# Fast point Rcnn

## Abstract

提出了一种统一、高效的基于点云的三维目标检测框架。我们的两阶段方法利用体素表示和原始点云数据来利用各自的优势。第一阶段网络以体素表示为输入，仅由轻量卷积运算组成，产生少量高质量的初始预测。在初始预测中，每个点的坐标和索引卷积特征与注意机制有效融合，既保留了精确的定位信息，又保留了上下文信息。第二阶段利用融合后的内点特征进一步细化预测。我们的方法在KITTI数据集上进行了三维和鸟瞰（BEV）检测评估，并以15FPS的检测率达到了目前的水平。

## Introduction

本文提出了一种统一、快速、有效的两阶段三维目标检测框架，该框架利用了体素表示和原始点云数据。我们网络的第一个阶段，称为VoxelRpn，直接利用点云的体素表示。计算经济的卷积层被用于高效率和高质量的检测

在第二阶段，我们应用一个轻量级的pointnet来进一步细化预测。有了少量的初始预测，第二阶段的速度也非常快。设计了具有注意机制的模块，将第一阶段的卷积特征与每个内点的坐标进行有效融合。它使每个点都知道其上下文信息。

我们的方法的一个特点是它受益于点云的体积表示和原始密集坐标表示。三维体积表示为处理点云提供了一种可靠的方法。

第二阶段的轻量pointnet再次检测点的坐标，以获得更多的定位信息，同时扩大了接收野，得到了较好的结果。由于该方法利用了点云上每个区域的卷积特征，具有较高的效率，因此将其命名为fast pointR-CNN。

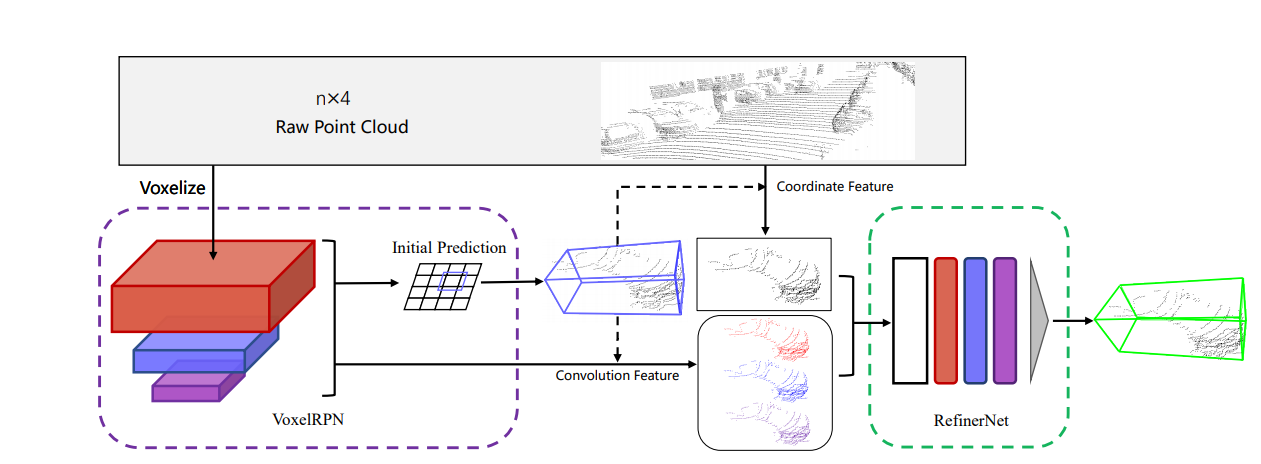
通过这种概念上简单的结构，我们实现了高效率，同时也获得了不错的三维检测精度，达到了最先进的效果。它甚至比以前以RGB和点云作为输入的方法更有效。本文的主要贡献有三个方面。

提出了一种快速实用的基于点云（不含RGB图像）的两级3D目标检测框架，该框架利用了点云的体积表示和原始密集输入。

•我们的系统包括二维和三维卷积以保存信息。我们融合卷积特征和点坐标进行盒细化。

•我们的系统以每秒15帧的速度运行，在BEV和3D检测方面达到了最先进的性能，特别是在高质量目标检测方面。

## 3.Our Method

在本文中，我们提出了一个简单快速的两阶段框架，用于点云数据的三维目标检测，如图1所示。第一阶段以体素表示为输入，产生一组初始预测。为了补偿体素化和连续卷积过程中精确定位信息的丢失，第二阶段将原始点云与第一阶段的上下文特征相结合，产生细化结果

3.1 motivation

激光雷达捕获的点云是一组结构不规则、分布稀疏的点。利用强大的CNN对点云数据进行训练和推理并非易事。将点离散为体素化输入[43，20]或将其投影到具有紧凑形状（如RGB图像[40，23]）的BEV形成一组解决方案，其中可以产生抽象和丰富的特征表示。然而，离散化过程不可避免地会引入量化伪影，其分辨率会降低到体素图中的单元数。此外，连续卷积和下采样操作也可能削弱原本存在于点云中的精确定位信号。

像PointNet[26]这样的方法是专门为直接处理点云数据而设计的。将这些方法直接应用于自主驾驶场景中规模较大的点云，可以得到更多的位置信息。但是它们需要大量的GPU内存和计算量，几乎不可能达到很高的检测速度。

为此，我们的方法是新的利用混合体素和原始点云，而不依赖于RGB图像。两个有效的阶段是将体素表示输入到体素预测神经网络中，以获得一组高速的初始预测值，再通过细化网络融合原始点云和提取的上下文特征，以获得更好的定位质量。下面将详细介绍这两个部分。

### 3.2 VoxelRPN

voxelRPN接受三维体素输入并产生三维检测结果。它是一个单级物体探测器。

输入表示voxelrpn的输入是体素化点云，实际上是一个规则网格。网格中的每个体素都包含了位于局部区域的原始点的信息。具体来说，我们将三维空间划分为空间排列的体素。假设点云的感兴趣区域是一个大小为（L，W，H）的长方体，并且每个体素的大小为（vl，vw，vh），则三维空间可以划分为大小为（L/vl，W/vw，V/vh）的三维体素网格。

体素中可能有多个点。在voxelnet[43]中，保留35个点并将其输入VFE层以提取特征。然而，我们的发现是，仅仅在每个体素中使用6个点跟随8通道MLP层就已经足够在经验上实现合理的性能。通过这种紧凑的表示形式，我们可以很容易地利用CNN的强大功能进行信息特征提取。

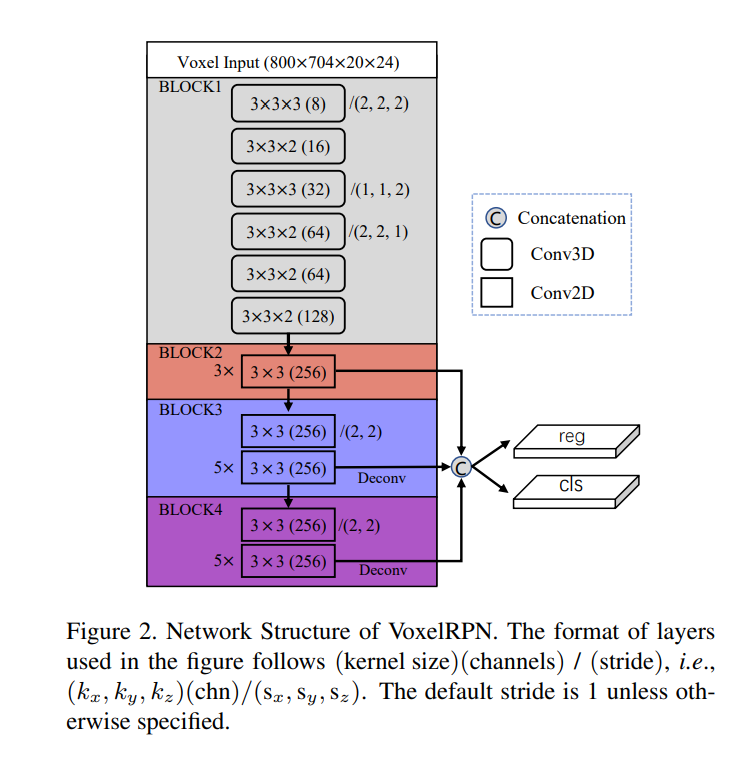
Network structure

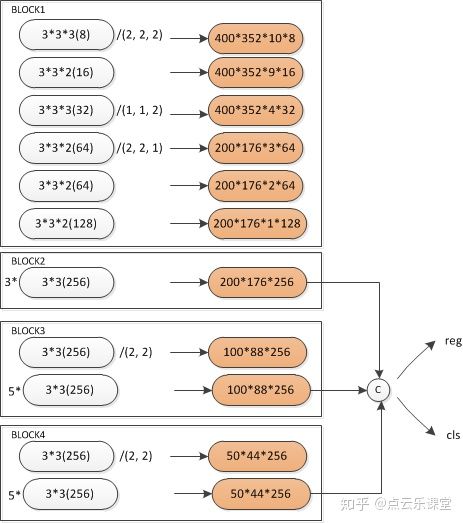
针对三维检测，我们的网络需要从（X，Y，Z）维度清晰地过滤信息。在[40，23]中，在生成体素表示时，Z维度被简单地转换为通道。

然后应用几个二维卷积。这样，沿Z维的信息很快消失。因此，只能在BEV上进行检测成为可能。不同的是，体素网[43]在生成体素后再进行三次三维卷积时保持三个独立的维度。注意到效率降低了。

沿着一个更合适的方向，我们发现许多连续的三维卷积对于保持三维结构是非常有效的。基于这一观察，我们的主干网由2D和3D卷积组成，实现了与PIXOR一样的高效率[40]，甚至比体素网更高的性能[43]。

我们在图2中显示了骨干网的详细信息。





第一部分由六个三维卷积层组成，其中只有少量的滤波器来保持时间预算。

为了更好地融合和保存信息，我们没有使用步长为2、核大小为3的滤波器对Z维特征进行大幅度的降采样，而是在Z维中插入核大小为2的三维卷积层。下面是三个二维卷积块，用于进一步提取和扩大感受野。

三维场景中同一类别的物体通常具有相似的尺度。因此，与2D图像中流行的多尺度对象检测器[21]不同，它根据不同的尺度将对象建议分配给不同的层，我们注意到hypernet[14]结构更合适。

具体来说，我们通过反卷积从区块2、3和4的最后一层提取特征图，如图2所示。然后将它们连接起来，在较低的层次上收集丰富的位置信息，在较高的层次上收集较强的语义信息。预定义的锚点[22]在这个融合特征图上与特定的比例和角度一起使用。然后分类头和回归头分别运行在该特征图上，对每个锚点进行分类，并对已有对象的位置进行回归。

3.2 RefinerNet

虽然VoxelRPN取得了良好的性能，但是由于体素化过程和第一块连续的跨步卷积仍然会丢失大量的定位信息，因此我们可以通过直接处理原始点云来进一步提高预测质量，但是我们的算法可以通过进一步的特征增强来补充这些信息精炼网。

RefinerNet利用点云的坐标。

F-PointNet[25]是利用PointNet从2D检测结果中回归3D amodal边界框的先驱性工作。只使用内部点进行推理，而不知道上下文信息。相反，我们的方法也受益于重要的上下文信息。

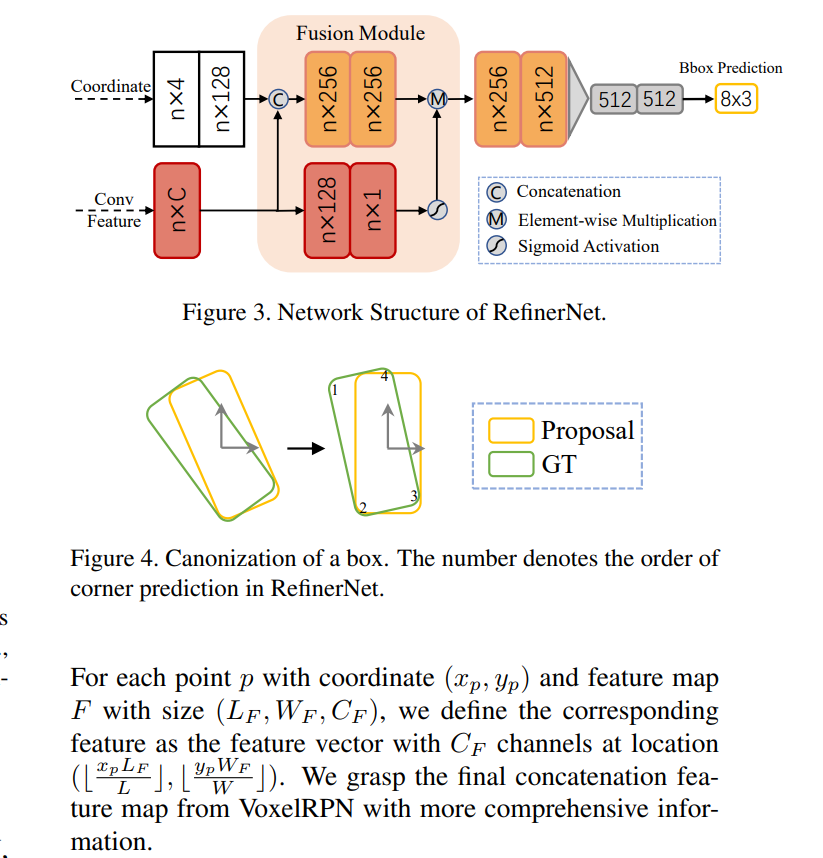
**Box feature**

利用voxelrpn的每个边界盒中的点来生成盒特征。与文献[25]中使用的两个独立网络不同，我们不仅将坐标作为输入，还将从voxelrpn中提取的特征作为输入。基于voxelrpn的卷积特征映射能够捕捉到目标的局部几何结构，并以层次化的方式逐步聚集，从而获得更大的利润预测接受域。然后利用pointnet将每个点映射到高维空间，通过最大池运算融合点的表示，以获取各点之间的信息及其上下文。

对于每个从voxelrPN预测的边界框，我们首先将其投影到BEV。然后使用BEV框区域周围的所有点作为输入，如图1所示。

对于具有坐标（xp，yp）的每个点p和具有大小（LF，WF，CF）的特征映射F，我们将相应的特征定义为具有位置（bxplf L c，bypwf W c）的CF通道的特征向量。我们从体素网络中获得了更为全面的拼接特征图。

在将每个点的坐标输入到后续网络之前，我们首先对它们进行规范化，以保证平移和旋转不变性。建议框周围0.3米范围内的点坐标通过旋转和平移进行裁剪和规范化。如图3所示，我们将坐标特征定义为通过MLP层获得的高维（128D）表示。



利用这两种网络结构源的特征，我们找到了一种有效融合它们的方法。我们设计了一个具有注意机制的综合特征生成模块，代替了简单的连接。如图3所示，我们首先将高维坐标特征与卷积特征连接起来。然后它与注意力相乘，由卷积特征产生。下面是一个由两个MLP层组成的轻量级点网，最大池将所有信息聚合到一个框中。

最后通过两个MLP层来实现盒子的细化，并在此基础上预测所有盒子角点的细化位置。如图4所示，在计算回归目标时，地面真值框和点云通过给定建议框的旋转和平移进行规范化。该操作将地面真值箱的角点按特定的顺序组织起来，可以减少旋转引起的角点顺序的不确定性。我们的实验证明了经典角损失的优越性。

这种轻量级的细化网络在不需要鸣钟和哨声的情况下，已经能够有效地提高盒预测的精度，特别是在考虑Z维和边界盒的情况下，无论是在3D还是BEV中都具有较高的预测精度。

训练我们的快速点R-CNN包括两个步骤。我们首先训练体素pn直到收敛。然后根据提取的特征和推断出的边界盒对细化网络进行训练。

VoxelRPN在VoxelRPN中，锚定分布在全局特征图的每个位置上。如果一个锚在BEV中的IoU值大于0.6，则认为它是一个正样本。回归目标是具有最高IoU值的地面真实边界框。如果一个锚点与所有地面真值框的IoU值低于0.45，则认为该锚点为负。我们将具有多任务损失的体素PN训练为loss=Lcls+Lreg，（1）其中Lcls是分类二进制交叉熵损失，Lcls=1 Npos X i Lcls（p pos i，1）+γNneg X i Lcls（p neg i，0），（2）Lcls（p，t）=（tlog（p）+（1−t）log（1−p））。（3） 在我们的实验中，我们使用γ=10。由于正负样本的分布不平衡，我们分别对它们的损失进行了归一化处理。OHEM[30]适用于分类损失的负项。每个锚点参数化为（xa，ya，za，ha，wa，la，θa），地面真值框参数化为（xg，yg，zg，hg，wg，lg，θg）。对于回归，我们采用以下参数化[43，9]，即∆1x=xg−xa da，∆1y=yg−ya da，∆1z=zg−za ha，∆1h=log（hg ha），∆1w=log（wg wa），∆1l=log（lg la），∆1θ=θg−θa。

（4） 回归损失定义为Lreg（x）=0.5（σx）2的平滑L1损失，如果| x |<1/σ2 | x |−0.5/σ2，否则（5），在我们的实验中σ设置为3。

RefinerNet注意到，在鸟瞰图（BEV）的前30个预测框中，我们在0.5 IoU thresh上的体素PN召回率超过95%。我们的精炼网是为了提高预测箱的质量。我们只训练它在积极的建议箱，其借据与地面真相是高于0.5贝弗利。

回归目标定义为从提议中心（XP，YP，ZP）到目标框的8个规范化角（Xi，G，Yi，G，Zi，G，I＝1，…，8）的偏移，如图4所示：席席2=席，G -XP，2Yi＝Yi，G YP，Y2Zi= ZI，g−zp（6）此参数化是直接在点坐标上进行处理的RefinerNet的通用自然设计。

