PV-RCNN阅读笔记

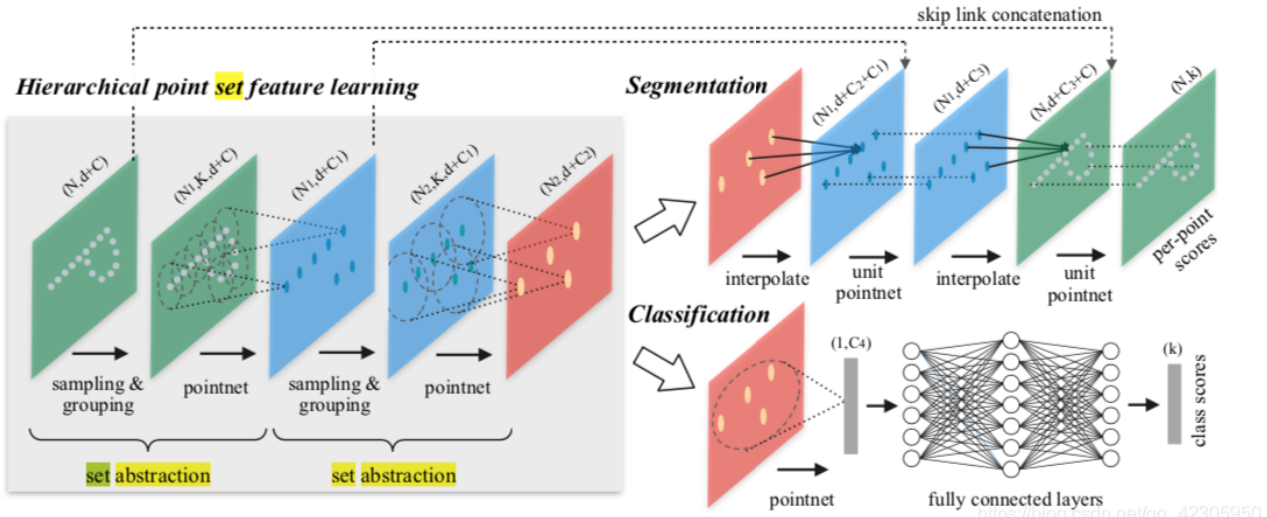
目前在kitti数据集榜单第一名，收录在CVPR2020。与pointRCNN同作者。

1 简介

文章称该方法把point-based和voxel-based两种方法的优势结合起来，提高了3D目标检测的表现。基于体素的操作可以高效的编码多尺度特征表示并生成高质量3D提案框，基于点操作有可变的感受野故可以保留更精确的位置信息。

voxel-based（grid）优缺点：高效、但信息损失降低定位细粒度的精度（finegrained localization accuracy）
point-based优缺点：计算成本高、但可以得到更大的感受野（by the point set abstraction）

> set abstraction操作：（1）取样用最远点采样FPS（2）grouping构建局部特征，不用KNN而用邻域球（3）用pointnet提取局部特征



2个创新操作：

- voxel-to-keypoint场景编码
- point-to-grid ROI特征提取

2 相关工作

这个不多说了。

3 检测框架

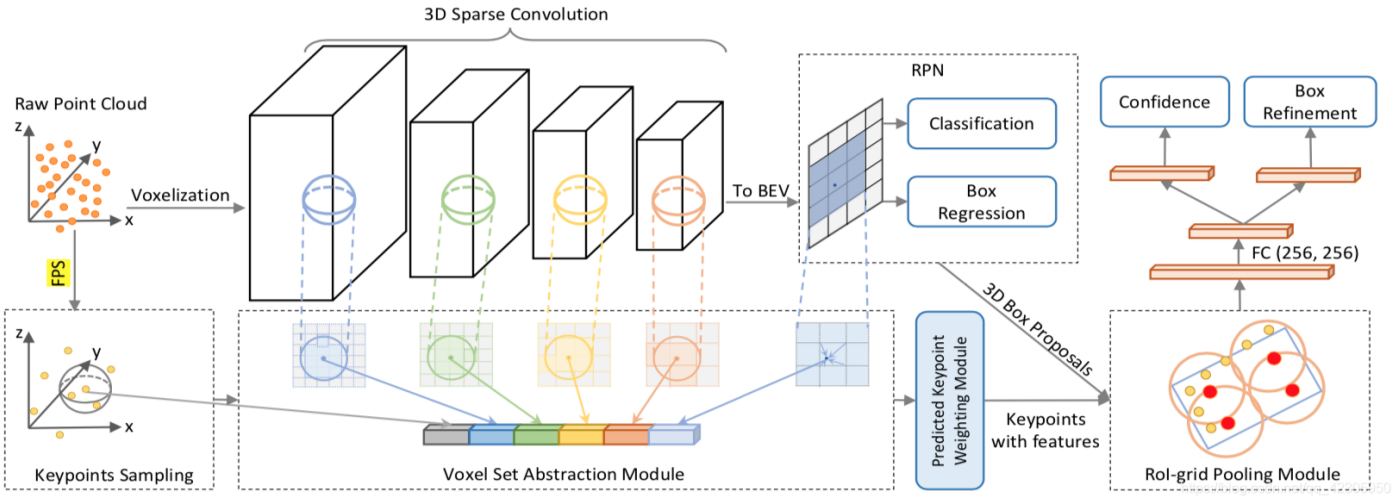
3.1 3D voxel CNN高效特征编码和提案生成

3D voxel CNN把场景划分为L*W*H的体素，非空体素的特征为点特征的均值（x,y,z,r）

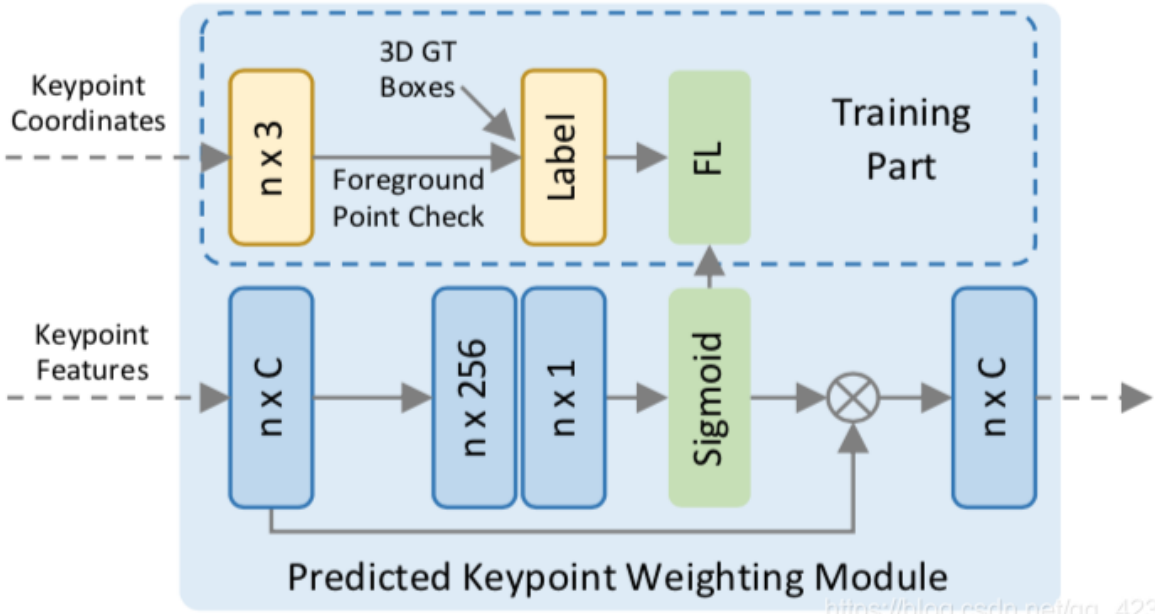
3D提案生成把8倍降采样后的3D特征体（volume，其实就是特征向量的集合）转换成2D鸟瞰特征图，用基于anchor的方法生成提案框。每个类有2 * L/8 * W/8个提案框（标签的平均尺寸，0度和90度）。此方法有更高的召回率。

dicussion目前2stage的框架需要pooling ROI来优化提案，但8倍下采样使空间分辨率很低，如果上采样得到更大尺寸的特征体/图，便会很稀疏。在传统RoI pooling或ROI align的时候通常会用双线性插值，这就使得在3D稀疏的表达上可能得到几乎都是0的特征表示。
pointnet系列提出的set abstraction操作可在可变邻域上编码点特征，由此提出了整合3D voxel CNN和一系列set abstrction操作。先提出关键点，然后再利用关键点编码voxel卷积过程的多尺度特征。

3.2 Voxel-to-keypoint Scene Encoding via Voxel Set Abstraction



关键点抽样 用最远点采样FPS获取2048个关键点（KITTI数据集）。此方法可使关键点均匀分布在非空体素，可代表整个场景。
Voxel Set Abstraction Module 体素特征提取VSA模块，把卷积得到的多尺度voxel特征结合到关键点，而pointnet用的是邻域点的特征。
扩展VSA模块 如上图，增加了8倍降采样后的2D鸟瞰特征图。把关键点pi投影到鸟瞰图，双线性插值得到BEV特征。由此，VSA模块提取的特征由原始点云特征、体素特征、鸟瞰图特征三部分组成。
关键点权重预测



全场景的信息由少数关键点表达，它们在后续的stage用来框优化。但有些FPS得到的关键点可能只表达的背景区域，所以要对前景点和背景点分配权重。由此，我们提出了Predicted Keypoint Weighting（PKW）如上图所示。
是否是前景点可由该点是否在groundtruth框内得到，在用focal loss进行训练。

3.3 Keypoint-to-grid RoI Feature Abstraction for Proposal Refinement

对3D voxel CNN产生的每一个3D提案(RoI)，用每个RoI的特征由多尺度关键点特征聚合而成进行框优化。由此提出了keypoint-to-grid RoI feature abstraction模块，如图

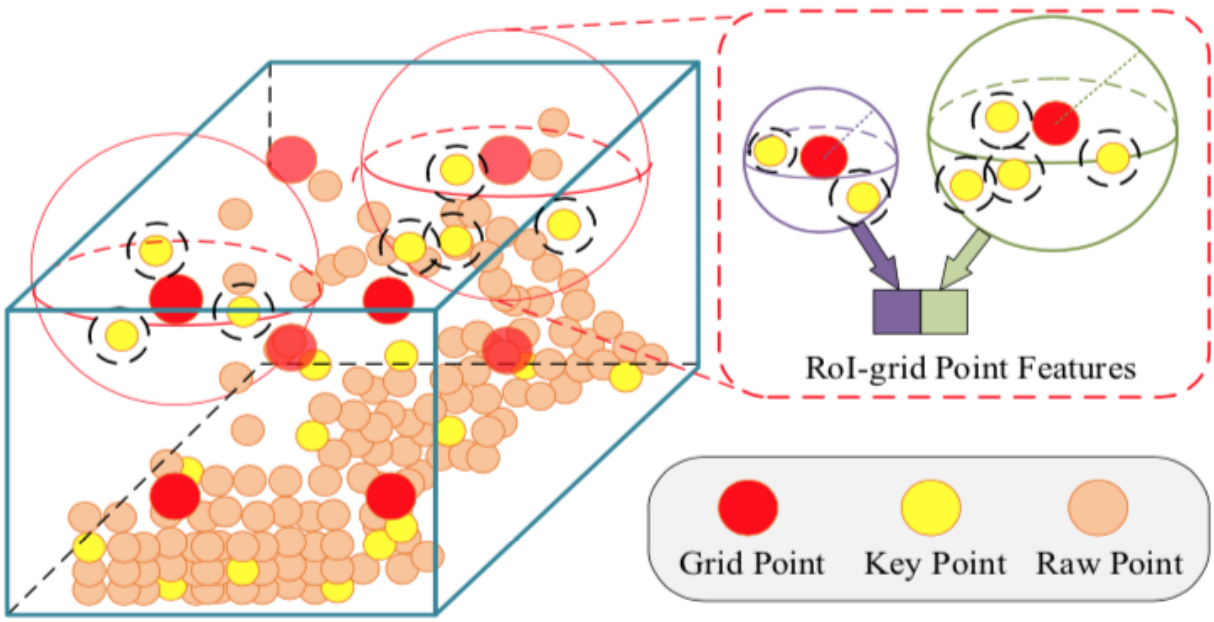


Figure 4. Illustration of RoI-grid pooling module. Rich context information of each 3D RoI is aggregated by the set abstraction operation with multiple receptive fields.

RoI-grid Pooling via Set Abstraction 对每个3D ROI，提出了RoI网格pooling，将关键点和不同尺度感受野的网格点（grid point）特征聚合起来，每个框采样6*6*6个网格点。再用set abstraction操作，在网格点的邻域球找关键点，再用pointnet模块将二者整合。再通过2层MLP将特征转换为256维，作为提案的特征。相比于之前提出的3D RoI pooling，此方法（RoI-grid pooling）可以在可变感受野上捕获更多的上下文信息（contextual information），甚至可以大过提案框，捕获周围的关键点信息。

3D框优化和置信度预测 优化网络使用2层MLP，分为置信度预测和框优化两个分支（看总图）。
置信度预测分支使用前人的3D Intersection-over-Union (IoU)，在提案框和gt框中训练。第k个提案框，置信度yk归一化到（0，1）如下计算，

$$y_k = \min\left(1,\ \max\left(0,\ 2\mathrm{IoU}_k - 0.5\right)\right),$$

loss也很常规，取交叉熵

$$L_{\text{iou}} = -y_k \log(\tilde{y}_k) - (1 - y_k) \log(1 - \tilde{y}_k),$$

框优化分支也用了前人的residual-based method（回头看，尤其是该作者的part a2），loss用smooth-L1。

3.4 训练losses

端到端训练 region proposal loss **Lrpn**、keypoint segmentation loss **Lseg**、proposal refinement loss **Lrcnn**，分别由以下公式得到

$$L_{\text{rpn}} = L_{\text{cls}} + \beta \sum_{\mathrm{r} \in \{x,y,z,l,h,w,\theta\}} \mathcal{L}_{\text{smooth-L1}}(\widehat{\Delta \mathrm{r}^a}, \Delta \mathrm{r}^a),$$

（其中**Lcls**由focal loss计算得到。）
Lseg也是由focal loss得到。

$$L_{\text{rcnn}} = L_{\text{iou}} + \sum_{\mathrm{r} \in \{x,y,z,l,h,w,\theta\}} \mathcal{L}_{\text{smooth-L1}}(\widehat{\Delta \mathrm{r}^p}, \Delta \mathrm{r}^p),$$

最终的loss由这三个loss相加得到。

4 小结

实验细节和结果文中很详细，等之后用到了再看。

刚入了这个坑不久。本以为2D检测基本走到头，3D检测发展空间会比较大，但！这个发paper的速度也太快了吧！KITTI榜单也是！本以为看的都是一些比较新的文章，但去KITTI榜单一看，排在前面的都是些陌生面孔。这谁顶得住！

希望自己能搞点东西出来。