# DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries

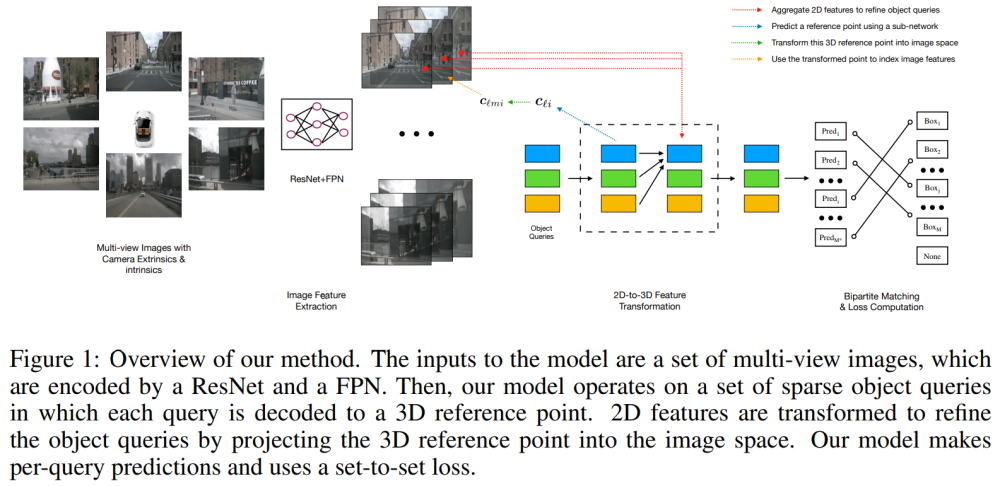
## 论文地址

<https://arxiv.org/abs/2110.06922>

## 代码地址

<https://github.com/wangyueft/detr3d>

## 模型架构



### Overview

模型架构输入一组投影矩阵已知的相机的RGB图像，并为场景中的目标输出一组3D Box相关参数。

* 将3D信息合并到架构中的中间计算，不在图像平面中执行纯2D计算

* 不估计密集的3D场景几何，避免相关的重构误差

* 避免诸如NMS的后处理步骤

### Model design

论文配套的代码已经开源，参考：[DETR3D](https://github.com/WangYueFt/detr3d) , inference流程有4个Step

1. 从nuscenes数据集中加载某个timestamp的6张RGB图像，组织成NCHW的 (6, 3, 900, 1600)

1. 6作为batch维度，进行Resnet+FPN抽取视觉语义，得到Image space上的featuremap(6 camers, 4 scales)

1. 预测BEV视角下的3D bbox (lidar坐标系)，操作类似DETR中的decoder，依赖相机-Lidar转换参数-GT是从Lidar中获取得到

* 1. 结合Lidar2cam的转换矩阵将可学习的3d reference point转到image space上，并采样出step2中对应pixel位置的features

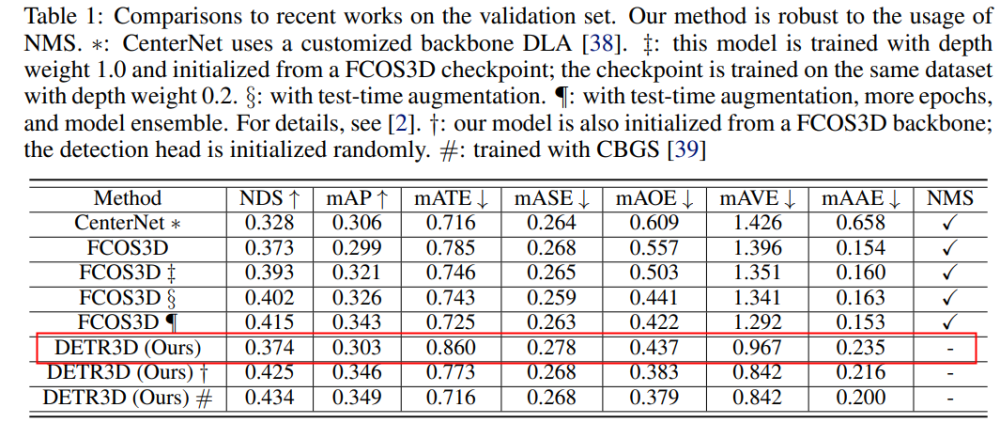
* 1. 另一组可学习的object queries先做self-attention， 然后结合a中的features做cross-attention

* 1. **备注:从论文中的描述看，step b中，是将features和对应的query相加，然后再经过Transformer的多头注意力机制(具体需对比代码中的实现)**

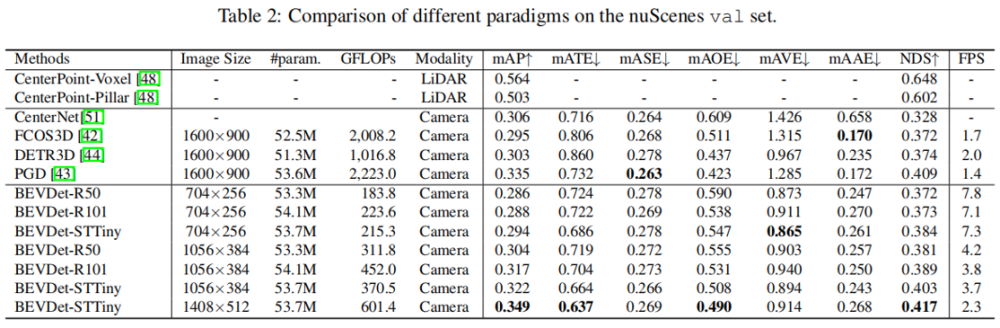
1. transformer的输出经过Linear后得到cls\_branch和bbox\_branch，作为3d bbox 的pred的结果

### 实验结果

不同的初始化方法对性能的提升比较明显，BEVDet对比的主要是下图红框中的map值，从另外一个角度来说，基于DETR3D这种方式的3D目标检测，应该是比较难优化好backbone的参数，



BEVDet释放出来的相关信息



### 总结:

基于DETR3D的这种方式，不需要对图像特征进行密集预测，直接使用Object Query来进行BEV空间下的目标搜索，再通过相机参数融合2D和3D特征，无需构建BEV语义地图信息，同样的，该方法如果需要在BEV视角下实现其他检测任务则比较困难，综合这两年的发展来看，似乎是从**2D->3D->BEV空间变换->多任务检测头**这种方式比较流行

