综述 | 3D目标检测多模态融合算法（附链接）

焉知智能汽车 2021-12-31 22:40

编者荐语：本篇文章主要想对目前处于探索阶段的3D目标检测中多模态融合的方法做一个简单的综述，主要内容为对目前几篇研究工作的总结和对这个研究方面的一些思考。

前言

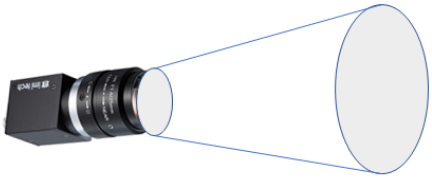
在前面的一些文章中，笔者已经介绍到了多模态融合的含义是将多种传感器数据融合。在3D目标检测中，目前大都是将lidar和image信息做融合。在上一篇文章中，笔者介绍到了目前主要的几种融合方法，即early-fusion,deep-fusion和late-fusion，并介绍了一种基于Late-fusion的融合方法。但是在大多数研究工作中，都是以deep-fuion的方法为主要的融合策略。

1 背景知识

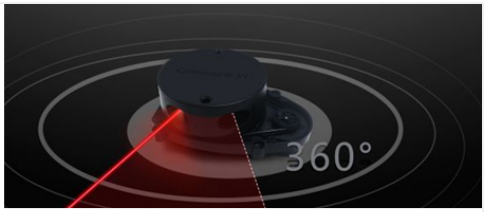
1.1 多模态融合的主要难点

难点一：传感器视角问题

3D-CVF（ECCV20）的研究提出的做fusion的对做融合工作最大的问题即是在视角上的问题，描述为如下图所示的问题，camera获取到的信息是“小孔成像”原理，是从一个视锥出发获取到的信息，而lidar是在真实的3D世界中获取到的信息。这使得在对同一个object的表征上存在很大的不同。



相机view-- 视锥



激光雷达-世界坐标系环视

难点二 数据表征不一样

这个难点也是所用多模态融合都会遇到的问题，对于image信息是dense和规则的，但是对于点云的信息则是稀疏的、无序的。所以在特征层或者输入层做特征融合会由于domain的不同而导致融合定位困难。

难点三 信息融合的难度

从理论上讲，图像信息是dense和规则的，包含了丰富的色彩信息和纹理信息，但是缺点就是由于为二维信息。存在因为远近而存在的sacle问题。相对图像而言，点云的表达为稀疏的，不规则的这也就使得采用传统的CNN感知在点云上直接处理是不可行的。但是点云包含了三维的几何结构和深度信息，这是对3D目标检测更有利的，因此二者信息是存在理论上的互补的。此外目前二维图像检测中，深度学习方法都是以CNN为基础设计的方法，而在点云目标检测中则有着MLP、CNN，GCN等多种基础结构设计的网络，在融合过程中和哪一种网络做融合也是比较需要研究的。

1.2 点云和imgae融合的纽带

既然做多模态特征融合，那么图像信息和点云信息之间必然需要联系才能做对应的融合。就在特征层或者输入层而言，这种联系都来自于一个认知，即是：对于激光雷达或者是相机而言，对同一个物体在同一时刻的扫描都是对这个物体此时的一种表征，唯一不同的是表征形式，而融合这些信息的纽带就是绝对坐标，也就是说尽管相机和激光雷达所处的世界坐标系下的坐标不一样，但是他们在同一时刻对同一物体的扫描都仅仅是在传感器坐标系下的扫描，因此只需要知道激光雷达和相机之间的位置变换矩阵，也就可以轻松的得到得到两个传感器的坐标系之间的坐标转换，这样对于被扫描的物体，也就可以通过其在两个传感器下的坐标作为特征联系的纽带。

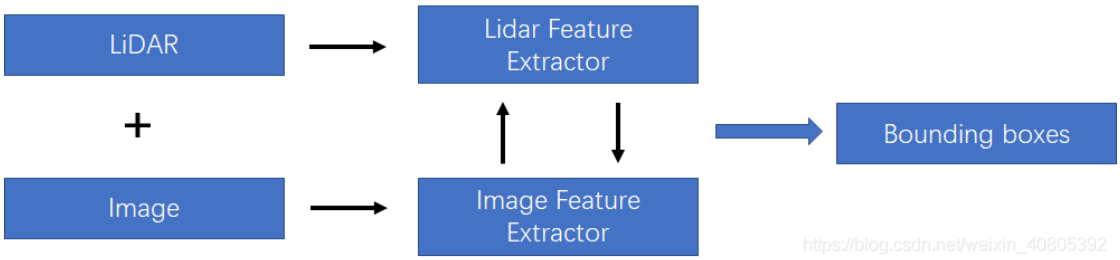
但是，就联系的纽带而言，由于在做特征提取过程中可能存在feature-map或者domain的大小的改变，所以最原始坐标也会发生一定的改变，这也是需要研究的问题。

2 目前存在的一些融合方法

如果硬要把目前存在的融合方法做一个划分的话，那么笔者从early-fuion，deep-fusion和late-fusion三个层面对最近的文章做一些简单介绍。这里把early-fusion和deep-fusion当做同一类融合方法介绍，late-fusion当做另外一种融合策略介绍。

2.1 early-fusian & deep-fusion

在上一篇文章中，笔者也提及到这种融合方法如下图所示。在后面，笔者将这两种方法都统称为特征融合方法，将late-fusion成为决策融合方法。如下图所示的融合方法，该融合需要在特征层中做一定的交互。主要的融合方式是对lidar 和image分支都各自采用特征提取器，对图像分支和lidar分支的网络在前馈的层次中逐语义级别融合，做到multi-scale信息的语义融合。



为了方便分析，在该种融合策略下，笔者按照对lidar-3D-detection的分类方法分为point-based的多模态特征融合和voxel-based的多模态特征融合。其差别也就是lidar-backbone是基于voxel还是基于point的。就笔者的理解是，基于voxel的方法可以利用强大的voxel-based的backbone（在文章TPAMI20的文章Part-A^2中有研究过point-based方法和voxel-based的方法最大的区别在于CNN和MLP的感知能力上，CNN优于MLP）。但是如果采用voxel-backbone的方法就会需要考虑点到图像的映射关系的改变，因为基于point的方法采用原始的点云坐标做为特征载体，但是基于voxel的方法采用voxel中心作为CNN感知特征载体，而voxel中心与原始图像的索引相对原始点云对图像的坐标索引还是存在偏差的。

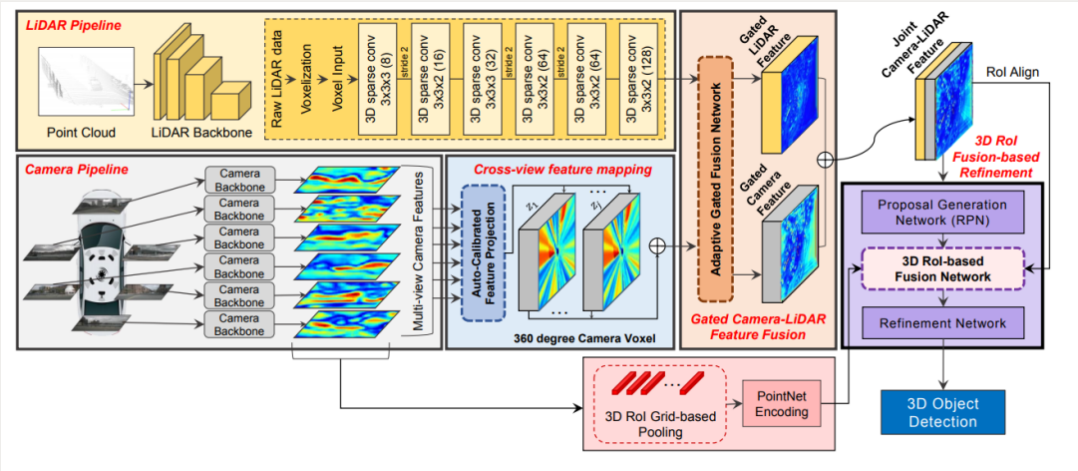
1）基于voxel-based的多模态特征融合

3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-View Spatial Feature Fusion for 3D Object Detection.

文章链接：

https://arxiv.org/pdf/2004.12636

这篇发表在ECCV20的多模态融合的文章网络结构图如下所示，该特征融合阶段为对特征进行融合，同时对于点云的backbone采用voxel-based的方法对点云做特征提取。所以这里需要解决的核心问题除了考虑怎么做特征的融合还需要考虑voxel-center作为特征载体和原始点云坐标存在一定的偏差，而如果将图像信息索引到存在偏差的voxel中心坐标上，是本文解决的另外一个问题。



3D-CVF特征融合方法

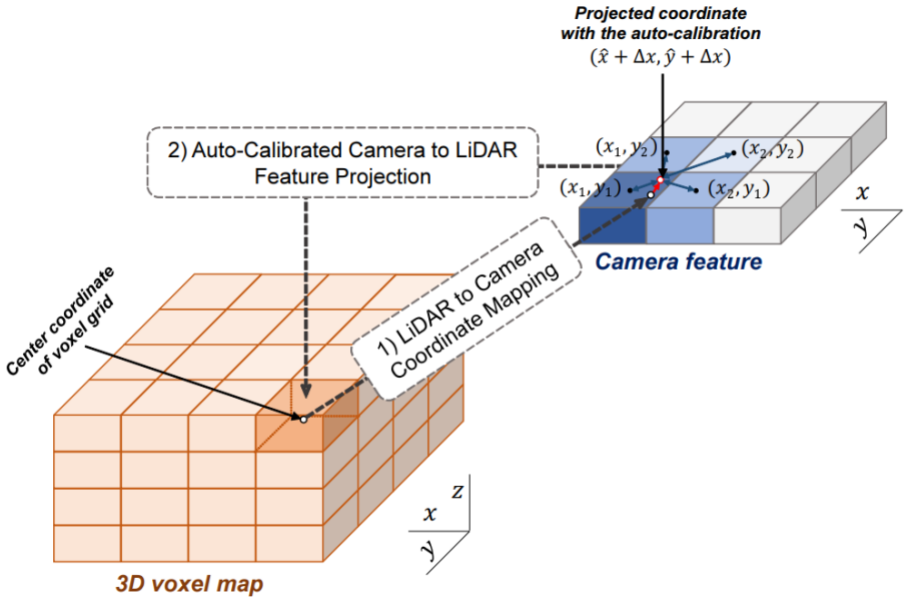
3D-CVF 将camera的pixel转化到点云的BEV视图上（voxel-feature-map）时，转化的大小是lidar-voxel-feature-map的x-y各自的两倍大小，也就是说整体的voxel个数是Lidar的四倍，即会包含比较多的细节信息。

以下表示的Auto-Calibrated Projection Method的设计方案，前面提到的是该结构是将image转化到bev上的网络结构，具体的做法是：

（1）投影得到一个camera-plane，该plane是图像特征到bev视角的voxel-dense的表达。

（2）将lidar划分的voxel中心投影到camera-plane上（带有一个偏移量，不一定是坐标网格正中心）

（3）采用近邻插值，将最近的4个pixel的image特征插值个lidar-voxel。插值的方式采用的是距离为权重的插值方法。



这样，作者就得到了了image信息的feature-map在lidar-voxel上的表示，值得提到的是前面说的偏移值是为了更好的使camera和lidar对齐。这里的第二步也就是为了解决上面提到的做标偏差的问题。
如下图所示，

2）基于point-based的多模态融合方法

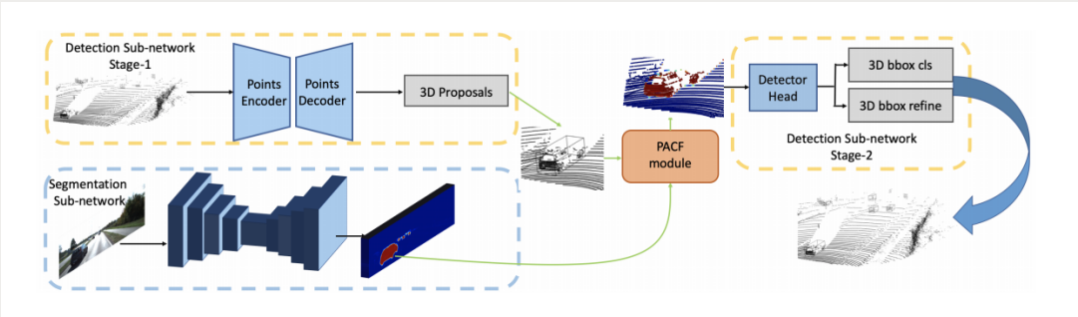
由于point-based的方法在特征提取过程也是基于原始点为载体（encoder-decoder的结构会点数先减少再增加但是点是从原始点中采样得到，对于GCN的结构则是点数不改变），所以在做特征融合时，可以直接利用前面提到的转化矩阵的索引在绝对坐标系上做特征融合，所以目前基本也都是基于point的方法比较好融合，研究工作也多一些。

PI-RCNN: An Efficient Multi-sensor 3D Object Detector with Point-based Attentive Cont-conv Fusion Module.

文章链接：

https://arxiv.org/pdf/1911.06084

发表在AAAI20，point分支和image分支分别做3D目标检测任务和语义分割任务，然后将图像语义分割的特征通过索引加到proposals的内部的点云上做二次优化。

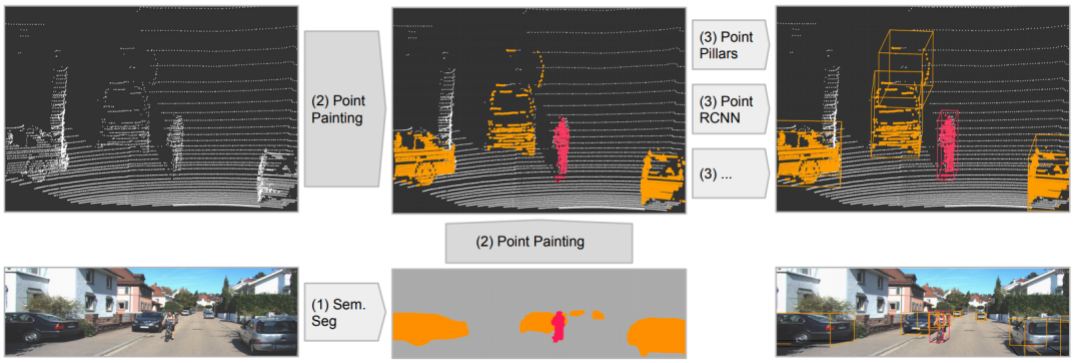


PointPainting: Sequential Fusion for 3D Object Detection

文章链接：

https://arxiv.org/pdf/1911.10150

发表在CVPR20，该工作的fusion方式是采用二维语义分割信息通过lidar信息和image信息的变换矩阵融合到点上，再采用baseline物体检测；可以理解为对于语义分割出的物体多了一些信息作为引导，得到更好的检测精度。和上面的pi-rcnn的不同之处是该融合是一个串联的网络结构，将语义分割后的特征和原始点云一起送入深度学习网络中。

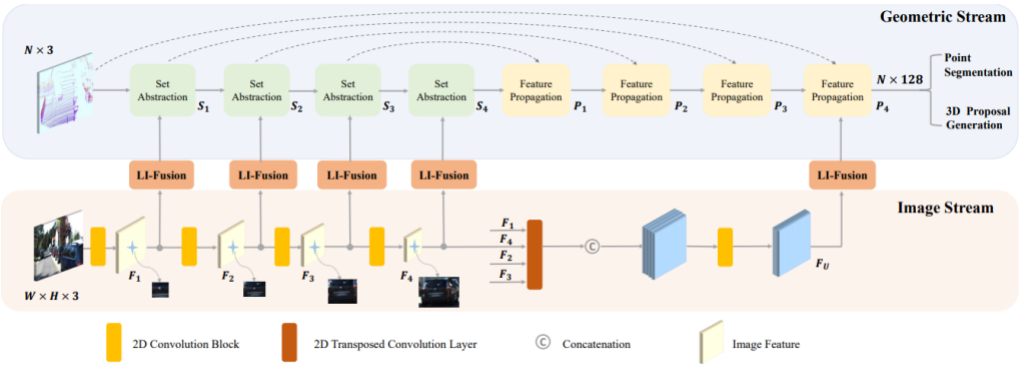


EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection

文章链接：

https://arxiv.org/pdf/2007.08856

发表在ECCV20的文章，如下图所示，其网络结构点云分支是point encoder-decoder的结构，图像分支则是一个逐步encoder的网络，并且逐层做特征融合。



2.2 late-fuion

在决策层面的融合相对简单很多，不需要考虑在信息层面的融合和互补，也就是说，只要是两种网络做同样的任务，那么在得到各自的结果后，对结果做决策上的选择融合。

CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection

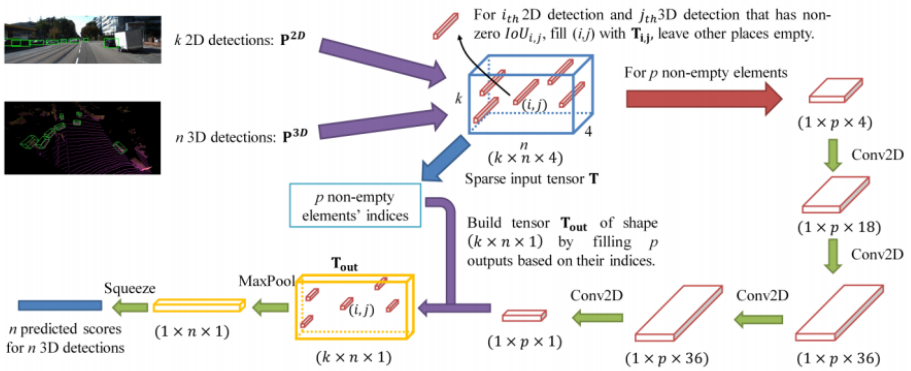
https://arxiv.org/pdf/2009.00784.pdf

这就是笔者上一篇分享的文章，从下图可以看出该网络经历了三个主要的阶段：

（1）2D和3D的目标检测器分别提出proposals

（2）将两种模态的proposals编码成稀疏张量

（3）对于非空的元素采用二维卷积做对应的特征融合。



3 笔者总结

本文主要就目前的几篇做多模态融合的文章做了一定的介绍，从大层面上的fusion的阶段讲起，然后进一步在deep-fuion阶段划分为基于point-based和voxel-based的融合方法，基于point的方法具有先天的优势

是具有和image 的索引和不具有空间变化，而voxel的方法可以更有效的利用卷积的感知能力。最后大家如果对自动驾驶场景感知的研究比较感兴趣和想要找文章的话，可以去以下链接找最新的研究。

https://github.com/LittleYuanzi/awesome-Automanous-3D-detection-methods

参考文献

[1] 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-View Spatial Feature Fusion for 3D Object Detection:
https://arxiv.org/pdf/2004.12636
[2]Part-A^2 Net: 3D Part-Aware and Aggregation Neural Network for Object Detection from Point Cloud :
https://arxiv.org/abs/1907.03670
[3] PI-RCNN: An Efficient Multi-sensor 3D Object Detector with Point-based Attentive Cont-conv Fusion Module. :
https://arxiv.org/pdf/1911.06084

People who liked this content also liked

永远退出机器学习界！从业八年，Reddit网友放弃高薪转投数学：风气太浮夸

新智元

手把手教你实现无监督学习的K-means算法

图灵教育

《自然》：机器视觉行为理解与脑神经有内在关联？上交卢策吾团队构建映射模型

机器之心