1. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection(2018)

   这是三维目标检测的经典论文，提出了体素特征编码，避免了过去需要人工进行特征工程的弊端，相比逐点表征，大大减小了开销。

   - 将空间中的点分组，沿三个维度划分成体素（小格子），经过VFE(Voxel Feature encoding)层，可以得到每个体素的特征表示。VFE先逐点提取特征，再使用最大池化聚合逐点特征，得到局部聚合特征。将局部聚合特征与逐点特征连接起来后，输入FCN得到体素的特征表示。
   - 将体素的特征表示连接起来，*C*D*H*W，输入中间卷积层，再输入RPN。锚框的采样频率为，沿H和W，每隔两个体素取样两个，最终rpn预测出每个点（特征图大小为H/2*W/2）的正类和负类的得分，和每个点的两个锚框的7个特征（与锚框的偏移，三个中心点坐标，三个尺寸，一个航向角）

2. Frustum PointNets for 3D Object Detection from RGB-D Data(2018)

   也是一篇3维目标检测的经典论文。本文围绕3D点云，结合2D目标检测，达到了很好的效果。本文利用了RGB值和深度特征，与我们的项目刚好符合，之后可以继续调研一下这篇文章的后续发展。

   - 先将RGB图输入2D目标检测网络，得到区域框和类别；再利用深度数据和相机的参数，将框内的点映射为一个视锥体内的点云；利用3D分割网络(pointnet)，对视锥体点云内的点进行分割，获得目标类别的点；将分割出的点输入3d框估计模块，其中，T-Net预测目标中心距离点云质心的残差，另一个框估计网络，输入为T-net目标中心坐标系下的点云，预测真实中心与T-net目标中心的残差，与NS个预设框尺寸的3的维度的残差值，NS个尺寸的得分，NH个与预设的航向角的残差和得分，共3+4*NS*+*2*NH个输出。最终可以得到3d框的中心坐标，尺寸和航向角。

3. SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation(2020)

   第一篇单阶段单目3D目标检测的工作，省略了预测2D框的步骤，看知乎上工业界至今仍常用。

   - 主干网络用的DLA，感觉不少用这个。之后分两个分支，关键点分支预测目标分类和关键点坐标(给每个点打分，具体方式见Centernet)，关键点为3D中心投影到2D后的点，回归分支预测3D信息。回归分支中，对每个关键点，回归预测一组值，三个关键点坐标的偏差值，三个尺寸的偏差值（是个指数比例的），观测角的sin和cos（进而计算yaw轴角）。
   - 关键点分支使用的focal loss，回归分支将预测的3d信息，先使用激活函数约束转换一下，再转为3d框的8个角点的坐标，计算与GT角点坐标的L1损失，算是使用统一的损失函数对它们进行回归。
   - 训练时，这两个分支是并行进行的，而预测时应该是先后进行的。
   - 训练时将梯度解耦了，这对预测3D信息是有利的：对于坐标的偏差值，使用GT投影中心点的坐标xy，结合预测的坐标偏差，回归GT坐标值。对于角度，除了角度之外用的都是GT值（主要是坐标），对于尺度也是如此。

4. Centernet:Objects as Points(2019)

   本文通过预测关键点的方法进行检测，省去了NMS后处理，并且避免了对大量冗余锚框进行训练和预测，该方法也在之后继承发展，如SMOKE

   - 以3D边界框的中心（实际上是投影2D框的中心）为该目标的关键点。对于输入的图片H*W，输出H/4*W/4*C的热图。最终对每个点预测H/4*W/4*(C+3+1+8)个值，不同模态的预测使用独立的分割头（4个），其中，C维为C个类别的得分，分析得到整张热图的peaks，作为待选关键点，3维为3d框的尺寸，1维为深度，8维为对角度的预测（这里是用了一种较冗余的方法 Multi-Bin based

method，将2 Pi的角度范围平分成两个bins，对于每个bins，预测两个bins的得分，得到偏差角（与当前bin中心角度的差值）的sin和cos）。
    - 在这里搞懂了怎么计算热图的损失，SMOKE与这个一样。热图的真值不只是关键点的GT为1，而是以关键点为中心的高斯分布，对于每个非关键点，选取所有分布中的最大值作为真值。

5. RTM3D: Real-time Monocular 3D Detection from Object Keypoints for Autonomous Driving(2020)

本文提出了一种，直接在2D图上提取3D bbox关键点（2D边界框中心）和顶点（焦点的投影和中心的投影），再通过几何约束，规范化3D重投影的单目3D目标检测方法。

- 主干网络（特征点检测网络）和Centernet相似，也用的DLA。先预测出来一个C维的热图，找到关键点和类别。同时，预测出9维的顶点热图，预测每个点是顶点的概率；预测出18维的热图，为每个顶点的2维offset，用来回归顶点坐标。
- 预测出9个顶点后，构建了一个能量函数，包括重投影误差（找到投影到2D图像面时，与预测点最近的3d框），深度误差，角度误差。这是一个非线性优化。

6. AVP-SLAM: Semantic Visual Mapping and Localization for Autonomous Vehicles in the Parking Lot(2020) 使用SLAM进行泊车的一篇文章，大致流程看懂了，其中精度的关键是对BEV图的拼接和语义分割。但我对SLAM的视觉里程计和局部定位等环境一无所知，还需要继续学习。