自动驾驶感知开发理解（二）：数据标定与算法方案选择

焉知智能汽车 2022-03-15 21:54

The following article is from 汽车电子与软件 Author 小蝎子

汽车电子与软件
每天分享一篇技术文章！

来源 | 汽车电子与软件
知圈 | 进"域控制器群"请加微13636581676,备注域

上一章《谈谈对自动驾驶感知开发理解》中我主要是分享了数据采集的各种形式，当获取大量的实车原始数据后，首要的工作就是为数据进行真值标注。在这章中我们讨论下数据标定的相关内容。

现在所有的CV领域公司应该都有庞大的人工标注团队，有些企业前期会选择外购数据来快速实现初版产品，但长期看自有数据仍是CV公司的核心竞争力，甚至是有成为护城河的能力。

人工标注样本这条路直接有效，但增效的瓶颈是十分明显的。简单算下效率：对于CV感知任务中标定相对容易的目标框标定（含状态、类别等信息勾选），假设1个标定人员平均一个目标框需要5秒，每天8个小时的标定量大概上限是5800个目标；在自动驾驶场景，单张图像内目标数量平均10个左右，那么一天一个标定人员的标定上限就是大概600张（这是把人当做机器不会疲劳的计算，实际标定中是达不到这个速度的)。这种效率支撑每年百万级的标定任务可能比较现实，但如果是应对数据驱动这种亿万级别的样本数量，将会是重大负担，以一亿张样本算、大概需要600人一年的工作量，成本将成企业严重负担且不具备可持续性。

很多公司会使用预标注来提高效率，通常都是使用一个计算量夸张的离线模型在待标定的样本上得到感知结果，再由人工完成标签清洗或者直接用这个伪标签进行后续开发。但这个模式逃脱不了离线模型对新增样本的适配问题，当相机发生成像、架设等重大变化时，离线模型依然需要喂送大量的人工真值标签。在纯CV领域，有非常多技术领域在研究这个样本使用问题，从应用角度考虑半监督（离线模型的伪标签使用也算是一种半监督算法）和迁移学习算法算是比较靠谱的方法，但目前看依然有很强的定制性，不同任务之间甚至同一任务不同数据量之间很难形成一套通用或长期有稳定有效的工具。

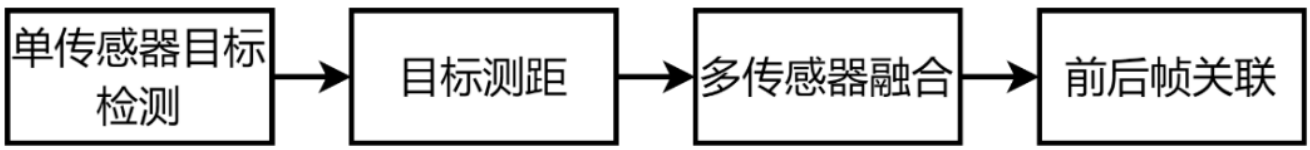
在自动驾驶感知（特别是纯视觉）任务上；其实可以从其它角度完成数据筛选和标注工作。首先需要讨论一下什么是感知真值：最先明确一点的是任何方案都没有绝对的真值，即使是人工仔细标注那也一定会有偏差，所以只要是能提供一套各方面性能均优于当前方案的结果理论上都可以作为真值使用。其次真值是与任务需求相关的，当前自动驾驶感知任务的需求一般包含动态障碍物感知（机动车、行人等)和静态障碍物（车道线、车

位、可行驶区域、通用障碍物等），而不同于其它CV感知任务，自动驾驶里面有个额外且非常重要的需求是对每个感知目标均要输出距离信息。

用于数据驱动的真值还有一个前置依赖是算法方案。如果算法方案中有很多模块是基于人工逻辑实现的，那么数据驱动的效率将会大幅降低，一个比较折中的做法是将端到端的真值降级成模块真值。

动态障碍物：

对于不含激光传感器的动态障碍物感知任务，以多一套标准的目标级融合算法流程来分析，基本的模块就是如下几个步骤（模块的顺序可能有变化，比如如果测距在检测前面可以理解为视觉点云再聚类的方案、如果多传感器融合在目标检测前可以理解为前融合，这种理解不严谨但能帮助大家快速入门)：



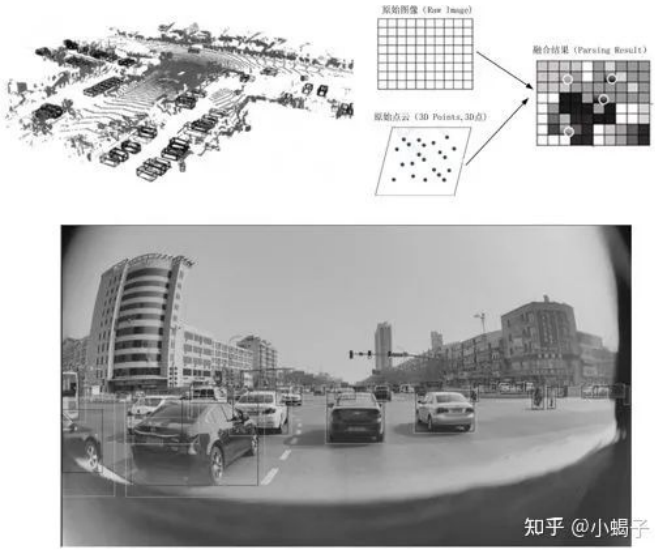
如前述分析动态目标感知的最终目标是输出一个世界坐标系下具有属性的目标队列，如果整个算法流程是由一个网络方案端到端实现，且能获取到视频级别目标队列真值那么整个闭环迭代就能运转起来；如果网络模型只能覆盖到前一级如单帧的多传感器融合或者单传感器测距，那真值需求就退化到单帧时刻的世界坐标系目标队列；如果方案中只有检测任务是基于网络实现，那真值需求就变成了图像的bounding box。

大批量世界坐标系的真值来源可以依靠高线束的激光，激光的测距性能对于非激光方案来说可以认为是天花板了，而且激光视角是单一的全局视角（多激光也可以融合成一个点云图)，不存在多个视角冗余标定的情况；在有可靠的同步和传感器架设信息的前提下，基于激光点云得到世界距离真值也可以转换到各个传感器上。目标的ID属性可以由标定人员借助视频信息完成标注。

模块	输出	理想伪真值来源
目标检测	目标框	激光点云标注结果投影
测距	世界坐标系下三维坐标（部分区域）	激光点云人工标注或3D检测算法伪标签
融合	世界坐标系下三维坐标（全部区域）	激光点云人工标注或3D检测算法伪标签
关联	目标ID及轨迹	人工标注

从算法选择上看，量产方案上直接做到多传感器融合+前后关联的还是比较少见的，当前单传感器测距+关联的网络方案和多传感器测距融合的网络方案在逐渐成为热门研究，那选择这几种方案的真值应该是逃不出上面的分析的底层逻辑。仅仅在目标检测上使用网络的方案已经比较成熟了，结合激光来生成图像标签的方法相信各

个自动公司也会是主流方案。比如北京地平线信息技术有限公司在2021.05.27申请了一个专利《对图像数据进行标注的方法、装置及电子设备》，就是一种结合激光点云数据，通过点云的3D检测获得目标的世界距离真值，通过投影矩阵完成图像真值的计算：



静态障碍物：

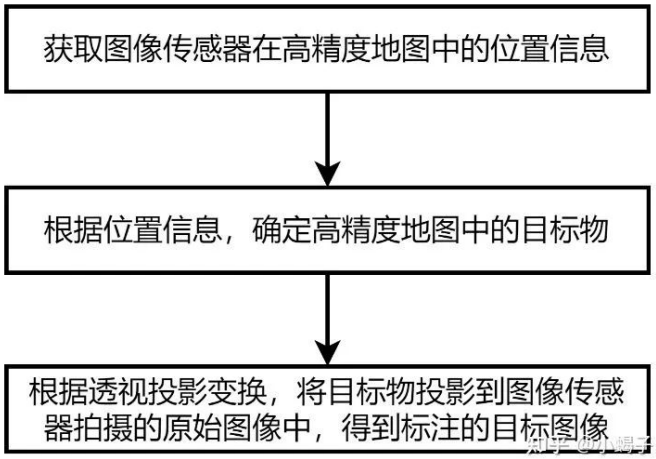
静态障碍物可以分为两种：一种是道路内的立体目标，如路沿、墙壁、立杆等；一种是道路路面标志信息，如车道线、车位线、箭头、斑马线等。第一类目标也是可以被激光有效感知的，在实际中通常以可行驶区域来表征结果。第二类目标靠激光不是特别稳定，但如果有该场景的高精地图，那么可以借助定位信息获得车辆所在时刻周边的地图信息，这些地图信息可以提供高质量的真值数据。

静态障碍物的感知由于类别较多，所以方案也较多。一般而言要么是基于检测任务得到目标的bounding box,要么是基于逐像素分割技术再做后续的目标生成。在此只讨论真值生成模式，具体的方案不做详细讨论。

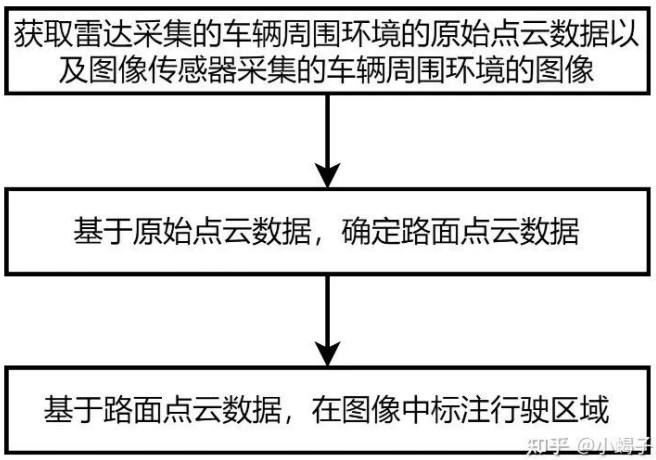
类别	输出	理想伪真值来源
立体障碍物	深度信息	激光点云
	可行驶区域	激光点云+高清地图
路面信息	线型目标	高清地图

结合激光和高清地图完成这类目标的真值标定也有较多专利，下面两篇是比较有代表性的：

第一篇是北京四维图新科技股份有限公司在2019.05.09申请的《图像标注方法、装置、系统及存储介质》，这篇专利就是介绍一种基于高清地图的标注方法，在获取图像传感器在高精度地图中的位置信息后，根据所述位置信息，确定所在高精度地图中的目标物；根据透视投影变换，目标物投影到图像传感器拍摄的原始图像中，得到标注的目标图像结果：

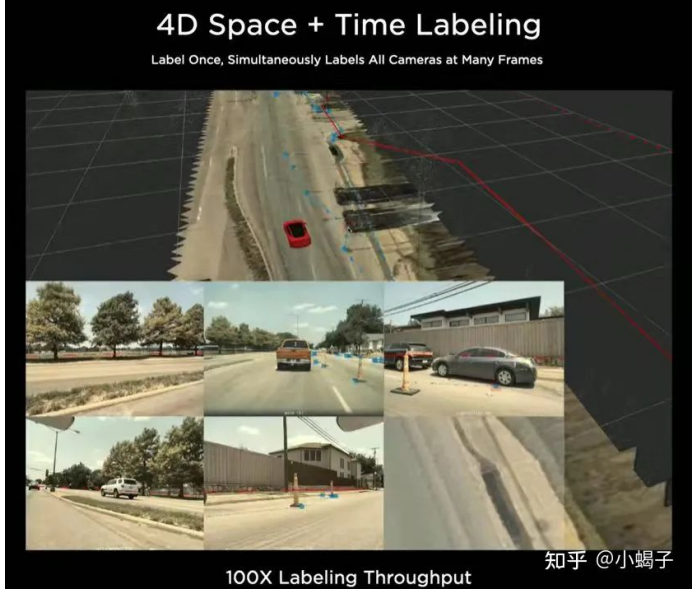


第二篇是驭势科技（北京）有限公司在2018.12.27申请的《一种图像行驶区域标汪方法、装置，车载设备及存储介质》。这篇专利介绍的方案如下：获取雷达采集的车辆周围环境的原始点云数据以及图像传感器采集的车辆周围环境的图像基于原始的激光点云数据，确定出路面点云数据；再基于路面点云数据，在图像中标注可行驶区域。



特斯拉4D标注：

当然即使没有这些高精度传感器，也可以有其它方法来提升多镜头的标注效率，比如Tesla自研的一套4D的标注设备，也能极大提升了数据产出的效率。



其思路是将获取的多路视频做三维重建，并结合时序信息形成一个4D空间，在这个空间中进行目标标注，可以通过投影直接反馈到各个相机的图像上，标注一次即可。这种标注方式相较传统的2D标注预期能提升100倍的效率。

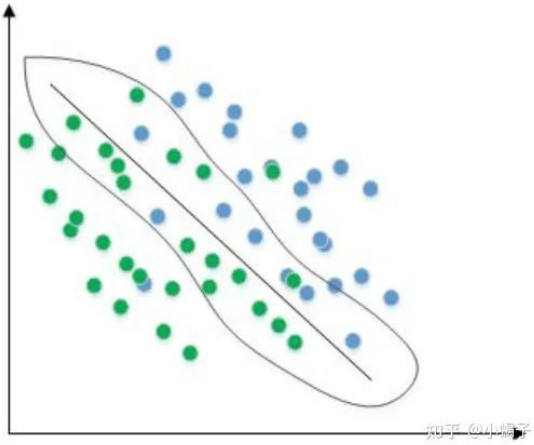
对于自动驾驶领域，相机是一个偏低端的传感器，因此借助高精度传感器来提升视觉的性能上限是一个非常有前景的工作。相信很多相关领域的同行都知道量产车的激光之争，很多不了解这块工作的人都会觉得不用激光是件非常疯狂的事情，但事实上如果你有大量额外的高精度传感器数据和一个合适的算法，纯视觉方案是能够不断逼近激光的感知能力。

下面就将分享一些这类视觉端到端的算法。

数据量与算法选择

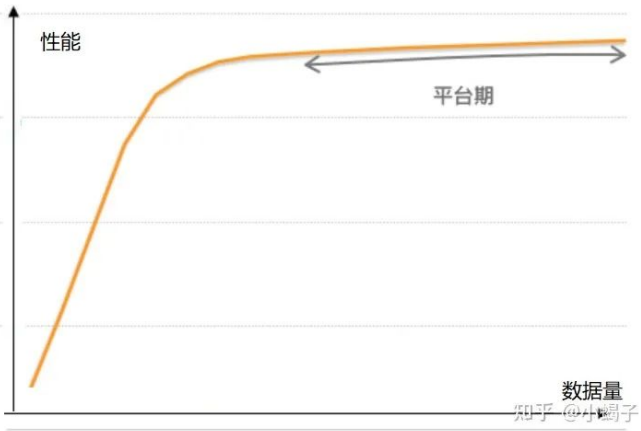
首先先承上启下聊一聊数据量和算法性能的关系。深度学习网络是机器学习的升级版，都可以理解成特征提取和任务判别两个部分。特征提取是将原始信息转化到一个易于判别的特征空间；任务判别是在特征空间中按给定的规则进行划分。

以最简单的分类任务来看（如下图），靠近分界面的点越多，对寻找出最合适的分界面越有利，而远离分界面的点再多，对结果也不起作用（具体的可以去研究支持向量机的原理）。因此从理论上看数据并不是越多模型性能就一定会更好。那我们还有必要如前文一样，费劲去收集大量的样本么？

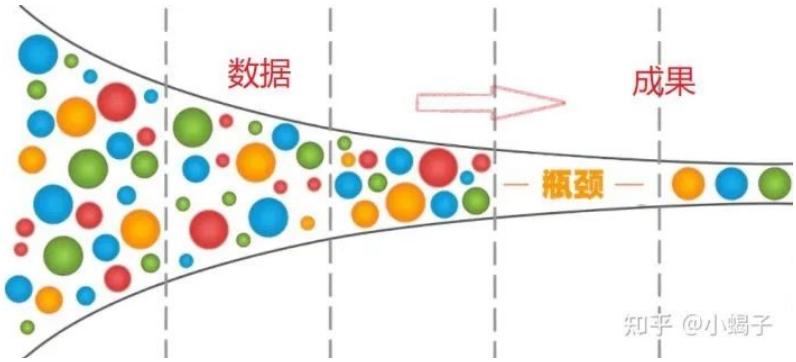


答案是肯定，简单分析下：首先支持向量机是一个封闭性、后置性的理论。它讨论的是当数据集给定，特征空间固定下的数据规律。而在实际问题中，我们面对的是一个完全开集的场景，有限维特征空间只是我们一种逼近真实世界的理解，其实并不存在。其次即使是在有限维度下研究，我们也无法提前获取支持向量的选择方式，也就是无法精准的描述出应该要去获取什么样的样本，这需要海量的样本去确定初始边界，然后才能去研究边界的支持向量问题，类似一个先有鸡还是先有蛋的螺旋证明问题。

在算法开发中，如果一个模型参数固定（特征维度也就固定），那在数据采集随机的情况下，性能与数据量的关系会遵循以下曲线关系：前期随数据量增加性能会逐渐提升，到了一个阶段后大量的数据都是非支持向量，性能提升缓慢，进入平台期。后续如果要突破这个平台期，要么去研究支持向量的挖掘，要么提高特征维度（如扩大模型计算量等）。

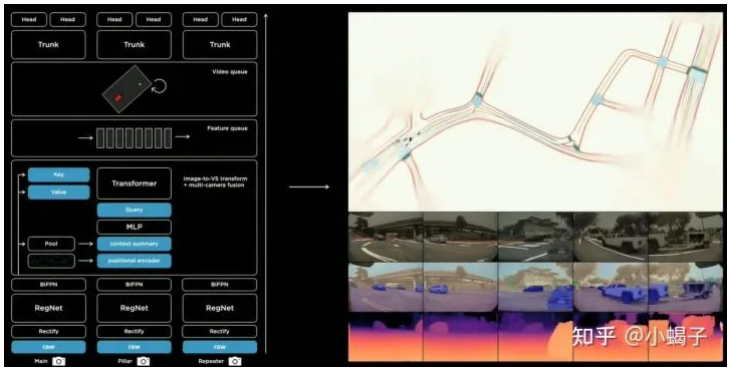


在实际的开发过程中，选择一个算法方案需要考虑的因素非常多，包括开发周期、芯片选型、性能需求、技术储备等等。如果从算法数据双轮滚动模型出发，最佳的方案是尽量选用端到端的纯网络设计。上一章聊过，如果一个算法方案的网络化比较靠前，那必然需要对数据做更多的处理同时还需要更多的人工参与，降低闭环的效率，最终成为整个环节的瓶颈。



特斯拉算法方案

特斯拉在21年展示了自己一套完整的感知框架，整个神经网络模型可以分成四大部分：



1. **图像特征提取网络：**特征提取使用了RegNet,多层次特征融合使用了BiFPN

2. **测距+多摄像头特征融合：**使用Transformer结构，这个也是这套架构的核心，既解决了目标测距的网络实现问题，也解决多个镜头的融合问题；

3. **时序特征网络：**在投影到世界坐标空间后，采用RNN的结构提取时序特征；在这一步中还使用了车辆的位姿信息做为输入。

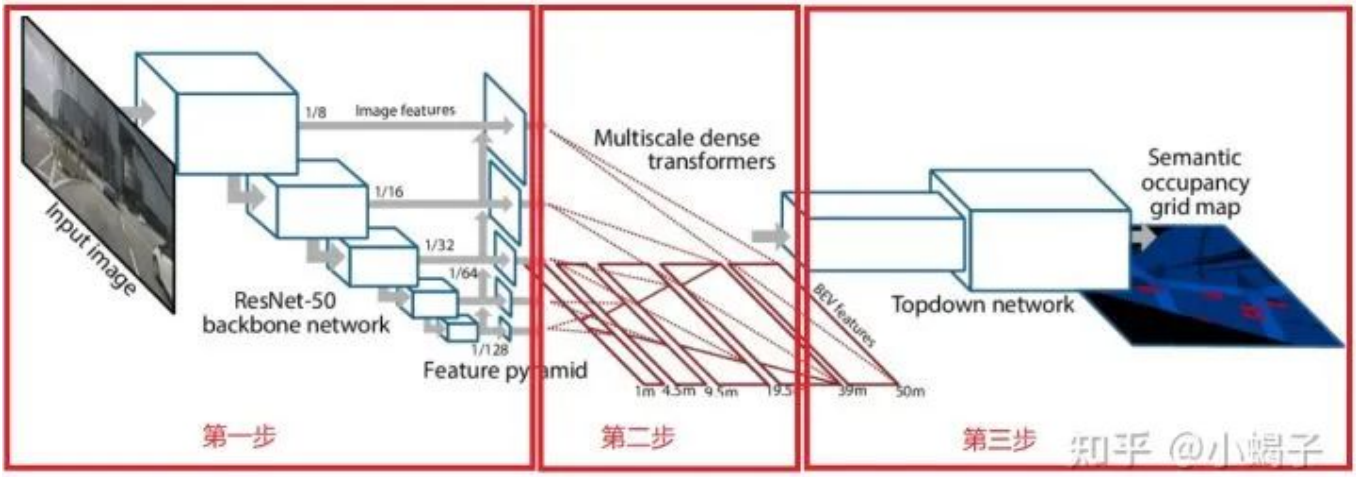
4. **多任务联合学习**：获取到时序的特征后，不同任务可以直接利用这个特征进行自身的任务学习，比如车道线检测、道路边缘检测、车辆检测等等。

整个框架未完全实现端到端（按马斯克和Andrej Karpathy在AI Day的问答环节的说法，他们不迷信深度学习，会在工作中逐步发现深度学习在某些功能上的优势后，才进行升级，然后再逐步探寻端到端的可能)，但是基本上感知各个模块都已完成了网络化。

除了特斯拉，也有其他厂商使用这套思想去设计自己的算法方案。从原理上看感知的端到端就是要使用网络逐步覆盖检测、测距、融合，再下一步就是场景建模；在学术界也有很多研究在支撑这种端到端技术的向前推进：

静态障碍物——PON(图像坐标到世界坐标投影的网络建模)

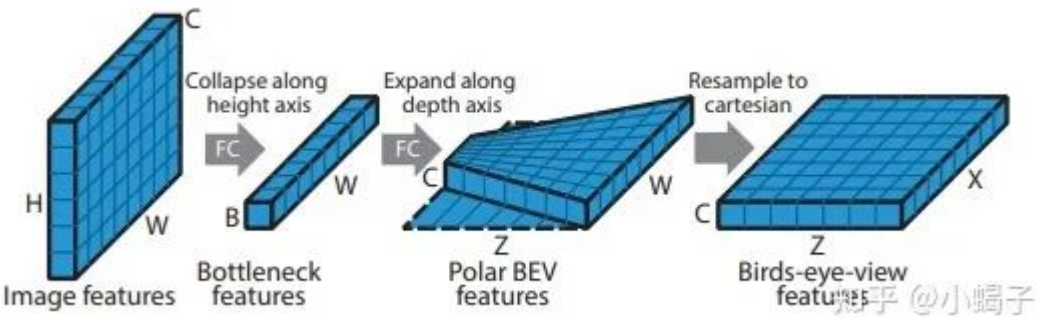
近些年做BEV分割的有很多，比如BEV-seg、Lit-Splat-Shoot框架等。这其中，我个人认为相对比较实用的是剑桥大学在2020年提出的Pyramid Occupancy Networks结构。这篇论文同时也使用贝叶斯概率来实现多相机融合，有创新性但没有解决完全的端到端。



PON结构主要分三步，如上图所示：

第一步：一个金字塔式的多尺度特征编码结构，每一层特征图将对应后面不同距离下的BEV特征，这种设计兼顾了远处小目标的感知能力以及近处分辨率过大引起的计算量问题。

第二步：第一步得到的多层特征图，分别使用Dense transformers结构来生成对应距离的BEV特征。先将特征图压缩，再沿入射角度展开特征，最后resize成目标特征尺寸。

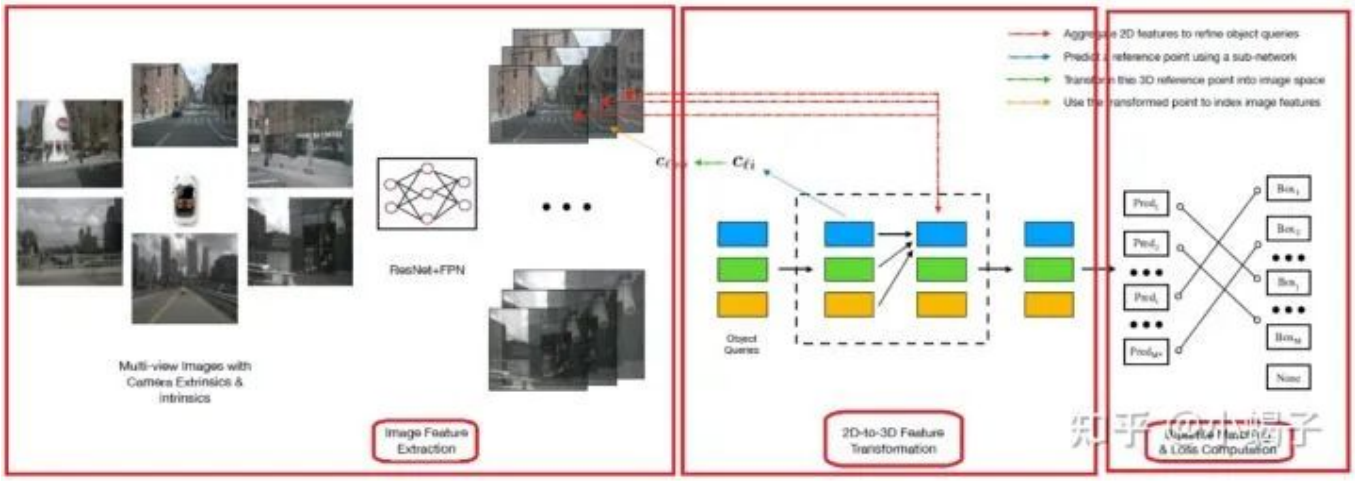


第三步：将第二步的多个特征图做解码得到最终的掩模图。从作者给的效果图看没有明显的分块效应，所以多个距离特征图存在一些重叠加权操作。

我认为这篇文章偏实用的一个主要原因是这个方法的核心就是聚焦在图像坐标系到俯视图栅格的转换上，而现在很多BEV方法比如Lift-Splat-Shoot框架，它们的落脚点偏向于在2D图像中直接提取深度，网络的复杂度会提高。而本文这种转换关系的设计的可移植性和可解释性就要高很多。

动态障碍物——DETR3D(多镜头融合的网络建模)

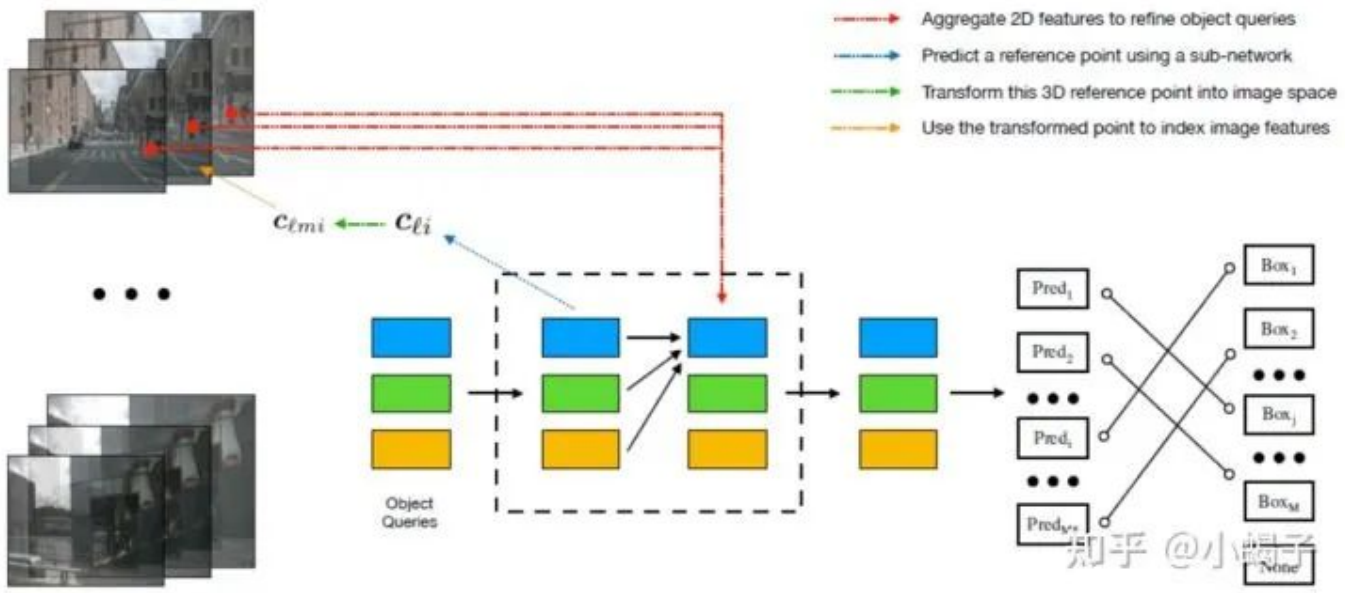
在这篇文章之前，很少有讨论多个相机检测并直接预测3D结果的端到端方案的。前文特斯拉的测距+多摄像头特征融合基本上与本文思想类似。



DETR3D的核心步骤也是分三步：

第一步：通过金字塔式的多尺度ResNet 主干来进行特征编码，每一个视角图像使用的主干权值共享。

第二步：在这一步中利用相机变换矩阵和几何逆投影将提取的2D特征和3D目标检测联系在一起，我们可以借助图像上的颜色线来理解这个步骤。



先初始化一个3D object queries

通过子网络对object queries每个目标预测一个参考点(图中蓝线)

将参考点利用相机参数投影到图像，找到其在原始图像中对应的位置（图中绿线）

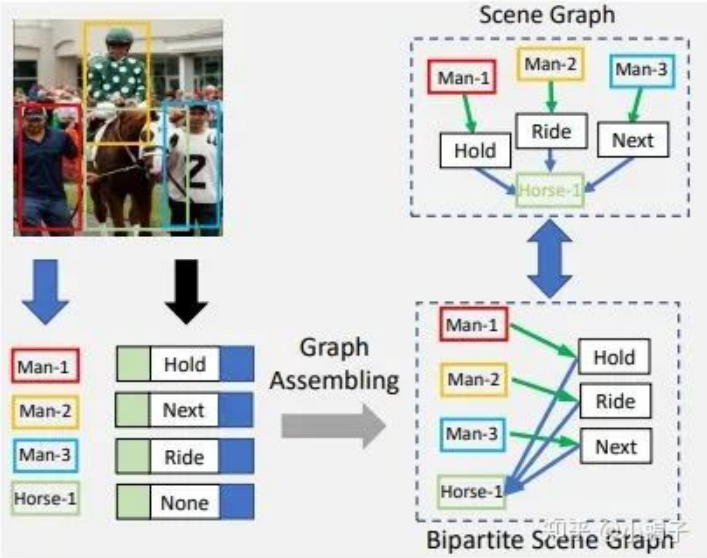
利用图像参考点索引到对应的目标图像特征，这个图像特征即第一步生成的某一个尺度下的特征（图中黄线）

使用多头注意力机制，利用2D特征对queries进行迭代优化（图中红线）

第三步：使用端到端的loss来监督整个训练过程。

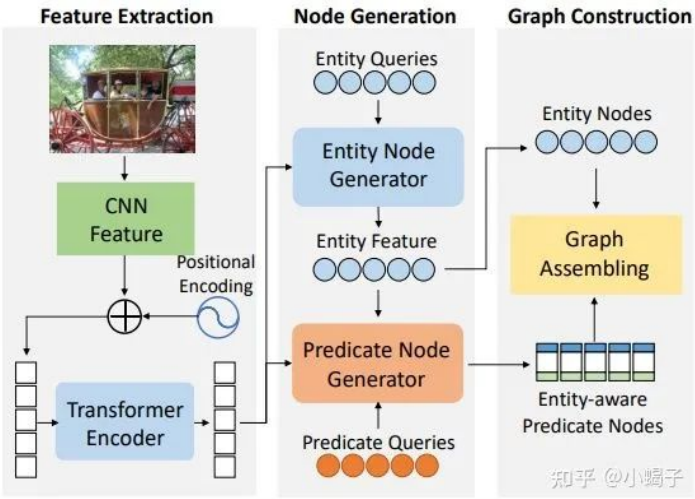
场景建模——SGTR(目标关系的网络建模)

场景建模的任务是描述图片中的目标和他们之间的相互关系，是个识别图中各个元素关系的任务。在SGTR方法中，将这个任务划分成两步：第一步先提取图像中的目标和谓语动词，然后第二步寻找目标和谓语词之间的联系，最终形成场景图，比如下面这个例子。



图像中目标有三个人：人1、人2、人3和一匹马，动作有三种：牵、挨着、骑；然后再有关系人1牵马、人2骑马、人3挨着马，最终得到场景描述

具体实现上，第一步提取目标（即图中的节点）和谓语动词：对原图做CNN卷积和positional encoding后，使用Transformer的编码结构得到两个输出，一个输出使用原始的transformer的解码结构得到目标结果，这个目标结果与上述transformer编码结构另一个输出一起形成查询列表，用于谓语动词的解码和生成。结合目标结果来生成谓语结果是本文的一个创新，因为谓语肯定与目标的位置和姿态都存在密切关联。



对于第二步计算目标与谓语的关系，主要就是寻找不同的距离空间，计算主语谓语和谓语宾语的距离矩阵。具体的训练方法论文中描述有点概念化，需要后续读代码和实操进行摸索。

到现在感知闭环中的几个关键模块都做了一些分享和讨论，下一章我会对开发中技术和研发做些理解分享。

Read more

People who liked this content also liked

我们研究了特斯拉、毫末智行「自动驾驶算法」的秘密

焉知智能汽车