多模态 3D 目标检测：

以前的工作：

单模态：基于图片：
基于 LiDAR：基于 LiDAR 又可以分为 voxel-based、point-based（pointnet 等）和 voxel-point based。

多模态：大多是基于 image 和 LiDAR 的双模态目标检测。

主要问题：

1.传感器视角问题

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。

2. 数据表征不一样

这个难点也是所有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。

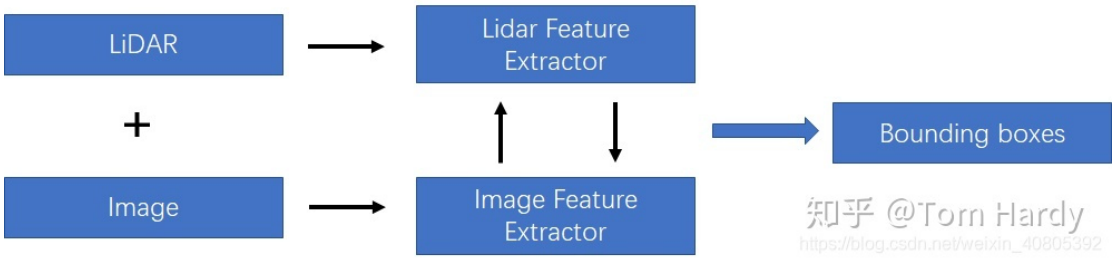
点云和 imgae 融合的纽带

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。

目前存在的一些融合方法

大致可分为三类：early-fusion、deep-fusion 和 late-fusion。

其中 early-fusion 和 deep-fusion 可分为一类，其网络结构如图：3D-CVF、

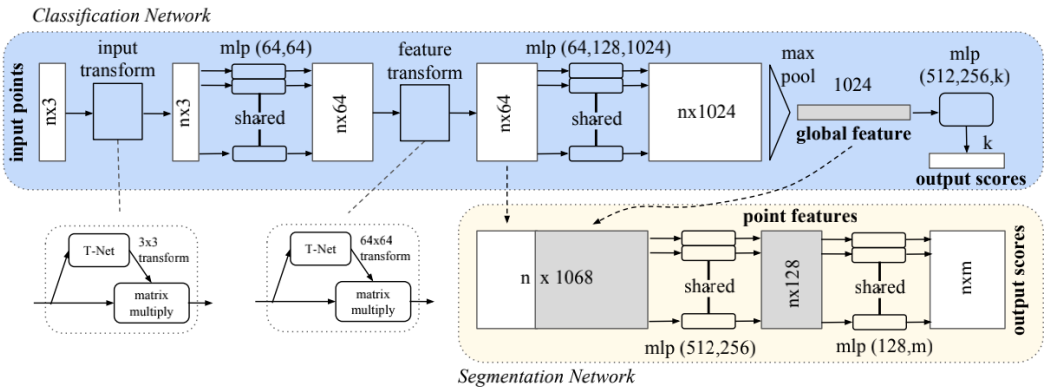


首先对 LiDAR 和 Image 分别提取特征，然后再对其进行特征融合。融合过程可以在分层特征中持续执行。

Late-fusion：

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。

PointNet：



Pointnet 是 point-based 的 3D 分割网络。如网络结构所示，pointnet 的输入为点的三维坐标（x，y，z）。

网络的两个亮点：

1.空间变换网络解决旋转问题：三维的 STN 可以通过学习点云本身的位姿信息学习到一个最有利于网络进行分类或分割的 DxD 旋转矩阵（D 代表特征维度，pointnet 中 D 采用 3 和 64）。至于其中的原理，我的理解是，通过控制最后的 loss 来对变换矩阵进行调整，pointnet 并不关心最后真正做了什么变换，只要有利于最后的结果都可以。pointnet 采用了两次 STN，第一次 input transform 是对空间中点云进行调整，直观上理解是旋转出一个更有利于分类或分割的角度，比如把物体转到正面；第二次 feature transform 是对提取出的 64 维特征进行对齐，即在特征层面对点云进行变换。

2.maxpooling 解决无序性问题：网络对每个点进行了一定程度的特征提取之后，maxpooling 可以对点云的整体提取出 global feature。

其中，mlp 是通过共享权重的卷积实现的，第一层卷积核大小是 1x3（因为每个点的维度是 xyz），之后的每一层卷积核大小都是 1x1。即特征提取层只是把每个点连接起来而已。经过两个空间变换网络和两个 mlp 之后，对每一个点提取 1024 维特征，经过 maxpool 变成 1x1024 的全局特征。再经过一个 mlp（代码中运用全连接）得到 k 个 score。分类网络最后接的 loss 是 softmax。