Weakly Supervised 3D Object Detection from Point Clouds

Zengyi Qin
Massachusetts Institute of Technology
qinzy@mit.edu

Jinglu Wang
Microsoft Research
jinglu.wang@microsoft.com

Yan Lu
Microsoft Research
yanlu@microsoft.com

论文地址： https://arxiv.org/abs/2007.13970
代码地址： https://github.com/Zengyi-Qin/Weakly-Supervised-3D-Object-Detection
论文信息： ACM MM 2020；麻省理工学院；微软研究院；

动机： 使用弱监督学习规避 3D 检测任务中，标签信息较难获得且不准确的问题。整体方案分为两步：第一步根据归一化点云密度无监督产生多个候选框，第二步，利用跨模态的知识蒸馏策略，将在图像上学习到的检测知识用在 3D 检测问题上，生成最终的候选框。

如何无监督提候选框：

论文提出根据点云分布密度和是否存在物体是相关的思路，先初步找到目标物体位置。为了避免 anchor 中点云密度受深度影响，提出了一种归一化密度的方式。

作者将点云投影至前视图，得到 pixel-wise map，每个 pixel 对应三维坐标有三个特征值。在 map 上以间隔 0.2m 截取大小为 32*32 的 patch，每个 patch 是一个四棱台范围内的数据投影得来的，所以每个 patch 在深度上可以对应到多个 3D box，同时随着深度增加，3D box 的体积越大，密度=落在 box 内的点数量/32*32，这样就规避了随着深度加深，密度降低的问题。这部分因为没用 GPU 并行化计算，所以耗时作者是没有算进整体网络的耗时的。

根据每个 3D box 算出来的归一化密度，和预先设置的阈值 0.5 做比较，低于阈值的认为不包含物体，进一步，将保留下的 3D box 向外扩张 0.2m，避免候选框中只有一部分的物体。

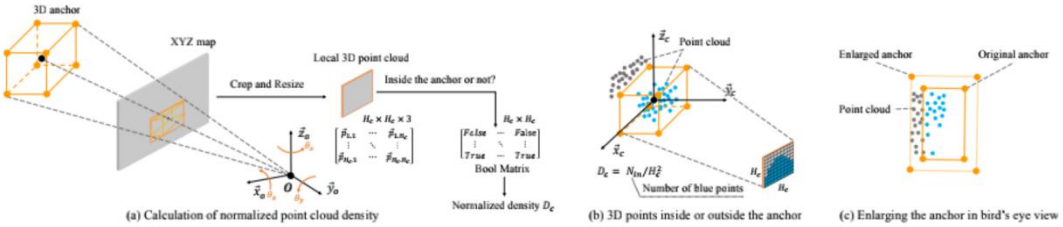

Figure 2: Normalized point cloud density. The point cloud density inside a volume is influenced by two factors that are 1) whether the volume contains an object and 2) the distance of the volume to the sensor. The density increases when an object is present but decreases as the distance grows. Our normalization strategy eliminates the influence of distance. (a) The preset 3D anchor is projected to the XYZ map, where its projection is cropped out and scaled to a square patch with fixed size $H_c \times H_c$ that is distance-irrelevant. The square patch represents H_c^2 3D points, where one pixel corresponds to one point. (b) Among the 3D points, N_{in} points are inside the 3D anchor. The normalized point cloud density D_c is calculated as N_{in}/H_c^2. (c) The grey and blue points are on the same object. If the original anchor fails to bound the object and only contains a part of it, the enlarged version would contain more points, i.e., the grey ones.

图像和点云之间的迁移学习：

作者提了一种跨模态的知识蒸馏策略，点云做 3D 检测的网络作为 student，在大型图像数据集上训练得到的模型作为 teacher。

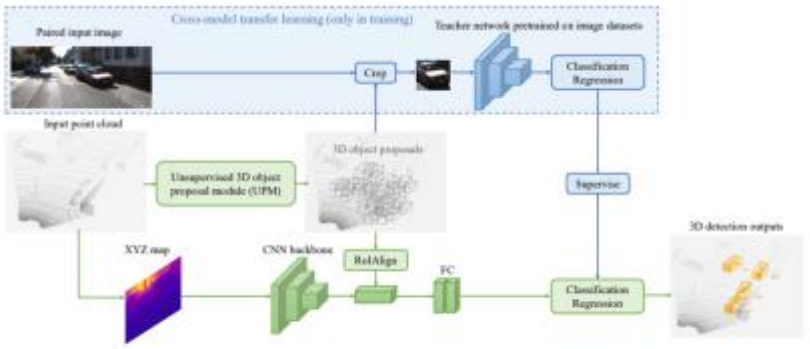

Figure 1: Overview of the proposed weakly supervised 3D object detection framework. The first key component is the unsupervised 3D object proposal module (UPM) that selects 3D anchors based on the normalized point cloud density. The second component is the cross-modal transfer learning module that transfers the knowledge, including object classification and rotation regression, from image datasets into the point cloud based 3D object detector.

Teacher(蓝色部分)：模型解决两个任务，第一个是输入一个物体，输出该物体种类；第二个是朝向预测，将 360°分为 16 个 bin，将朝向问题转换为 multi-bin classification 的问题。

Student(绿色部分)：由 backbone, RoIAlign 和 FC 组成, 输入 backbone 的是 pixel-wise map，将由无监督学习产生的候选框对应到 pixel-wise map 和图像数据（作者选用 kitti 数据中提供了和点云数据对应的图像数据）中，裁剪出候选框覆盖的区域，将裁剪出的图像数据送入 teacher 网络中做分类预测，将 student backbone 提取的候选框特征利用 RoIAlign 得到固定维度，后送入 FC 做分类预测，通过交叉熵损失函数保证 student 给出的预测结果和 teacher 给出的结果是一致的。

跨模态的知识蒸馏会有以下两点问题

1 如下图 a，teacher 给出某些样本的预测结果在 sl=0.4 和 sh=0.6 区间中，表示他对这些样本分类不是很有把握，这种时候，student 是学习不到什么知识的。

2 对于 teacher 预测的结果，往往会设置一个阈值，将结果二分类，这样就会导致 student 学习没有区分度，比方说阈值设置为 0.6 时，teacher 对正样本预测 0.9，0.7，转为 student 监督标签时都是 1，如下图 b。

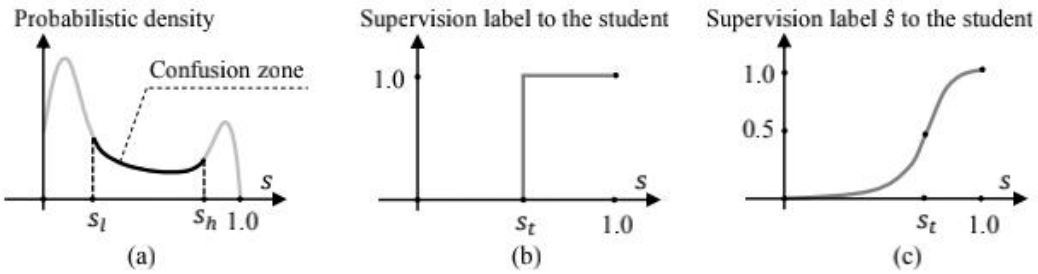

Figure 3: Supervision rectification. Given an object proposal to be classified, if the confidence of the teacher network is within the confusion zone, the loss value will be masked out.

根据以上两点，做以下修改：

1 现利用公式 $\hat{s} = \frac{1 + e^{(s_t - 1)k}}{1 + e^{(s_t - s)k}}$ 将 teacher 输出结果曲线平滑为上图 c

2 将 student 和 teacher 之间的损失函数改为，即不计算落在 sl 和 sh 区间中的样本。其中^是平滑化后的 teacher 输出，~是 student 输出。

$$\mathcal{L}_r = -\left[\hat{s}\log(\tilde{s}) + (1-\hat{s})\log(1-\tilde{s})\right] \cdot \mathbb{1}(s \notin [s_l, s_h]) \quad (2)$$

实验结果：

实验对比挑选了三个 SOTA 的弱监督学习方法：PCL[Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Loddon Yuille. 2018. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence* (2018).], OICR[Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *CVPR*.]和 MELM[Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. 2018. MinEntropy Latent Model for Weakly Supervised Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1297–1306.],因为该论文提出的方法是不能预测 3D 检测框的，所以比较是在 2 维检测结果上进行的，结果显示在 recall、AP 都是有很大的提升。

个人总结：

提供了一种跨模态知识蒸馏的思路，但是限定需要是成对的数据，还是有些不容易的。论文中关于如何用点云密度无监督筛选候选框还是有借鉴意义的，可以用在一些后处理中。