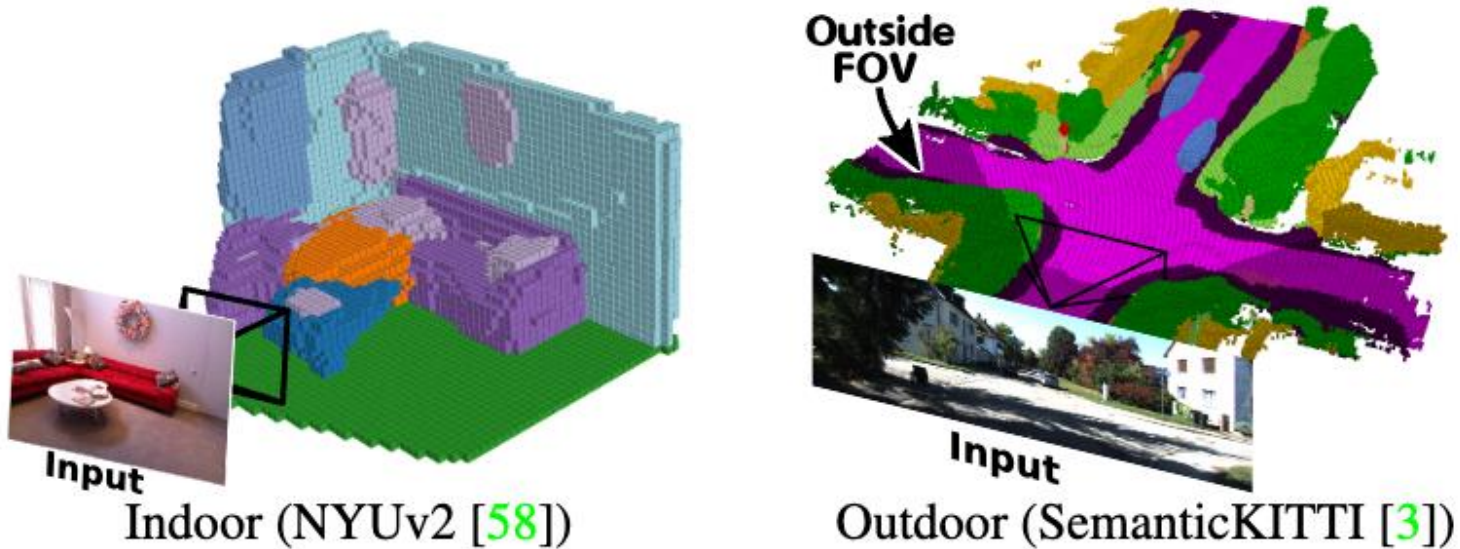
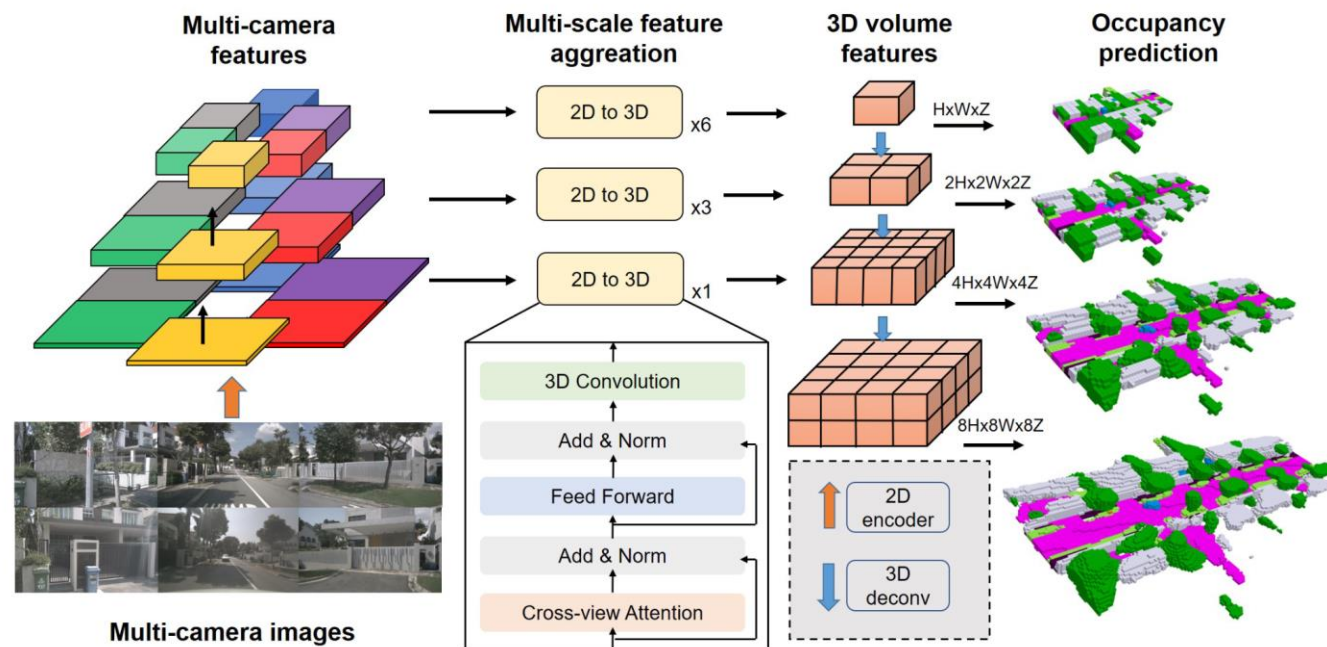
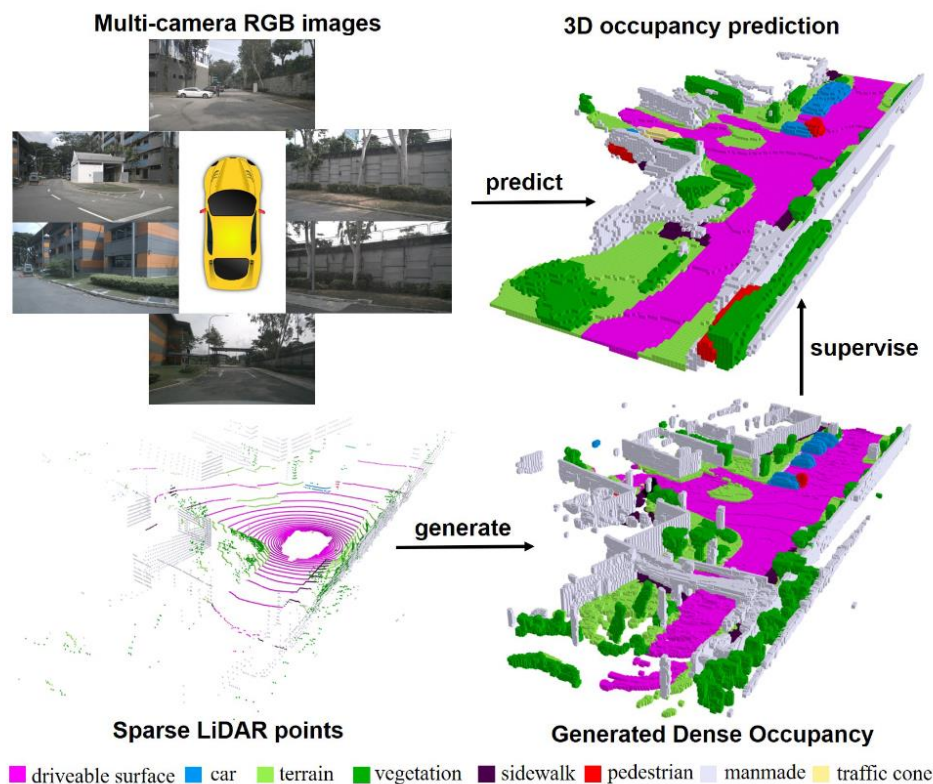


Figure 1. **RGB Semantic Scene Completion with MonoScene.** Our framework infers dense semantic scenes, hallucinating scenery outside the field of view of the image (dark voxels, right).

03. BEV感知应用-Occupancy

学术界Occupancy

- 将空间划分0/1两种类别的voxel
- Occupancy可以看做 深度估计网络在BEV感知下的形式
- 分辨率主要是问题

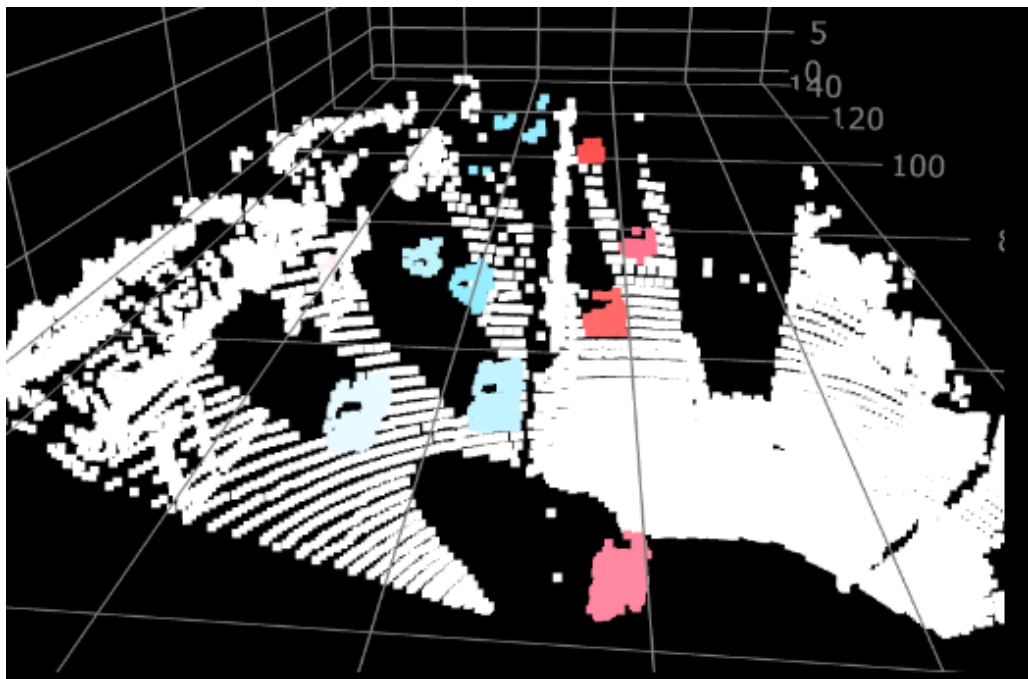


03. BEV感知应用-Occupancy

Occupancy Flow /SceneFlow (场景流)



Emerged Forward Flow



EmerNeRF Scene Flow可视化

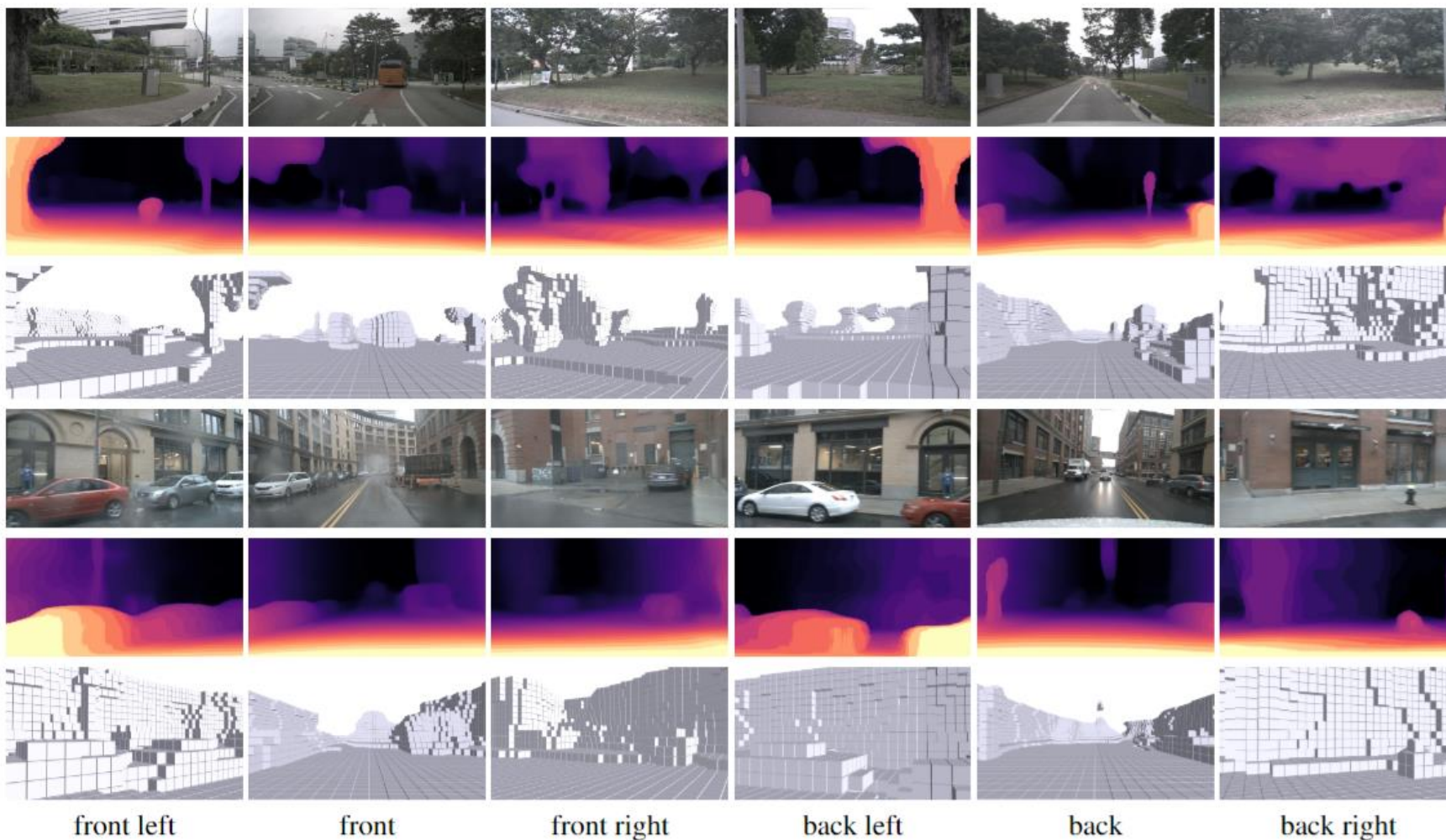


Tesla Occupancy Flow (CVPR2022)

03. BEV感知应用-Occupancy

基于视觉occupancy的本质是单目深度估计从2D感知(PV)到3D感知(BEV)的自然延伸

难点：大量，准确，丰富的训练数据（和深度估计一样）



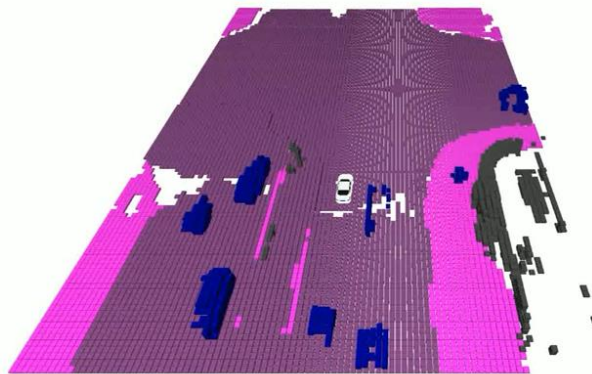
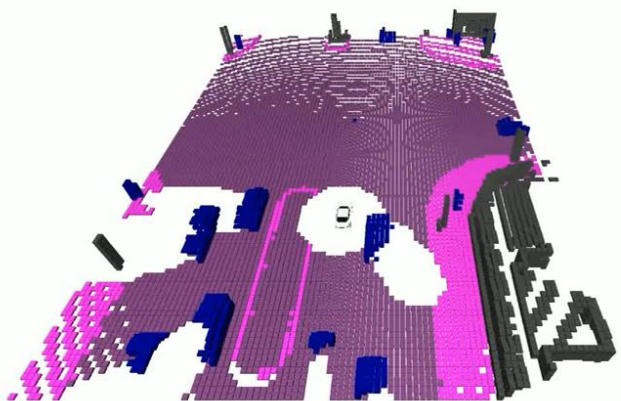
03. BEV感知应用-Occupancy

CARLA仿真数据集构建Occupancy GT



gt

pred



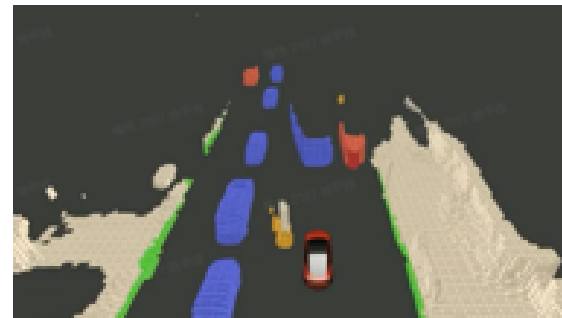
03. BEV感知应用-总结



静态感知 (路面要素、锥桶)



动态感知 (车辆、行人)



通用障碍物感知 (Occupancy)



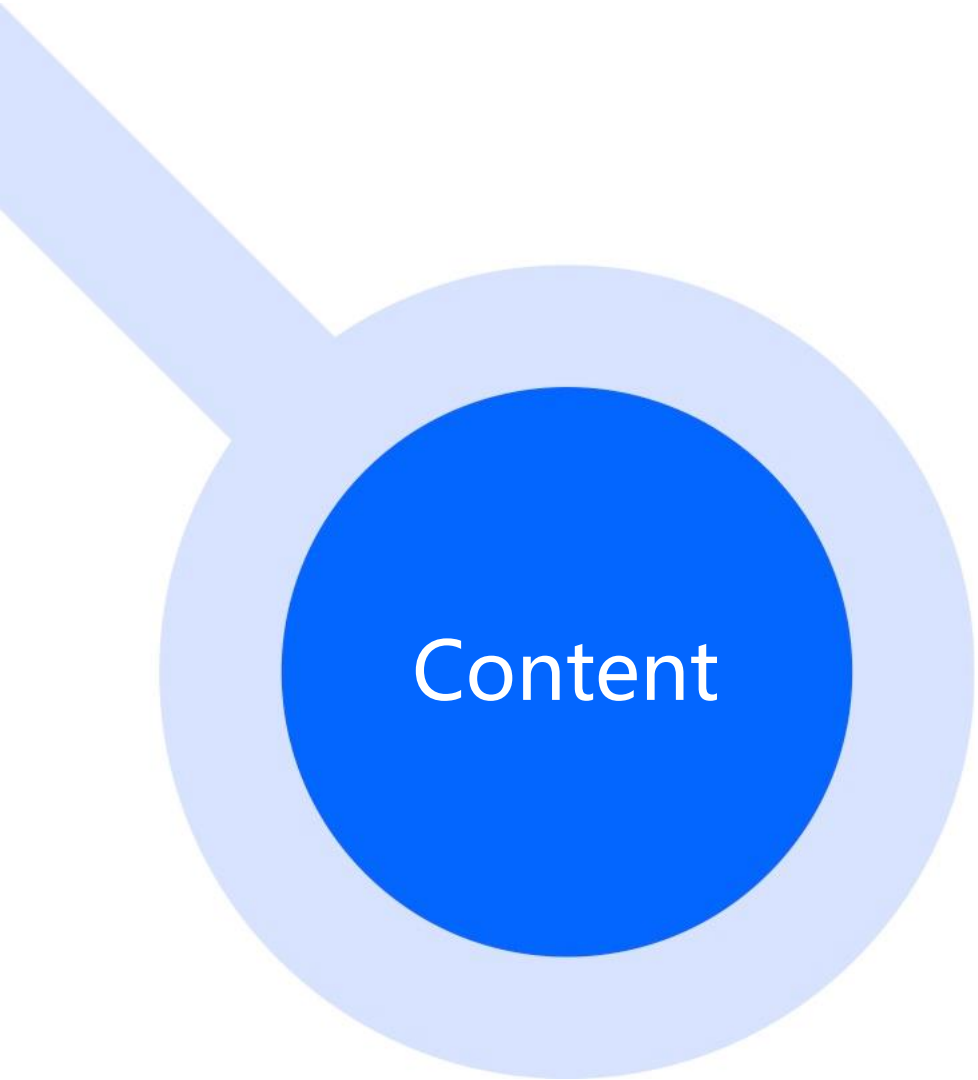
白名单感知

- 稀疏表达
- 容易建模 (速度, 加速度, jerk)
- 结合语义理解



非名单通用感知

- 稠密表达
- 包罗万象
- 感知范围精度平衡



Content

上半节

- 01. 自动驾驶感知技术发展
- 02. BEV感知框架
- 03. BEV感知应用

下半节

- 04. 云端自动标注
- 05. 仿真系统
- 06. 端到端自动驾驶