【论文阅读】【三维目标检测】PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection

原创 麒麒哈尔 2020-02-10 17:44:57 4746 收藏 15 版权
分类专栏： 论文阅读

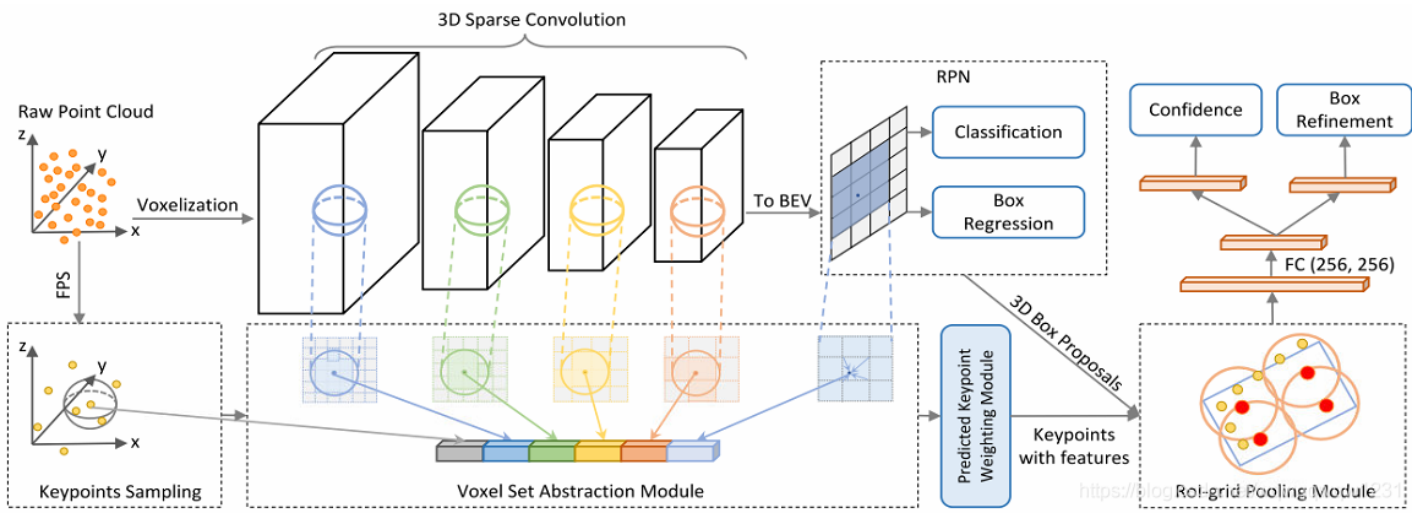
文章目录

本论文目前是KITTI排名第一，香港中文大学和商汤出品，该作者还提出了PointRCNN和Part-A^2Net。

PV-RCNN

本文将Grid-based（我一般常称为Voxel-based）的方法和Point-based的方法优缺点结合了起来。本文首先说明了Grid-based和Point-based的方法的优缺点：
“Generally, the grid-based methods are more computationally efficient but the inevitable information loss degrades the fine- grained localization accuracy, while the point-based methods have higher computation cost but could easily achieve larger receptive field by the point set abstraction.”

网络的结构图如下：



RPN

Backbone: 3D Sparse Convolution

在本文中没有介绍太多，但在作者之前的一篇文章“Part-A^2 Net: 3D Part-Aware and Aggregation Neural Network for Object”中介绍的比较详细，由于是backbone，其实也比较通用。那作者为什么要用3D Sparse Convolution呢，作者在文中提到：*“Because of its high efficiency and accuracy”*

Classification & Regression Head

将3D的feature map转为俯视图，高度变为通道，然后使用每个cell每个类别设置两个anchor，角度分别为0和90度。

实验表明使用这种backbone和anchor的设置方式，Recall高：*“As shown in Table 4, the adopted 3D voxel CNN backbone with anchor-based scheme achieves higher recall perfor- mance than the PointNet-based approaches [25, 37]”*

但这里有个问题是anchor的角度是0或者90度，那-90度是怎么处理的？这相当于是怎么处理相反方向的车？车辆朝向的这个量这个在Proposal生成的过程中是否考虑？如果考虑，则怎么回归相反方向的车，这种anchor设置看起来不合理；如果不考虑，那么在通过Proposal生成6x6x6的grids的时候的顺序怎么确定，难道就一直不考虑？这个得通过具体Loss或者代码中看了。
这一点，Part-A^2Net也没有细讲，但引文可以追溯到SECOND: Sparsely Embedded Convolutional Detection。这样子其实也可以理解，就相当于在图像处理中，网络要学会对左右翻转的鲁邦性。

Voxel Set Abastraction Module（VSA）

Discussion

有了Proposal，就要提取Proposal中的feature，形成一个固定大小的feature map了，本文将Proposal分成了6x6x6的栅格。那么如何计算6x6x6的每个cell的feature呢？

然后作者提出了对目前方法不足的地方的讨论：
“(i) These feature volumes are generally of low spatial resolution as they are downsampled by up to 8 times, which hinders accurate localization of objects in the input scene.
(ii) Even if one can upsample to obtain feature volumes/maps of larger spatial sizes, they are generally still quite sparse.”
也就是说使用差值的方法，类似于图像中的目标检测那样的RoI Align的方法不太好。

作者就提出了一种思路，使用PointNet++中的SA层，对每个cell，使用SA层，综合这个cell一定范围内的BackBone输出的feature map中的feature。但作者提出，这种方法，计算量太高。
“A naive solution of using the set abstraction operation for pooling the scene feature voxels would be directly aggregating the multi-scale feature volume in a scene to the RoI grids. However, this intuitive strategy simply occupies much memory and is inefficient to be used in practice. For instance, a common scene from the KITTI dataset might result in 18, 000 voxels in the 4×downsampled feature volumes. If one uses 100 box proposal for each scene and each
box proposal has 3 × 3 × 3 grids. The 2, 700 × 18, 000 pairwise distances and feature aggregations cannot be efficiently computed, even after distance thresholding.”
为了解决这个问题，作者提出了VSA Module，来减少要聚集的feature的总数量，也就是上例子中的18000。

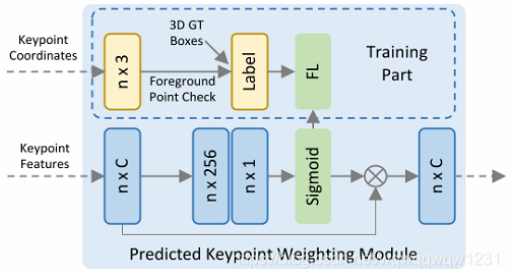
VSA Module

VSA Module在示例图中已经画的非常形象了。过程如下：（公式1,2,3）
1）在原始点钟用Furthest Point Sampling选n个点
2）在每一层中的feature map中，使用SA Module综合每个点一定邻域内的feature
3）然后把所有feature concat起来
Extended VSA Module还多两种feature：

- 在Backbone输出的feature map转到的BEV图中，用2D bilinear interpolation计算的feature
- 使用原始点云通过SA Module计算的feature

PKW Module（Predicted Keypoint Weighting）

问题是n个点中，有些点事前景点，比较重要，有些点是背景点，不重要。这就要区分一下，通过这n个点的feature，可以计算n个weight，weight由真实的mask做监督训练，然后用这weight乘以点的feature，得到每个点的最终的feature。（公式5）这个过程被称为PKW module。

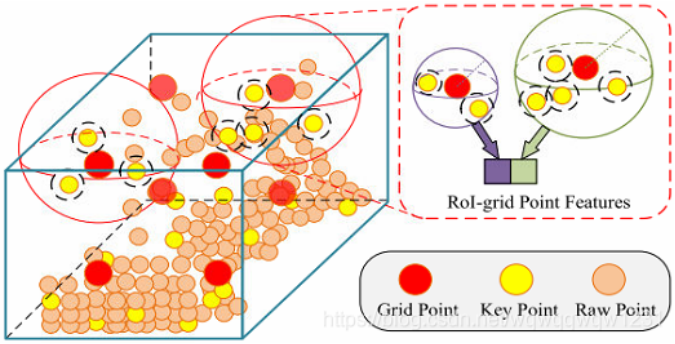


上述过程是使用n个点来表示整个场景，文章中叫做voxel-to-keypoint scene encoding，n个点叫做key-points

到此，我们有了Proposal和n个点的坐标和对应的feature。

RCNN

有了Proposal，就可以生成6x6x6个cell，对于每个cell的中心点，可以在之前得到的n个点中选取那些在其邻域的点，然后使用SA Module综合特征，得到cell的特征。（公式6,7）



得到了Proposal的固定大小的特征，就可以做confidence prediction和box refinement了。这里要注意的是confidence prediction的真实值是由IOU给出的。

Experiments

在KITTI上和Waymo Open Dataset上效果都很好。

Ablation Studies：

- 验证了voxel-to-keypoint scene encoding的有效性，与RPN和朴素的想法做了对比。
- 验证了different features for VSA module。

- 验证了PKW module的有效性。
- 验证了RoI-grid pooling module比RoI-aware pooling module（PointRCNN中的方法）的有效性。

思考

本论文的作者之前的论文还有PointRCNN，Part-A^2Net。本文主要引用STD，这里就做一下对比，看看网络的发展脉络。

PointRCNN完全使用PointNet++做特征提取的module，包括RPN中的backbone和RCNN中的特征提取部分。

STD相比于PointCNN，加入了RoI-grid的部分。由于RCNN中使用voxel表示的，RCNN中的特征提取也变成了3D Convolution。

Part-A^2Net，相比于STD，一开始就是用Voxel的表示方法，将RPN中的主干网络也换成3D Convolution。（当然还有提出了Part location的表示等等）抛开细节的特征表示不谈，我认为其实Part-A^2Net就是本文中朴素的想法。

PV-RCNN解决了本文提出的Part-A^2Net计算效率低的问题。

本文阅读起来非常舒服，就像在讲一个故事，**写论文的方法也可以很好的借鉴本文。**