文章编号：2095-2163(2021)09-0016-06 中图分类号：U461.91 文献标志码：A

联合 3D 实例分割和目标检测用于自动驾驶

赵 璐，宋新萍，姚振鑫
（上海工程技术大学 机械与汽车工程学院，上海 201600）

摘 要：为了提高自动驾驶中目标检测的精度与效率问题，本文中提出了一个简单而实用的检测框架，预测 3D BBox 和实例分割，在精度和效率之间实现了良好的平衡。首先，通过融合采样获得点云，骨干网络提取了局部特征和全局上下文信息；其次，设计了两个分支网络来预测语义标签和偏移量，将每个点移向其各自的实例中心，基于简单的聚类策略生成对象建议，对于每个集群仅生成一个建议，不再需要非最大抑制（NMS）过程；最后，应用提出的基于关键点的方法来优化每个提案的 3D BBox。通过将相同实例上的点投票为其目标关键点，将最小二乘拟合算法应用于预测的关键点。在公开的 KITTI 数据集上的实验结果表明，与其它基于特征嵌入的方法相比，本文所提出的方法可以显着改善实例分割结果，优于 KITTI 测试基准上的大多数 3D 对象检测器。

关键词：实例分割；3D BBox；自动驾驶；最小二乘拟合；KITTI

Joint 3D instance segmentation and target detection for autonomous driving

ZHAO Lu, SONG Xinping, YAO Zhenxin
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201600, China)

【Abstract】 In order to improve the accuracy and efficiency of target detection in autonomous driving, a simple and practical detection framework is proposed in this paper to jointly predict 3D BBox and instance segmentation, achieving a good balance between accuracy and efficiency. First, the point cloud is obtained by fused sampling and the backbone network is used to extract local features and global context information. Then, two branch networks are designed to predict the semantic label and offset to move each point to its respective instance center. Based on the results of the previous stage, object suggestions can be generated based on a simple clustering strategy. For each cluster, only one suggestion is generated and the non-maximum suppression (NMS) process is no longer needed here. Finally, our proposed method based on key points is applied to optimize the 3D BBox of each proposal. This is done by voting the points on the same instance as their target key points, and then applying the least squares fitting algorithm to the predicted key points. The experimental results on the public KITTI dataset show that compared with other methods based on feature embedding, the proposed method can significantly improve the instance segmentation results.At the same time, it is also better than most 3D object detection on the KITTI test benchmark.

【Key words】 instance segmentation; 3D BBox; automatic driving; least squares fitting; KITTI

0 引 言

3D 物体检测是 3D 感知中的一项具有挑战性的任务，是自动驾驶汽车（AV）的基本组成部分。与普通 2D 物体检测不同，AV 需要从现实世界中估计更多的 3D 边界框信息，以完成诸如路径规划、避免碰撞之类的高级任务。3D 物体检测的最新方法利用了不同类型的数据，包括传感器采集的数据，单眼图像，立体图像。在自动驾驶中，LiDAR 捕获的点云是更通用，信息量更大的数据格式[1]。LiDAR 点云的 3D 实例分割和对象检测示例如图 1 所示，图像分别显示了原始点云和 3D 检测结果，真实情况和预测结果分别以绿色和红色绘制，其中的红点是前景点的预测的对象中心，RGB 图像仅在此处用于可视化。

目标检测通常将对象表示为 2D 或 3D BBox，并带有多个参数，如 Bbox 的中心、尺寸和方向等，这种简单的表示形式适合于深度学习框架，同时也有一些局限性。例如，对象的形状信息已被完全丢弃，对于某个 BBox，不可避免要包含来自背景或其它对象的一些像素；在遮挡的情况下，这种情况变得更加严重；此外，BBox 表示不够准确，无法描述对象的确切位置。为了克服此问题，已为每个 BBox 使用了一个附加的实例 mask，以消除其它对象或背景的影响，通常实例 mask 是二进制的，以描述像素是否属于此对象。

作者简介：赵 璐（1992-），男，硕士研究生，主要研究方向：无人驾驶感知与定位；宋新萍（1967-），女，博士，副教授，主要研究方向：人工智能、图像处理；姚振鑫（1995-），男，硕士研究生，主要研究方向：图像处理、视觉 slam。

收稿日期：2021-02-16

图 1　来自 LiDAR 点云的 3D 实例分割和对象检测示例

Fig. 1　Example of 3D instance segmentation and object detection from LiDAR point cloud

本文设计了一个自下而上的联合 3D 实例分割与目标检测的端到端框架，其主要目标是更好地进行 3D 目标检测。

1　相关工作

1.1　来自多个传感器的 3D 物体检测

MV3D 是一项开创性的工作，融合了视觉和雷达点云信息，将点云投射到鸟瞰图和前视图中，这样既能减少计算量，又不至于丢失过多的信息[2]。受 MV3D 启发，AVOD 提出一种特征提取方式，借助 FPN 的想法，从点云输入和 RGB 图像得到全分辨率的特征图送入 RPN 网络，弥补了 mv3d 网络在小目标物体检测上的不足[3]。

1.2　基于 LIDAR 的 3D 检测

已被数个领先的 3D 检测器证明，LiDAR 传感器具有抗干扰能力强、对光照不敏感以及测距精度高等特点。学者们提出了两种类型的方法，即单阶段方法和两阶段方法。相比于单阶段的方法，两阶段的方法检测精度更高，但需要更多的网络推理时间。

单阶段方法：VoxelNet 代替了人工提取的特征，将点云划分为等距的 3D 体素，并通过类似 PointNet 的网络来应用 VFE 层，通过神经网络学习低级几何特征，从而表现出良好的性能[4]。HVNet 通过引进混合体素网格来提高检测精度[5]。

两阶段方法：与直接生成 3D 边界框的单阶段方法不同，两阶段方法旨在通过在第二阶段来优化第一阶段生成的 3D 提议，以生成更准确的目标检测框。最近，PointRCNN 利用具有 SA 和 FP 层的 PointNet++来提取每个点的特征，提出了区域提议网络（RPN）来生成提议，并应用了改进模块来预测边界框和类标签[6]。

1.3　3D 实例分割

当前的 3D 方法可以分为两种：第一种基于检测方法提取 3D 边界框，并在每个框内利用 mask 学习分支来预测对象 mask。SGPN 基于提取的特征为每个点建立了一个相似度矩阵，通过相似矩阵计算点在特征空间的距离来判断两个点是否属于同一对象[7]；第二种基于分段的方法预测语义标签，并利用点嵌入将点分组为对象实例，与 SGPN 不同，新提出的 GSPN 采用综合分析策略来生成实例细分的提议。

2　本文的方法

联合实例分割和 3D BBoxes 回归框架如图 2 所示，可以分为两部分：第一部分是基于聚类策略的对象建议，第二部分是基于最小二乘拟合算法的 BBoxes 改进。第一部分包括融合采样模块、语义分割与中心偏移模块以及基于聚类的提案生成模块；第二部分包括 3D 关键点选择模块和最小二乘拟合模块。

2.1　骨干网络

融合采样：通过融合采样[8]，删除了基于点的方法中必不可少的 Feature Propagate（FP）层，从而极大地减少了框架的运行时间[9]。

点云特征提取：对于采用融合采样策略后的输入点云，使用具有多尺度采样和分组操作的常用 PointNet++（无 FP 层以增加网络的运行速度）网络作为骨干网络。特别是，设计的框架具有独立的模块，特征提取模块可以使用任何点特征提取网络代替。

2.2　语义分割与中心偏移

利用语义分割模块来提取描述性特征，并预测每个点的语义标签，与语义分割平行，采用偏移分支来学习相对偏移，将每个点移至其各自的真实实例中心。通过这种方式，可以将同一对象实例的点移

向相同的中心,并将其聚集得更近,从而可以将点更好地分组为对象,并且可以将附近的同一类对象分离。

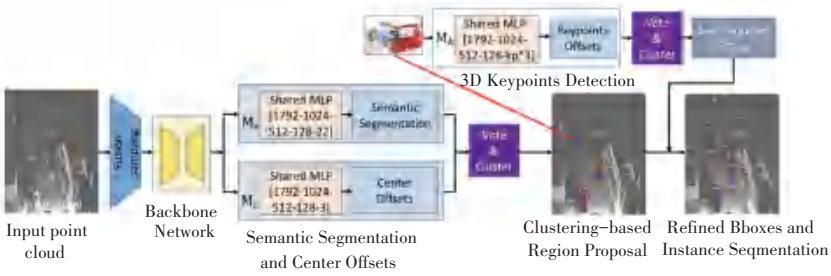

图 2 联合实例分割和 3D BBoxes 回归框架图

Fig. 2 Joint instance segmentation and 3D BBoxes regression frame diagram

2.2.1 语义分割

以逐点特征作为输入,设计了一个分割分支 M_s 模块用于语义类预测。由于采用了多尺度采样和分组策略,局部结构和全局上下文信息都已在每个点状特征向量中进行了编码,这有利于处理不同大小的对象。为了很好地解决分类中的类不平衡问题,使用焦点损失作为评测标准,式(1):

$$L_{cls} = -\sum_{i=1}^{C}(y_i\log(p_i)(1-p_i)^{\gamma}\alpha_i + (1-y_i)\log(1-p_i)(p_i)^{\gamma}(1-\alpha_i)) \quad (1)$$

其中,C 表示类别数,如果真相属于第 i 类,则 y_i 等于 1,否则等于 0;p_i 是第 i 类的预测概率;$\gamma \in (0,+\infty)$ 是聚焦参数;$\alpha_i \in [0,1]$ 是第 i 类的加权参数。

2.2.2 中心偏移

中心偏移模块 M_c 用于预测每个点到其对象中心的偏移量,只要所有点都拉到其所属物理中心的同一对象,则可以将其直接分为不同的实例。因此,偏移量分支对骨干网络提取的特征进行编码,以针对 N 个前景 (FG) 点生成 N 个偏移向量 $O = \{o_1, \cdots, o_N\} \in R^{N\times 3}$。对于属于同一实例的点,通过 L1-distance 回归损失将其学习偏移量约束,式(2):

$$L_{o\_reg} = \frac{1}{\sum_i m_i}\sum_i \| O_i - (\hat{C}_i - U_i) \| \cdot m_i \quad (2)$$

其中,$m = \{m_1, \cdots, m_N\}$ 是一个二进制掩码,如果点 i 在实例上,则 $m_i = 1$,否则 $m_i = 0$;3D 坐标 $U_i = (x_i, y_i, z_i)$,其中 $i \in \{1, \cdots, N\}$;$\hat{C}_i$ 是点 i 所属的实例的质心,式(3):

$$\hat{C}_i = \frac{1}{N^I_{g(i)}}\sum_{j\in I_{g(i)}} U_j \quad (3)$$

其中,$g(i)$ 将点 i 映射为其对应的真实实例的索引,即包含点 i 的实例,$N^I_{g(i)}$ 是实例 $I_{g(i)}$ 中的点数。

从点到其实例质心的距离通常具有较小的值(0~1 m),考虑到不同类别的不同对象大小,发现网络很难回归精确的偏移量,尤其是对于大型对象的边界点距离实例质心相对较远。为了解决这个问题,制定了方向损失,来约束预测偏移矢量的方向,式(4):

$$L_{o\_dir} = -\frac{1}{\sum_i m_i}\sum_i \frac{o_i}{\| o_i \|_2} \cdot \frac{\hat{C}_i - U_i}{\| \hat{C}_i - U_i \|_2} \quad (4)$$

2.3 基于聚类的提案生成

通过语义分割与中心偏移的操作,所有 FG 点都将汇总到其相应对象的中心。可以通过简单的聚类算法(即 k 均值[10])进行实例分割,k-means 是较经典的聚类算法之一,该算法的效率高,在对大规模数据进行聚类时被广泛应用。聚类后,通过平均前 k 个预测值,还为每个实例生成了平均值 BBox。

2.4 BBox 优化

尽管从第一阶段开始的 BBox 预测非常精确,但仍有一些改进空间。类似于其它基于两阶段的方法,将相同实例上的点投票为其目标关键点,将最小二乘拟合算法应用于预测的关键点,通过得到的 3D 旋转 $R \in SO(3)$ 和平移 $t \in R^3$ 来优化第一阶段得到的 BBox。

(1)3D 关键点检测。对于生成的每个提案,使用 3D 关键点检测模块 M_K 来检测每个对象的 3D 关键点,M_K 会预测从可见点到目标关键点的每点欧几里德平移偏移,这些可见点以及预测的偏移量将投票给目标关键点,通过聚类算法收集投票点,并选择聚类中心作为投票关键点。

(2)最小二乘拟合。给定对象的两个点集,一个来自 LIDAR 坐标系中 M 个检测到的关键点 $\{k\ V_j\}_{j=1}^{M}$,另一个来自对象坐标系中其对应点 $\{k\ V_j$

$\}_{j=1}^{M}$，BBoxs 优化模块计算参数 (R,t) 使用最小二乘拟合算法，该算法通过最小化平方损失式(5) 来找到 R 和 t：

$$L_{least-squares} = \sum_{j=1}^{M} \| k V_j - (R \cdot k V_j' + t) \|^2 \quad (5)$$

其中，M 是对象的选定关键点数。

2.5　多任务损失

多任务损失被用于训练的网络，包括语义分割损失 L_{cls}、中心偏移损失 L_{center}、关键点平移偏移损失 $L_{keypoints}$ 和 3D BBox 回归损失 L_{reg}，式(6)：

$$L = L_{cls} + L_{center} + L_{keypoints} + L_{reg} \quad (6)$$

中心偏移损失，在训练过程中，直接为每个 FG 点生成监督信号，并将损失函数表述为式(7)：

$$L_{center} = L_{o\_reg} + L_{o\_dir} + L_{size} + L_{\theta} \quad (7)$$

其中，L_{o_reg}、L_{o_dir}、L_{size} 和 L_{θ} 分别是偏移量约束损失、方向损失、BBox 尺寸和方向角的平滑度 L1-distance 损失。

BBox 回归损失：每个建议都编码为 7 维向量，包括对象中心 (c_x,c_y,c_z)，对象尺寸 (h,w,l) 和头方向角 θ。旋转后的 3D 联合损失在这里用作评测标准，式(8)：

$$L_{reg} = 1 - IOU(B_g,B_d) = \frac{B_g \cap B_d}{B_g \cup B_d} \quad (8)$$

其中，B_d 和 B_g 分别代表预测的 BBox 和真实的 BBox。

3　实验

3.1　数据集

本文利用公共 KITTI 数据集评估了 3D 实例分割和对象检测的框架。整个数据集分为训练集和测试集两个子集，分别由 7 481 和 7 518 帧组成。

3.2　3D 实例分割

在 KITTI 中，已为 3 类对象汽车、行人和骑自行车的人提供了 3D BBox 注释，只需提取每个 BBox 内部的点，即可为每个对象生成实例 mask。3D 实例地面真实情况的示例如图 3 所示，其中不同的颜色表示该图像底部的不同对象，绿色的 BBox 是真实结果，其它颜色 BBox 是预测结果，底部图像仅用于可视化。

图 3　基于 KITTI 3D BBox 注释生成的实例分割

Fig. 3　Based on the basic facts of instance segmentation generated by KITTI 3D BBox annotations

3.3　KITTI 上的 3D 对象检测

将本文的 3D 点云目标检测与其它最新方法进行了比较，结果见表 1。在 3D 和 BEV 检测任务中，本文的方法在所有竞争对手中均获得了出色的性能，在不同的召回率设置方面优于最新方法，具有更好的检测覆盖率和准确性，如图 4 所示。将从 LiDAR 检测到的 3D 边界框投影到 RGB 图像，以实现更好的可视化，如图 5 所示。

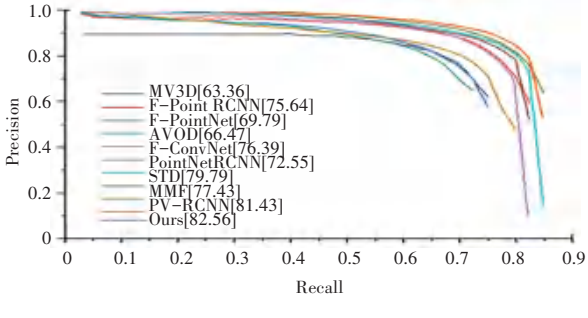

图 4　KITTI 3D 物体检测测试集上两阶段方法的评估结果

Fig. 4　Evaluation results of different two-stage methods on the KITTI 3D object detection test set

3.4　可视化检测到的关键点

在实例分割的基础上，可视化了检测到的 3D 关键点，如图 6 所示。绿色点是真实 3D 关键点，红色点是预测的 3D 关键点，实验结果表明本文的方法能够选择合适的关键点。

表 1 与 KITTI 测试服务器上其它方法的性能比较

Tab. 1 The performance comparison with other methods on the KITTI test server

Method	Modality	BEV(0.5)			3D(0.7)			FPS
		Easy	Moderate	Hard	Easy	Moderate	Hard	
One-stage								
VoxelNet	LiDAR	87.95	78.39	71.29	77.82	64.17	57.51	4.4
ConFuse	LiDAR+RGB	94.07	85.35	75.88	83.68	68.78	61.67	16.7
SECOND	LiDAR	89.39	83.77	78.59	83.34	72.55	65.82	20
PointPillars	LiDAR	90.07	86.56	82.81	82.58	74.31	68.99	42
SA-SSD	LiDAR	95.03	91.03	85.96	88.75	79.79	74.16	25
HVNet	LiDAR	92.83	88.82	83.38	87.21	77.58	71.79	31
Point-GNN	LiDAR	93.11	89.17	83.9	88.33	79.47	72.29	2
3DSSD	LiDAR	94.64	90.81	86.35	88.36	79.57	74.55	25
Two-stage								
Mv3D	LiDAR+RGB	86.49	78.98	72.23	74.97	63.63	54.00	2.8
F-PointNet	LiDAR+RGB	91.17	84.67	74.77	82.19	69.79	60.59	5.9
AVOD	LiDAR+RGB	89.75	84.95	78.32	76.39	66.47	60.23	10
PointRCNN	LiDAR	92.13	87.39	82.72	86.96	75.64	70.70	-
F-ConvNet	LiDAR+RGB	91.51	85.84	76.11	87.36	76.39	66.69	2.1
FastPointRCNN	LiDAR	90.87	87.84	80.52	85.29	77.40	70.24	15.4
MMF	LiDAR+RGB	93.67	88.21	81.99	88.40	77.43	70.22	12.5
STD	LiDAR	94.74	89.74	86.42	87.95	79.79	74.16	25
PV-RCNN	LiDAR	94.98	**90.65**	86.14	**90.25**	81.43	76.82	12.5
Ours	LiDAR	**95.15**	90.55	**87.30**	90.10	**82.56**	**77.25**	10.5

图 5 在 KITTI 基准上联合实例分割和 3D 对象检测的 3 个示例

Fig. 5 Three examples of joint instance segmentation and 3D object detection on the KITTI benchmark

图 6 3D 关键点可视化

Fig. 6 3D key point visualization

4　结束语

本文提出了用于联合 3D 对象检测和实例分割的统一框架。第一阶段设计了一个语义分割模块 M_s 与平行的中心偏移模块 M_c，将属于同一对象的所有前景点拉到其所属物理中心；第二阶段，提出了一种简单有效的基于关键点的方法来优化第一阶段得到的 BBox。本文所提出的框架仅需几个区域提议就能获得最新的效果，这对于在实际应用中进行实时感知非常重要。当前，使用 PointNet ++作为骨干网，是实时检测率的瓶颈。将来，希望设计一个更高效的骨干网络，以使系统实时运行，以便在 360° 视点上进行目标检测。

参考文献

[1] HE C, ZENG H, HUANG J, et al. Structure aware single-stage 3d object detection from point cloud[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11873-11882.

[2] CHEN X, MA H, WAN J, et al. Multi-view 3d object detection network for autonomous driving[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 1907-1915.

[3] KU J, MOZIFIAN M, LEE J, et al. Joint 3d proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 1-8.

[4] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4490-4499.

[5] YE M, XU S, CAO T. Hvnet: Hybrid voxel network for lidar based 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1631-1640.

[6] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. arXiv preprint arXiv: 1706.02413, 2017.

[7] PHAM Q H, NGUYEN T, HUA B S, et al. Jsis3d: Joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8827-8836.

[8] YANG Z, SUN Y, LIU S, et al. 3dssd: Point-based 3d single stage object detector[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11040-11048.

[9] WANG W, YU R, HUANG Q, et al. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2569-2578.

[10] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep clustering for unsupervised learning of visual features[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 132-149.

(上接第 15 页)

消费者信心指数的预测效果达到最好。该方法有效预测了消费者信心指数，可以将其应用到其它经济指标的预测，从而更好的掌握经济指标的变化趋势。

参考文献

[1] 王微，刘涛. 以强大国内市场促进国内大循环的思路与举措[J]. 改革，2020(9)：5-14.

[2] 杨娜，王静雅. 基于 ARIMA 模型的消费者信心指数分析与预测[J]. 企业导报，2012(3)：7，33.

[3] 董现垒，Bollen Johan，胡蓓蓓. 基于网络搜索数据的中国消费者信心指数的测算[J]. 统计与决策，2016(5)：9-13.

[4] 刘伟江，李映桥. 基于网络搜索数据的消费者信心指数预测研究——以台湾地区为例[J]. 浙江学刊，2015(2)：180-186.

[5] 邹鸿飞，王建州. 一种基于差分灰狼算法的消费者信心预测指数的设计[J]. 数量经济技术经济研究，2019，36(2)：120-134.

[6] 唐晓彬，董曼茹，张瑞. 基于机器学习 LSTM&US 模型的消费者信心指数预测研究[J]. 统计研究，2020，37(7)：104-115.

[7] Odendaal Hanjo, Reid Monique, Kirsten Johann F. Media - Based Sentiment Indices as anAlternative Measure of Consumer Confidence[J]. South African Journal of Economics, 2020, 88(4)：409-434.

[8] 徐映梅，高一铭. 基于互联网大数据的 CPI 舆情指数构建与应用——以百度指数为例[J]. 数量经济技术经济研究，2017，34(1)：94-112.