论文阅读-《Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving》(2019CVPR, Cornell University)-No.2

Author of This Article Analysis：魏鑫烯；Time：2019-07-14

Open Source

Code

Motivation

1. 在三维目标检测领域，相比于基于激光雷达的检测方法，基于单目或者双目相机的算法的准确率非常低，目前 KITTI 上的结果：仅使用图像的算法最高精度 10%，基于激光雷达的最高精度 66%，基于图像+激光雷达的最高精度 73%，在以前人们认为这一问题的原因是基于图像得到的深度数据不够准确；
2. 在三维目标检测的应用中（如无人驾驶领域），目前主要依靠激光雷达数据，激光雷达较为昂贵，而且这些应用中应该使用额外的传感器数据来保持系统的鲁棒性，图像数据就是较为理想的候选传感器数据。

Ps：CVPR2019 目标检测相关论文共 43 篇

Contribution

1. 解释了基于双目图像和激光雷达数据的三维目标检测算法之间精度差异的主要原因：不是双目图像方法的精度问题，二是双目图像中深度数据的表示方法；
2. 提出了一个新的用于基于双目图像的三维目标检测算法的深度表示方法：类似于激光雷达中深度表示的方法，显著提高了算法精度；

Content

三维目标检测中的深度表示

在基于激光雷达的方法中：深度信息表示为三维点坐标中的一部分；在基于图像的方法中：深度数据被单独作为 RGB 之外的一个额外通道，如在论文《Multi-level fusion based 3d object detection from monocular images》（2018CVPR）中。

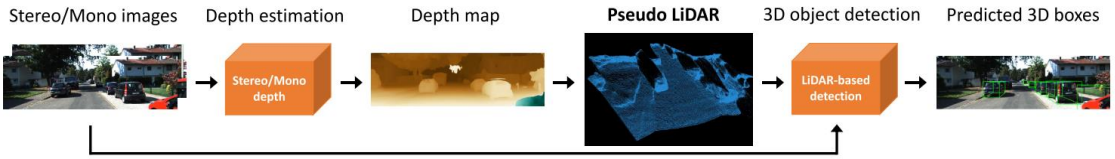
论文中的 pseudo-LiDAR 深度表示方法

即参考激光雷达方法中的表示方式，构造类似的点云数据，具体操作如下：

1. 双目相机得到的深度数据+图像上该像素点的 uv 坐标->基于图像坐标系的三维点云数据；

2. 由于激光雷达得到的点云是有扫描范围限制的，丢弃第一步得到的三维点云中超出一定范围的点（如垂直方向上高出激光雷达设备 1m 的点云）；
3. 激光雷达数据中还包含表示反射强度的值（intensities，0~1），在由图像获取的点云中全部表示为 1。

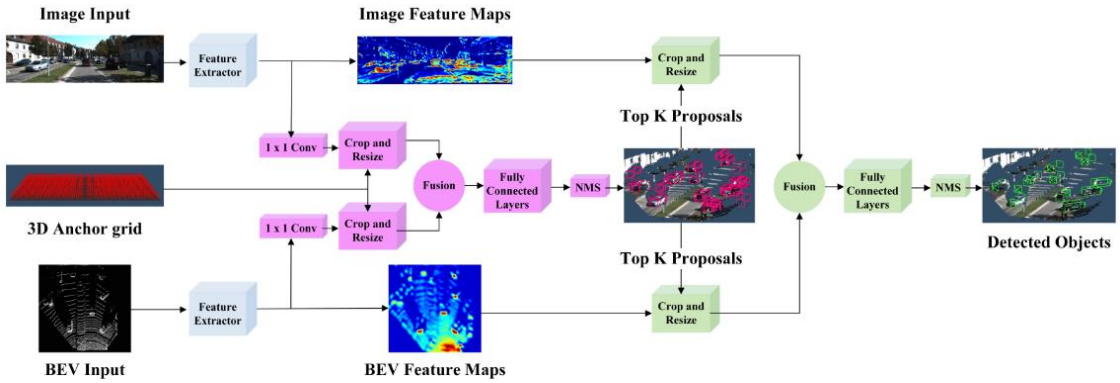
算法框架



主要步骤：

1. 获取图像深度信息，使用 PSMNet 网络（《Pyramid stereo matching network》, CVPR2018）；
2. 将深度信息+像素 uv 坐标转换成激光雷达方法中的三维点云数据（多视几何方法，重投影）；
3. 使用两种方法来进行三维目标检测：
   - 基于激光雷达的三维目标检测方法处理点云数据，使用 Frustum PointNet（《Frustum pointnets for 3d object detection from rgb-d data》, CVPR2018, Code)。
   - 在俯视图视角下将点云数据重新转换为 2D 图像数据，同时其剩下一维（深度）数据作为与图像数据对应的额外一维数据，使用 AVOD 算法处理（《Joint 3d proposal generation and object detection from view aggregation》, 2018IROS, Code）。

AVOD: 输入 RGB 图像以及激光点云数据，利用 FPN(Feature Pyramid Networks)网络得到二者全分辨率的 feature map，然后通过 crop&resize 提取两个 feature map 对应的 feature crop 并融合，最后挑选出 3D proposal 以实现 3D 物体检测。整个过程是 two-stage detection，可以理解为 MV3D 的加强版，网络结构：

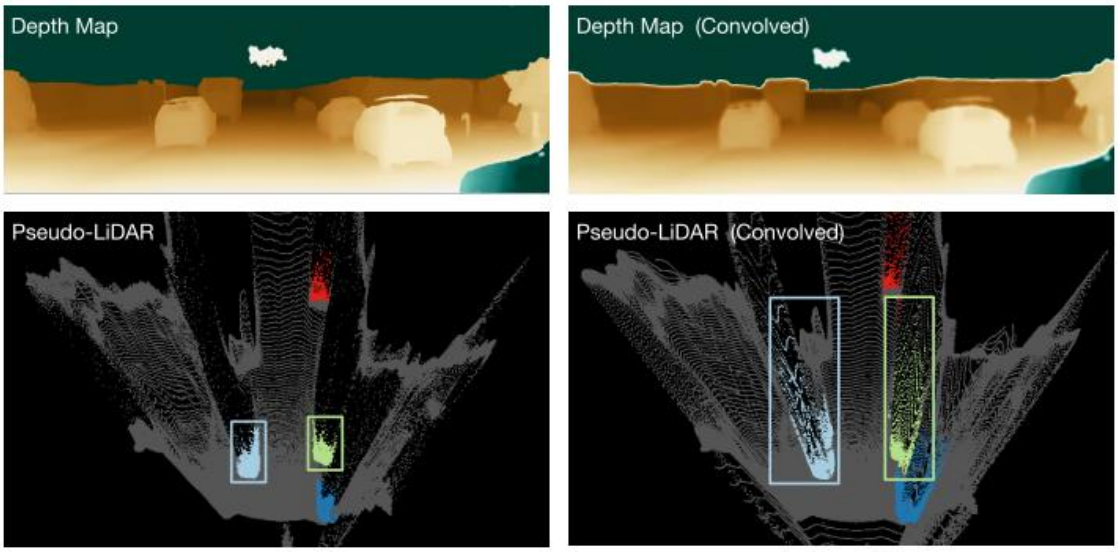


理论分析

对深度图的前视图做二维卷积，然后把结果重投影会三维空间，并查看其俯视图，可以看出相较于原图点的深度有了较大偏差，越远的点偏差越大，作者认为这是由于卷积神经网络在基于图像的学习过程中有两个前提假设：

1. 点的近邻点是有意义的；
2. 可以使用统一的方式来处理某个点的所有近邻点。

而这两个假设仅当图像中全部点在同一三维平面上才有意义，物体边缘的近邻点可能存在深度上的断层，这些信息不利于神经网络的学习。



Experiments

Detection algorithm	Input signal	IoU = 0.5			IoU = 0.7		
		Easy	Moderate	Hard	Easy	Moderate	Hard
MONO3D [4]	Mono	30.5 / 25.2	22.4 / 18.2	19.2 / 15.5	5.2 / 2.5	5.2 / 2.3	4.1 / 2.3
MLF-MONO [30]	Mono	55.0 / 47.9	36.7 / 29.5	31.3 / 26.4	22.0 / 10.5	13.6 / 5.7	11.6 / 5.4
AVOD	Mono	61.2 / 57.0	45.4 / 42.8	38.3 / 36.3	33.7 / 19.5	24.6 / 17.2	20.1 / 16.2
F-POINTNET	Mono	70.8 / 66.3	49.4 / 42.3	42.7 / 38.5	40.6 / 28.2	26.3 / 18.5	22.9 / 16.4
3DOP [5]	Stereo	55.0 / 46.0	41.3 / 34.6	34.6 / 30.1	12.6 / 6.6	9.5 / 5.1	7.6 / 4.1
MLF-STEREO [30]	Stereo	-	53.7 / 47.4	-	-	19.5 / 9.8	-
AVOD	Stereo	89.0 / 88.5	77.5 / 76.4	68.7 / 61.2	74.9 / 61.9	56.8 / 45.3	49.0 / 39.0
F-POINTNET	Stereo	89.8 / 89.5	77.6 / 75.5	68.2 / 66.3	72.8 / 59.4	51.8 / 39.8	44.0 / 33.5
AVOD [16]	LiDAR + Mono	90.5 / 90.5	89.4 / 89.2	88.5 / 88.2	89.4 / 82.8	86.5 / 73.5	79.3 / 67.1
F-POINTNET [23]	LiDAR + Mono	96.2 / 96.1	89.7 / 89.3	86.8 / 86.2	88.1 / 82.6	82.2 / 68.8	74.0 / 62.0

Conclusion

个人觉得这份工作对领域内的意义在于分析了影响基于图像的三维目标识别算法精度的主要原因：不是基于图像的深度分析有天然缺陷，二是之前算法中深度信息的表示方法存在问题。