OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Perception - Vision

Dr. Hongyang Li

Shanghai AI Lab
Mar 27 2024

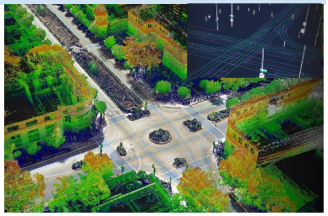
OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

BEV Perception

One Page Summary

BEV感知 | Introduction



Front/perspective view
前视/放射视角

⇨

Bird's eye view BEV
鸟瞰图视角

任务：given环视camera input, 生成3D框

1. 多传感器融合背景下的一种**视角表达形式**，统一在相同视角下（BEV）表达；BEV本身**并不是新热点**

2. **BEV感知**：如何创新算法/pipeline, 将来自不同sensor的feature 最优表达，**是热点、是趋势**

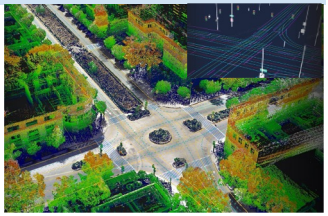
3. BEV视角下**优势：**
   a. 没有scale/occlusion问题
   b. 有利于后续规划控制模块开发

有效解决检测中：
- 遮挡
- 侧向来车

4. BEV感知**核心/难点：**
   a. 视角变化后的深度估计
   b. GT数据获取
   c. 不同传感器特征融合
   d. 不依赖相机参数/Domain adaptation

One Page Summary

BEV感知 | Introduction



Front/perspective view
前视/放射视角

⇨

Bird's eye view BEV
鸟瞰图视角

任务：given环视camera input, 生成3D框

1. 多传感器融合背景下的一种**视角表达形式**，统一在相同视角下（BEV）表达；BEV本身**并不是新热点**

2. **BEV感知**：如何创新算法/pipeline, 将来自不同sensor的feature 最优表达，**是热点、是趋势**

3. BEV视角下**优势：**
   a. 没有scale/occlusion问题
   b. 有利于后续规划控制模块开发

有效解决检测中：
- 遮挡
- 侧向来车

4. BEV感知**核心/难点：**
   a. 视角变化后的深度估计
   b. GT数据获取
   c. 不同传感器特征融合
   d. 不依赖相机参数/Domain adaptation

BEV感知 | Motivation

Significance

- 相机感知性能究竟能否beat LiDAR?
  - 现在是AP 40 vs 70 差距
  - 如果能，大幅降低量产部署成本
  - 如果不能，why? where's the gap?

Space

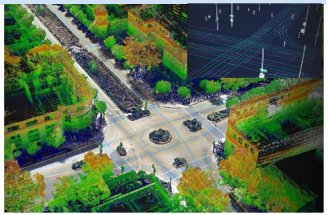
- 核心是相机feature如何描述3D空间信息
- 三维场景重建、认知决策、协同感知

Readiness

- High-quality Benchmark **(Waymo/nuScenes)**
  - large-scale, diverse
- Waymo Challenge 2022
  - let community work on together
- Inspired by Transformer/ViT/MAE
- 工业界/学术界各种方案、尝试 **重点**

One Page Summary

BEV感知 | Introduction



Front/perspective view
前视/放射视角

⇨

Bird's eye view BEV
鸟瞰图视角

任务：given环视camera input, 生成3D框

1. 多传感器融合背景下的一种**视角表达形式**，统一在相同视角下（BEV）表达；BEV本身**并不是新热点**

2. **BEV感知**：如何创新算法/pipeline, 将来自不同sensor的feature 最优表达，**是热点、是趋势**

3. BEV视角下**优势：**
   a. 没有scale/occlusion问题
   b. 有利于后续规划控制模块开发

有效解决检测中：
- 遮挡
- 侧向来车

4. BEV感知**核心/难点：**
   a. 视角变化后的深度估计
   b. GT数据获取
   c. 不同传感器特征融合
   d. 不依赖相机参数/Domain adaptation

BEV感知 | Motivation

Significance

- 相机感知性能究竟能否beat LiDAR?
  - 现在是AP 40 vs 70 差距
  - 如果能，大幅降低量产部署成本
  - 如果不能，why? where's the gap?

Space

- 核心是相机feature如何描述3D空间信息
- 三维场景重建、认知决策、协同感知

Readiness

- High-quality Benchmark **(Waymo/nuScenes)**
  - large-scale, diverse
- Waymo Challenge 2022
  - let community work on together
- Inspired by Transformer/ViT/MAE
- 工业界/学术界各种方案、尝试 **重点**

Why us?

BEVFormer

arXiv:2203.17270

Unwatch 57 | Fork 67 | Starred 727

- 受到Tesla方案启发，改进版Transformer
- 加入时序信息，更好融合历史特征

nuScenes Top 1 **重点**

BEVFormer++

- 尝试各种backbone/head/时序/ensemble
- Pushing Performance to the Next Extreme

15% 34% 56%

Waymo 1st place **重点**

LidarSeg Track as well

Future Work

- 产业界量化部署/MDC/J5等
- 3D真值获取/新融合方案

BEV Perception

目录

1 BEV感知：背景与动机

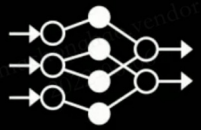
- Camera-only Setting
- Core issues in 3D Perception
- BEV Perception: Next Paradigm for AD

Problem Setup: Multi-Camera 3D Perception

INPUT

8 Cameras

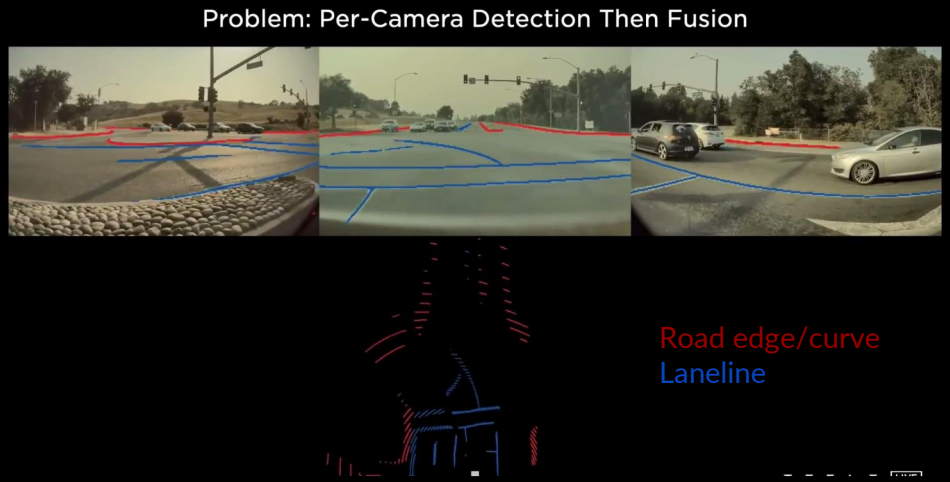
AI/DL/NN

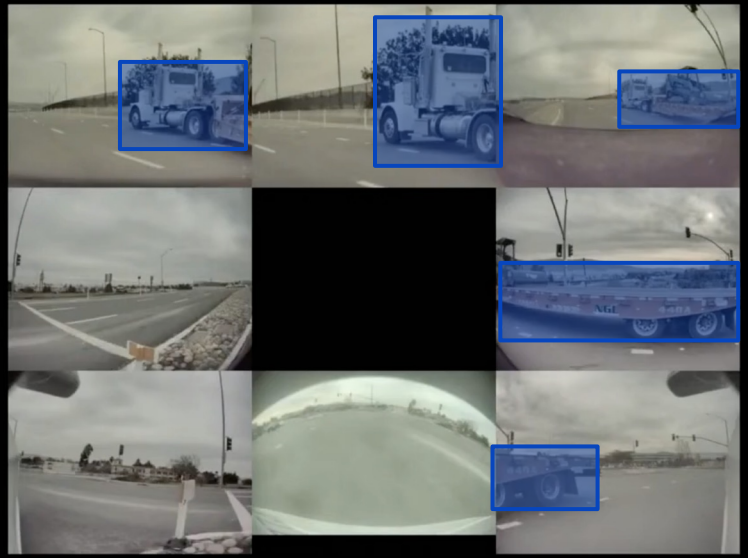


OUTPUT

3-Dimensional “Vector Space”

Problem: Per-camera detection then fusion (nV)





Traditional method:

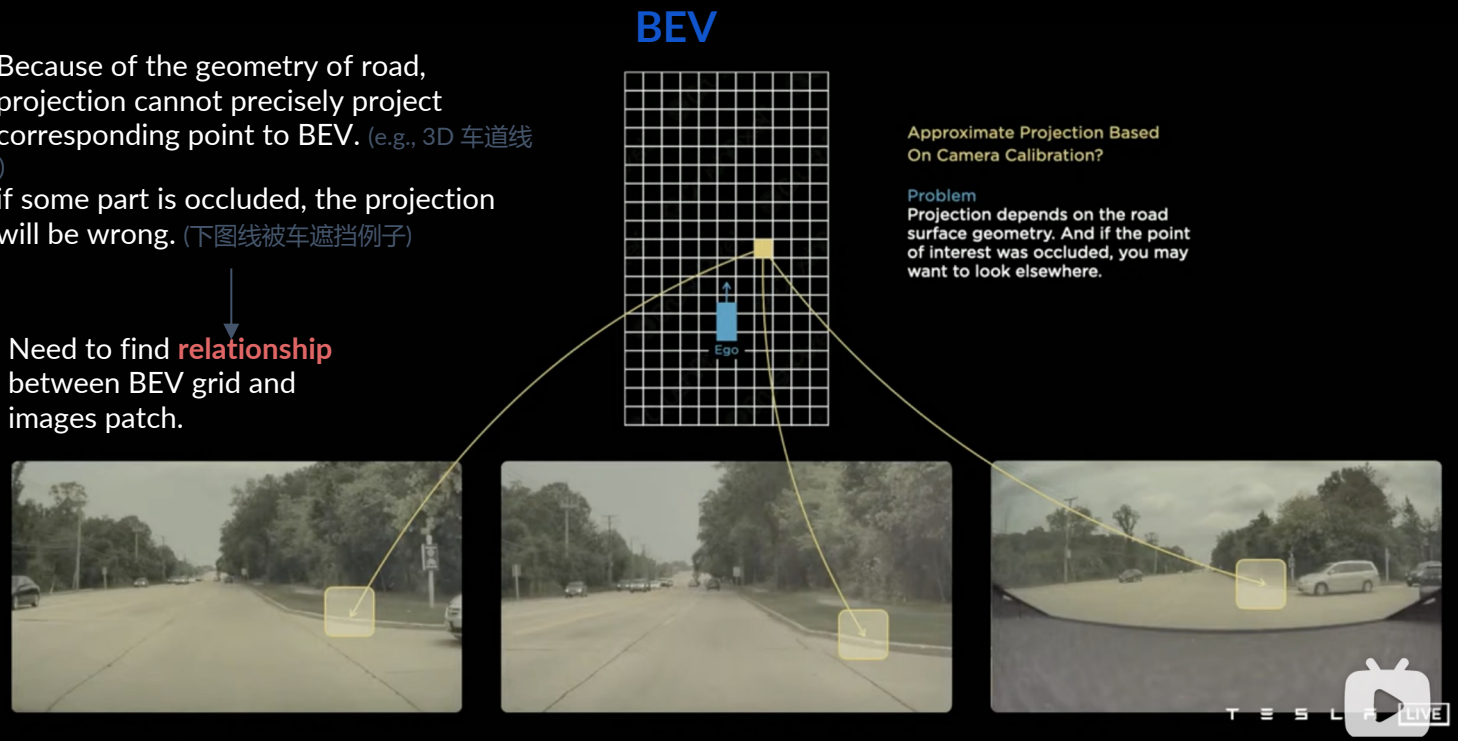
- project from image plane(Front view) to vector space (BEV). → Don't have depth per pixel
- assumption ground is horizontal. → which is not ture

Fusion is difficult as objects span **differently** across images.

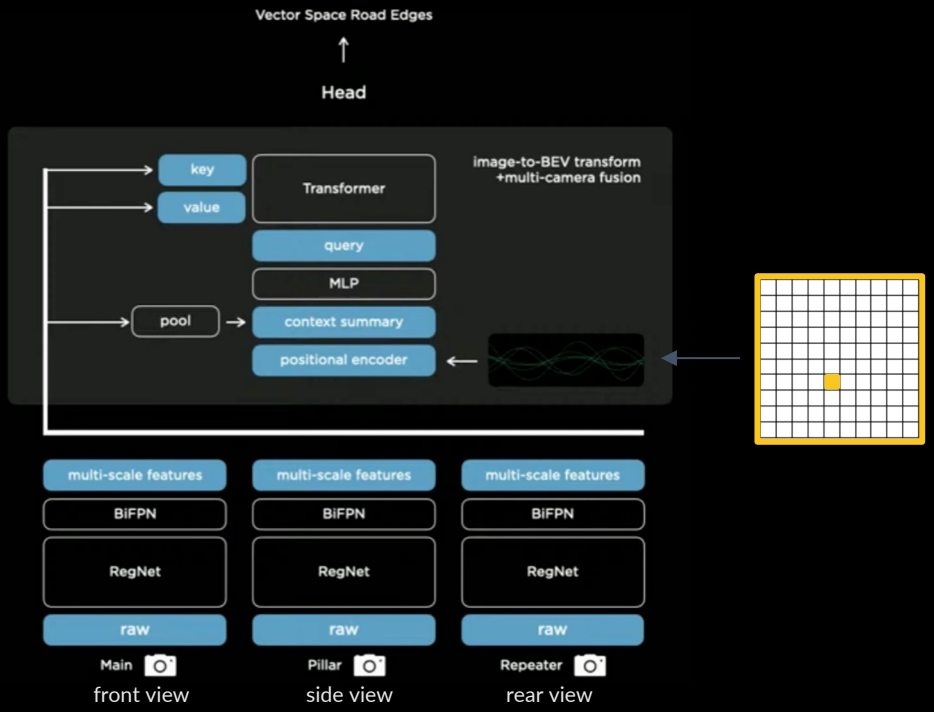
Solution: BEV-based End-to-End Perception

- Because of the geometry of road, projection cannot precisely project corresponding point to BEV. (e.g., 3D 车道线)
- if some part is occluded, the projection will be wrong. (下图线被车遮挡例子)

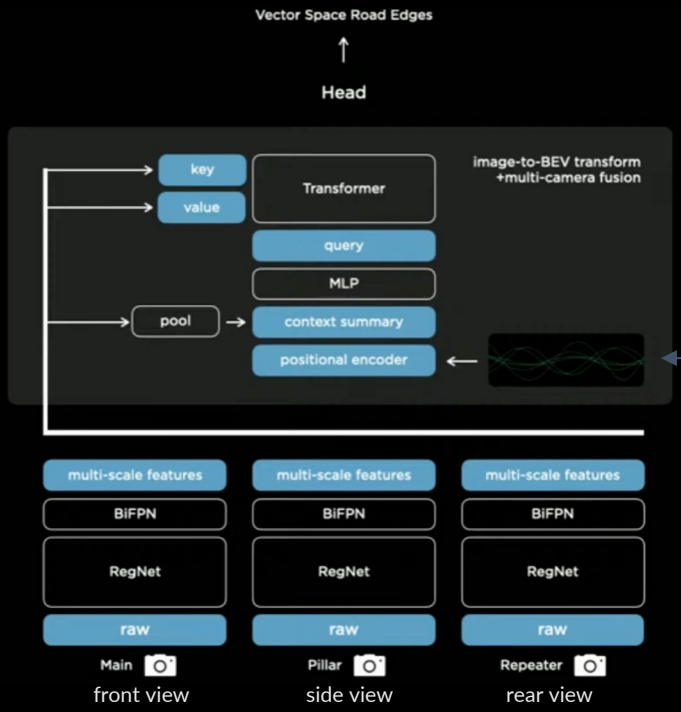
Need to find **relationship** between BEV grid and images patch.

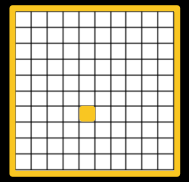


Implementation: Transformer-based View Transformation

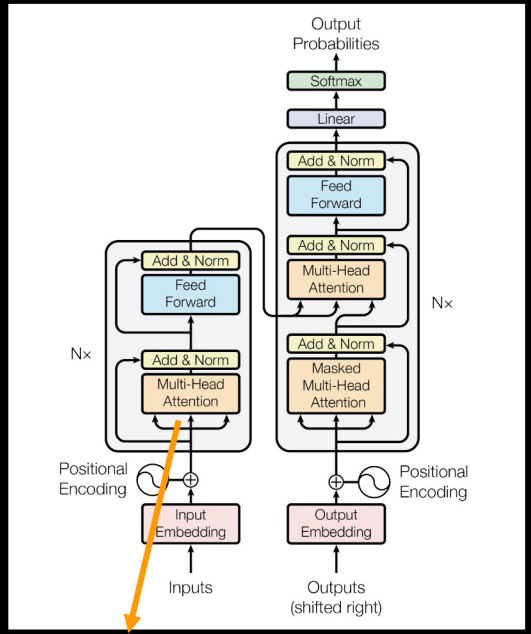


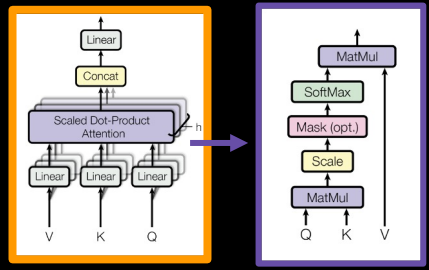
Implementation: Transformer-based View Transformation





$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$





Multi-head attention Dot-product Attention

Background on Transformer

- What: a query and a set of key-value pairs to an output
- The output: a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

Attention is all you need,
https://arxiv.org/pdf/1706.03762.pdf
2017

Core Issues in 3D Perception

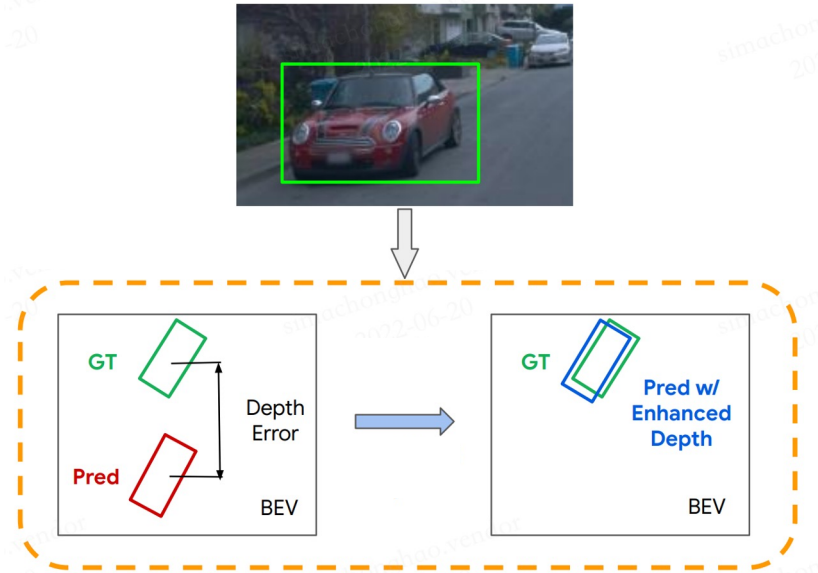
1. Depth Estimation: Gap between Camera and Lidar. **Need Accurate Depth**
1. Sensor Fusion: Early Fusion v.s. Late Fusion
1. Domain Adaptation: More robust 3D Perception across Different Scenes
1. Dataset: 3D Data Generation

Core Issues in Camera 3D Perception

Accurate *Depth*: Bridging the gap between Camera method and LiDAR method

Core Issues in Camera 3D Perception

Accurate *Depth*: Bridging the gap between Camera method and LiDAR method



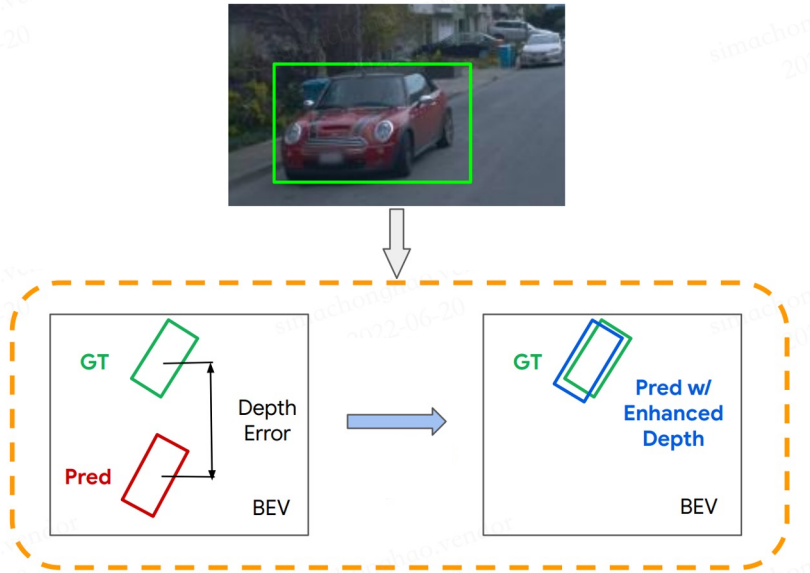
[1] Depth Estimation Matters Most: Improving Per-Object Depth Estimation for Monocular 3D Detection and Tracking, *arxiv:2206.03666*

Core Issues in Camera 3D Perception

Accurate *Depth*: Bridging the gap between Camera method and LiDAR method

解决思路

- Pseudo-LiDAR Track
  - 利用深度估计，将图像处理成伪点云
- Center-point Track
  - 根据预测的heatmap回归localization
- Depth Pretrain
  - 让backbone编码深度信息
- ***BEV视角变化***
  - ***回避直接做depth，在BEV 空间下做检测***

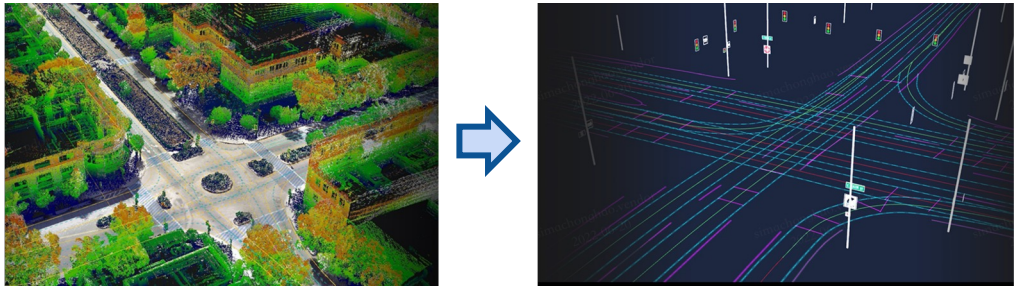


[1] Depth Estimation Matters Most: Improving Per-Object Depth Estimation for Monocular 3D Detection and Tracking, *arxiv:2206.03666*

BEV Perception: Next Paradigm for Autonomous Driving

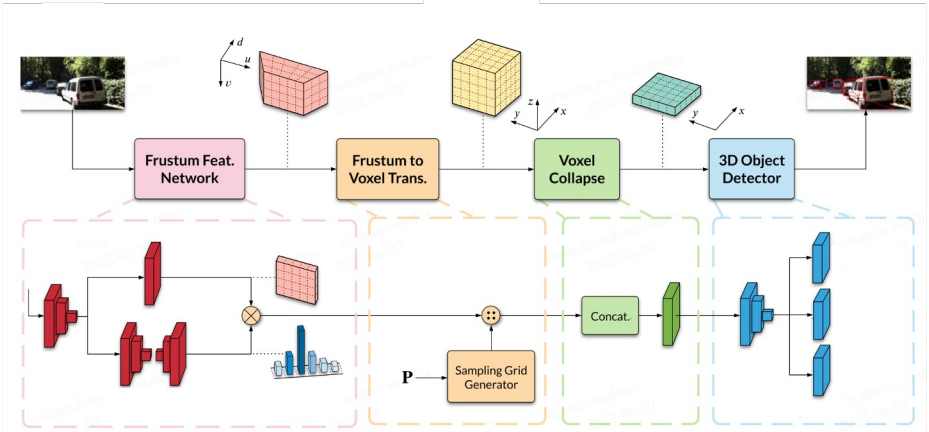
工业界

- 模型设计
- 传感器信息融合
- 聚焦三维真值获取
- 模型计算效率提升(车端，剪枝，量化，部署)



学术界

- 模型设计
  - BEV感知 / 建图 / 多任务
  - BEV空间中的端到端感知决策一体化
- 数据集Benchmark
  - 支持BEV感知 / 多任务



2 融合感知算法业界方案

Trending in **Academia 学术界**

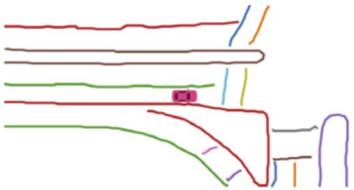
Trending in Academia: 泛BEV Perception

2021.7

- HDMapNet

- 输出BEV下高精地图
- 提出BEV下相机和激光雷达 特征的融合

- 输出BEV建图



Vectorized HD map

Trending in Academia: 泛BEV Perception

2021.7

- HDMapNet

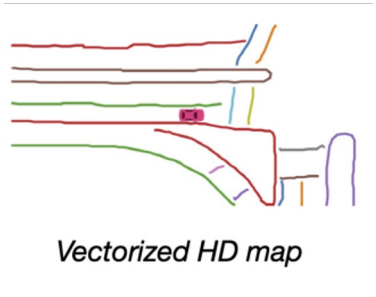
- 输出BEV下高精地图
- 提出BEV下相机和激光雷达 特征的融合

2021.10-12

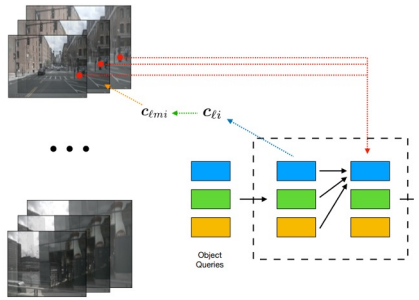
- DETR3D
- BEVDet

- 环视相机在BEV下融合的目标检测

- 输出BEV建图



- 隐式做BEV feature



Trending in Academia: 泛BEV Perception

2021.7
- HDMapNet

- 输出BEV下高精地图
- 提出BEV下相机和激光雷达 特征的融合

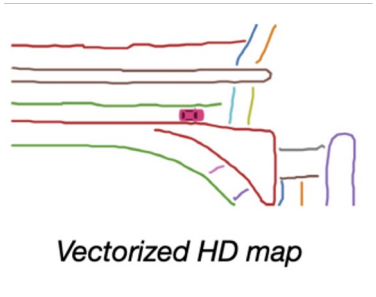
2021.10-12
- DETR3D
- BEVDet

- 环视相机在BEV下融合的目标检测

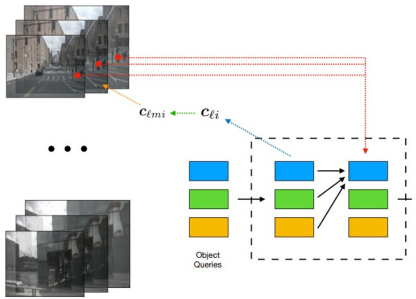
2022.3
- **BEVFormer**
- **PersFormer**

- **通过相机参数显示构建BEV表征**

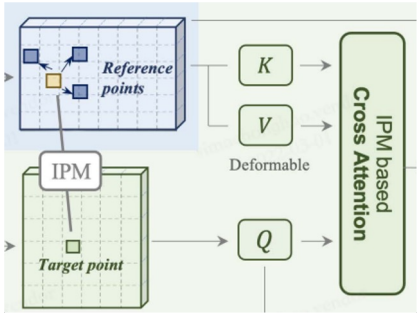
- 输出BEV建图



- 隐式做BEV feature



- 显式做BEV feature



Trending in Academia: 泛BEV Perception

2021.7
- HDMapNet

- 输出BEV下高精地图
- 提出BEV下相机和激光雷达 特征的融合

2021.10-12
- DETR3D
- BEVDet

- 环视相机在BEV下融合的目标检测

2022.3
- **BEVFormer**
- **PersFormer**

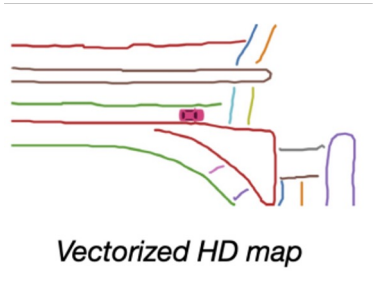
- **通过相机参数显示构建BEV表征**

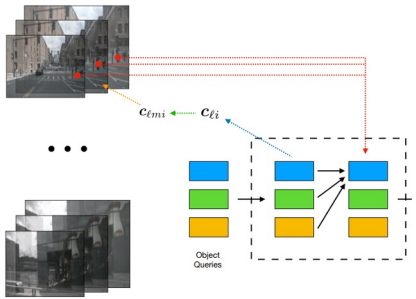
2022.5
- BEVFusion （达摩院）
- BEVFusion （MIT）
- FUTR3D

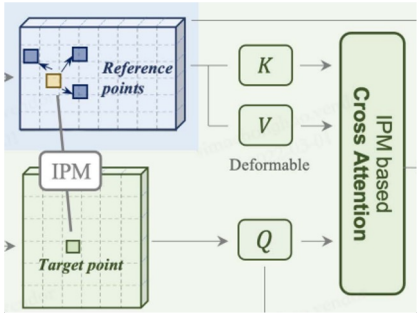
- **BEV下做多模态特征融合**

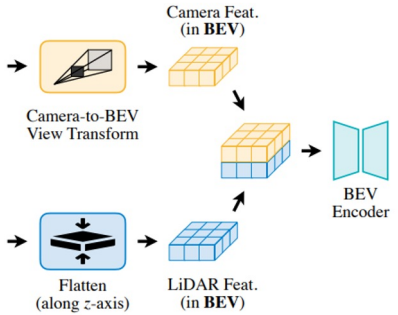
核心问题: 如何建模从front view转移到BEV (View Transformation) 得到更有效Feature?

- 输出BEV建图
- 隐式做BEV feature
- 显式做BEV feature
- 在BEV-feature维度做融合

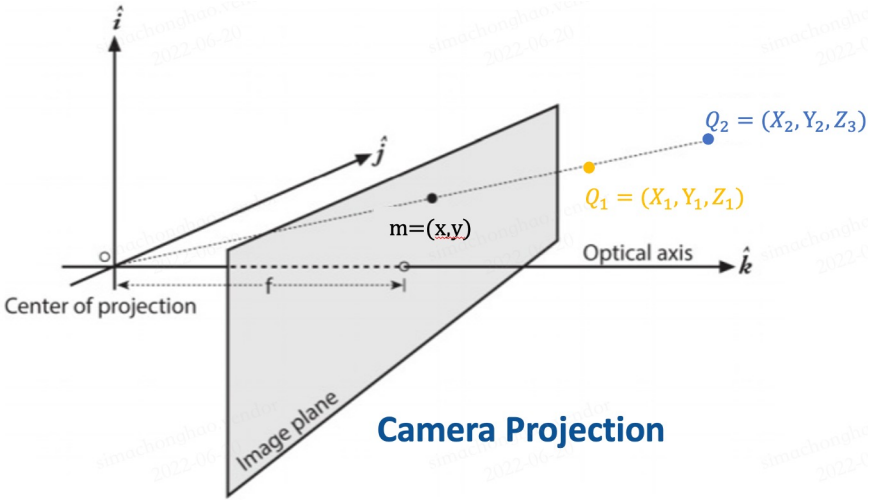








View Transformation

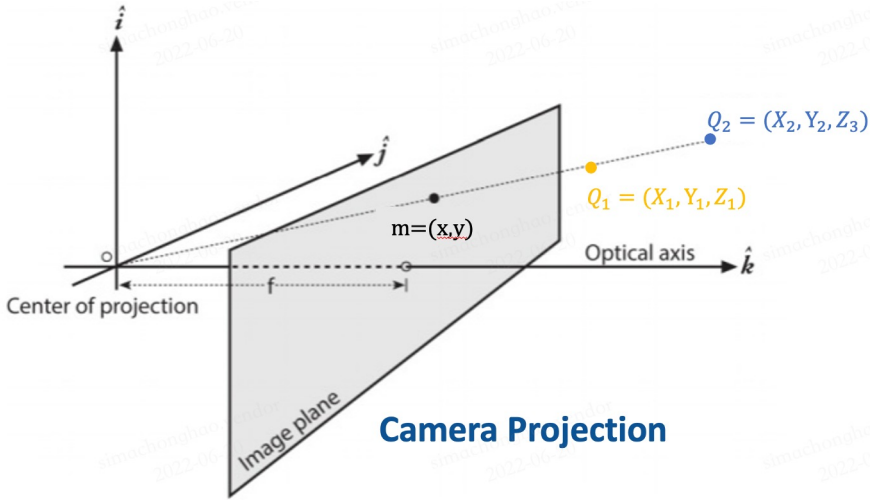


Camera Projection

Issues:

- From 3D to 2D:

View Transformation

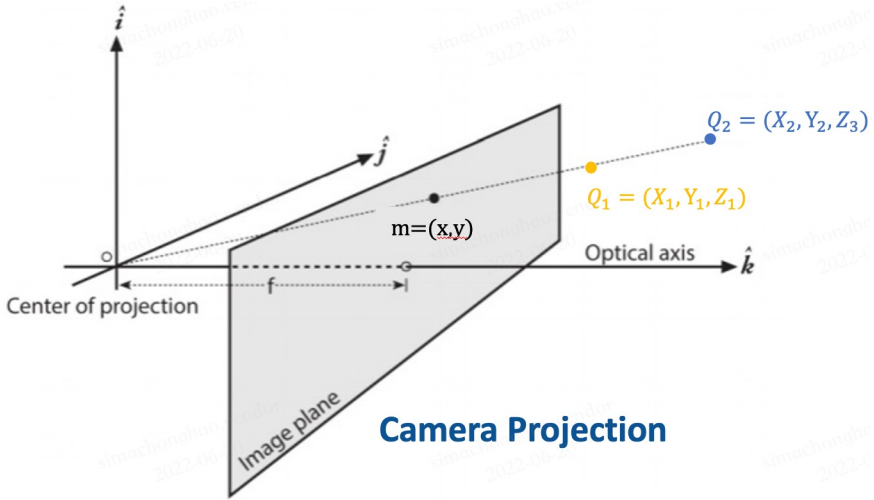


Camera Projection

Issues:

- From 3D to 2D:
  - Multiple 3D points will hit the ***same 2D pixel***

View Transformation

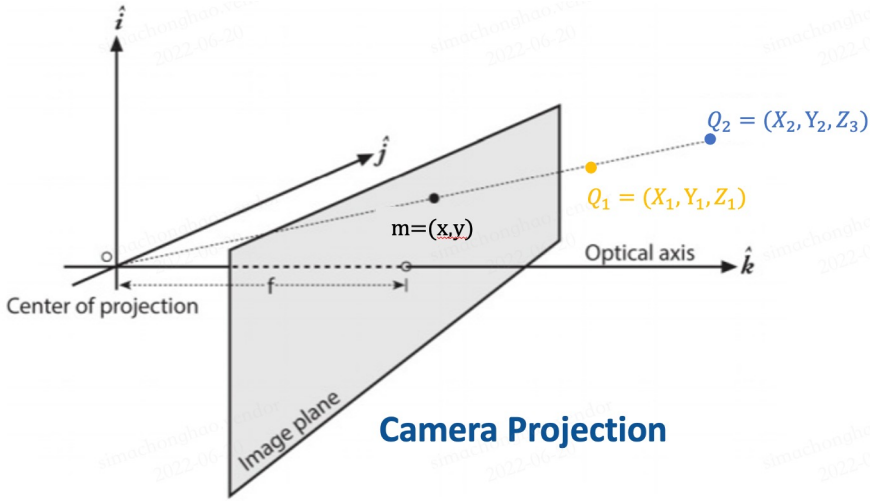


Camera Projection

Issues:

- From 3D to 2D:
  - Multiple 3D points will hit the ***same 2D pixel***
- From 2D to 3D:

View Transformation

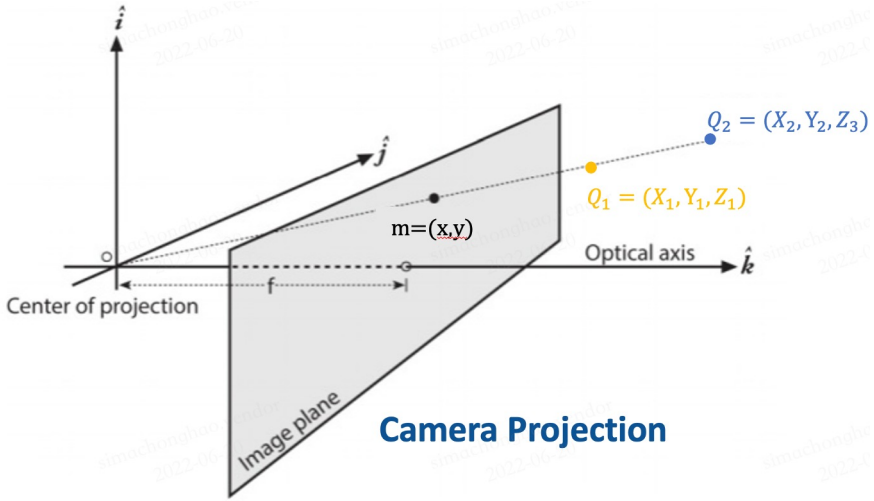


Camera Projection

Issues:

- From 3D to 2D:
  - Multiple 3D points will hit the ***same 2D pixel***
- From 2D to 3D:
  - ***Depth*** is unknown

View Transformation



Camera Projection

Issues:

- From 3D to 2D:
  - Multiple 3D points will hit the ***same 2D pixel***
- From 2D to 3D:
  - ***Depth*** is unknown

No matter what, the transformation is ill-posed

Two Ways to Address View Transformation

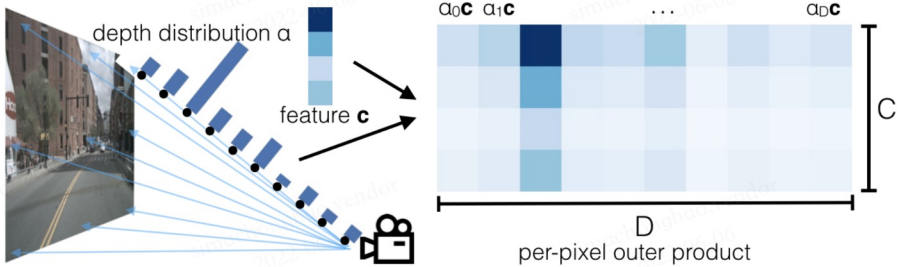
Way 1: From-2D-to-3D prior

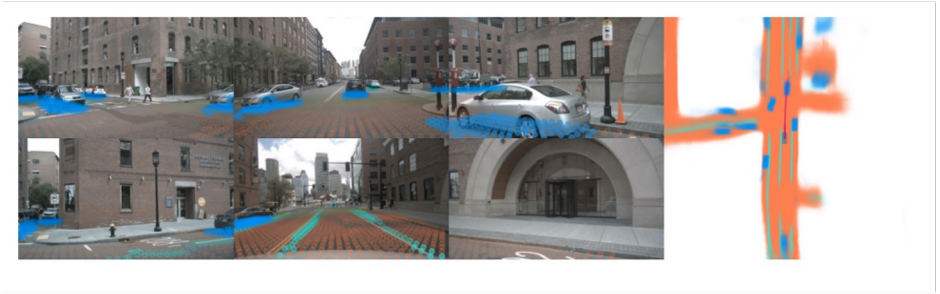
- 既然深度未知，那就预测深度
  i. Lift, Splat, Shoot and its derivant
  ii. Pseudo-LiDAR Family

Way 2: From-3D-to-2D prior

- 根据3D到2D的投影，index局部特征
  i. DETR3D and its derivant
  ii. Explicit BEV feature
- Implicit 3D Positional Embedding

2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant

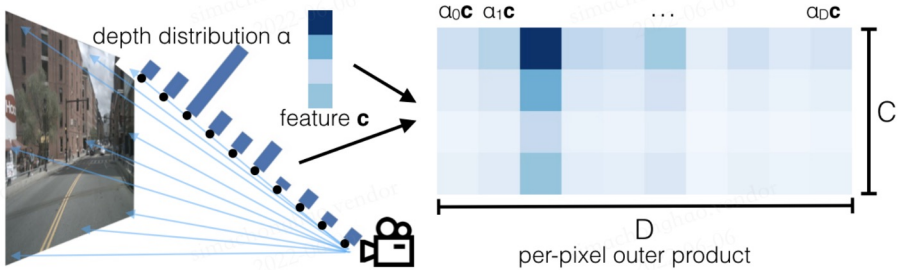


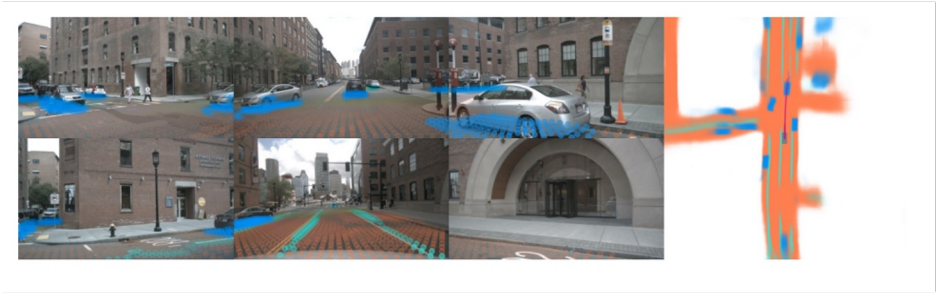


- Lift-Splat-Shoot(LSS) [1]：使用按bin分类的深度分布代替连续深度估计

[1] Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D, *ECCV 2020*.

2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant





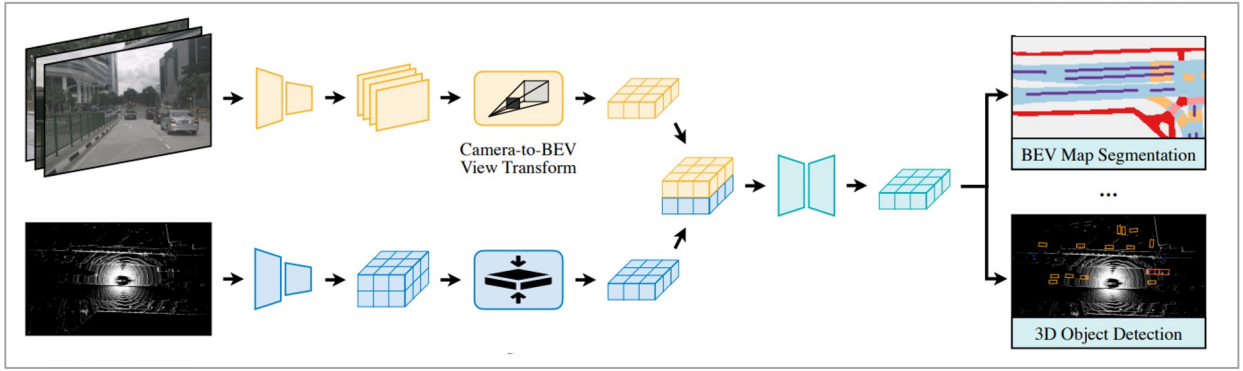
- Lift-Splat-Shoot(LSS) [1]：使用按bin分类的深度分布代替连续深度估计
- **优势**：
  - 深度分布相对容易生成
- **劣势**：
  - 生成的分布是离散且稀疏的，与真实场景差距较大
  - 物体边界不易处理
- **后续工作**：
  - CaDDN [2]
  - FIERY [3]
  - **BEVDet / BEVFusion / BEVDepth**

[1] Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D, *ECCV 2020*.

[2] Categorical Depth Distribution Network for Monocular 3D Object Detection, *CVPR 2021*.

[3] FIERY: Future Instance Prediction in Bird's-Eye View from Surround Monocular Cameras, *ICCV 2021 (Oral)*.

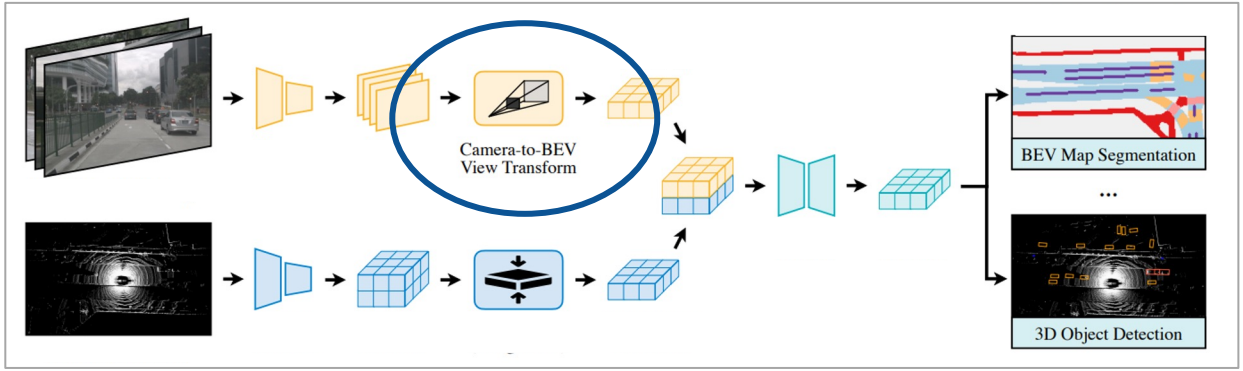
2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant



- BEVFusion [1]: LSS + VoxelNet

[1] BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, *arxiv:2205.13542*.

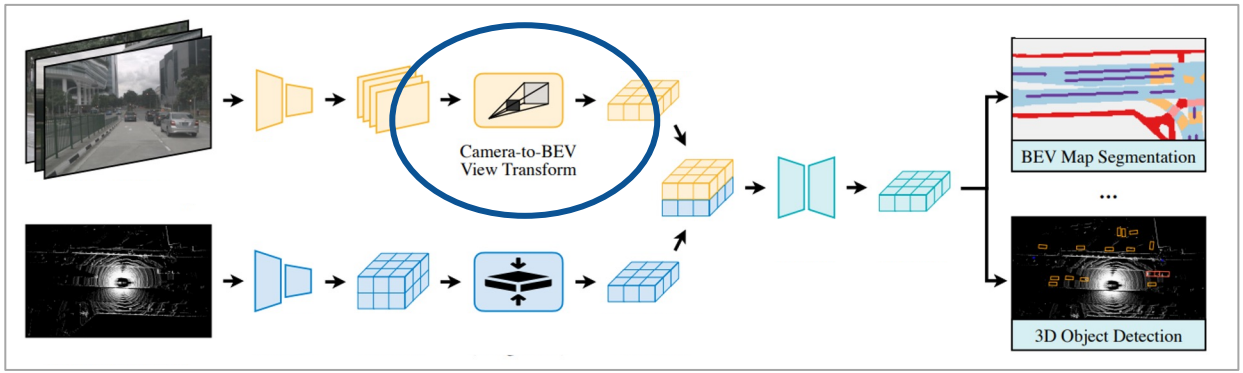
2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant



- BEVFusion [1]：LSS + VoxelNet
- **Camera-to-BEV View Transform**参照LSS的做法，加速BEV pooling
- 参照TransFusion，**在BEV下融合**Camera和LiDAR feature

[1] BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, *arxiv:2205.13542*.

2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant

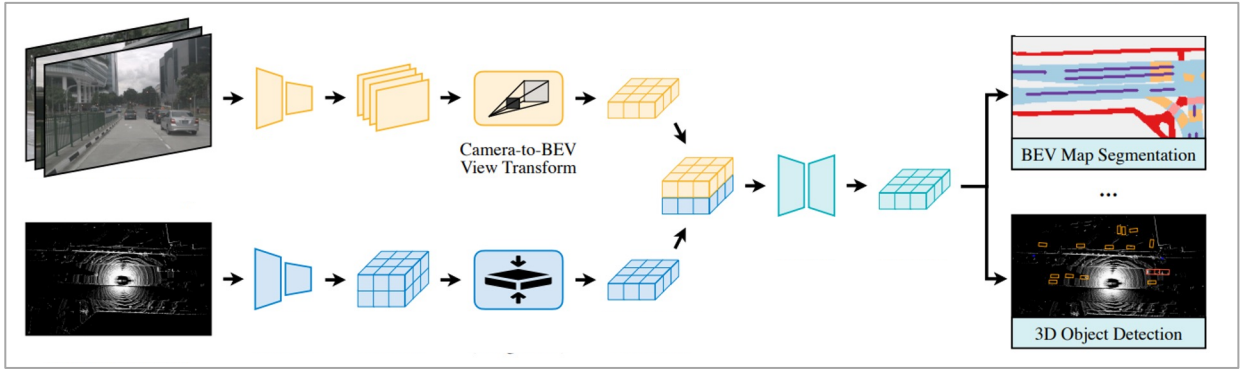


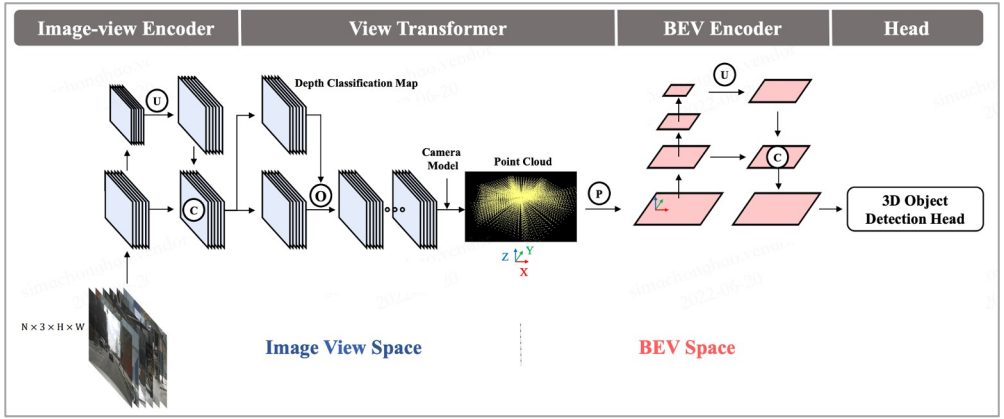
- BEVFusion [1]：LSS + VoxelNet
- **Camera-to-BEV View Transform**参照LSS的做法，加速BEV pooling
- 参照TransFusion，**在BEV下融合**Camera和LiDAR feature
- nuScenes NDS: 0.761

Current SOTA
Any Modality

[1] BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, *arxiv:2205.13542*.

2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant





- BEVFusion [1]：LSS + VoxelNet
- **Camera-to-BEV View Transform**参照LSS的做法，加速BEV pooling
- 参照TransFusion，**在BEV下融合**Camera和LiDAR feature
- nuScenes NDS: 0.761

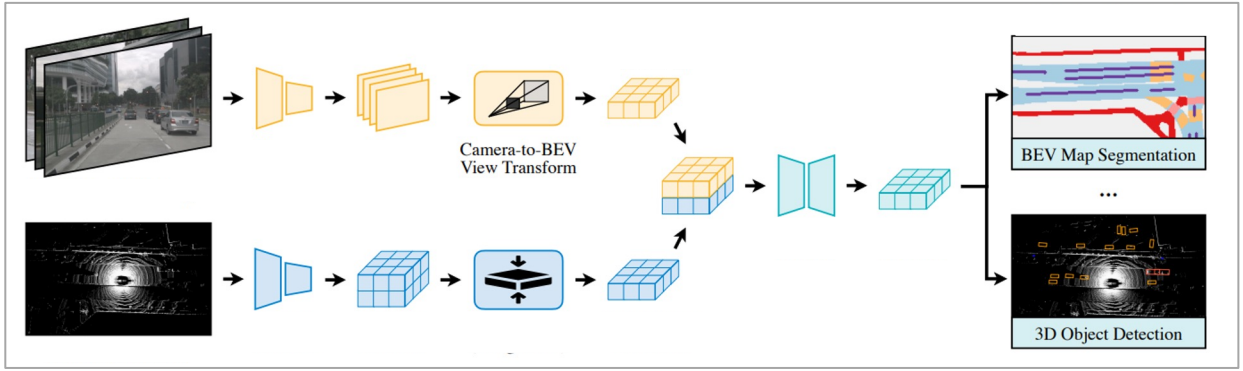
Current SOTA
Any Modality

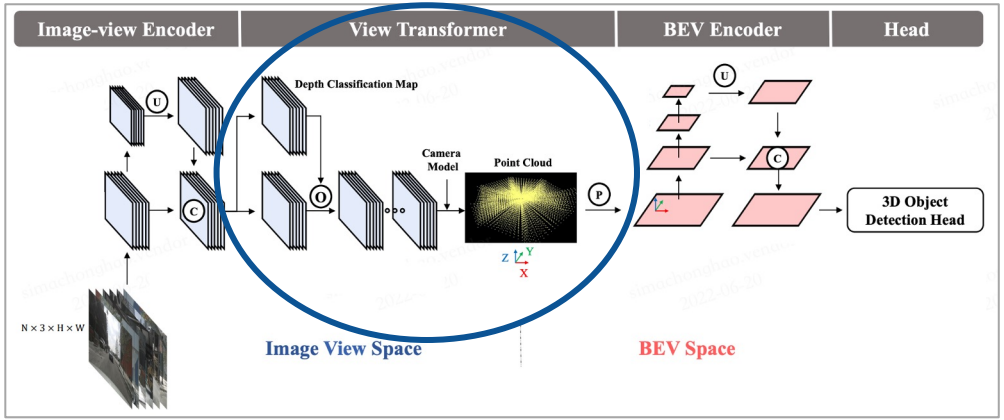
- BEVDet [2]： LSS + CenterPoint

[1] BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, *arxiv:2205.13542*.

[2] BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View, *arxiv:2112.11790*.

2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant





- BEVFusion [1]：LSS + VoxelNet
- **Camera-to-BEV View Transform**参照LSS的做法，加速BEV pooling
- 参照TransFusion，**在BEV下融合**Camera和LiDAR feature
- nuScenes NDS: 0.761

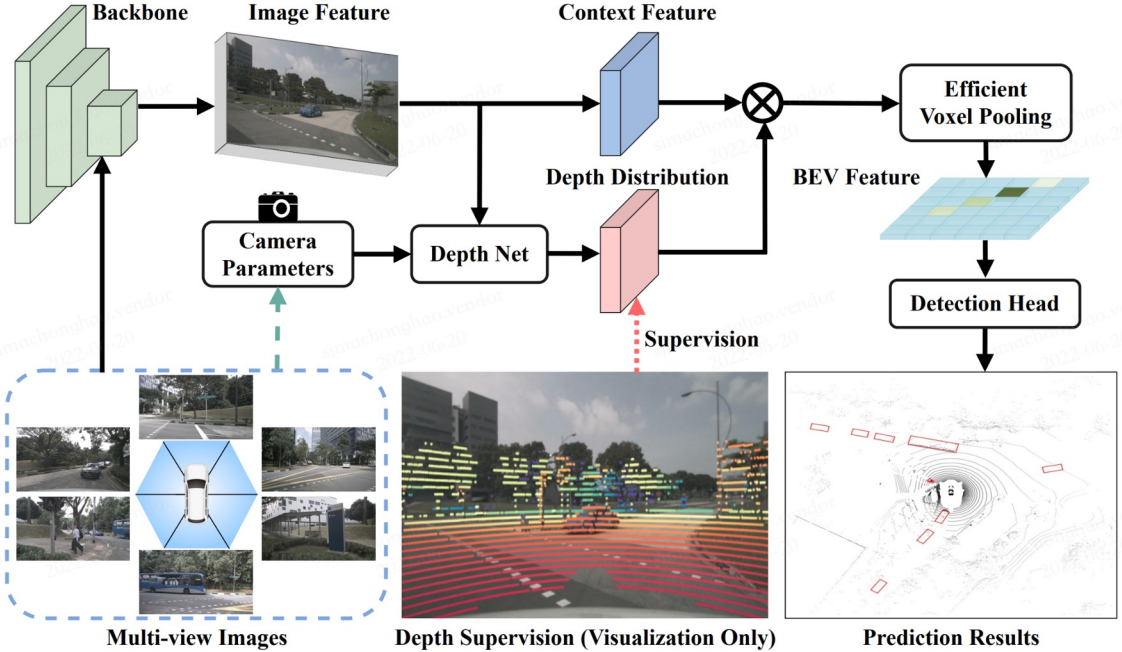
Current SOTA
Any Modality

- BEVDet [2]： LSS + CenterPoint
- **View TransFormer** 在 LSS基础上改进，增加了 BEV 空间下的数据增广
- nuScenes NDS: 0.569

[1] BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, *arxiv:2205.13542*.

[2] BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View, *arxiv:2112.11790*.

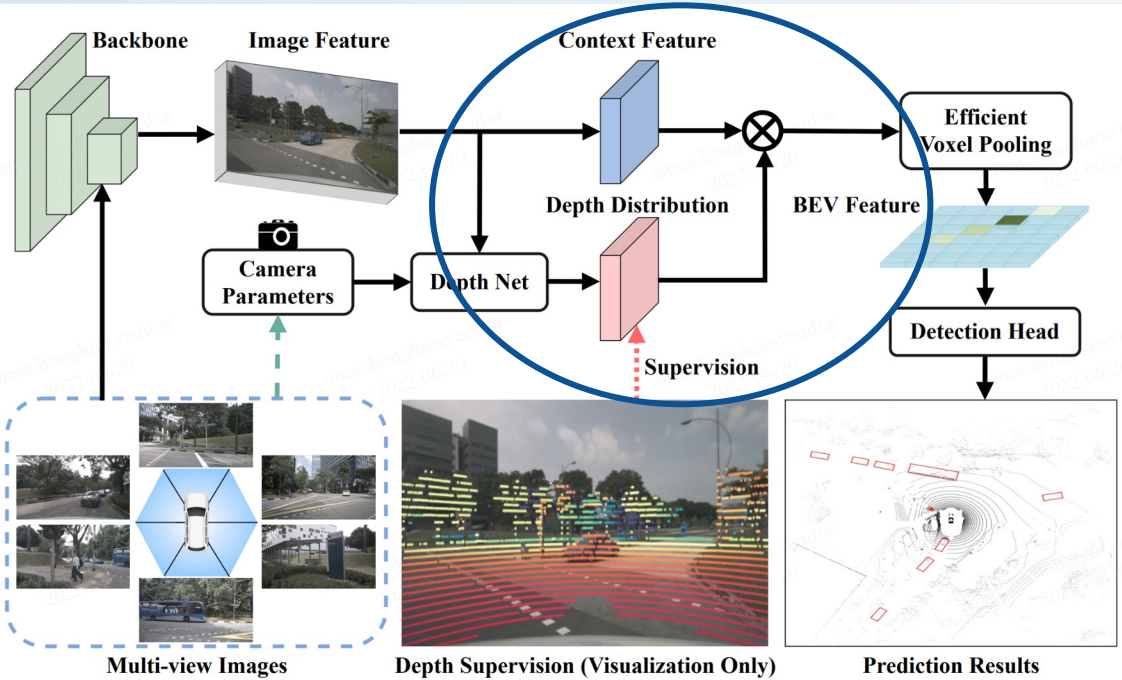
2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant



- BEVDepth [1]: LSS + Depth supervision

[1] BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection, *arXiv:2206.10092*.

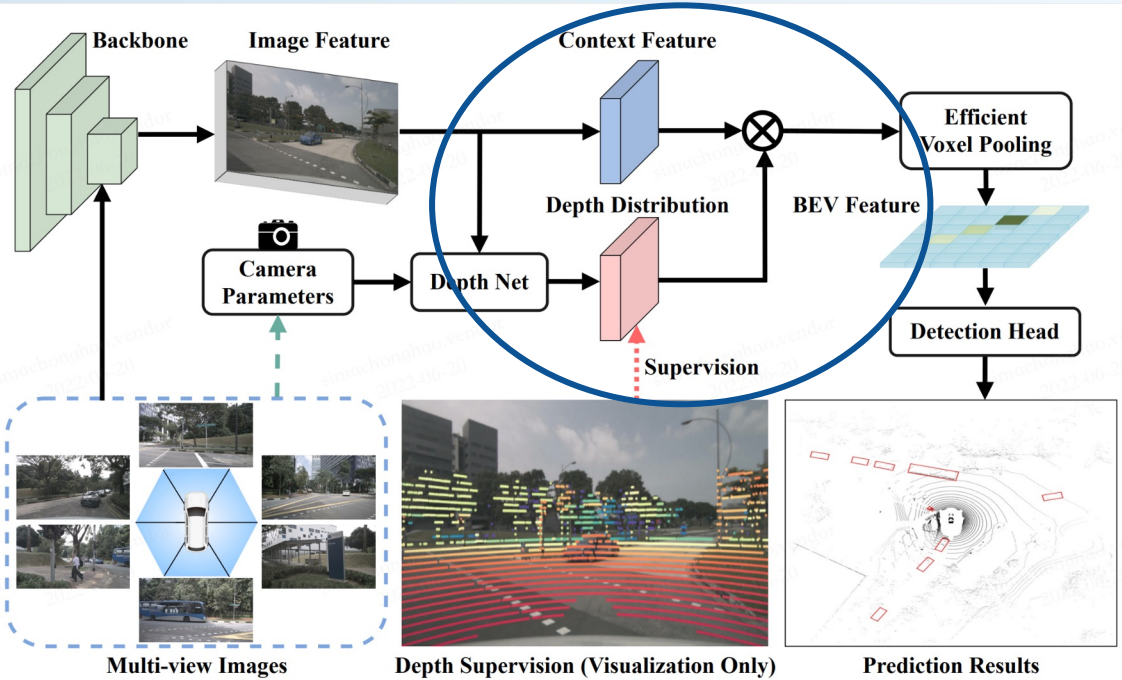
2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant



- BEVDepth [1]：LSS + Depth supervision
- **BEV Feature**参照LSS的做法，加入LiDAR作为深度分布的监督信号

[1] BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection, *arXiv:2206.10092*.

2D to 3D: Lift-Splat-Shoot (LSS) and its Derivant

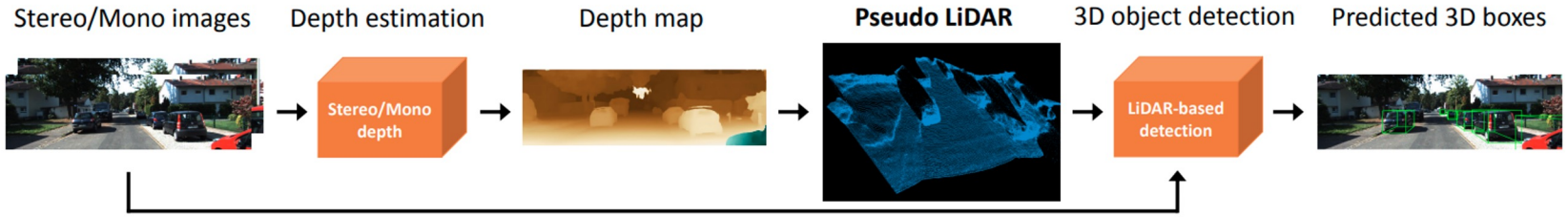


- BEVDepth [1]：LSS + Depth supervision
- **BEV Feature**参照LSS的做法，加入LiDAR作为深度分布的监督信号
- nuScenes NDS: 0.600 MEGVII 旷视

Current SOTA Camera-only

[1] BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection, *arXiv:2206.10092*.

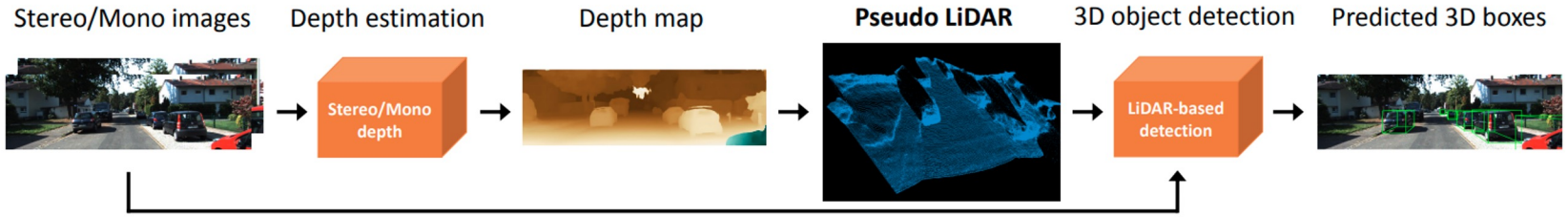
2D to 3D: Pseudo Lidar Family



- Pseudo-LiDAR [1]：使用像素级别的深度估计，把图像拓展成伪点云

[1] Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving Representation, *CVPR 2019*.

2D to 3D: Pseudo Lidar Family



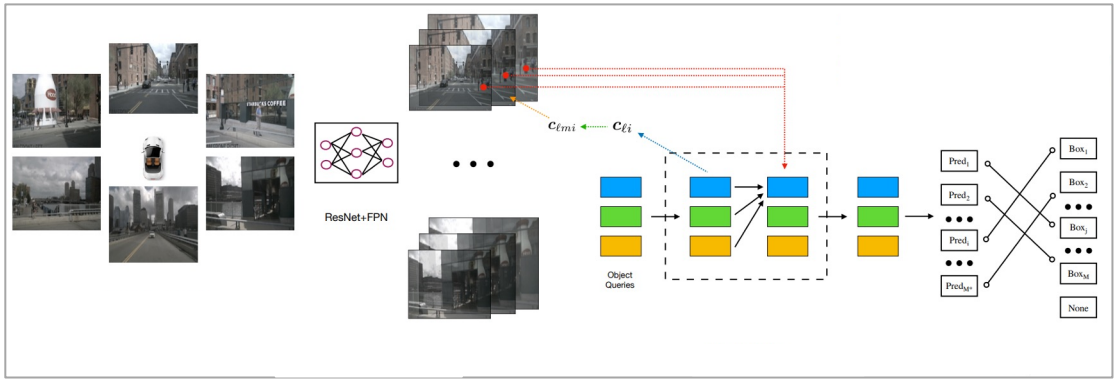
- Pseudo-LiDAR [1]：使用像素级别的深度估计，把图像拓展成伪点云
- **优势**：
  - 深度图是连续的，可以接入点云检测
- **劣势**：
  - 检测质量强依赖于深度估计，目前通常是不准的
  - 室外场景像素级别的绝对深度真值不易获取
- **后续工作**：
  - Pseudo-LiDAR++ [2]
  - Patch-Net [3]

[1] Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving Representation, *CVPR 2019*.

[2] Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving, *ICLR 2020*.

[3] Rethinking Pseudo-LiDAR Representation, *ECCV 2020*.

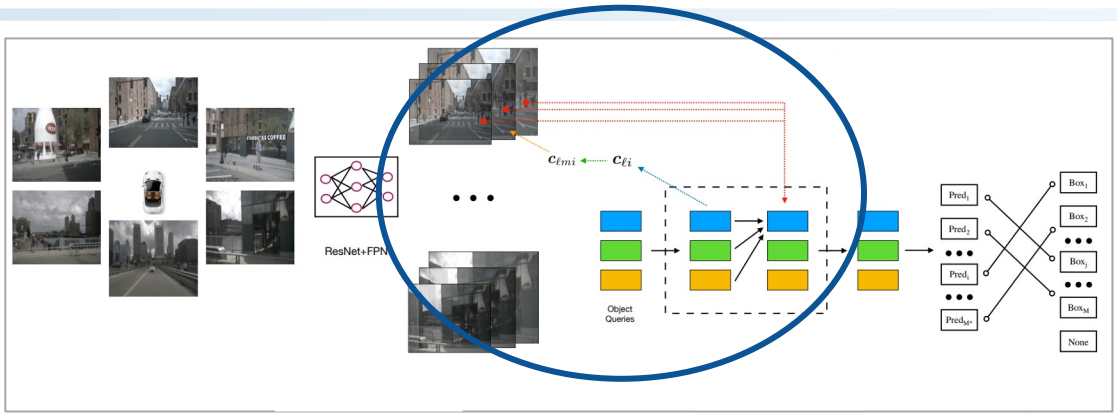
3D to 2D: DETR3D and its Derivant



- DETR3D [1]：环视camera feature上，根据BEV到front view的关系，采样front view feature，输出3D目标检测

[1] DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries, *CoRL 2021*.

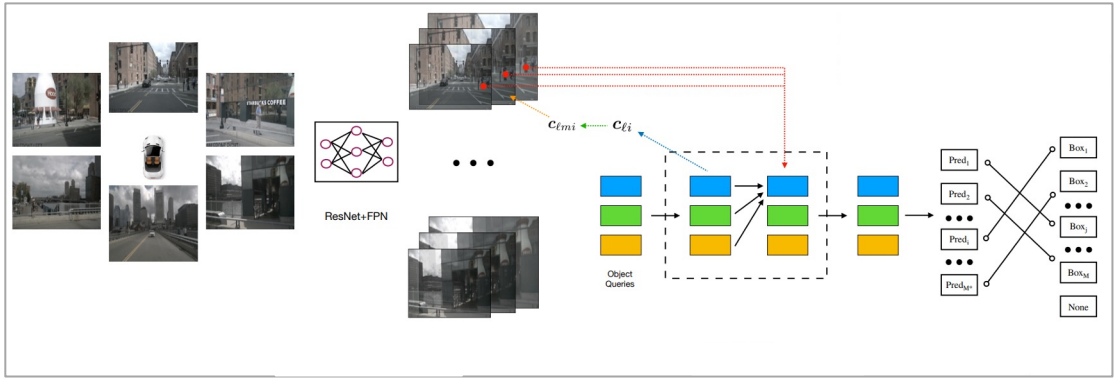
3D to 2D: DETR3D and its Derivant



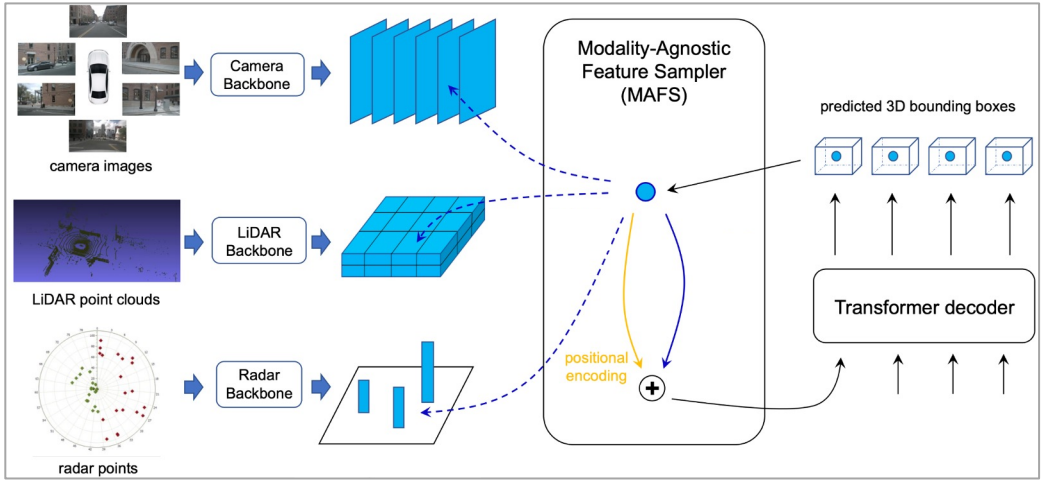
- DETR3D [1]：环视camera feature上，根据BEV到front view的关系，采样front view feature，输出3D目标检测
- **Feature Transformation**： 3D query投影到2D（front view）平面，查询2D feature，decode成object bbox
- nuScenes NDS: 0.479

[1] DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries, *CoRL 2021*.

3D to 2D: DETR3D and its Derivant



- DETR3D [1]：环视camera feature上，根据BEV到front view的关系，采样front view feature，输出3D目标检测
- **Feature Transformation**： 3D query投影到2D（front view）平面，查询2D feature，decode成object bbox
- nuScenes NDS: 0.479

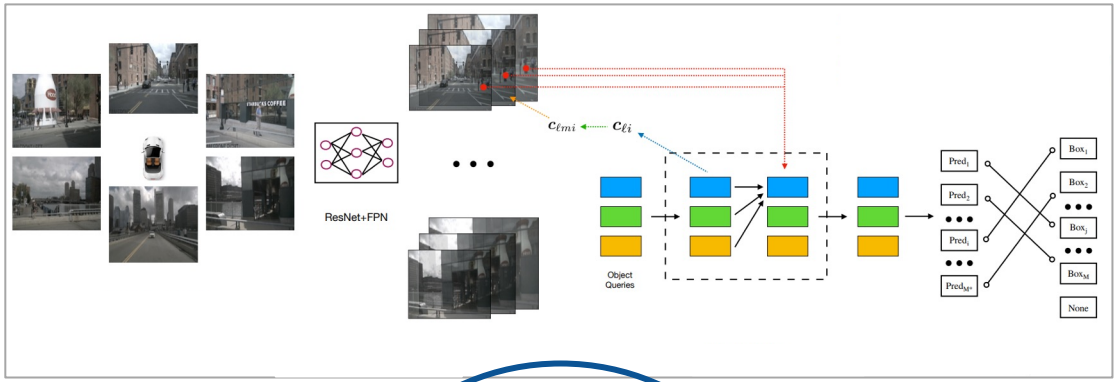


- FUTR3D [2]：sensor fusion of DETR3D

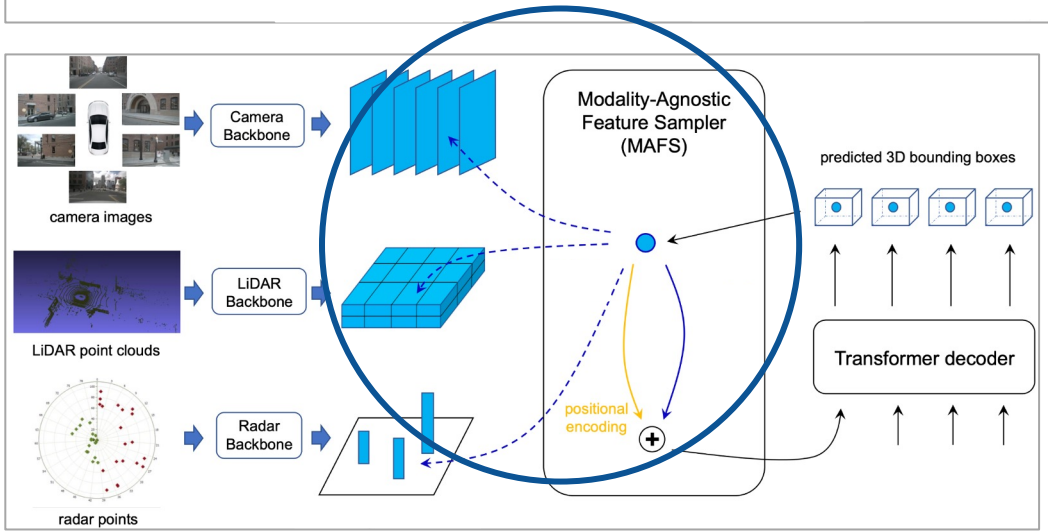
[1] DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries, *CoRL 2021*.

[2] FUTR3D: A Unified Sensor Fusion Framework for 3D Detection, *arXiv:2203.10642*.

3D to 2D: DETR3D and its Derivant



- DETR3D [1]：环视camera feature上，根据BEV到front view的关系，采样front view feature，输出3D目标检测
- **Feature Transformation**： 3D query投影到2D（front view）平面，查询2D feature，decode成object bbox
- nuScenes NDS: 0.479

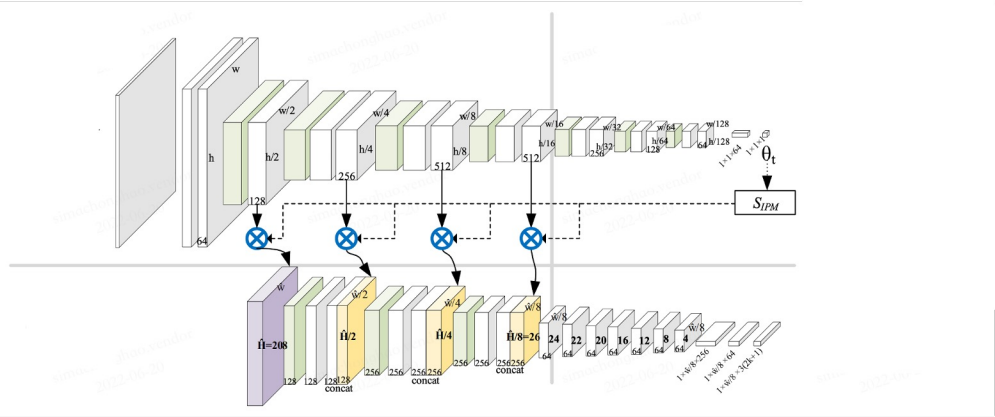


- FUTR3D [2]：sensor fusion of DETR3D
- **MAFS**： 3D query分别投影到2D平面，voxel，radar去查询对应feature，decode成object bbox
- nuScenes NDS: 0.680

[1] DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries, *CoRL 2021*.

[2] FUTR3D: A Unified Sensor Fusion Framework for 3D Detection, *arXiv:2203.10642*.

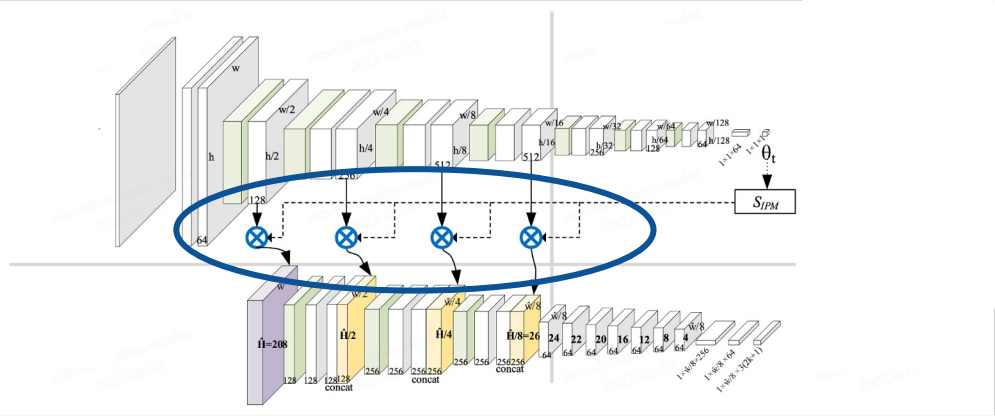
3D to 2D: Explicit BEV Feature



- 3D-LaneNet [1]：BEV下3D车道线检测

[1] 3D-LaneNet: End-to-End 3D Multiple Lane Detection, *ICCV 2019*.

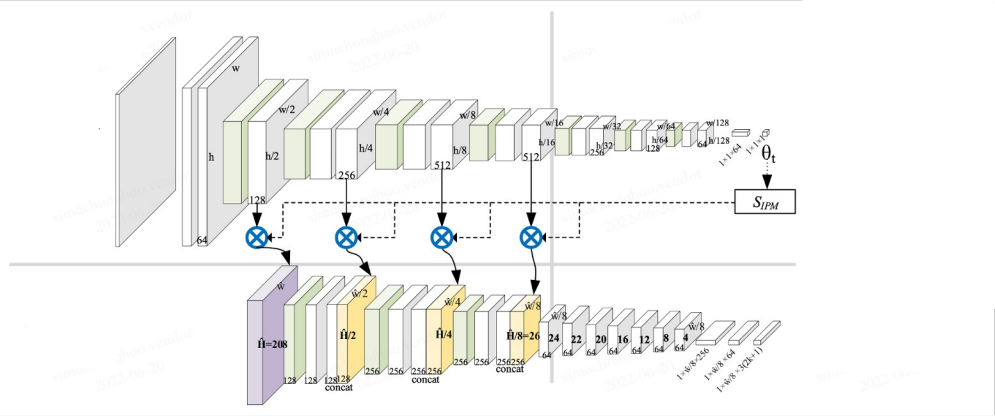
3D to 2D: Explicit BEV Feature



- 3D-LaneNet [1]：BEV下3D车道线检测
- **Projection to Top view**： feature从front view投影到BEV，***基于IPM***，用grid sampler采样得到BEV feature
- OpenLane F1: 40.2

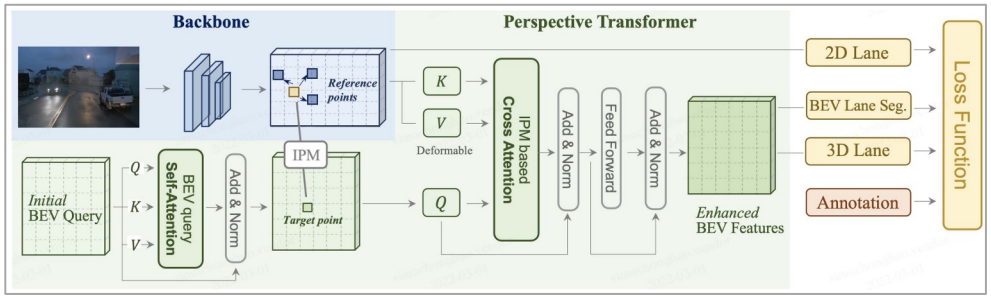
[1] 3D-LaneNet: End-to-End 3D Multiple Lane Detection, *ICCV 2019*.

3D to 2D: Explicit BEV Feature



- 3D-LaneNet [1]：BEV下3D车道线检测
- **Projection to Top view**： feature从front view投影到BEV，***基于IPM***，用grid sampler采样得到BEV feature
- OpenLane F1: 40.2

OpenDriveLab https://github.com/OpenDriveLab/OpenLane

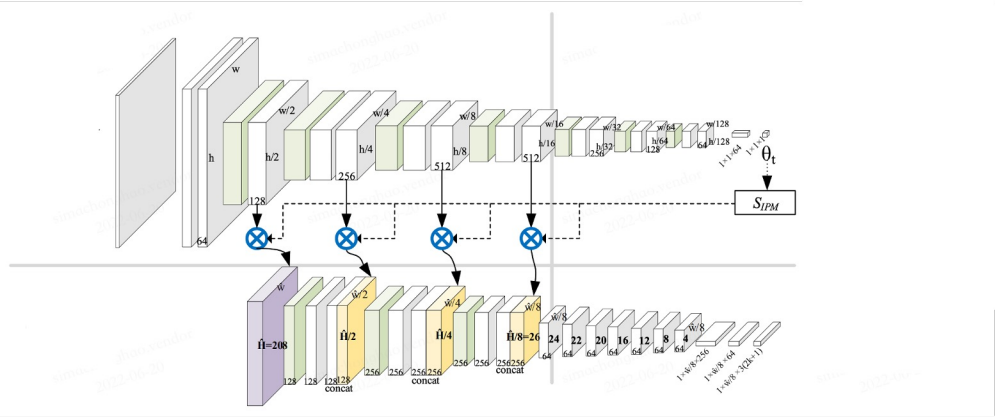


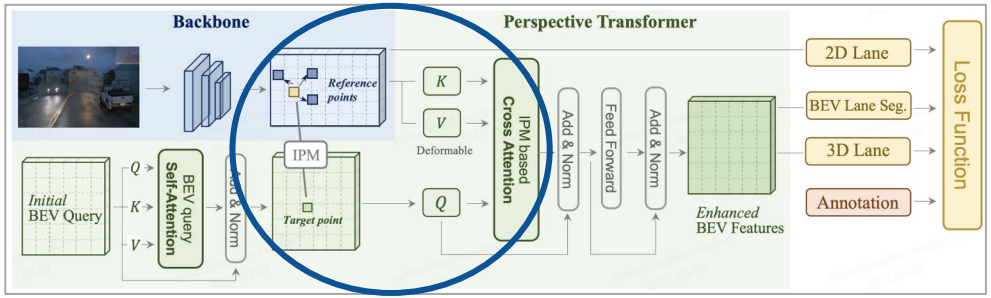
- PersFormer [2]：2D-3D车道线联合检测

[1] 3D-LaneNet: End-to-End 3D Multiple Lane Detection, *ICCV 2019*.

[2] PersFormer: 3D Lane Detection via Perspective Transformer and the OpenLane Benchmark, *arXiv:2203.11089*.

3D to 2D: Explicit BEV Feature





- 3D-LaneNet [1]：BEV下3D车道线检测
- **Projection to Top view**： feature从front view投影到BEV，***基于IPM***，用grid sampler采样得到BEV feature
- OpenLane F1: 40.2

Current SOTA

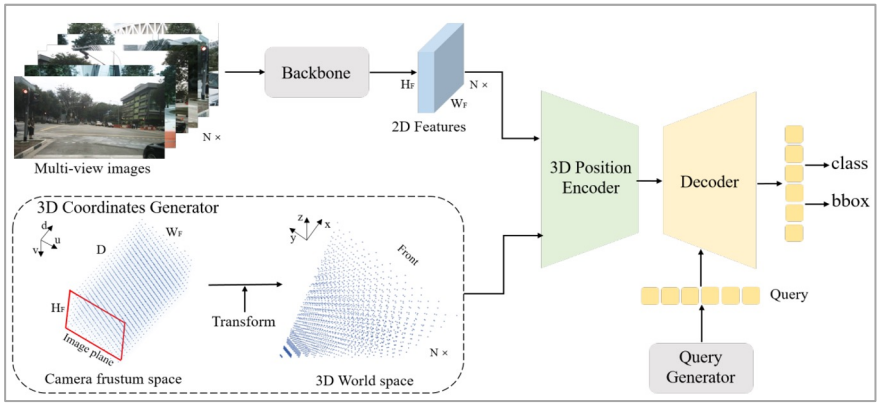
PerceptionX https://github.com/OpenPerceptionX/OpenLane

- PersFormer [2]：2D-3D车道线联合检测
- **IPM-based Cross Attention**： 以front view feature为key和value，通过***IPM***得到2D-BEV的参考点，使用BEV query去查询得到BEV feature
- OpenLane F1: 49.0

[1] 3D-LaneNet: End-to-End 3D Multiple Lane Detection, *ICCV 2019*.

[2] PersFormer: 3D Lane Detection via Perspective Transformer and the OpenLane Benchmark, *arXiv:2203.11089*.

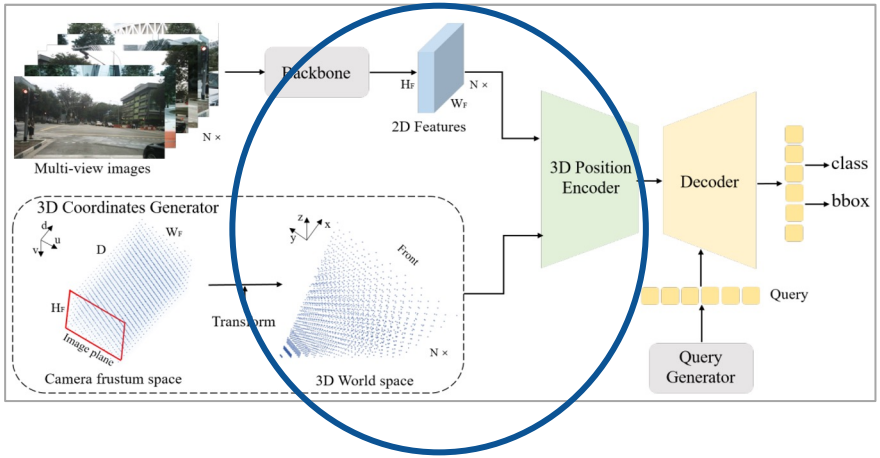
3D to 2D: Implicit 3D Positional Embedding



- PETR [1]：基于3D位置编码的3D物体检测

[1] PETR: Position Embedding Transformation for Multi-View 3D Object Detection, *arxiv:2203.05625*.

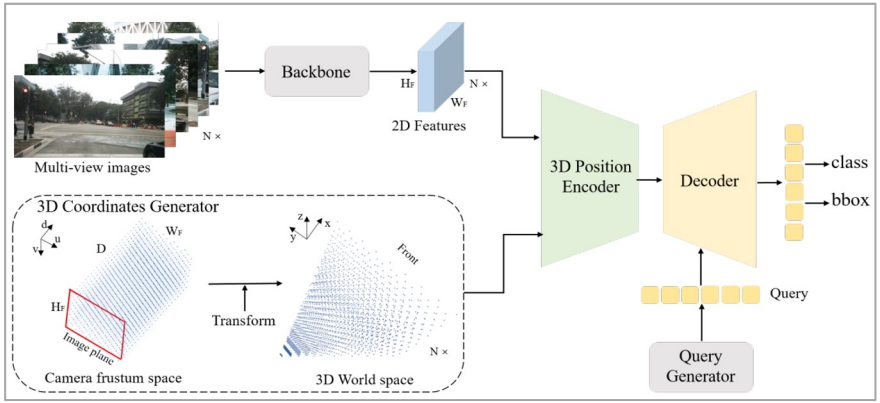
3D to 2D: Implicit 3D Positional Embedding

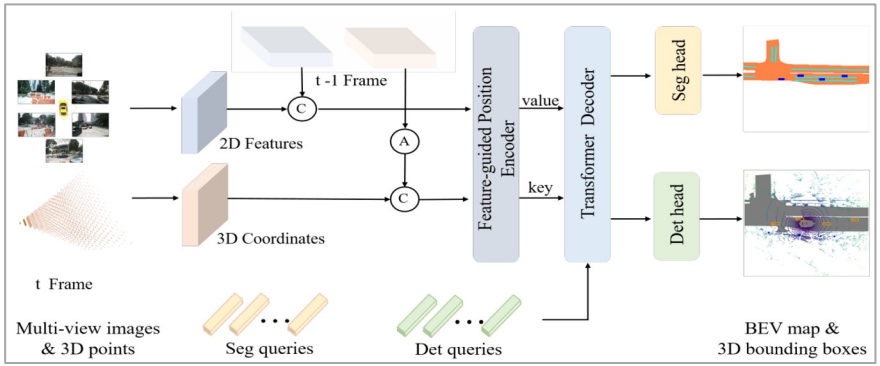


- PETR [1]：基于3D位置编码的3D物体检测
- **3D Coordinates Generator & 3D Position Encoder**：生成基于3D坐标的位置编码，编码2D特征送入Encoder，输出特征具有3D空间信息
- nuScenes NDS: 0.481

[1] PETR: Position Embedding Transformation for Multi-View 3D Object Detection, *arxiv:2203.05625*.

3D to 2D: Implicit 3D Positional Embedding





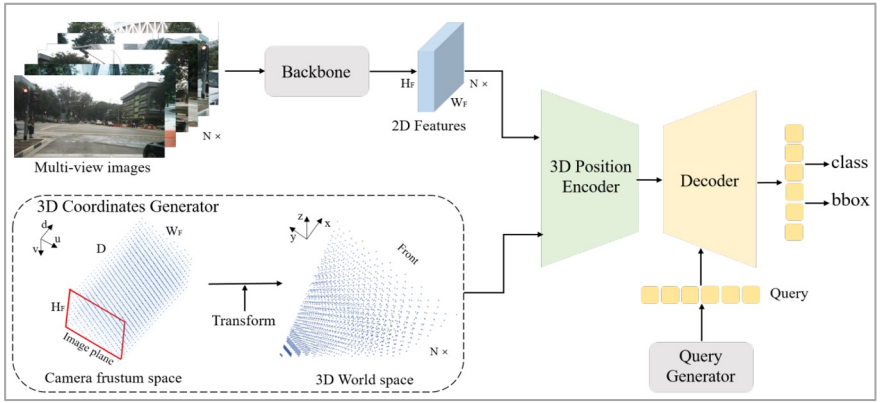
- PETR [1]：基于3D位置编码的3D物体检测
- **3D Coordinates Generator & 3D Position Encoder**：生成基于3D坐标的位置编码，编码2D特征送入Encoder，输出特征具有3D空间信息
- nuScenes NDS: 0.481

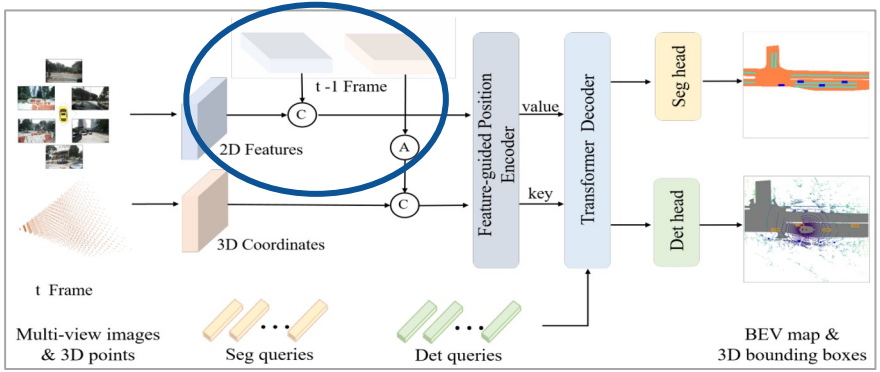
- PETRv2 [2]在PETR的基础上，加入了时序信息

[1] PETR: Position Embedding Transformation for Multi-View 3D Object Detection, *arxiv:2203.05625*.

[2] PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images, *arxiv:2206.01256*.

3D to 2D: Implicit 3D Positional Embedding





- PETR [1]：基于3D位置编码的3D物体检测
- **3D Coordinates Generator & 3D Position Encoder**：生成基于3D坐标的位置编码，编码2D特征送入Encoder，输出特征具有3D空间信息
- nuScenes NDS: 0.481

- PETRv2 [2]在PETR的基础上，加入了时序信息
- **时序操作**：在image space下融合历史帧的信息
- nuScenes NDS: 0.582

Second Best Camera-only

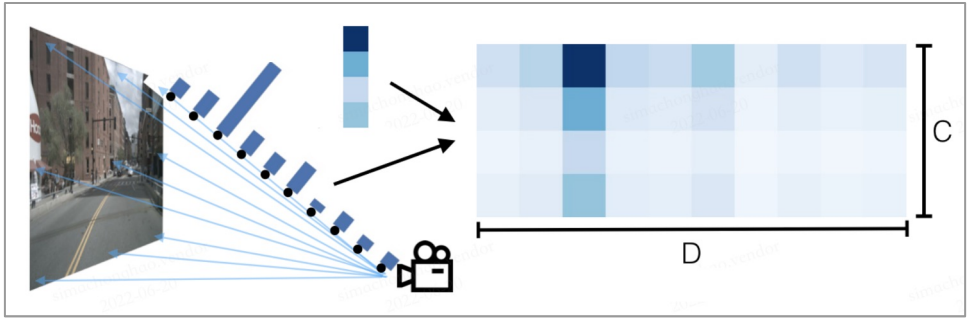
MEGVII 旷视

[1] PETR: Position Embedding Transformation for Multi-View 3D Object Detection, *arxiv:2203.05625*.

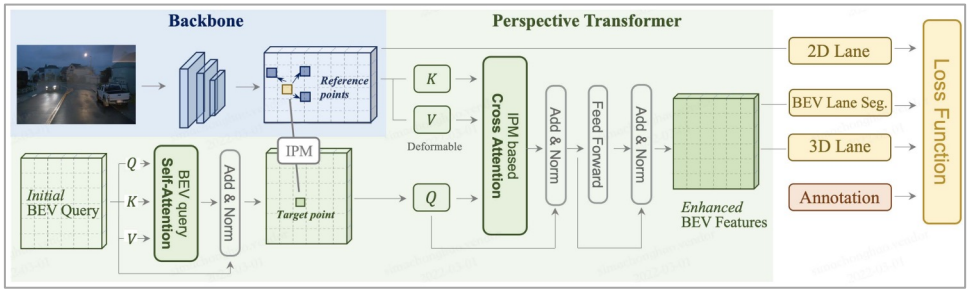
[2] PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images, *arxiv:2206.01256*.

Summary of BEV Perception in Academia

- From 2D-to-3D prior
  - 预测深度
    i. Lift, Splat, Shoot and its derivant
    ii. Pseudo-LiDAR Family



- From 3D-to-2D prior
  - 根据3D到2D的投影，index局部特征
    i. DETR3D and its derivant
    ii. Explicit BEV feature
  - Implicit 3D Positional Embedding



Both perspectives are promising on nuScenes / Waymo leaderboard

3 BEVFormer 及相关工作

BEVFormer and its Variant

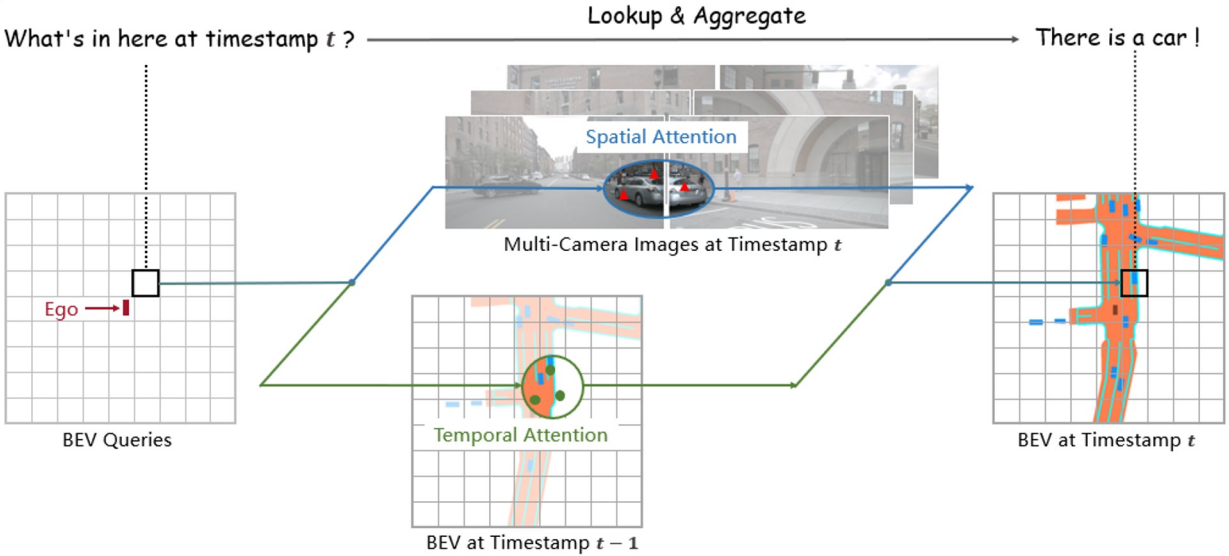
BEVFormer: Demo

BEVFormer: Motivation

BEVFormer

基于Deformable Attention模型实现了一种融合多视角相机 (multi-camera) 和时序特征的端到端框架，适用于多种自动驾驶感知任务

关键模块



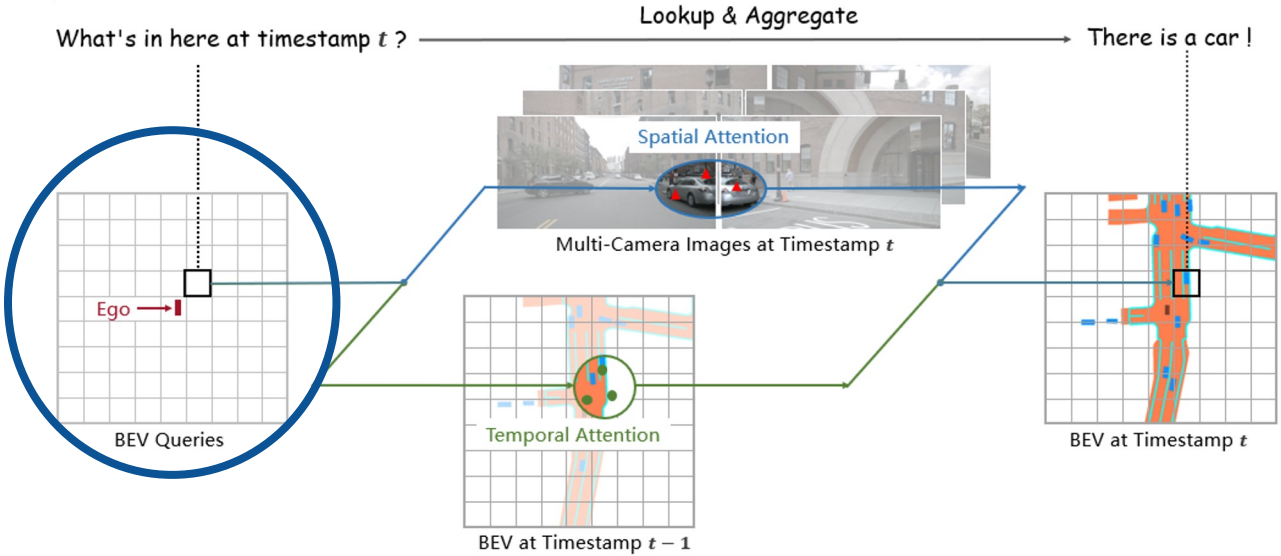
BEVFormer: Motivation

BEVFormer

基于Deformable Attention模型实现了一种融合多视角相机 (multi-camera) 和时序特征的端到端框架，适用于多种自动驾驶感知任务

关键模块

- **BEV Queries Q**：用于查询得到BEV特征图



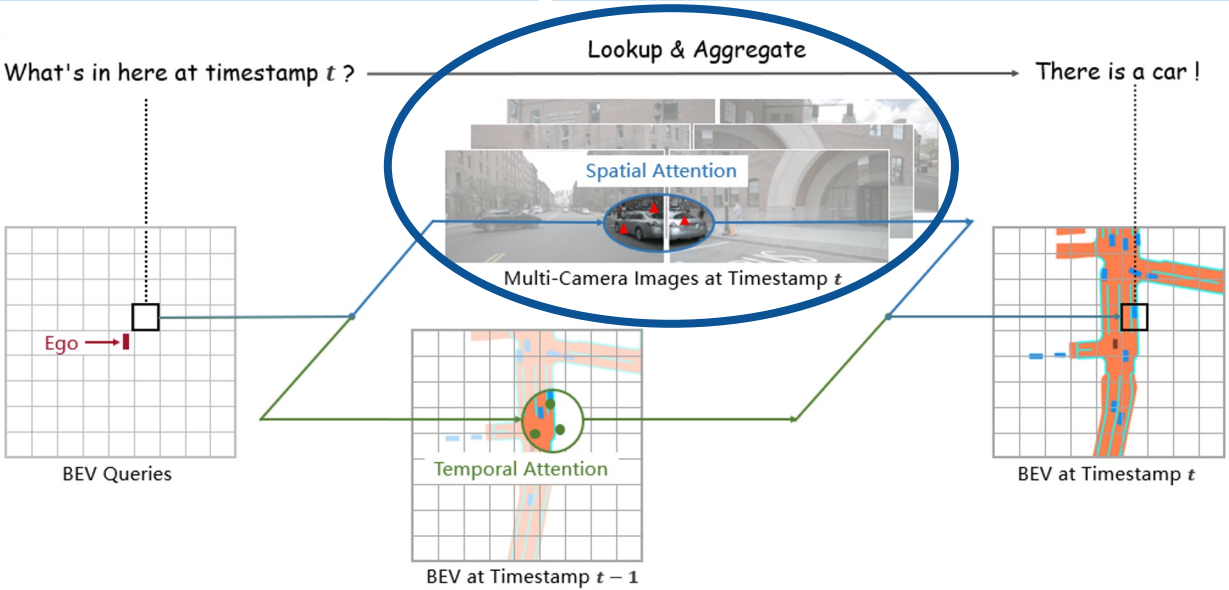
BEVFormer: Motivation

BEVFormer

基于Deformable Attention模型实现了一种融合多视角相机 (multi-camera) 和时序特征的端到端框架，适用于多种自动驾驶感知任务

关键模块

- **BEV Queries Q**：用于查询得到BEV特征图
- **Spatial Cross-Attention**：用于融合多视角特征



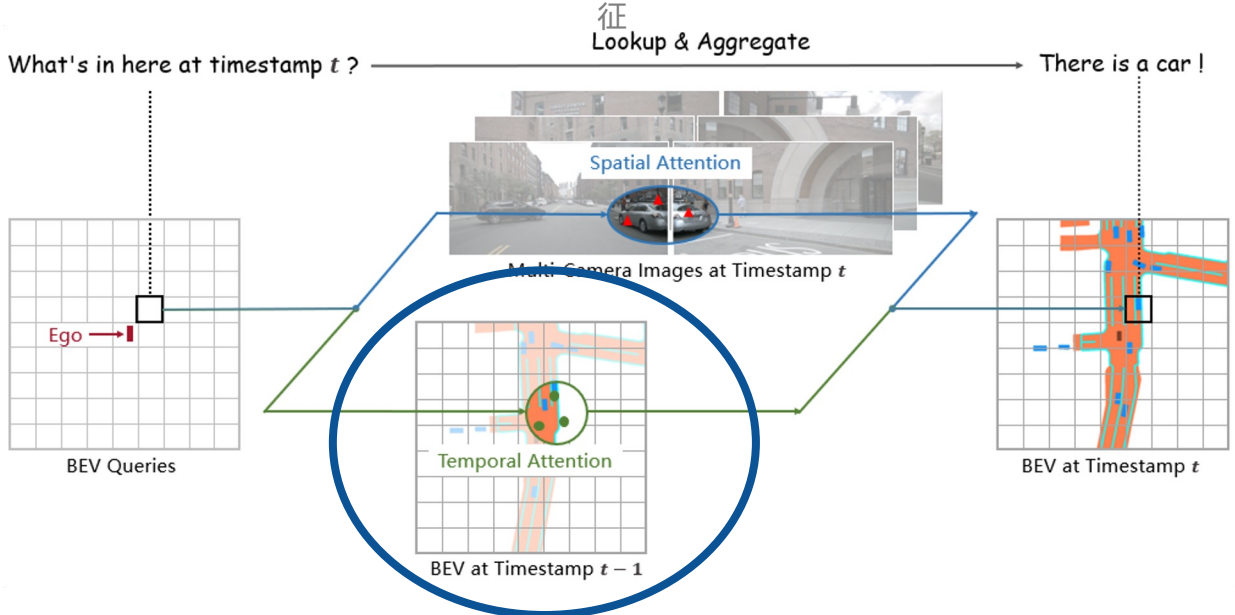
BEVFormer: Motivation

BEVFormer

基于Deformable Attention模型实现了一种融合多视角相机 (multi-camera) 和时序特征的端到端框架，适用于多种自动驾驶感知任务

关键模块

- **BEV Queries Q**：用于查询得到BEV特征图
- **Spatial Cross-Attention**：用于融合多视角特征
- **Temporal Self-Attention**：用于融合时序BEV特征



BEVFormer: Motivation

BEVFormer

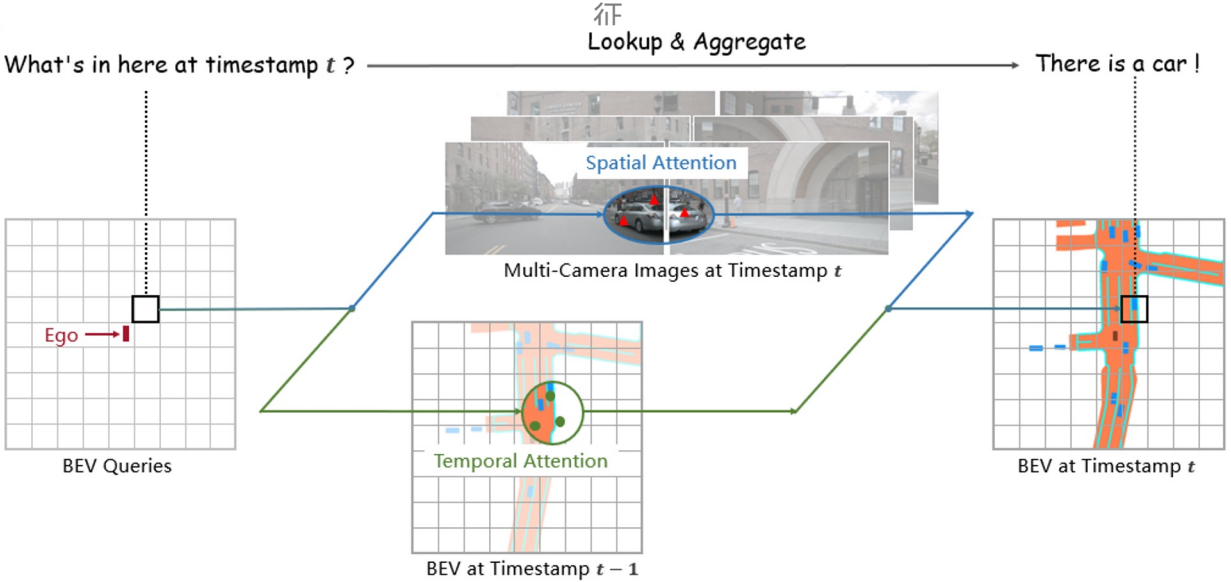
基于Deformable Attention模型实现了一种融合多视角相机 (multi-camera) 和时序特征的端到端框架，适用于多种自动驾驶感知任务

关键模块

- **BEV Queries Q**：用于查询得到BEV特征图
- **Spatial Cross-Attention**：用于融合多视角特征
- **Temporal Self-Attention**：用于融合时序BEV特征

关键点

- Using **learnable** queries to represent real world from BEV view
- Look up spatial features in images and temporal features in previous BEV map, aka **Spatial-temporal**



BEVFormer: Overall Architecture

BEVFormer

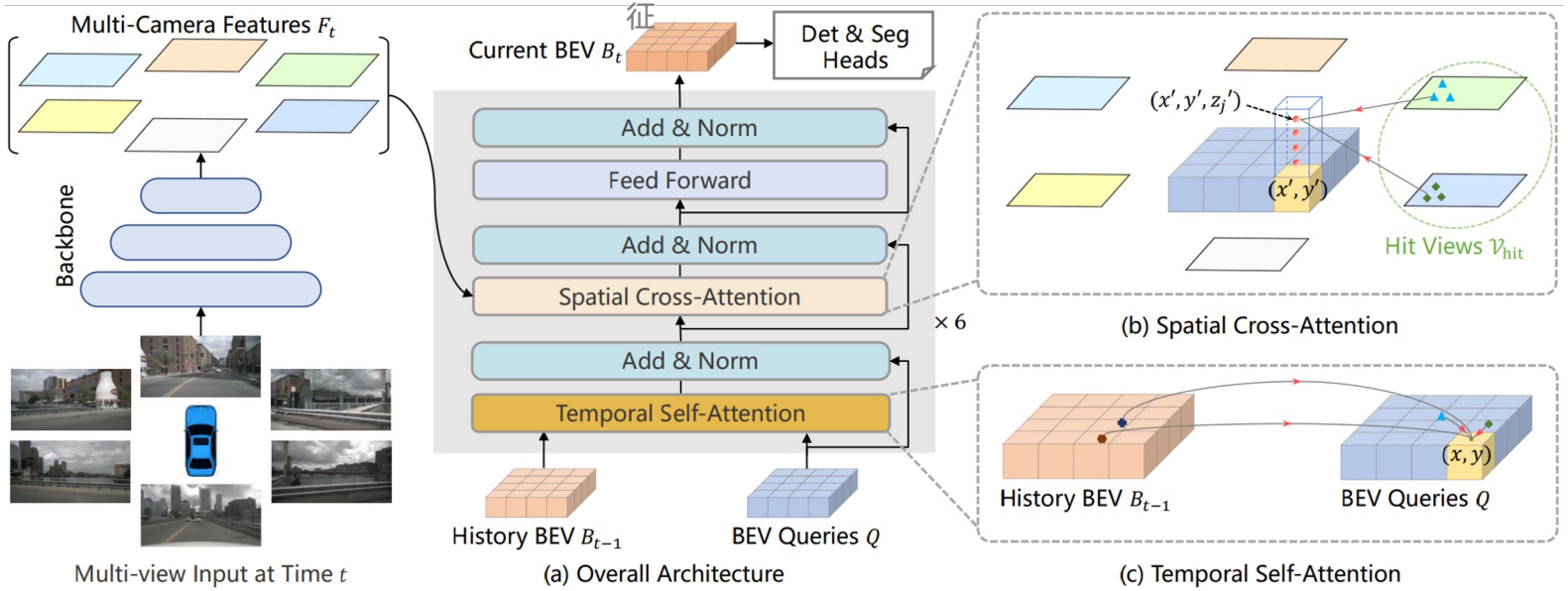
基于Deformable Attention模型实现了一种融合多视角相机 (multi-camera) 和时序特征的端到端框架，适用于多种自动驾驶感知任务

关键模块

- **BEV Queries Q**：用于查询得到BEV特征图
- **Spatial Cross-Attention**：用于融合多视角特征
- **Temporal Self-Attention**：用于融合时序BEV特征

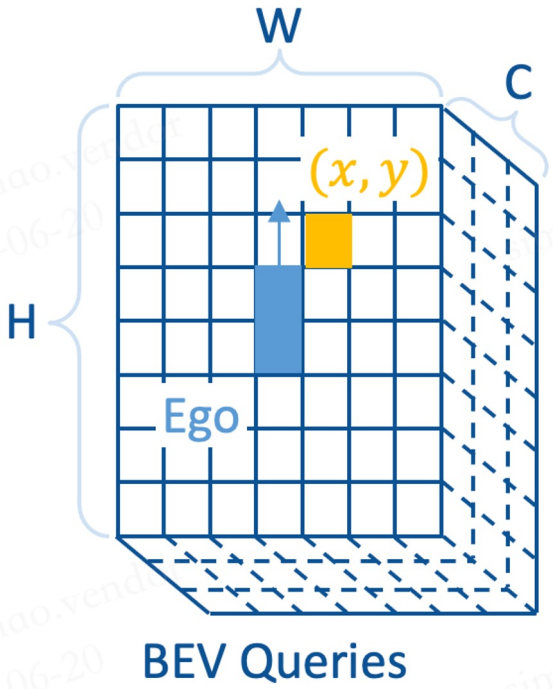
关键点

- Using **learnable** queries to represent real world from BEV view
- Look up spatial features in images and temporal features in previous BEV map, aka **Spatial-temporal**

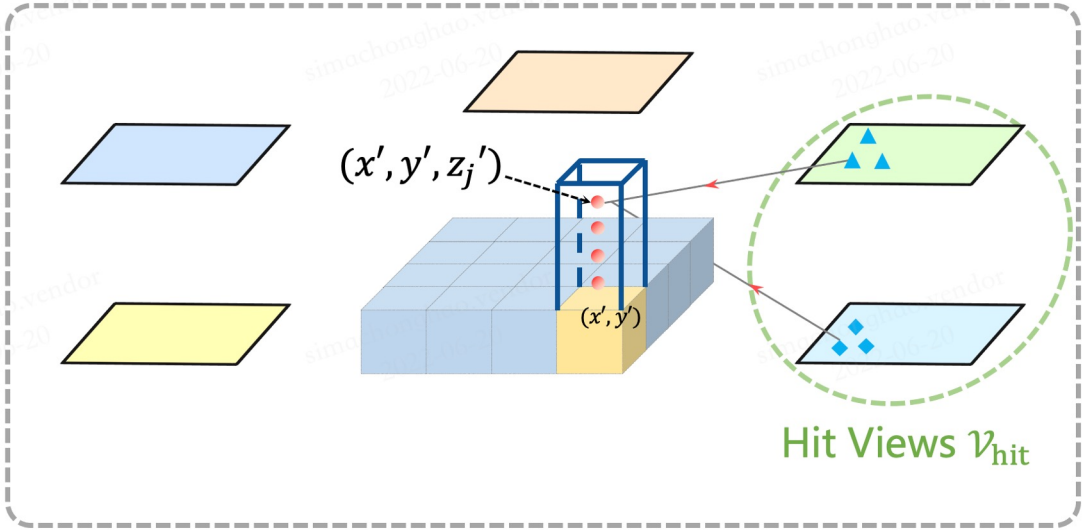


BEVFormer: BEV Queries

- BEV queries为H*W*C的***可学习参数***, 用来捕获围绕***自车***的BEV特征。
- 每个位于(x, y)位置的query都仅负责***表征其对应的小范围区域***。
- 轮番查询***spatial***和***temporal信息***，生成BEV特征图



BEVFormer: Spatial Cross-Attention



Spatial Cross-Attention

查询spatial信息

具体步骤

【Step 1】 Lift each BEV query to be a ***pillar***

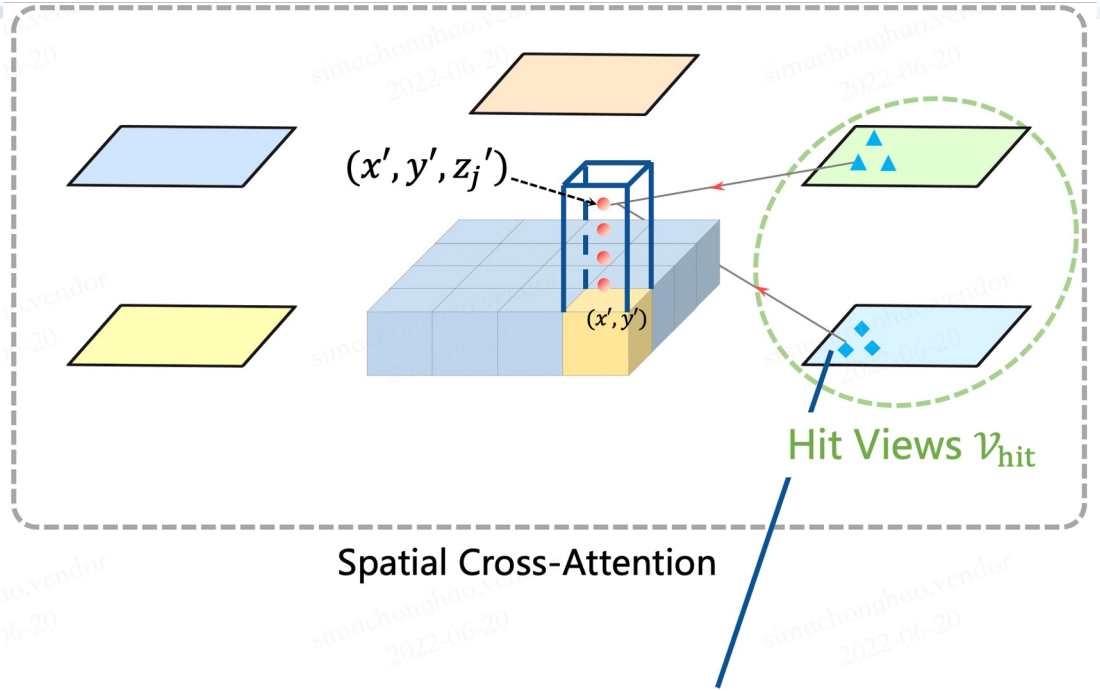
【Step 2】 Project the ***3D points*** in pillar to ***2D points*** in views

【Step 3】 Sample features from regions in ***hit views***

【Step 4】 Fuse by weight

[1] Deformable DETR: Deformable Transformers for End-to-End Object Detection, *ICLR 2021 Oral*

BEVFormer: Spatial Cross-Attention



Sparse Attention, e.g., Deformable Attention [1]

查询spatial信息

具体步骤

【Step 1】 Lift each BEV query to be a ***pillar***

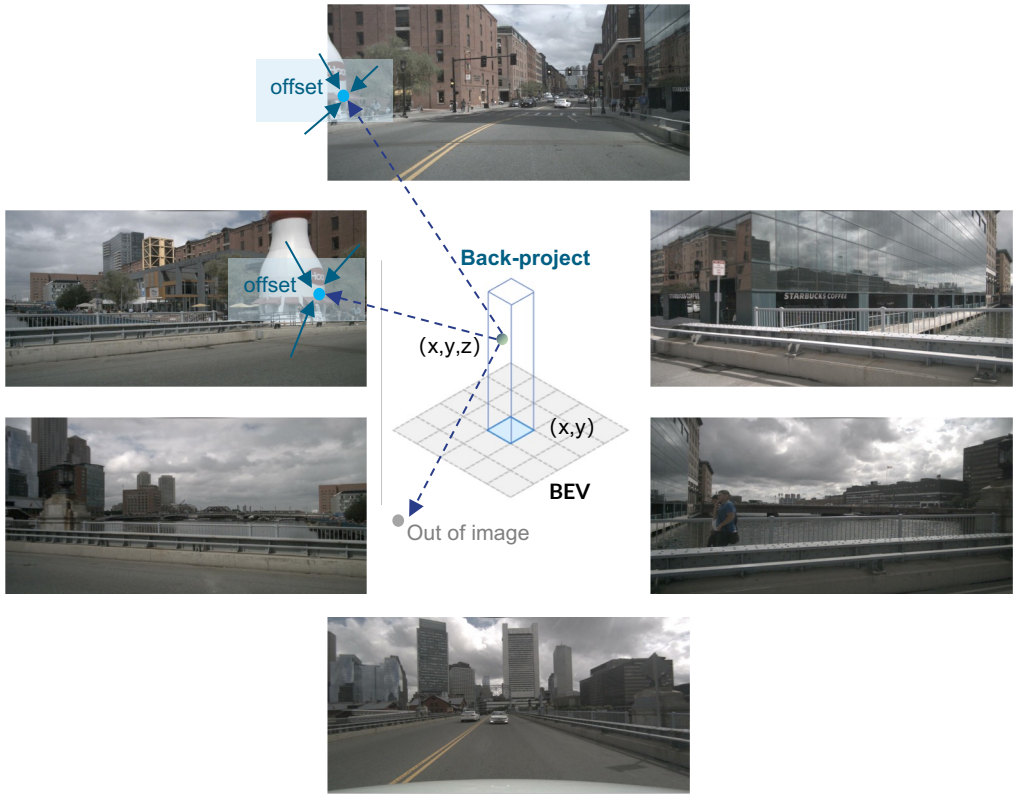
【Step 2】 Project the ***3D points*** in pillar to ***2D points*** in views

【Step 3】 Sample features from regions in ***hit views***

【Step 4】 Fuse by weight

[1] Deformable DETR: Deformable Transformers for End-to-End Object Detection, *ICLR 2021 Oral*

BEVFormer: Spatial Cross-Attention



查询spatial信息

具体步骤

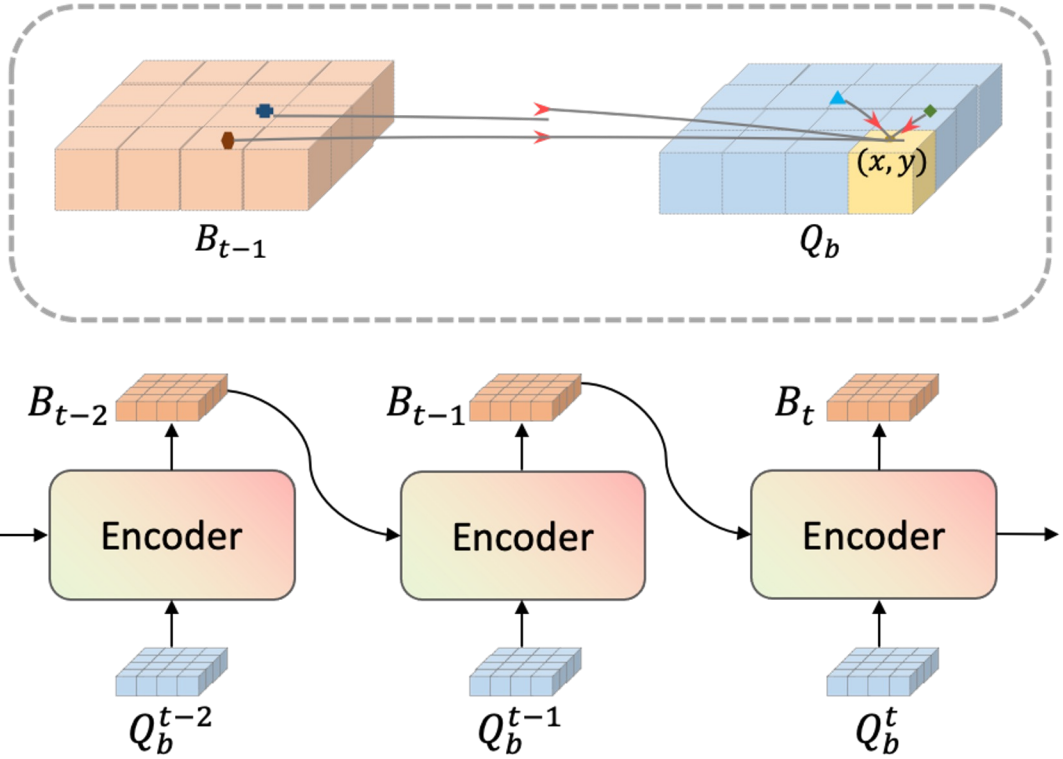
【Step 1】 Lift each BEV query to be a ***pillar***

【Step 2】 Project the ***3D points*** in pillar to ***2D points*** in views

【Step 3】 Sample features from regions in ***hit views***

【Step 4】 Fuse by weight

BEVFormer: Temporal Self-Attention



查询temporal信息

具体步骤

【Step 1】 ***Align two BEV maps*** according to the ego motion

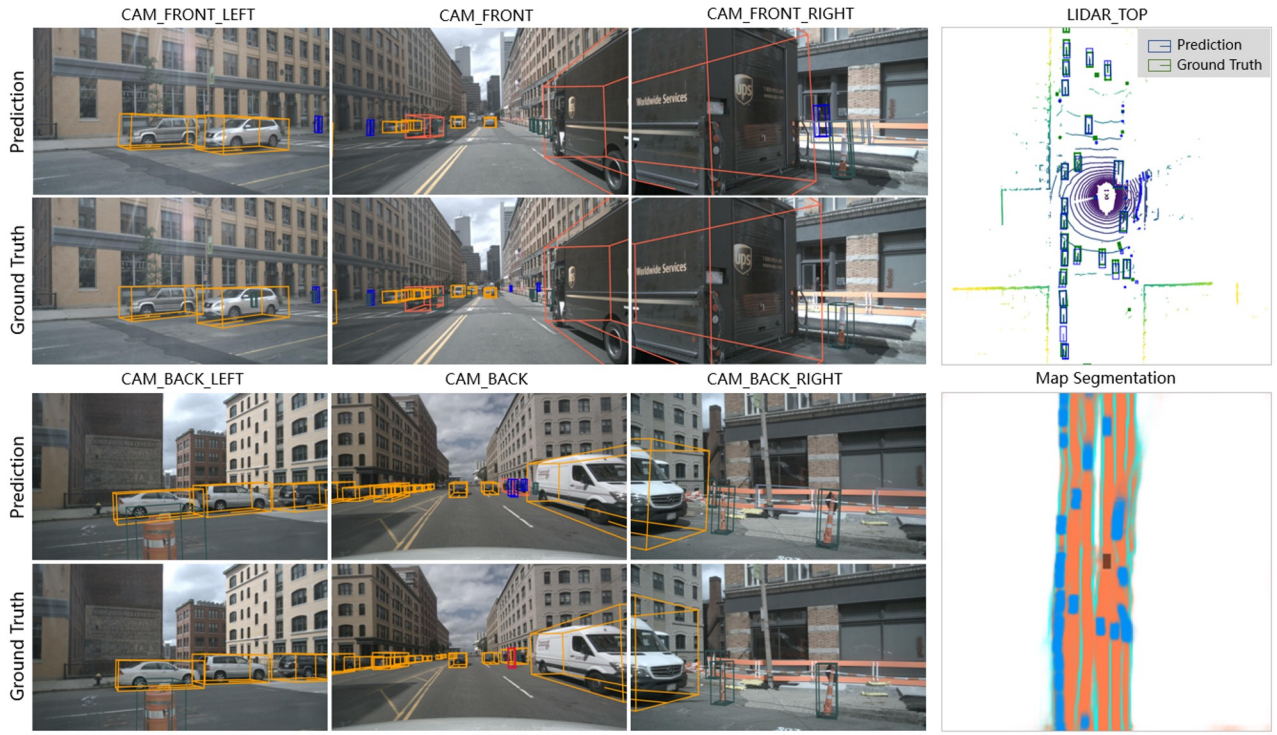
【Step 2】 Sample features from ***both past and current***.

【Step 3】 ***Weighted summation*** of sampled features from past and current BEV maps.

【Step 4】 Use ***RNN-style*** to iteratively collect history BEV features

BEVFormer: Explicit BEV feature

- 多任务学习：3D目标检测和地图语义分割
- 可迁移性：常用的2D检测头，都可以通过很小的修改迁移到3D检测上



BEVFormer: Performance on nuScenes & Waymo 1.2

nuScenes *test set*, NDS: ***56.9 v.s. 47.9***

Waymo 1.2 *val set*, L1/APH: ***28.0 v.s. 22.0***

Table 1: **3D Detection Results on nuScenes `test` set.** ∗ notes that VoVNet-99 (V2-99) [21] was pre-trained on the depth estimation task with extra data [31]. "BEVFormer-S" does not leverage temporal information in the BEV encoder. "L" and "C" indicate LiDAR and Camera, respectively.

Method	Modality	Backbone	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
SSN [54]	L	-	0.569	0.463	-	-	-	-	-
CenterPoint-Voxel [51]	L	-	0.655	0.580	-	-	-	-	-
PointPainting [43]	L&C	-	0.581	0.464	0.388	0.271	0.496	0.247	0.111
FCOS3D [45]	C	R101	0.428	0.358	0.690	0.249	0.452	1.434	**0.124**
PGD [44]	C	R101	0.448	0.386	**0.626**	**0.245**	0.451	1.509	0.127
BEVFormer-S	C	R101	0.462	0.409	0.650	0.261	0.439	0.925	0.147
BEVFormer	C	R101	**0.535**	**0.445**	0.631	0.257	**0.405**	**0.435**	0.143
DD3D [31]	C	V2-99*	0.477	0.418	**0.572**	**0.249**	**0.368**	1.014	**0.124**
DETR3D [47]	C	V2-99*	0.479	0.412	0.641	0.255	0.394	0.845	0.133
BEVFormer-S	C	V2-99*	0.495	0.435	0.589	0.254	0.402	0.842	0.131
BEVFormer	C	V2-99*	**0.569**	**0.481**	0.582	0.256	0.375	**0.378**	0.126

Table 3: **3D Detection Results on Waymo `val` set under Waymo evaluation metric and nuScenes evaluation metric.** "L1" and "L2" refer "LEVEL_1" and "LEVEL_2" difficulties of Waymo [40]. *: Only use the front camera and only consider object labels in the front camera's field of view (50.4°). †: We compute the NDS score by setting ATE and AAE to be 1. "L" and "C" indicate LiDAR and Camera, respectively.

Method	Modality	Waymo Metrics				Nuscenes Metrics				
		IoU=0.5		IoU=0.7						
		L1/APH	L2/APH	L1/APH	L2/APH	NDS†↑	AP↑	ATE↓	ASE↓	AOE↓
PointPillars [20]	L	0.866	0.801	0.638	0.557	0.685	0.838	0.143	0.132	0.070
DETR3D [47]	C	0.220	0.216	0.055	0.051	0.394	0.388	0.741	**0.156**	0.108
BEVFormer	C	**0.280**	**0.241**	**0.061**	**0.052**	**0.426**	**0.440**	**0.679**	0.157	**0.101**
CaDNN* [34]	C	0.175	0.165	0.050	0.045	-	-	-	-	-
BEVFormer*	C	0.308	0.277	0.077	0.069	-	-	-	-	-

BEVFormer: Performance on nuScenes & Waymo 1.2

nuScenes *test set*, NDS: ***56.9 v.s. 47.9***

Waymo 1.2 *val set*, L1/APH: ***28.0 v.s. 22.0***

Table 1: **3D Detection Results on nuScenes `test` set.** ∗ notes that VoVNet-99 (V2-99) [21] was pre-trained on the depth estimation task with extra data [31]. "BEVFormer-S" does not leverage temporal information in the BEV encoder. "L" and "C" indicate LiDAR and Camera, respectively.

Method	Modality	Backbone	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
SSN [54]	L	-	0.569	0.463	-	-	-	-	-
CenterPoint-Voxel [51]	L	-	0.655	0.580	-	-	-	-	-
PointPainting [43]	L&C	-	0.581	0.464	0.388	0.271	0.496	0.247	0.111
FCOS3D [45]	C	R101	0.428	0.358	0.690	0.249	0.452	1.434	**0.124**
PGD [44]	C	R101	0.448	0.386	**0.626**	**0.245**	0.451	1.509	0.127
BEVFormer-S	C	R101	0.462	0.409	0.650	0.261	0.439	0.925	0.147
BEVFormer	C	R101	**0.535**	**0.445**	0.631	0.257	**0.405**	**0.435**	0.143
DD3D [31]	C	V2-99*	0.477	0.418	**0.572**	**0.249**	**0.368**	1.014	**0.124**
DETR3D [47]	C	V2-99*	0.479	0.412	0.641	0.255	0.394	0.845	0.133
BEVFormer-S	C	V2-99*	0.495	0.435	0.589	0.254	0.402	0.842	0.131
BEVFormer	C	V2-99*	**0.569**	**0.481**	0.582	0.256	0.375	**0.378**	0.126

Table 3: **3D Detection Results on Waymo `val` set under Waymo evaluation metric and nuScenes evaluation metric.** "L1" and "L2" refer "LEVEL_1" and "LEVEL_2" difficulties of Waymo [40]. *: Only use the front camera and only consider object labels in the front camera's field of view (50.4°). †: We compute the NDS score by setting ATE and AAE to be 1. "L" and "C" indicate LiDAR and Camera, respectively.

Method	Modality	Waymo Metrics				Nuscenes Metrics				
		IoU=0.5		IoU=0.7						
		L1/APH	L2/APH	L1/APH	L2/APH	NDS†↑	AP↑	ATE↓	ASE↓	AOE↓
PointPillars [20]	L	0.866	0.801	0.638	0.557	0.685	0.838	0.143	0.132	0.070
DETR3D [47]	C	0.220	0.216	0.055	0.051	0.394	0.388	0.741	**0.156**	0.108
BEVFormer	C	**0.280**	**0.241**	**0.061**	**0.052**	**0.426**	**0.440**	**0.679**	0.157	**0.101**
CaDNN* [34]	C	0.175	0.165	0.050	0.045	-	-	-	-	-
BEVFormer*	C	0.308	0.277	0.077	0.069	-	-	-	-	-

消融实验总结

- Still, ***strong backbone*** matters.

BEVFormer: Performance on nuScenes & Waymo 1.2

nuScenes *test set*, NDS: ***56.9 v.s. 47.9***

Waymo 1.2 *val set*, L1/APH: ***28.0 v.s. 22.0***

Table 1: **3D Detection Results on nuScenes `test` set.** ∗ notes that VoVNet-99 (V2-99) [21] was pre-trained on the depth estimation task with extra data [31]. "BEVFormer-S" does not leverage temporal information in the BEV encoder. "L" and "C" indicate LiDAR and Camera, respectively.

Method	Modality	Backbone	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
SSN [54]	L	-	0.569	0.463	-	-	-	-	-
CenterPoint-Voxel [51]	L	-	0.655	0.580	-	-	-	-	-
PointPainting [43]	L&C	-	0.581	0.464	0.388	0.271	0.496	0.247	0.111
FCOS3D [45]	C	R101	0.428	0.358	0.690	0.249	0.452	1.434	**0.124**
PGD [44]	C	R101	0.448	0.386	**0.626**	**0.245**	0.451	1.509	0.127
BEVFormer-S	C	R101	0.462	0.409	0.650	0.261	0.439	0.925	0.147
BEVFormer	C	R101	**0.535**	**0.445**	0.631	0.257	**0.405**	**0.435**	0.143
DD3D [31]	C	V2-99*	0.477	0.418	**0.572**	**0.249**	**0.368**	1.014	**0.124**
DETR3D [47]	C	V2-99*	0.479	0.412	0.641	0.255	0.394	0.845	0.133
BEVFormer-S	C	V2-99*	0.495	0.435	0.589	0.254	0.402	0.842	0.131
BEVFormer	C	V2-99*	**0.569**	**0.481**	0.582	0.256	0.375	**0.378**	0.126

Table 3: **3D Detection Results on Waymo `val` set under Waymo evaluation metric and nuScenes evaluation metric.** "L1" and "L2" refer "LEVEL_1" and "LEVEL_2" difficulties of Waymo [40]. *: Only use the front camera and only consider object labels in the front camera's field of view (50.4°). †: We compute the NDS score by setting ATE and AAE to be 1. "L" and "C" indicate LiDAR and Camera, respectively.

Method	Modality	Waymo Metrics				Nuscenes Metrics				
		IoU=0.5		IoU=0.7						
		L1/APH	L2/APH	L1/APH	L2/APH	NDS†↑	AP↑	ATE↓	ASE↓	AOE↓
PointPillars [20]	L	0.866	0.801	0.638	0.557	0.685	0.838	0.143	0.132	0.070
DETR3D [47]	C	0.220	0.216	0.055	0.051	0.394	0.388	0.741	**0.156**	0.108
BEVFormer	C	**0.280**	**0.241**	**0.061**	**0.052**	**0.426**	**0.440**	**0.679**	0.157	**0.101**
CaDNN* [34]	C	0.175	0.165	0.050	0.045	-	-	-	-	-
BEVFormer*	C	0.308	0.277	0.077	0.069	-	-	-	-	-

消融实验总结

- Still, ***strong backbone*** matters.
- ***Local attention*** is better for global attention (~4.4 in NDS)
- ***Temporal clues*** matters (higher recall, more accurate velocity)
- ***Multi-tasks learning*** benefits 3D object detection but hurts BEV map segmentation

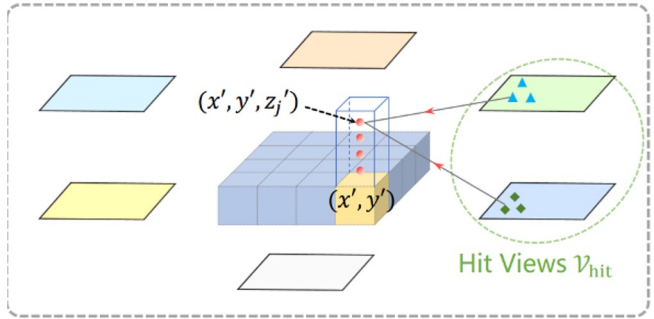
BEVFormer: Ablation on Attention Module

Table 5: The detection results of different methods with **various BEV encoders** on nuScenes val set.

Method	Attention	NDS↑	mAP↑	mATE↓	mAOE↓	#Param.	FLOPs	Memory
VPN* [30]	-	0.334	0.252	0.926	0.598	111.2M	924.5G	~20G
List-Splat* [32]	-	0.397	0.348	0.784	0.537	74.0M	1087.7G	~20G
BEVFormer-S†	Global	0.404	0.325	0.837	0.442	62.1M	1245.1G	~36G
BEVFormer-S‡	Points	0.423	0.351	0.753	0.442	68.1M	1264.3G	~20G
BEVFormer-S	Local	**0.448**	**0.375**	**0.725**	**0.391**	68.7M	1303.5G	~20G

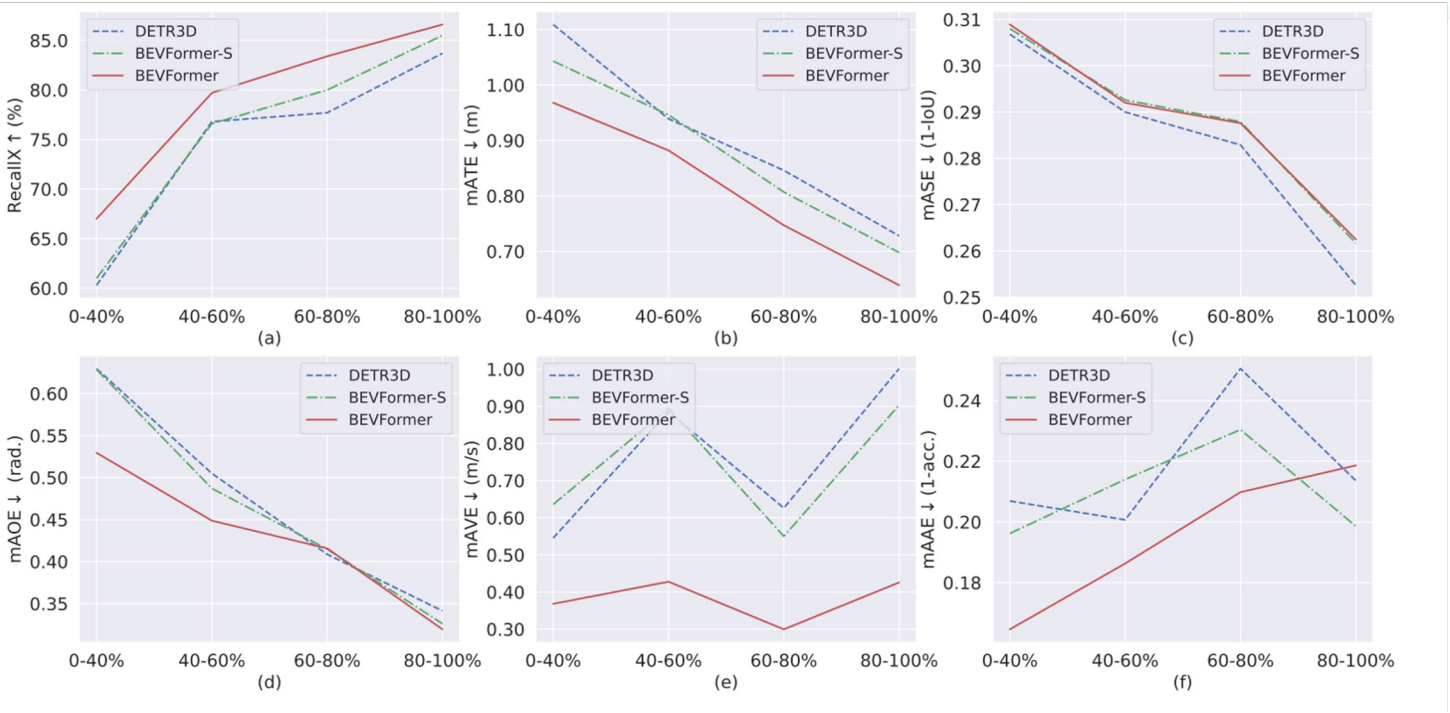
Ablation Sum-up

- Global attention consumes too much resource.
- Point interaction has limited receptive field.
- Deformable attention can balance the computational cost and receptive field.



(b) Spatial Cross-Attention

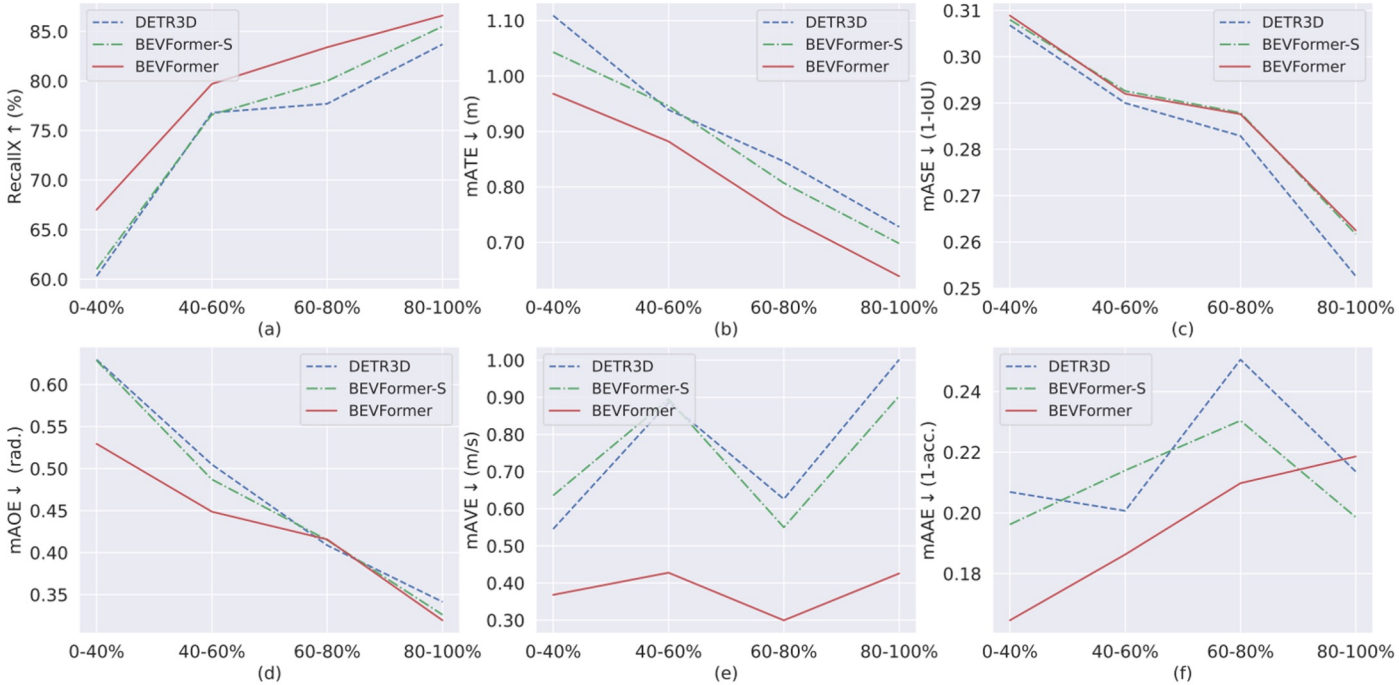
BEVFormer: Ablation on Temporal Clues



visibility that {0-40%, 40-60%, 60-80%, 80-100%} of objects can be visible

- mATE: mean Average Translation Error
- mASE: mean Average Size Error
- mAOE: mean Average Orientation Error
- mAVE: mean Average Velocity Error
- mAAE: mean Average Attribute Error

BEVFormer: Ablation on Temporal Clues



With temporal clues,

we obtain:

- **Higher recall,** especially for **low-visible** objects
- More accurate **location estimation**
- Very accurate estimation of **velocity**

visibility that {0-40%, 40-60%, 60-80%, 80-100%} of objects can be visible

BEVFormer: Ablation on Multi-task Learning

Table 4: **The Results on 3D detection and map segmentation task.** Comparison of training segmentation and detection tasks jointly or not. *: We use VPN [30] and Lift-Splat [32] to replace our BEV encoder for comparison, and the task heads are the same. †: Results from their paper.

Method	Task Head		3D Detection		BEV Segmentation (IoU)			
	Det	Seg	NDS↑	mAP↑	Car	Vehicles	Road	Lane
Lift-Splat† [32]	✗	✓	-	-	32.1	32.1	72.9	20.0
FIERY† [18]	✗	✓	-	-	-	38.2	-	-
VPN* [30]	✓	✗	0.333	0.253	-	-	-	-
VPN*	✗	✓	-	-	31.0	31.8	76.9	19.4
VPN*	✓	✓	0.334	0.257	36.6	37.3	76.0	18.0
Lift-Splat*	✓	✗	0.397	0.348	-	-	-	-
Lift-Splat*	✗	✓	-	-	42.1	41.7	77.7	20.0
Lift-Splat*	✓	✓	0.410	0.344	43.0	42.8	73.9	18.3
BEVFormer-S	✓	✗	0.448	0.375	-	-	-	-
BEVFormer-S	✗	✓	-	-	43.1	43.2	**80.7**	21.3
BEVFormer-S	✓	✓	0.453	0.380	44.3	44.4	77.6	19.8
BEVFormer	✓	✗	0.517	**0.416**	-	-	-	-
BEVFormer	✗	✓	-	-	44.8	44.8	80.1	**25.7**
BEVFormer	✓	✓	**0.520**	0.412	**46.8**	**46.7**	77.5	23.9

With multi-tasks training, we obtain:

- Higher NDS with multiple task heads
- Lower Road and Lane IoU

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

BEVFormer++

BEVFormer++： Waymo Challenge Camera-Only Track

>1,700 valid submissions

>15 countries

>400 Community discussions

BEVFormer++：Waymo Challenge Camera-Only Track

不同面积大小代表不同收益

核心指标APL相比于BEVFormer baseline改进带来性能提升21%

Method	LET-mAPL	LET-mAP	LET-mAPH
Waymo Baseline	14.77	22.61	18.17
BEVFormer Baseline	34.6	50.2	46.1
Our Solution	**56.16**	**70.69**	**65.93**

BEVFormer++： Waymo Challenge Camera-Only Track

| View Transformation优化

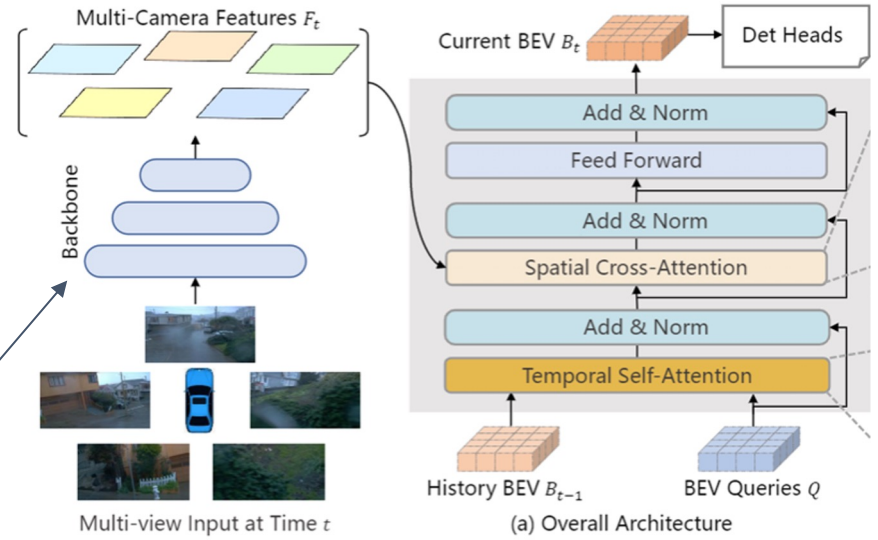
Recall

核心问题: 如何建模从front view转移到BEV (View Transformation) 得到更有效Feature?

BEVFormer++： Waymo Challenge Camera-Only Track

| View Transformation优化

更大backbone

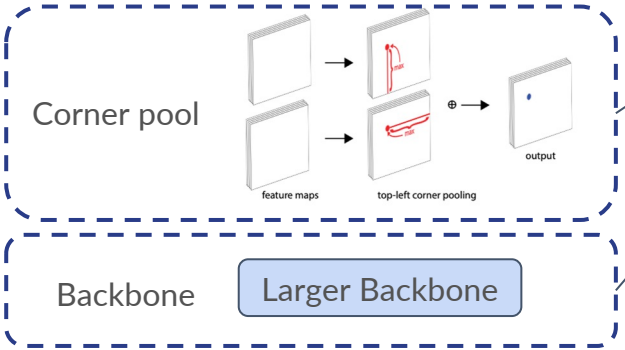


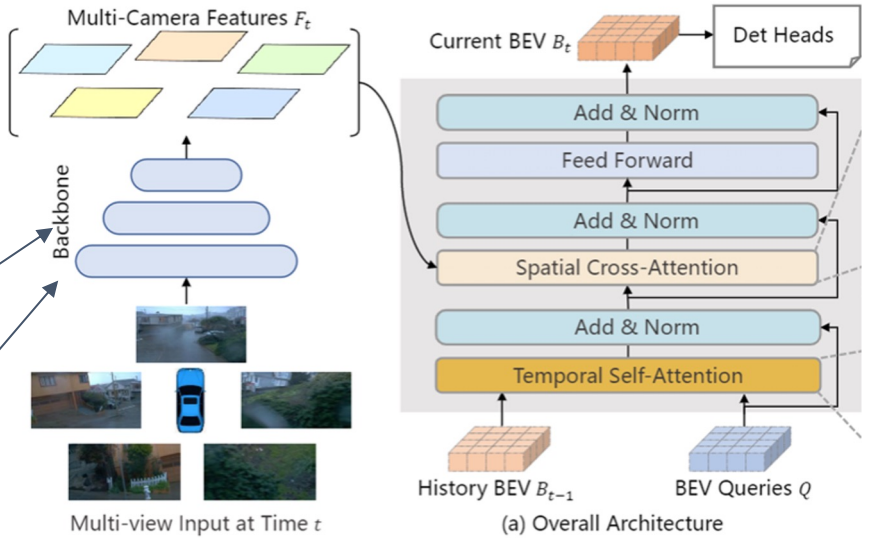
Backbone → Larger Backbone

BEVFormer++： Waymo Challenge Camera-Only Track

| View Transformation优化

在图像特征中加入corner pooling，增加特征感受野，从而提升BEV特征质量

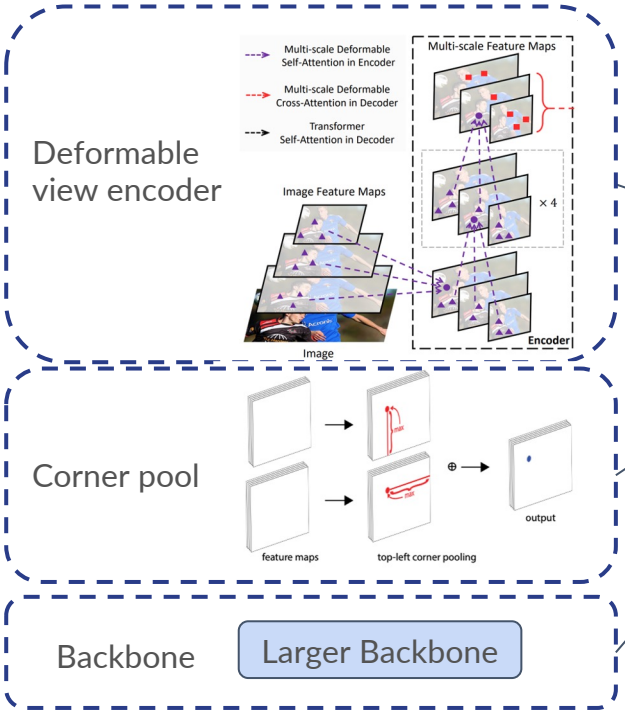


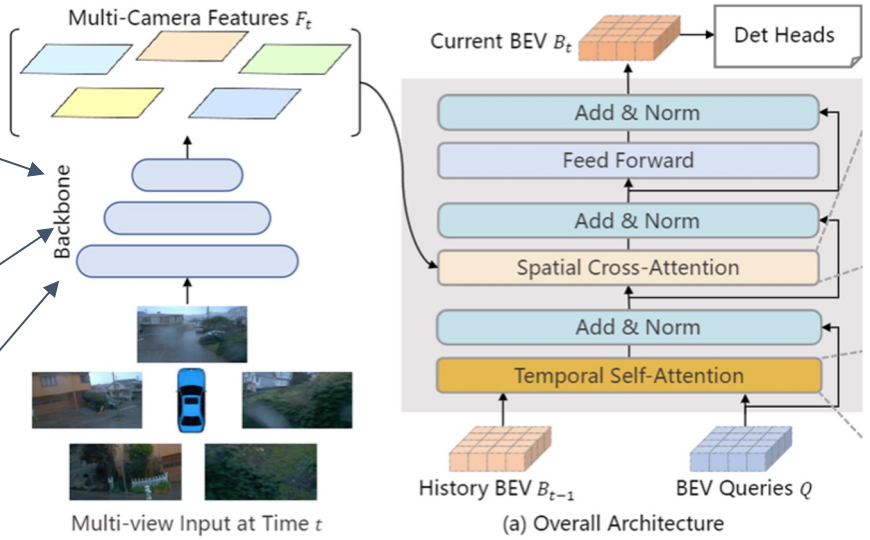


BEVFormer++： Waymo Challenge Camera-Only Track

| View Transformation优化

利用deformable view encoder增强特征的多尺度融合

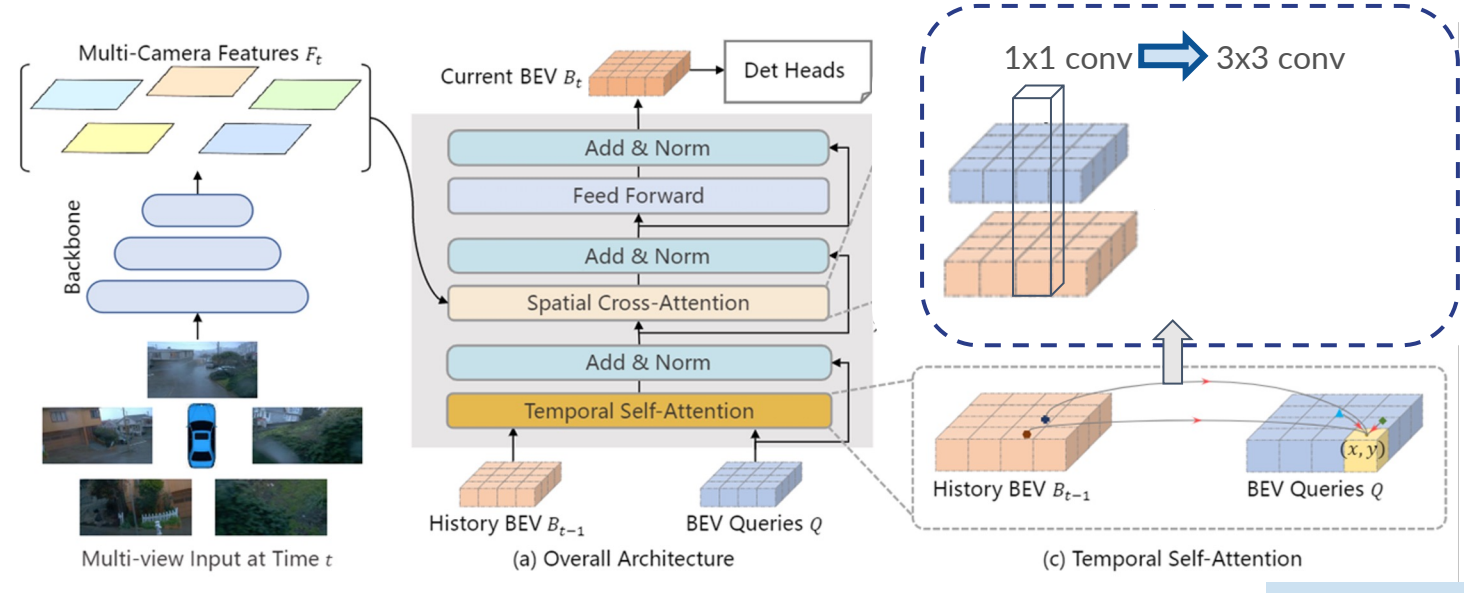




BEVFormer++：Waymo Challenge Camera-Only Track

| 时序特征优化

用3x3 conv***代替***时序注意力中的linear offset预测，提升对于移动物体和相机**参数不准等情况的鲁棒性**

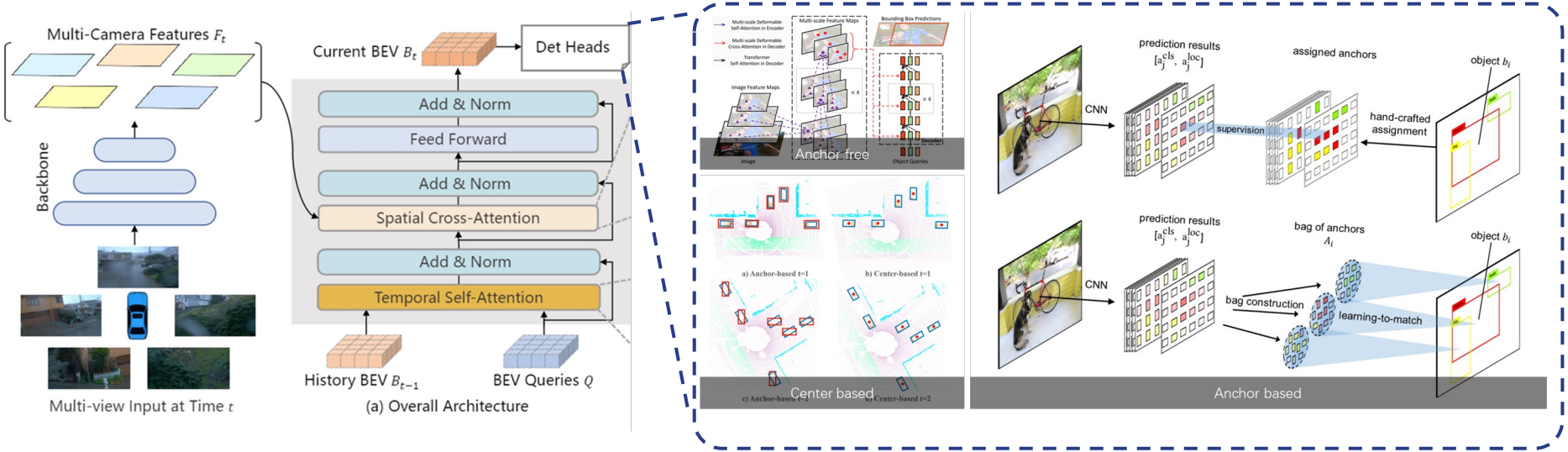


Gain: +1.3%

BEVFormer++： Waymo Challenge Camera-Only Track

| BEV下多种检测器应用

在BEV特征上采用了三种不同的检测头，并将所有模型用于ensemble以提升最终性能

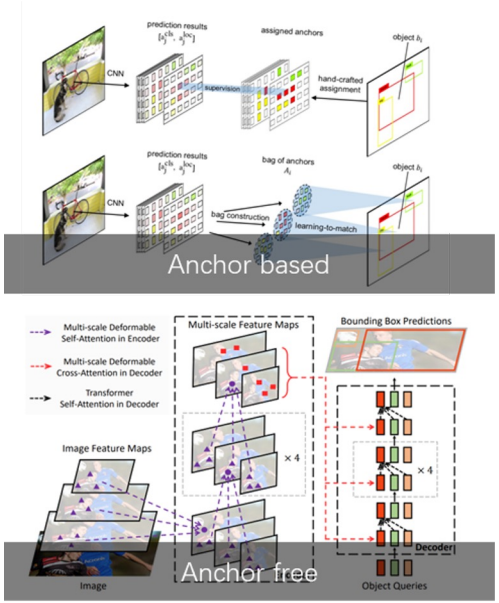


Gain: +3.7%

BEVFormer++：Waymo Challenge Camera-Only Track

| BEV下多种检测器应用

在BEV特征上采用了三种不同的检测头，并将所有模型用于ensemble以提升最终性能



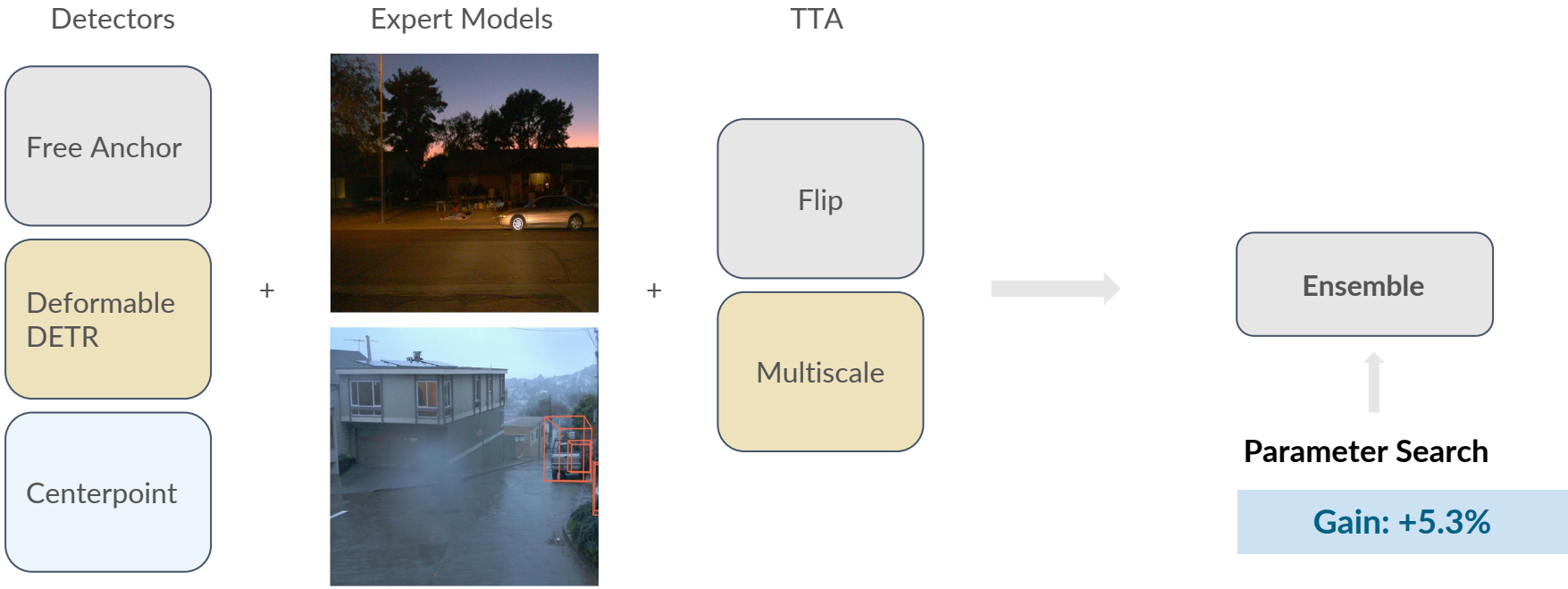
- anchor based head收敛快，BEV下缓解遮挡
- ***anchor based在车类等大物体上指标更好***

- DETR head基于Hungarian算法assign，对BEV下小物体更友好
- ***anchor free在行人等小物体上指标更好***

BEVFormer++：Waymo Challenge Camera-Only Track

| 多种模型的Ensemble策略

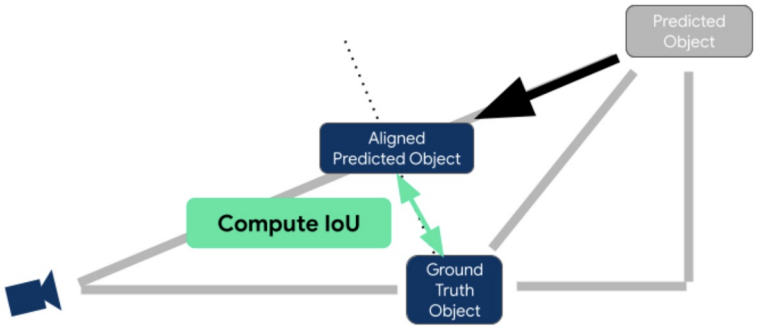
充分利用基于不同场景/不同类别数据的expert model + 不同结构的模型 + TTA (Test Time Augmentation)，
利用遗传算法搜索得到最优ensemble参数

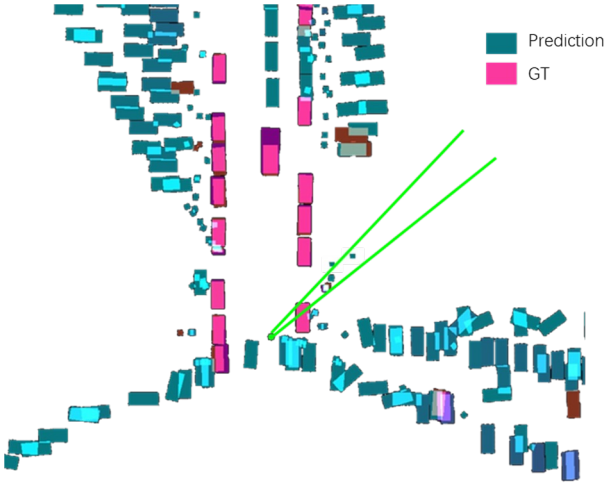


BEVFormer++：Waymo Challenge Camera-Only Track

| LET IoU [1] based assignment & NMS

- Assignment: 更偏向于径向分布的BEV特征点，更集中于图像上物体对应特征
- NMS：对于行人等小物体更加友好





LET IoU更易于将小物体拉近

Gain: +2.9%

[1] LET-3D-AP: Longitudinal Error Tolerant 3D Average Precision for Camera-Only 3D Detection, *arXiv:2206.07705*.

BEVFormer: Performance on Waymo Challenge 2022

2022 Waymo Challenge
Camera-only Track

第一名：*BEVFormer*
Shanghai AI Lab/上海人工智能实验室
56.2

Leaderboard

* disqualified from the 2022 Waymo Open Dataset Challenge

Method Name	Object Type	Camera	Frames [-p]	LET-3D-APL	LET-3D-AP	Date (Pacific Daylight Time)
	ALL_NS	ALL				
1 BEVFormer-Shanghai AI Lab	ALL_NS	ALL	0	0.5616	0.7069	2022-05-23 23:21
2 MV-FCOS3D++	ALL_NS	ALL	0	0.5127	0.6627	2022-05-24 00:04
3 FCOS3D-MVDet3D	ALL_NS	ALL	0	0.5106	0.6603	2022-05-23 22:51
4 DETR4D	ALL_NS	ALL	0	0.4927	0.6656	2022-05-24 02:07
5 3DMVT*	ALL_NS	ALL	0	0.4876	0.6550	2022-05-23 21:00
6 CMKD	ALL_NS	ALL	0	0.3790	0.5101	2022-05-23 22:53
7 Anonymous Submission*	ALL_NS	ALL	0	0.3576	0.5162	2022-05-23 06:12
8 CenterLS	ALL_NS	ALL	0	0.2914	0.4117	2022-05-23 01:53
9 MVImages2PointsDet	ALL_NS	ALL	0	0.2387	0.3853	2022-05-23 08:39
10 MonoWatch	ALL_NS	ALL	0	0.2163	0.2833	2022-05-23 22:15

Page 1 of 2 10 Rows

Waymo榜单：As of 2022.6.9
https://waymo.com/open/challenges/2022/3d-camera-only-detection/

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

其他流行工作

CONTENTS

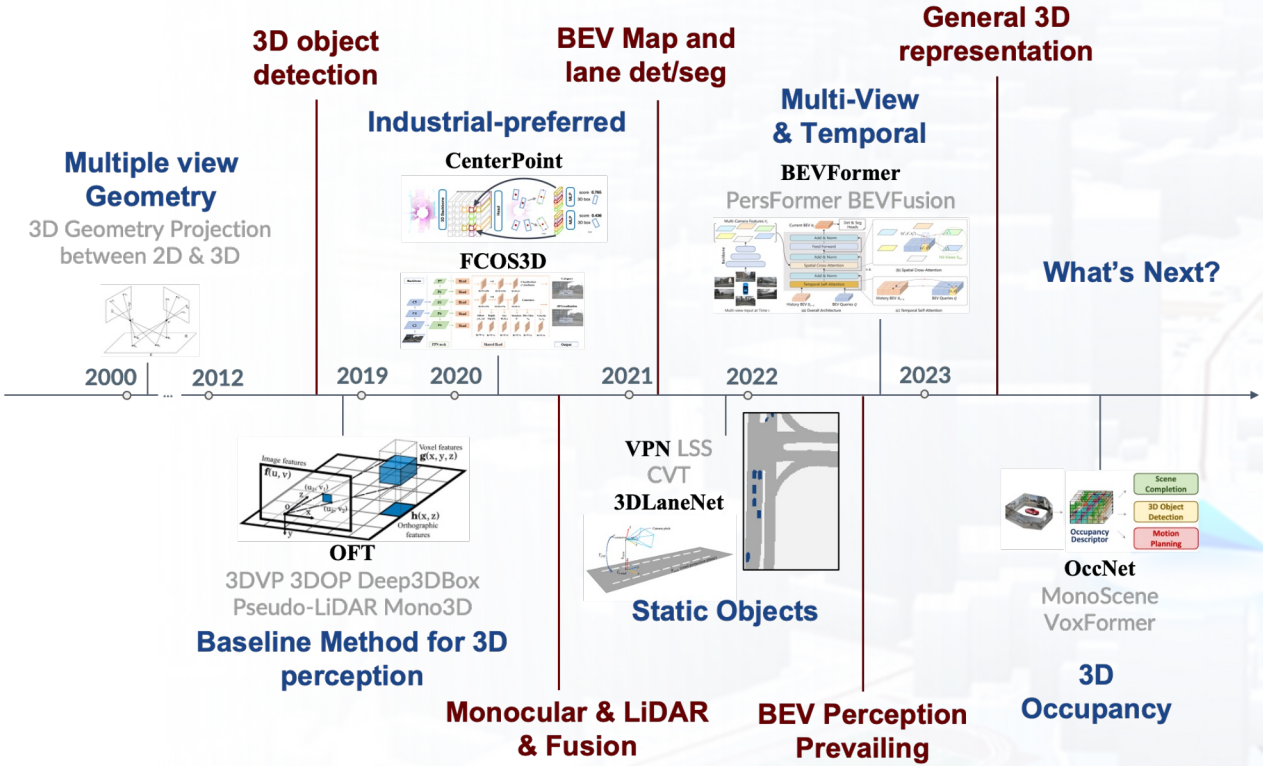
ICCV23
PARIS

Scene as Occupancy

Github: https://github.com/OpenDriveLab/OccNet

arXiv: https://arxiv.org/abs/2306.02851

How does 3D perception evolve into 3D occupancy?



3D Occupancy Prediction

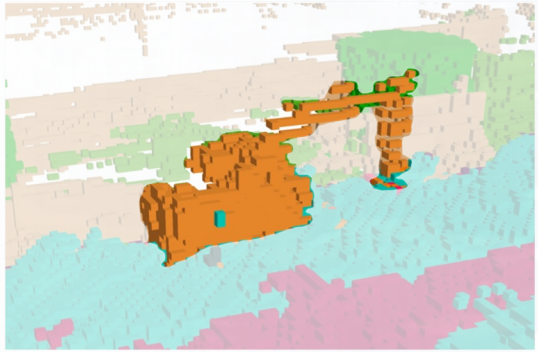
(a)

What's the problem of current 3D perception representation?

- **3D bbox (a)** ignores the **detailed geometric** of an irregular object.

3D Occupancy Prediction

(a)



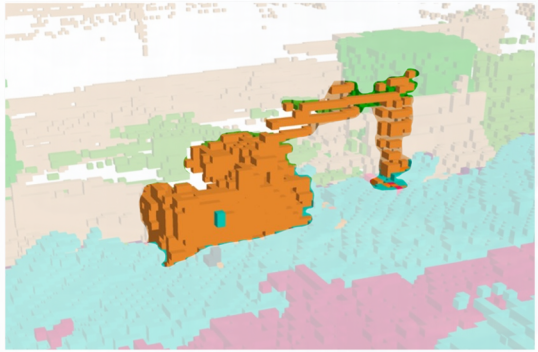
(b)

What's the problem of current 3D perception representation?

- **3D bbox (a)** ignores the **detailed geometric** of an irregular object.
- while **3D occupancy (b)** catches the geometric shape well.

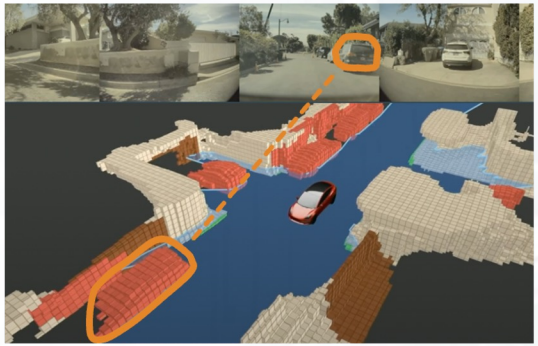
3D Occupancy Prediction

(a)



(b)

(c)

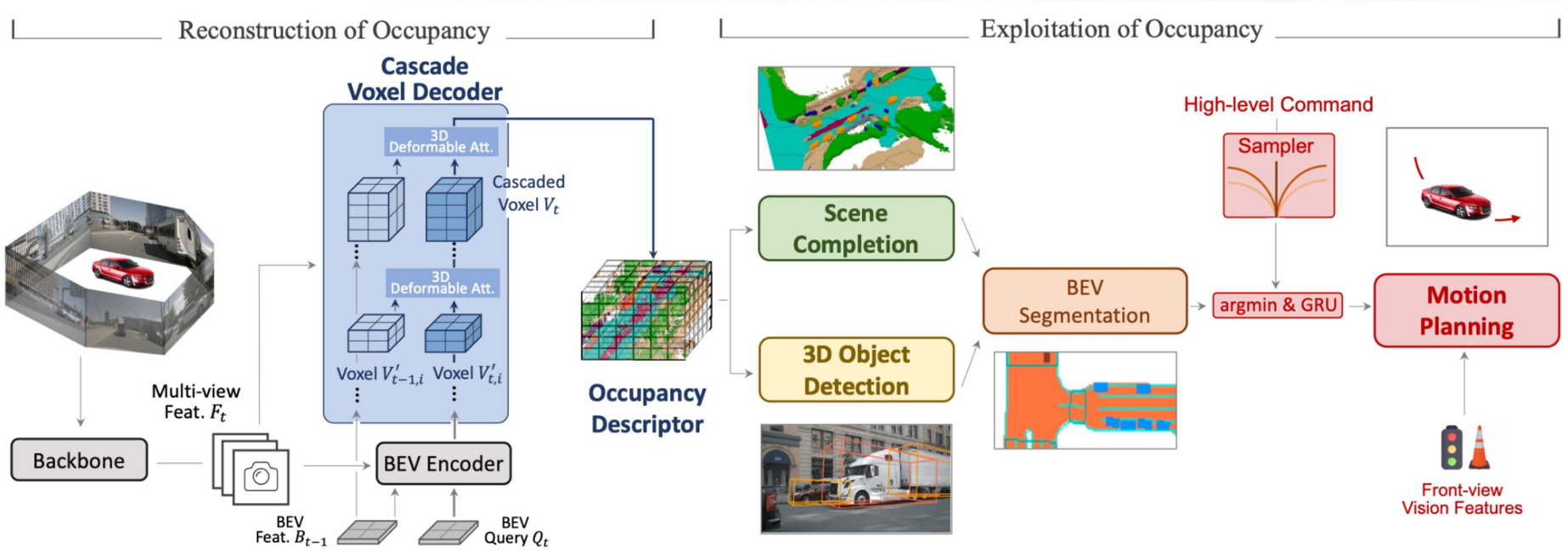


(d)

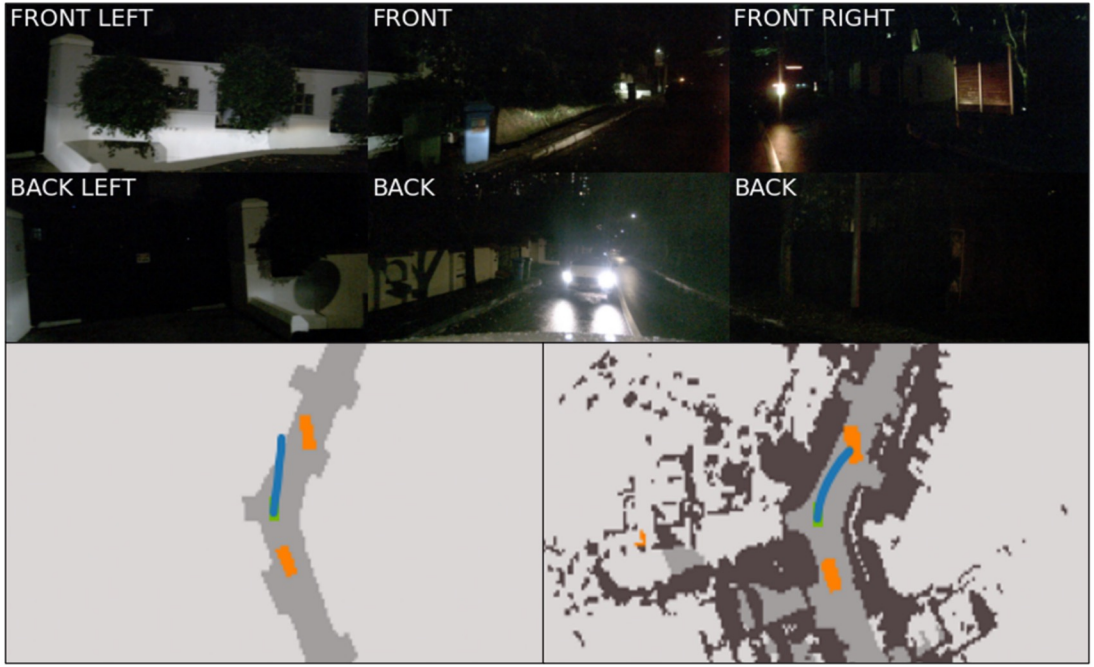
What's the problem of current 3D perception representation?

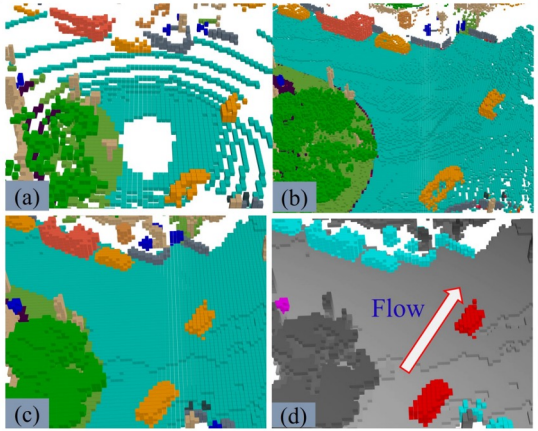
- **3D bbox (a)** ignores the **detailed geometric** of an irregular object.
- while **3D occupancy (b)** catches the geometric shape well.
- **Mobileye (c) and Tesla (d)** also adore such a representation.

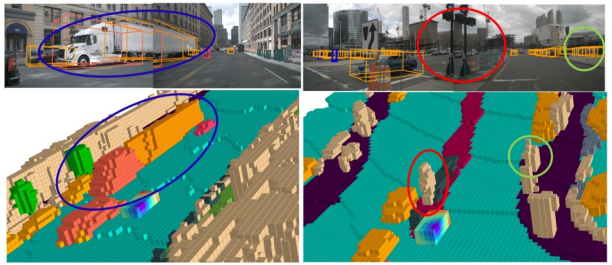
Scene as Occupancy - Pipeline



Scene as Occupancy - Experiments







Occupancy - Future Work

- **3D Visual Pre-training**
  - given **large-scale paired point cloud and images**, can we pre-train **ONE** strong 3D backbone that facilitates **MANY** 3D tasks such as detection, depth, and so on.

- **Open Questions**
  - **Sparse / deformable / long-range representation.**
  - **Open-vocabulary to address unseen objects.**
  - **Explicit guidance from occupancy to planning**

4 BEV感知｜思考与讨论

BEV感知工作：未来思考

1

- 以纯视觉为基础，继续提升物体环视检测性能，弥补**纯视觉**的3D检测与**LiDAR**物体检测的性能差异
  - Depth/Pseudo-LiDAR

2

- 多模态信息融合，视觉与LiDAR特征**融合**
  - 如何高效、优雅的融合
  - Waymo leaderboard前几名都以LiDAR-only为主，why?

3

- 对部署更为**友好的**BEV模型
  - 工业界如何应用起来Transformer等heavy结构?

BEV感知工作：未来思考

1

- 以纯视觉为基础，继续提升物体环视检测性能
  - LiDAR与纯视觉物体检测性能存在明显差距

Lidar	no	no	0.637	0.256	0.233	0.321	0.216	0.122	0.704	0.682	n/a	
Camera	no	no	0.445	0.631	0.257	0.405	0.435	0.143	0.535	1.096	n/a	

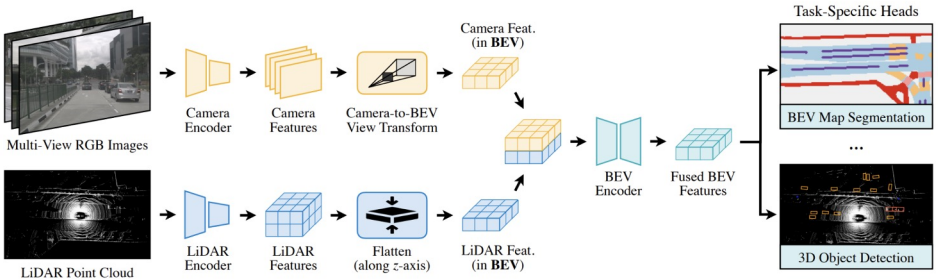
*nuScenes，截止到三月

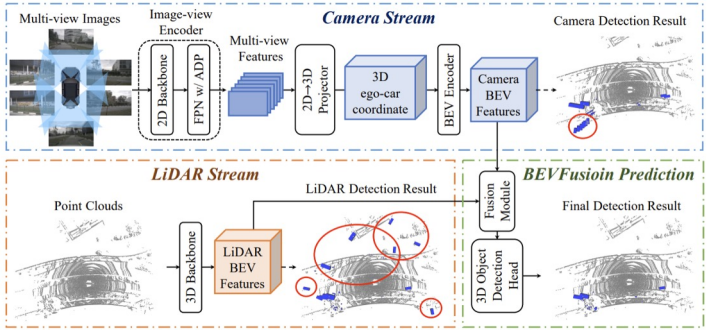
  - 目前发现一些比较有潜力的点
    - Depth Pretrain: 目前在不少文章中的实验结果都表明用上深度预训练能够明显提升3D检测性能
    - 时序信息：前后帧信息是解决深度问题的关键
    - 模型设计：更合理的模型设计从根本上提升BEV检测的性能

BEV感知工作：未来思考

2

- 多模态信息融合，视觉与LiDAR特征融合
  - BEV同时作为LiDAR下游任务常见的一种特征结构，可以更好地与图像BEV特征align和融合





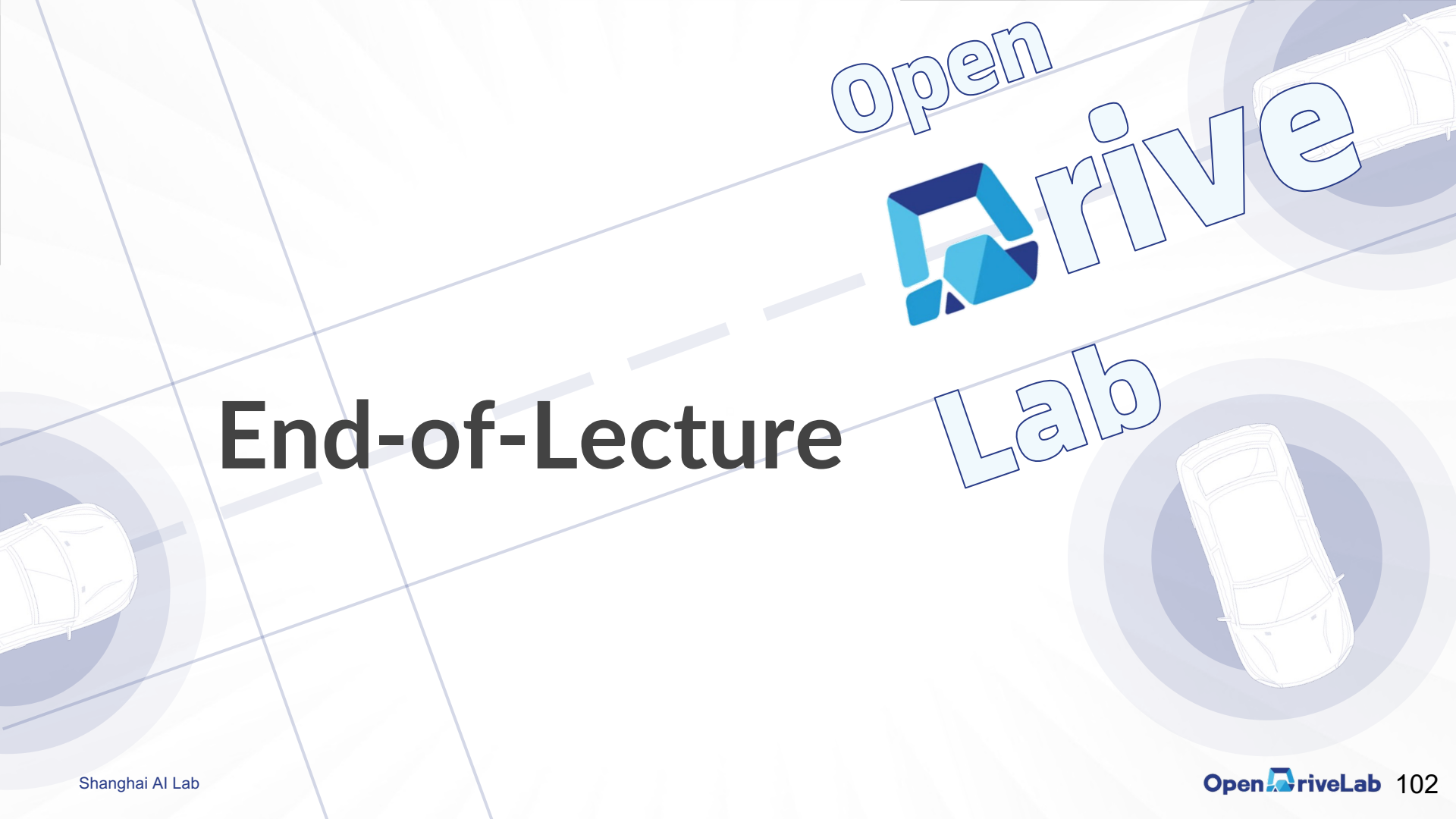
Left: BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, *arXiv:2205.13542*.
Right: BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework, *arXiv:2205.13790*.

BEV感知工作：未来思考

3

- 对部署更为友好的BEV模型
  - 通过对模型的精简让BEVFormer在实车上能够低功耗的实时运行
  - 实现Layer Normalization等一些OP
  - MDC/ TI /NV Xavier/ J5/ 自研芯片等

END



End-of-Lecture