# BMBEV：+3D Object Detection

## LSSDet

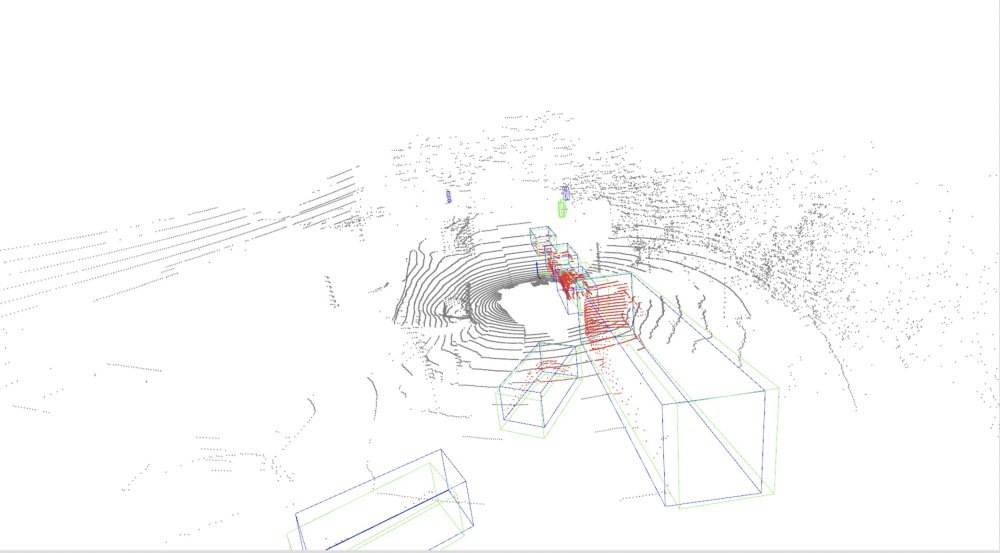
### Introduction

LSSDet是基于LSS模型的架构扩展到3D目标检测任务上的模型，LSS通过对图像进行特征提取之后，生成pseudo cloud point，并将伪点云通过Pointpillars中对点云的处理方式进行BEV视角特征变换，然后再在BEV视角特征下进行后续的任务，如3D目标检测，语义地图构建等BEV视角常见任务

* 模型输入：来自6个camera的image，以及camera to lidar的转换矩阵

* 模型输出：object在lidar坐标系下的3dbbox，project到image(上图)和Lidar点云(下图)中如下所示





### Model design

模型架构层面主要包括Image Encoder、View Transformer、BEV Encoder、3D Detection Head等几个部分，多模态模式下，模型架构增加Voxel Encoder、Points Middle Encoder等点云处理流程，输入为nuscenes数据集的环视摄像头采集的images，各功能模块主要功能如下:

#### 视觉感知:

* Image Encoder: 对nuscenes某个timestamp时刻的6张Image进行多尺度的特征提取，输入为[b, 6, 64, feature H, feature W]，同时对Feature Map上的每个像素点预测深度分布d，外积后最终得到特征编码后的伪点云**[b, n, 64, d, feature H, feature W]**，n为当前使用的图片数，d为离散深度分布

* View Transformer: 将前述步骤生成的伪点云变换为BEV视角下的特征，输出shape=**[b, 64, 512, 512]**，512为BEV Feature Map的尺度，64为类似的Voxel编码特征，此时就已经生成了BEV视角下的Feature Map

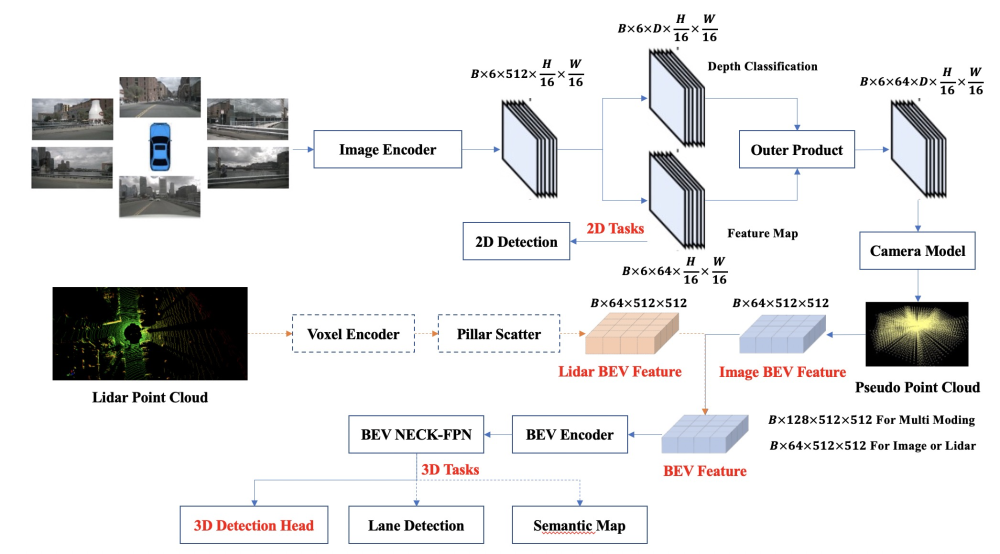
* BEV Encoder: 对生成的BEV视角特征进行多尺度的特征提取，进一步细化特征

* 3D Detection Head: 下游任务检测头，这里针对3D目标检测任务，因此接入3D Detection Head，可以根据不同的task接入不同的Head

#### 点云融合:

* Voxel Encoder: 将lidar点云划分成各Voxel块，并进行特征提取，参照PointPillars中对点云的处理方式

* Points Middle Encoder: 将经过特征编码后的点云特征，对应到BEV视角下，并进行特征融合，最终输出BEV Feature Map的shape=[b, 64, 512, 512]，再进行多模态预测下，与Image生成的BEV特征进行Concat后再送入BEV Encoder

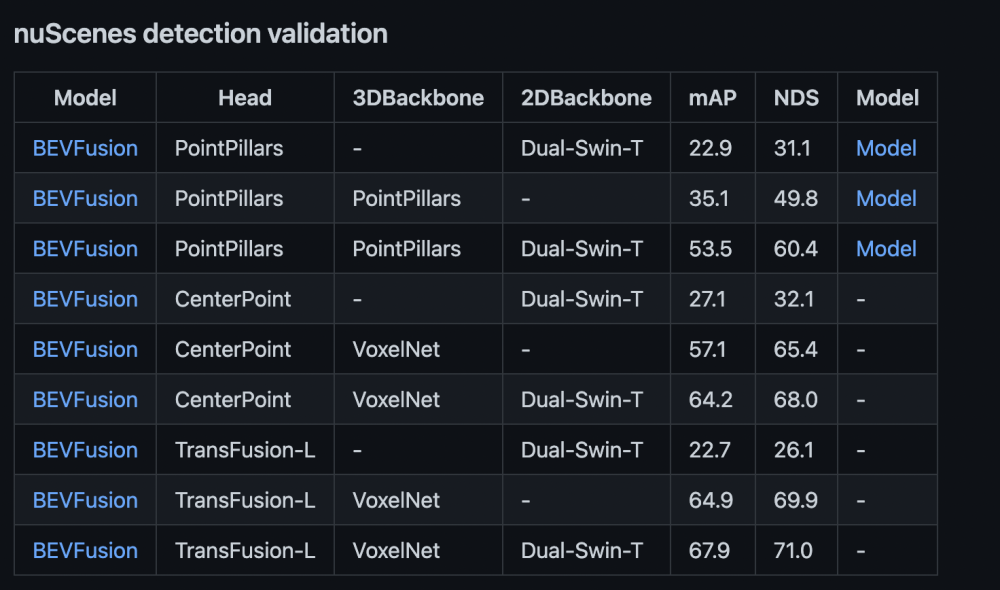


[📎LSSDet框架.pptx](https://yuque.antfin.com/attachments/lark/0/2022/pptx/21256453/1654503140969-71521932-4e5c-499a-8adc-a535d6a72c1e.pptx)

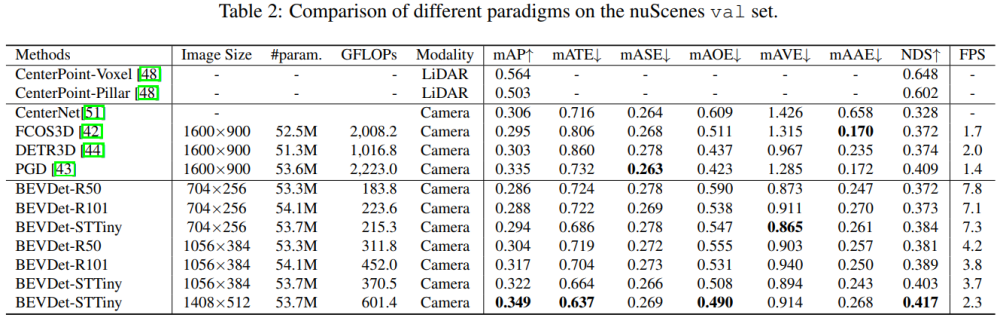
### Experiment

#### LSSDet-NuScenes

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Input Size** | **Image Encoder** | **View Transformer** | **Voxel Encoder** | **Points**  **Middle Encoder** | **BEV Encoder** | **BEV Neck** | **Lr-Config** | **MAP** | **NDS** | **MMDet3D-Config** | **MMDet3D-Log** |
| LSSDet | 256,704 | Efficientnet-b3 | Lift-Splat-64 | / | / | SECOND | SECFPN | cyclic（40） | 27.194 | 33.241 | [📎lssdet\_efficientnet-b3\_second\_secfpn\_circlenms\_256\_704\_cyclic\_40e\_nus.py](https://yuque.antfin.com/attachments/lark/0/2022/py/21256453/1654503140948-09772eda-24ac-4a67-91ad-3d94f81dbd1d.py) | [📎20220417\_172159.log.json](https://yuque.antfin.com/attachments/lark/0/2022/json/21256453/1654503141003-42c4a920-6092-49e8-9e7f-ede82fb9d144.json) |
| 256,704 | Efficientnet-b3 | Lift-Splat-64 | / | / | SECOND | SECFPN | cosine（20） | 27.281 | 33.572 | [📎lssdet\_efficientnet-b3\_second\_secfpn\_circlenms\_256\_704\_cosine\_20e\_nus.py](https://yuque.antfin.com/attachments/lark/0/2022/py/21256453/1654503140946-2ade58b1-54cf-4592-9966-46c4ab5fd607.py) | [📎20220418\_190517.log.json](https://yuque.antfin.com/attachments/lark/0/2022/json/21256453/1654503140950-506058ef-3807-45d5-9bbb-1ab956dd3c28.json) |
| 256,704 | Efficientnet-b5 | Lift-Splat-64 | / | / | SECOND | SECFPN | cyclic（40） | 27.821 | 33.730 | [📎lssdet\_efficientnet-b5\_second\_secfpn\_circlenms\_256\_704\_cyclic\_40e\_nus.py](https://yuque.antfin.com/attachments/lark/0/2022/py/21256453/1654503141169-55ea756f-1ff3-4c2c-a131-398a97fc1f87.py) | [📎20220421\_110610.log.json](https://yuque.antfin.com/attachments/lark/0/2022/json/21256453/1654503141245-fd526287-5bd8-49ab-8cac-c4f6be4f8758.json) |
| LSSDet-Mixup | 256,704 | Efficientnet-b3 | Lift-Splat-64 | PillarFeatureNet | PointPillarsScatter | SECOND | SECFPN | cyclic（25） | 46.216 | 55.023 | [📎lssdet\_efficientnet-b3\_pillarnet\_second\_secfpn\_circlenms\_256\_704\_cyclic\_40e\_nus.py](https://yuque.antfin.com/attachments/lark/0/2022/py/21256453/1654503141368-309498a7-b5f0-42ad-b86b-abab488d7343.py) | [📎20220424\_163644.log.json](https://yuque.antfin.com/attachments/lark/0/2022/json/21256453/1654503141227-e1dee745-138d-4399-813e-25e467b43020.json) |



#### 对比SOTA



#### 

#### 实验结论:

* 更换不同的学习率方式，对性能基本没有什么影响，即与SOTA存在差异的主要原因在于模型结构不同，重点放在模型结构调整上

* 更换性能更强的Backbone性能提升不明显，与BEVDet论文中的结果类似

### 可视化结果

#### 纯视觉感知:

[LSSDet训练可视化结果.pptx](https://yuque.antfin.com/attachments/lark/0/2022/pptx/21256453/1654503141297-fdbf1aeb-7a8e-4233-83cd-d3b5376d9853.pptx)

#### 纯视觉存在问题:

* 远处目标不太容易检测出来，分析原因在于图像分辨率较低，原图像1600\*900->704\*256，图像的缩放尺寸比较大，图中小目标不容易检测，解决方法是可以增大图像分辨率，但训练时间会变长

* 对于人等这些目标训练的效果不太好，一是由于多个目标在图像中混合在一起，人的形态各异，详细在三维空间中定位比较困难，二是这些样本相对比较少，解决方法是增加数据和借助Lidar点云来进行空间区分

* 对于某些场景下，模型对于高度的预测存在一些问题，一是这种场景的数据量比较小，二是在将特征变换到BEV视角下之后，高度细化特征丢失，解决方法是在BEV视角生成时，增加高度层面的区分和增加数据

* 需要实验进一步细化验证

#### 纯视觉感知改进优化方法:

#### Lidar+Camera融合感知效果

[LSSDet-Mixup可视化.pptx](https://yuque.antfin.com/attachments/lark/0/2022/pptx/21256453/1654503141369-1354c5ea-369b-47ea-9f4a-e0e39bf2c95c.pptx)