



上海交通大学学位论文

一种基于多源传感器信息融合的
路侧导航增强单元

姓 名：杨嘉业
学 号：519021910359
导 师：张欣
学 院：航空航天学院
学科/专业名称：航空航天工程
申请学位层次：学士学位

2023 年 5 月

A Dissertation Submitted to
Shanghai Jiao Tong University for Bachelor Degree

TRUSTED VEHICLE NAVIGATION BASED ON
MULTI-SENSOR INTEGRATION FROM
NEXT-GENERATION ROADSIDE UNIT

Author: Jiaye Yang
Supervisor: Xin Zhang

School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, P.R.China
May 14th, 2023

上海交通大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：
日期： 年 月 日

上海交通大学 学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

公开论文

内部论文，保密 1 年 / 2 年 / 3 年，过保密期后适用本授权书。

秘密论文，保密 ____ 年（不超过 10 年），过保密期后适用本授权书。

机密论文，保密 ____ 年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打“√”）

学位论文作者签名： 指导教师签名：
日期： 年 月 日 日期： 年 月 日

摘要

在自动驾驶的落地过程中，车路一体化系统是尤为重要的一个环节。车路一体化是指利用无线网络，将车端与路端紧密相连，实现车端与路端的信息交换、信息共享。目前的车路一体化或车路协同至多实现超视距态势通知，相关信息没有进入车辆自主导航的闭环回信路，未能对保障车辆全天候、全场景下的导航，即目前导航领域最关心的“可信导航”有所贡献。本研究力求突破这一瓶颈。目前，车辆终端导航定位主要依赖于全球卫星导航系统（GNSS），但该系统受限于卫星相关误差、传播途径相关误差、接收机相关误差等限制，对车辆的定位、测速精度有限。

针对这一问题，我们希望在路侧单元装备激光雷达与相机，通过基于激光雷达与视觉融合的目标跟踪方法，完成车辆的路侧定位。然后将定位结果发回车端，结合车端的 GNSS 定位结果，通过基于优化的方法得到车辆位置的预测值，补偿 GNSS 在脆弱场景下的定位误差，达到导航增强的目的。

我们希望采用 EagerMOT 作为目标跟踪方法，并在 KITTI^[1] 数据集上验证了其效果。我们在由清华大学与百度公司发布的 DAIR-V2X 公开数据集上训练了二维与三维目标检测器。其中，二维目标检测器采用了 YOLOv4 框架，三维目标检测器采用了 PointRCNN 框架。最后我们提出了基于因子图优化的车辆定位补偿方法。

关键词：GNSS，新型路侧单元，目标跟踪，目标检测，因子图，可信导航

ABSTRACT

In the process of autonomous driving, Vehicle-Road Integration System is a particularly important part. Vehicle-Road Integration System refers to the use of wireless networks to closely connect the vehicle end and the road end to realize information exchange and information sharing. At present, the navigation and positioning of vehicle mainly depends on the global satellite navigation system (GNSS). But the system is limited by satellite-related errors, propagation path related errors and receiver relevant errors, thus limiting the accuracy of vehicle positioning and speed measurement.

In response to this problem, we hope to equip roadside units with LIDARs and cameras. And positioning the vehicles on the road end by object tracking method based on the fusion of LIDAR and vision. Then the positioning results will be sent back to the vehicle end, combined with the GNSS positioning results on the road end, predicting the vehicle position based on optimization method. This approach will compensate the positioning errors of GNSS in fragile scenes, and achieve the purpose of navigation enhancement.

We hope to adopt EagerMOT as the object tracking method. Its performance is evaluated on KITTI dataset. We trained 2D and 3D object detectors on the DAIR-V2X public dataset released by Tsinghua University and Baidu Inc. The 2D object detector adopts YOLOv4 framework and the 3D object detector adopts PointRCNN framework. Finally, we propose a vehicle positioning compensation method based on factor graph.

Key words: GNSS, roadside unit, object tracking, object detection, factor graph, trusted navigation

目 录

摘要	I
ABSTRACT	II
第一章 绪论	1
1.1 研究背景与研究意义	1
1.2 国内外研究现状	2
1.2.1 车路协同研究现状	2
1.2.2 激光雷达目标跟踪研究现状	3
1.2.3 视觉目标跟踪研究现状	4
1.2.4 激光雷达与视觉融合研究现状	5
1.3 研究内容与研究路线	6
1.3.1 研究路线	6
1.3.2 研究内容	6
1.4 本章小结	7
第二章 公开数据集框架	8
2.1 车路协同数据集综述	8
2.2 数据集标注与标定	9
2.2.1 KITTI 数据集介绍	9
2.2.2 KITTI 数据集标注格式	11
2.2.3 KITTI 数据集标定格式及坐标转换	11
2.2.4 DAIR-V2X 标注与标定	13
2.3 本章小结	14
第三章 目标跟踪器搭建	15
3.1 EagerMOT 原理	15
3.1.1 检测器融合	16
3.1.2 数据关联	16
3.1.3 轨迹生命周期管理	17
3.2 评价指标	18
3.2.1 CLEAR MOT 指标	18

3.2.2 HOTA 指标	19
3.3 实验结果	19
3.4 本章小节	21
第四章 目标检测器训练	22
4.1 2D 目标检测器	22
4.1.1 YOLOv4 网络模型	23
4.1.2 YOLOv4 训练及结果	24
4.2 3D 目标检测器	26
4.2.1 PointRCNN 网络模型	26
4.2.2 PointRCNN 训练及结果	28
4.3 本章小结	29
第五章 基于优化的路侧导航增强原理	30
5.1 车辆状态估计的数学描述	30
5.2 基于因子图的状态估计理论	31
5.3 基于因子图的非线性最优化算法	33
5.4 本章小结	34
第六章 全文总结	35
6.1 全文工作总结	35
6.2 创新点	35
6.3 不足与展望	35
参 考 文 献	37
附录 A	42
攻读学位期间学术论文和科研成果目录	44
致 谢	45

第一章 绪论

1.1 研究背景与研究意义

“十三五”期间我国综合交通运输发展取得了显著成效，但与经济社会高质量发展的总体要求相比，仍存在智慧交通发展水平不高、交通基础设施数字化建设亟待加快、交通运输与新技术的融合尚不充分等问题，距离交通强国建设尚有一定差距。随着北斗三号全球卫星导航系统等核心空间基础设施的开通服务、新一代信息技术的发展，构建自主可信、国际领先的时空综合服务体系具备基本条件。2022国家重点研发计划《广域交通可信导航信号与时空服务系统关键技术》应运而生。本课题主要的研究内容响应了该项目的课题3“高精泛源时空感知网络及车路一体化信息融合技术”。

在自动驾驶的落地过程中，车路一体化系统是尤为重要的一环。车路一体化是指利用无线网络，将车端与路端紧密相连，实现车端与路端的信息交换、信息共享。目前，车辆终端导航定位主要依赖于全球卫星导航系统(GNSS)，但该系统受限于卫星相关误差、传播途径相关误差、接收机相关误差等限制，对车辆的定位、测速精度有限。以GPS为例，近年来，其定位精度(水平，圆概率精度，CEP)达到了2~3米，其测速精度达到了0.2m/sec(95%置信度)。路侧终端由于可以预先铺设，其位置，硬件设备都是预先获得精确信息的。且路侧终端由于视野开阔，硬件资源丰富，计算能力强大，可以很好的解决车端对自身定位能力不足的问题。

我们希望在路边设置一种新型路侧单元，利用激光雷达与相机作为传感器对路过车辆进行定位。将定位结果发送回车端，结合车端GNSS定位结果，基于优化方法，预测车辆位置，最终实现导航增强的功能。本课题对卫星导航脆弱场景下的车辆安全具有重要意义。

1.2 国内外研究现状

1.2.1 车路协同研究现状

车路协同技术是通过无线通讯技术将车端、路端有机结合起来，实现交通环境数据信息的交换共享，信息处理，从而可以为车辆提供更精确的感知环境信息。

1950 年代末，通用汽车在新泽西州打造了一条埋入大量通信设备的高速公路。这在当时引起轰动，也是车路协同产业发展的雏形。加州 PATH 计划 (Partner for Advanced Transit and Highways) 成立于 1986 年，由加州交通部和加州大学伯克利分校合作建立，是北美第一个专注于现在称为智能交通系统 (Intelligent Transportation Systems, ITS) 主题的组织。其目标是应用电子、通信与自动化等新兴先进科技，增加高速公路的容量与安全性，减少交通堵塞、空气污染和能源消耗。该计划一直参与自动化公路与自动驾驶车辆的研究、发展与测试之中。被连接的车辆可以与其他车辆或者交通设施，如交通信号灯，进行通信^[2]。

车用无线通信技术 (Vehicle to Everything, V2X) 是将车辆与一切事物相连接的新一代信息通信技术，其中 V 代表车辆,X 代表任何与车交互信息的对象，当前 X 主要包含车、人、交通路侧基础设施和网络。借助人，车，路，云平台的全方位连接与信息交互，V2X 可以提升行驶安全，提高交通效率，提供出行信息服务，支持实现自动驾驶等等。大约在 2016 年前后，美国基于 DSRC (Dedicated Short Range Communications) 的 V2X 协议栈基本制定完毕，并有丰田、通用先后量产支持 DSRC 的汽车。2022 年 12 月 13 日，代表整个美国智能交通行业的十大组织联合发出声明，重申了对快速部署 V2X 的支持，并认为 2023 年是 V2X 部署的关键年。

我国工信部早在 2018 年就明确将 5920MHz-5925MHz 划分给 C-V2X (Cellular Vehicle-to-Everything) 并明确表明 C-V2X 是我国唯一使用的技术路线。2020 年发改委联合工信部等其他 10 个单位发布了《智能汽车创新发展战略》^[3]，提到“结合 5G 商用部署，推动 5G 与车联网协同建设；开展特定区域智能汽车测试运行及示范应用，验证车辆“人-车-路-云”系统协同性等，支持优势地区创建国家车联网先导区”。

1.2.2 激光雷达目标跟踪研究现状

三维激光雷达是一种主动探测式传感器，它向外发出激光束，返回探测到物体的点云数据信息，从而精确地获得探测到物体的距离信息。激光雷达目标跟踪往往是采用基于检测的跟踪范式进行的，其步骤可以大致划分为目标检测，状态预测，数据关联，状态更新共四个部分。其中，目标检测往往用现有的 SOTA (state of the art) 业界最前沿检测器。状态预测往往用平滑与滤波方法，根据当前帧的目标的运动状态与位姿，预测下一帧中目标的运动状态与位姿。数据关联将预测的状态值与检测的目标状态匹配在一起，是最重要的步骤。状态更新则根据数据关联的结果，对当前帧目标的运动状态与位姿进行更新。

在目标检测阶段，国内外基于神经网络贡献了不少优秀的算法与思路，按照思路可以分为基于点云的方法，基于体素的方法，基于截面图的方法。^[4]

基于点云的方法直接输入点云进行目标分类分割任务。由于点云是三维不规律信息，典型的卷积网络不方便对点云直接处理。过去的方法大多是划分为规律的空间网格，或者投影到某一截面，从而利用卷积网络处理。但这样将引入冗余信息，并损害原始数据的自然特征。2017 年，Charles R. Qi 的团队提出了 PointNet 网络^[5]，直接处理点云，考虑到了点云的无序性、空间相关性与旋转不变性。用空间变换网络 (spatial transformer network) 将点云正则化预处理，也即在空间中旋转对齐。用多层感知机 (MLP) 将数据从低维投射到高维，避免池化操作中信息损失过多。再用对称性的 Max 池化函数解决点云输入的无序性，提取出一个高维向量作为全局特征，以此为基础进行后续的分割分类。同年，该团队还提出 PointNet++ 网络^[6]，在 PointNet 的基础上，进一步考虑了点云的局部信息。首先，利用最远点采样法 (FPS) 对整个点云进行局部采样，选出若干个中心点。再为每个中心点在一定半径的局部区域内选择 k 个临近点。最后对每一个局部区域都用 PointNet 提取局部特征，并以此为基础进行后续的分割分类任务。2019 年，香港中文大学的 Shi S 团队提出了 PointRCNN 模型^[7]，该模型是首个两阶段的基于点云的网络。在第一阶段，模型将整个场景分割为前景点与背景点，对前景点特征提取后生成预选框。在第二阶段通过置信度预测与预选框优化获得最终的检测结果。

基于截面图的方法的代表作是 Chen X 等人在 2017 年提出的 MV3DNet

模型^[8]，该模型同时融合了点云的剖视图特征与 RGB 图片特征，分别在点云的俯视图、前视图与 RGB 图像中提取特征。在俯视图特征中计算候选区域后投影到前视图和图像中，经过兴趣区域 (ROI) 整合到同一维度再输入网络中融合。

基于体素的方法的代表作是 2018 年 Y. Zhou 等人提出的 VoxelNet 模型^[9]。首先将点云划分为等间距的 3D 体素，经过了点的随机采样与归一化处理后，又引入了体素特征编码器 (Voxel Featured Encoding, VFE)，每个非空体素都进行了特征提取。然后，这些特征被输入 3D 卷积神经网络 (Convolutional Middle Layers) 进行特征抽象。最后通过提议区域生成网络 (Region Proposal Network, RPN) 进行目标检测。

在数据关联步骤中，常见的算法有全局最近邻法 (Global Nearest Neighbor, GNN)、多假设跟踪算法 (Multiple Hypothesis Tracking, MHT)、联合概率数据关联 (Joint Probability Data Association, JPDA) 等，该方案的目标匹配精度高，但匹配速度慢、计算成本高。2020 年，Weng X 等人提出了 AB3DMOT 算法^{[10] [11]}，该算法将匈牙利匹配算法与卡尔曼滤波估计结合，实现了快速的目标关联。

1.2.3 视觉目标跟踪研究现状

单目视觉跟踪早期主要采用传统的滤波方法，如卡尔曼滤波 (Kalman Filter)、粒子滤波 (Particle Filter)、均值漂移 (Meanshift) 等。近年来流行的研究框架可以分为相关滤波 (Correlation Filter) 框架与孪生网络 (Siamese Network) 框架两个大方向。

2010 年，Bolme 团队首次将相关滤波方法用在了跟踪领域，提出了误差最小平方和滤波器 (Minimum Output Sum of Squared Error filter, MOSSE)^[12]，用最小化均方误差的思路产生滤波器，进而获得跟踪目标的新位置。2012 年，Henriques 等人在 MOSSE 的基础上提出了循环结构检测方法 (Circulant Structure with Kernel)^[13]，一方面修改损失函数为岭回归形式，再引入核函数求解，另一方面引入循环矩阵，达到密集采样与提高运算效率的效果。2014 年，Henriques 等人提出了核相关滤波算法 (Kernel Correlation Filter, KCF)^[14]，该方法在 CSK 算法的基础上，用方向梯度直方图 (Histogram of Oriented Gradient, HOG) 多通道特征替换了原来的单通道灰度特征，并采用了高斯

核函数求解岭回归问题。同一年, Danelljan 等人提出了 DSST(Discriminative Scale Space Tracker)^[15]方法, 在 MOSSE 的基础上, 用两个滤波器分别应对尺度和位置的变化, 同时引入 HOG 特征。

孪生网络框架, 采用两个成对的结构一样的神经网络, 网络之间共享权值、参数等信息。该框架可以接受两个输入, 并对两个输入进行相同的变换, 通过比较输出的两个结果的欧氏距离判断输入之间的相似性。这一思路最早在 1993 年被 J Bromley 应用在美国支票的签名验证场景上^[16]。在目标跟踪问题里, 一个输入可以是初始帧的目标区域, 以此作为模板, 而将后一帧中的候选区域作为第二个输入。孪生网络要做的就是找到两帧间相似度最高的候选区域, 如这一系列的开山之作 SiamFC 框架^[17]。

1.2.4 激光雷达与视觉融合研究现状

传感器信息融合可以分为像素级、特征级和目标级融合^[18]。像素级融合直接融合原始数据, 获取的细节信息最丰富, 因而其准确性和鲁棒性最好。但是像素级融合的前提是信息来自于同类传感器或者传感器拥有同样的量级, 需要传感器之间进行高精度匹配, 因而实时性较低。特征级融合先提取原始数据的特征, 再融合多个特征, 根据目标已有特征对融合特征进行匹配, 获得目标的信息。目标级融合先提取原始数据中的目标信息, 然后融合多个目标信息, 得到最终完整信息。其只对目标信息进行融合, 不受传感器类别的限制, 能够保证实时性。

EagarMOT 是 Aleksandr Kim 等人 2021 年提出的融合框架^[19], 用现成的 2D 与 3D 检测器先得到目标的 2D 与 3D 检测结果, 再用交比 IoU 将同一个目标的两个结果关联在一起得到一个融合实例。在数据关联部分, 该方法采用两阶段进行, 第一阶段对 3D 检测结果和预测结果进行数据关联, 对于没有匹配上的实例和轨迹再进入第二阶段匹配, 利用 2D 检测结果进行数据关联。

1.3 研究内容与研究路线

1.3.1 研究路线

我们希望在路边设置一种新型路侧单元，上面装备了相机、激光雷达等传感器。利用激光雷达的测距测向功能获得深度信息，利用相机的色彩捕捉功能获得图像信息，通过 2d-3d 融合的多目标跟踪方法，获得车辆相对路侧传感器的相对位置姿态。然后，通过路侧与车端的可靠通信，将位姿传回车端，车端利用其作为新增的独立位置约束，结合误差检测与估计理论，提高车辆当前位置估计的精确度，提高车辆导航的可信性。如下图所示：

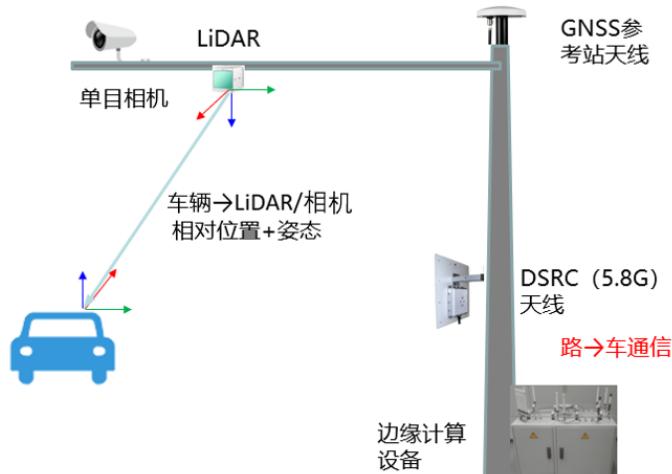


图 1-1 路侧单元架设示意图

但受到实验条件限制，目前所依托的重点研发项目实施仅 6 个月，硬件平台尚未搭建完全，作者也身处国外，所以只能依托公开数据集进行实验。

1.3.2 研究内容

本文主要依托现有公开数据集，搭建了目标跟踪器，训练了二维与三维检测器，并提出了一种基于因子图的误差补偿方法。本文主要内容如下所示：

第一章为绪论。对本课题的研究背景与意义进行论述，回顾了车路一体化概念的发展历史，对众多国内外基于激光雷达和基于视觉的目标跟踪方法进行阐述，分析了国内外基于传感器融合的目标跟踪方法，并确定了本文的研究路线。

第二章为现有公开数据集框架。首先总结了近年来适用于自动驾驶与车路协同的公开数据集。再介绍了 KITTI 数据集和 DAIR-V2X 数据集的传感器配置，标签与标定格式，其中包括了激光雷达与摄像头的标定原理及其坐标转换关系。

第三章为目标跟踪器搭建。首先介绍了 EagerMOT 目标跟踪框架的原理，再介绍了目标跟踪常用的评测方法，最后在 KITTI 数据集上复现了该框架，并与论文结果进行了对比。

第四章为目标检测器训练。首先介绍了 2D 检测器 YOLOv4 的原理，训练环境与网络配置，训练结果。再介绍了 3D 检测器 PointRCNN 的原理，训练环境与网络配置，训练结果。最后对检测器训练结果进行了分析。

第五章为基于优化的路侧导航增强原理，首先介绍了路侧增强的数学模型，再介绍了基于因子图的状态估计理论，各个因子节点的残差计算，最后介绍了因子图理论中的常用的优化方法。

第六章为总结与展望，首先总结本文的主要工作，然后展望后续的研究与改进工作。

1.4 本章小结

本章介绍了研究背景与意义，国内外研究现状，研究内容与研究路线。

第二章 公开数据集框架

2.1 车路协同数据集综述

高质量的路侧单元数据集具有重要的工业价值，可以加速路侧的车辆检测模型在车路协同中的迭代优化，在促进创新的学术研究方面也扮演着重要角色。近年来，基于激光雷达与相机的目标检测与跟踪数据集被发布，例如 KITTI^[20]，ApolloScape^[21]，Waymo^[22]，NuScenes^[23]等。

但是，这些数据集往往基于车载传感器，针对自动驾驶场景。而基于路侧单元视角收集的公开数据集则相对稀少，尤其是包含了 3D 点云数据的数据集，这也为我们的研究工作带来了一定的挑战。自去年以来，一些基于路侧激光雷达与相机的数据集陆续发布，这些较新的数据集被总结如下：^[24]

表 2-1 基于路侧传感器收集的公开数据集

数据集	年份	单位	激光雷达	相机	交通场景
IPS300+[25]	2022	清华大学	2×Robosense Ruby-Lite	2 RGB	城市
DAIR-V2X ^[26]	2022	清华大学 百度公司	1 300-beam LIDAR	1 RGB	城市高速公路
A9-Dataset ^[27]	2022	慕尼黑 工业大学	1 Ouster-OS1 64-beam LiDAR	1 RGB	Autobahn 高速公路

IPS300+：为推动路侧多模态感知研究在合作车辆基础设施系统，Wang 等人^[25]在 2022 年发布了一个双模态数据集。该数据集配备了路侧激光雷达和摄像头，收集场景是一个城市路口，占地面积 3000 平方米，覆盖半径 300 米。两个感知单元 (IPU) 被安装在交叉路口的对角线上，距离用于数据采集的区域地面 5.5 米。每个感知单元由一个 80 束 RoboSense Ruby-Lite LiDAR 和两个 Sensing-SG5 彩色摄像头组成。数据集包括涵盖不同时间的 14198 帧据。被两个 IPU 收集的每一帧点云数据，都被存储为单个 PCD 文件用于标注。每帧都有平均 319.84 个标签，包括行人、骑车人、三轮车、汽车、公共汽车、卡车和工程车辆等。

DAIR-V2X：为了加速车路协同自动驾驶的计算机视觉研究和创新，Yu

等人^[26]于 2022 年发布了 DAIR-V2X 数据集。该数据集采集自北京高级别自动驾驶示范区 10 公里的城市道路、10 公里的高速公路和 28 个路口。在 28 个路口各部署了四对 300 束路侧激光雷达和高分辨率摄像头。DAIR-V2X-I 是 DAIR-V2X 的子集，专用于路侧协同感知，包含 10084 帧图像，分别联合标注了图像和路边激光雷达点云数据。注释器详尽地标记了每个图像和点云帧中的 10 个对象类别中的每一个目标，包括不同的车辆、行人和骑车人。

A9-Dataset: 2022 年，Christian 等人^[27]展示了基于德国慕尼黑附近 3 公里长 Providentia++ 试验场路边传感器的 A9 数据集。传感器包括摄像头、雷达和 64 束 Ouster 激光雷达。它们被安装在龙门桥和桅杆上，提供道路景观。数据集提供标记图像和多个路段的激光雷达点云和白天 A9 高速公路上密集交通的不同角度记录。版本 R0 由 1098 个带标签的帧和 14,459 个带标签的 3D 对象组成，包括汽车、拖车、卡车、货车、行人、公共汽车、摩托车、自行车等九类对象。

以上数据集中，IPS300+ 数据集需要申请授权使用，作者较晚才从数据集作者处取得授权，因此没有来得及用于实验。DAIR-V2X 是一个目标跟踪数据集，其帧与帧之间没有时序联系，只能作为检测器的训练数据，不能验证跟踪器的性能。A9-Dataset 虽然为连续采集的数据集，但是激光雷达与相机采集的数据在时间上是分开的，并非对同一场景在同一时间段内采集，无法进行数据融合步骤，且作者所在地区无法注册数据集官网账号。

因此，我们暂时在 DAIR-V2X 数据集上训练检测器，在配套生态健全的 KITTI 数据集上验证跟踪器性能。

2.2 数据集标注与标定

2.2.1 KITTI 数据集介绍

KITTI^[1] (Karlsruhe Institute of Technology and Toyota Technological Institute) 是用于移动机器人和自动驾驶的最受欢迎的数据集之一。该数据集的主要目的是推动以自动驾驶为目标的计算机视觉和机器人算法的发展。它包含多种传感器模式记录的数小时交通场景，包括 2 个高分辨率 RGB、2 个灰度立体相机和激光雷达，高精度 IMU/GPS 导航系统。其示意图如2-2所示

示，由于我们仅关注激光雷达与相机传感器，所以仅画出这两个传感器的坐标系。

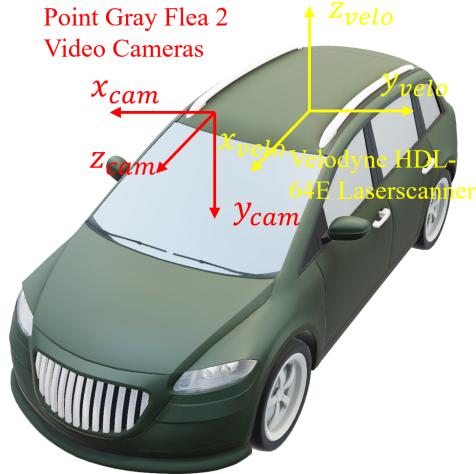


图 2-2 KITTI 数据集采集平台示意图

KITTI 数据集包括了目标检测数据集与目标跟踪数据集，前者为采集时间离散的目标数据集，后者包含了连续的目标信息。我们选择了后者作为跟踪框架的验证集。KITTI 数据集分为训练集与测试集。区别在于前者除了传感器数据与标定文件外还提供标注文件，即真值，用于检测器，跟踪器的训练，后者仅提供传感器数据与标定文件，其真值储存在官方服务器中。训练集包含了 20 个时间序列，测试集包含了 28 个时间序列。每个序列将每一帧的图片与点云信息分别储存在 `image_01`, `image_02` 与 `velodyne` 文件夹中。其中 `image_01` 为左侧彩色相机的采集数据，`image_02` 为右侧彩色相机的采集数据，二者结合可以进行双目视觉的研究。如图2-3所示为其中一帧的点云数据与图像。



图 2-3 KITTI 点云与图像可视化结果

2.2.2 KITTI 数据集标注格式

KITTI 数据集将每一帧的标注以 txt 形式储存在了 label 文件夹下，每行为一个目标信息。每一行以空格为分隔符，共 15 列，表示信息如表2-2所示：

表 2-2 KITTI 数据集标注格式

列数	意义	备注
1	类型	共有 Car, Van, Truck 等 8 种类型
1	截断程度	从 0 至 1，代表了目标被图片边框的截断程度
1	遮挡程度	取整数集 (0,1,2,3)，标志了目标被其他物体遮挡的程度
1	alpha	目标在激光雷达坐标系下位置向量与 x 轴的夹角
4	2D 包围框	目标在图像坐标系下的包围框对角坐标 (x_1, y_1, x_2, y_2)
3	3D 尺寸	目标的三维包围框的大小，即高度，宽度，长度
3	3D 位置信息	目标在激光雷达坐标系下的坐标 (x, y, z)
1	rotation_y	目标自身朝向与激光雷达坐标系 x 轴夹角

2.2.3 KITTI 数据集标定格式及坐标转换

数据集的标定包括了相机内参标定，外参标定，雷达-相机标定等部分。

如图2-4所示，在 KITTI 数据集中一共定义了 4 个坐标系，激光雷达坐标系 (velodyne coordinate)，相机坐标系 (camera coordinate)，修正的相机坐标系 (rectified camera coordinate)，图像坐标系 (image coordinate)。

假设空间中一个点在激光雷达坐标系中的坐标为

$$X_{\text{velo}} = (x_{\text{velo}}, y_{\text{velo}}, z_{\text{velo}}, 1)^T \quad (2.1)$$

在相机坐标系坐标下的坐标为

$$X_{\text{cam}} = Tr_{\text{velo_to_cam}} X_{\text{velo}} \quad (2.2)$$

其中， $Tr_{\text{velo_to_cam}} = [R|t]$ ， R 为坐标系旋转矩阵， t 为原点间的平移向量。

KITTI 数据集中，四个相机的图像平面原点的排列与激光雷达坐标系的 y 轴是平行的，但平面之间并非共面，为了让四个相机的图像平面共面，

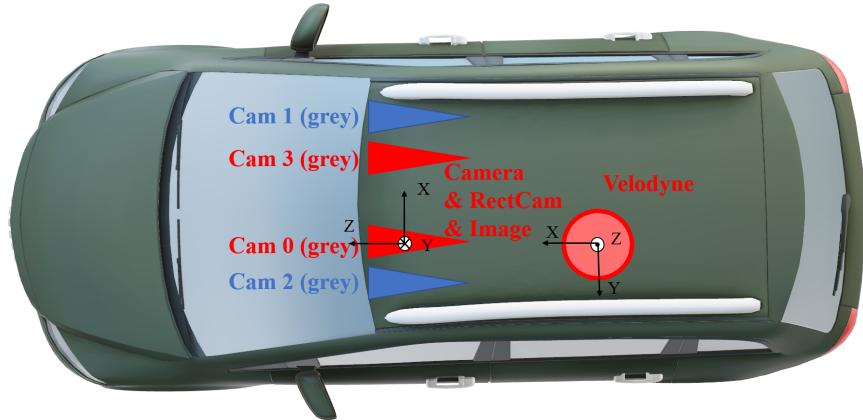


图 2-4 KITTI 数据集坐标系示意图

需要将每个相机坐标系进行旋转，旋转矩阵为 $R_{\text{rect}}^{(i)}$ ，其中 i 为四个相机的编号。旋转后，我们得到了修正的相机坐标系，在此坐标系中，该点坐标标记为 $X_{\text{rect_cam}}$ ，满足

$$X_{\text{rect_cam}} = R_{\text{rect}}^{(i)} X_{\text{cam}} \quad (2.3)$$

将 0 号修正相机坐标系中的三维坐标系投影至图像坐标系 Y 关系为

$$Y = P_{\text{rect}}^{(i)} X_{\text{rect_cam}} \quad (2.4)$$

其中 $P_{\text{rect}}^{(i)}$ 为第 i 个相机的投影矩阵

$$P_{\text{rect}}^{(i)} = \begin{pmatrix} f_u^{(i)} & 0 & c_u^{(i)} & -f_u^{(i)} b_x^{(i)} \\ 0 & f_v^{(i)} & c_v^{(i)} & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad (2.5)$$

投影矩阵 P_{rect} 是相机内参 K 与相机外参 $[R|t]$ 的乘积。

$$P_{\text{rect}} = K[R|t] \quad (2.6)$$

为了简洁起便，此处省略了相机编号 i 。

相机内参 K 是相机坐标系下一个点的坐标到相机图像坐标系的变换矩阵，包含了相机像素点的长宽 u, v ，相机坐标系原点在图像坐标系下的坐标 (u_0, v_0) ，相机焦距 f 等信息。

$$K = \begin{pmatrix} f_u^{(i)} & 0 & c_u^{(i)} & -f_u^{(i)}b_x^{(i)} \\ 0 & f_v^{(i)} & c_v^{(i)} & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad (2.7)$$

式中

$$f_u^{(i)} = \frac{f}{u} \quad f_v^{(i)} = \frac{f}{v} \quad c_u^{(i)} = u_0 \quad c_v^{(i)} = v_0 \quad (2.8)$$

相机外参 $[R|t]$ 是世界坐标系到相机坐标系的变换矩阵。在 KITTI 数据集中，由于相机坐标系已作过修正，外参中旋转矩阵 $R = I$, I 为单位矩阵。由于 4 个相机沿着相机坐标系 x 轴排列，平移向量 t 仅在 x 轴方向有分量 $-b_x^{(i)}$ ，即

$$t = (b_x^{(i)}, 0, 0)^T \quad (2.9)$$

结合式(2.6)可以发现第一列最右侧为 $-f_u^{(i)}b_x^{(i)}$ 。注意(2.6)中的相机内参 K 由 3×3 矩阵被补齐为了 4×4 矩阵， $K(4,4) = 1$ ，其余为 0。

2.2.4 DAIR-V2X 标注与标定

DAIR-V2X 数据集在路侧单元架设了 1 个 RGB 相机与 1 个激光雷达，如图2-5所示。

现实中激光雷达和相机与地面均有大约 11° 的夹角，这导致目标激光雷达坐标系中的高度与现实中的高度不对应，给研究带来了一定的麻烦。数据集为方便起见，沿着平行于地面的方向建立了虚拟的 x 轴和 y 轴，并以此建立了虚拟激光雷达坐标系，方便研究。

但论文作者并未将相机坐标系 $x-y$ 平面旋转至与地面平行，这使得相机坐标系与激光雷达坐标系之间不仅仅只有 $x-y$ 平面上的旋转。这一点与 KITTI 的基本设置不符合。由于主流的可视化工具，目标检测器，目标跟踪器不少在 KITTI 上开发，这一个小小的区别会导致可视化结果，检测器训

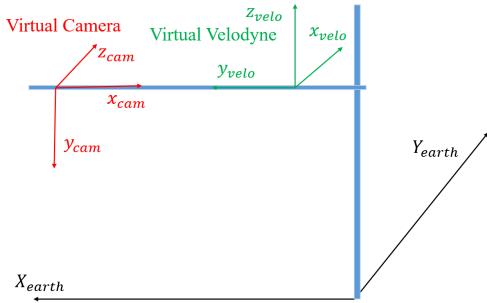


图 2-5 DAIR-V2X 数据集坐标系示意图

练习出现异常。如图2-6所示，直接在DAIR-V2X数据集进行可视化或检测器训练时，检测器接受的3D包围框是与真值有角度偏差的。



图 2-6 采用原始相机坐标系产生的偏角

因此我们利用标定文件，计算出相机坐标系 $x - y$ 平面与地面的夹角，将其旋转至水平，重新生成标注与标定文件。

2.3 本章小结

本章总结了近年来适用于自动驾驶与车路协同的公开数据集。再介绍了KITTI数据集和DAIR-V2X数据集的传感器配置，标签与标定格式，其中包括了激光雷达与摄像头的标定原理及其坐标转换关系。

第三章 目标跟踪器搭建

为了实现路侧单元对车辆的对目标跟踪任务，我们采用了 EagerMOT 框架。由于缺少合适的路侧目标跟踪数据集，我们在 KITTI 数据集上对该框架进行了性能测试。

3.1 EagerMOT 原理

EagerMOT 框架结合了互补的 2D 和 3D（例如 LiDAR）目标信息，这些信息是从预训练的物体检测器中获得的。框架的总体概述如图3-7所示。作为每一帧的输入，首先从预先训练好的检测器中获得一组 3D 边框检测 ${}^{3d}D_t$ 和一组 2D 边框检测 ${}^{2d}D_t$ 。然后，融合模块 (i) 将来自 2D 和 3D 的检测结果中相同的对象关联起来，(ii) 两阶段数据关联模块进行目标关联，并且该关联基于当前可用的检测结果信息（例如完整的 2D+3D 信息，仅含 2D 信息或仅含 3D 信息），然后更新跟踪状态和，(iii) 最后采用简单的跟踪管理机制初始化或终止轨迹。

这种跟踪方法允许所有检测到的对象与轨迹相关联，即使它们仅在图像域或 3D 传感器中被检测到。这样，当目标被短时间遮挡后可以被恢复。并且当检测器之一漏检时，目标也能保持原有的 3D 位置。而且重要的是，在目标进入 3D 传感器感知范围之前，这种方法就已经可以在图像域中跟踪远处的物体。一旦目标进入感应范围，我们对每个轨迹可以顺利地初始化一个 3D 运动模型。

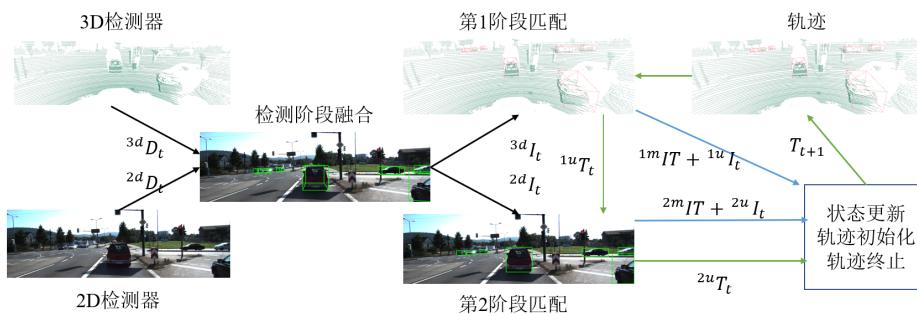


图 3-7 EagerMOT 框架流程图

3.1.1 检测器融合

获得来自相机的 2D 视频输入与来自激光雷达 3D 信息流的检测结果 ${}^{2d}D_t$ 与 ${}^{3d}D_t$ 。设这些来自多传感器的目标的融合结果为 $I_t = \{I_t^0, I_t^1, \dots, I_t^n\}$ ，其中，每一个实例 I_t^i 都对应于现实中的一个目标，它可以同时包含 2D 与 3D 的检测信息，也可以仅包含其中一种检测信息。特殊地，定义 ${}^{3D}I_t$ 为包含了 3D 检测信息的实例的集合，无论该实例是否包含 2D 检测信息。类似地，定义 ${}^{2D}I_t$ 为包含了 2D 检测信息的实例的集合，无论该实例是否包含 3D 检测信息。

对来自多传感器的检测结果进行数据关联，采用贪心算法，将 3D 检测结果投影到 2D 图像域上，计算二者的交比 (Intersection of Union)

$$\text{Intersection of Union} = \frac{\text{Area}_{3D} \cap \text{Area}_{2D}}{\text{Area}_{3D} \cup \text{Area}_{2D}} \quad (3.1)$$

将所有可能的检测配对按照交比进行排序，按照交比从高到低的顺序，当 2D 与 3D 检测结果均尚未配对成功，且交比大于预先设置的阈值 (threshold) θ_{fusion} 时，产生融合实例 ${}^{both}I_t^i$ 。以此得到融合的检测实例集合 ${}^{both}I_t$ 。其中，每个实例包含了来自 2D 检测器的 2D 边界框 (bounding box) 与来自 3D 检测器的 3D 位置与 3D 边界框，因此也属于两个单模态检测实例集合 ${}^{3D}I_t$ 与 ${}^{2D}I_t$ ，即 ${}^{both}I_t \in {}^{3D}I_t$ 且 ${}^{both}I_t \in {}^{2D}I_t$

将剩下的未匹配的检测结果 ${}^{3D}I_t^i$ 与 ${}^{2D}I_t^i$ 作为部分观测结果 (partial observation) 加入检测实例集合中与对应的中。

上面的思路虽然简单，但是过去的经验证明了它的鲁棒性。

3.1.2 数据关联

数据关联分为两阶段进行，在第一阶段进行 3D 目标的数据关联，将 3D 检测结果与已有的 3D 轨迹匹配。在第二阶段，在 2D 图像域进行目标关联，弥补目标遮挡时 3D 检测器漏检的场景。

轨迹参数同时包含 2D 与 3D 信息的轨迹为 ${}^{both}I_t$ ，包含 3D 信息的轨迹为，包含 2D 信息的轨迹为 ${}^{2D}I_t$ 。它们有如下关系： ${}^{both}I_t \in {}^{3D}I_t$ 且 ${}^{both}I_t \in {}^{2D}I_t$ 。其中， ${}^{3D}I_t$ 额外设置了一个速度参数 v 。

一阶段匹配首先用 3D 信息将 3D 检测结果 ${}^{3D}I_t$ 与已有的 3D 轨迹 ${}^{3D}T_t$

匹配，计算 ${}^3D I_t$ 与 ${}^3D T_t$ 间的缩放距离 (scaled distance)。缩放距离如下定义

$$d(B^i, B^j) = ||B_\rho^i - B_\rho^j|| * \alpha(B^i, B^j) \quad (3.2)$$

$$\alpha(B^i, B^j) = 2 - \cos\langle B_\gamma^i, B_\gamma^j \rangle \quad (3.3)$$

其中， $B_\rho^i = [x, y, z, h, w, l]$ 是包含了 3D 位置信息与包围框信息的向量，表示了包围框相对于地面纵轴的方向。这种缩放距离与传统的欧式距离或马氏距离相比在实验结果中更加鲁棒。

如同3.1.1中的贪心算法，按照缩放距离从小到大对“目标-轨迹”对进行排序，将匹配最佳且缩放距离小于阈值的“目标-轨迹”对作为成功的配对。输出配对的集合 ${}^{1m}IT_t = I_t^i, T_t^j, \dots$ 其余未匹配的结果标记为 ${}^{1u}I_t$ 与。此时所有的 3D 结果，或者成功配对，或者未配对且不参与后续匹配。

二阶段匹配在图像域用仅有 2D 信息的实例 ${}^{2D}I_t \setminus {}^{both}I_t$ 与剩下的 2D 轨迹 ${}^{1u}T_t \cap {}^{2D}T_t$ 匹配。匹配方式与3.1.1类似，计算目标与轨迹的 2D 交比，根据交比及阈值 θ_{2D} 用贪心算法匹配。值得注意的是，轨迹在包含 3D 信息时，将 3D 预测框投影到 2D 图像域作为边界框计算，仅在不含 3D 信息时，将最后一次记录的 2D 边界框参与计算。此阶段输出的结果为匹配对 ${}^{2m}IT_t = \{ \{I_t^i, T_t^j\}, \dots \}$

状态更新 2D 边界框直接采用匹配目标的 2D 边界框，如果该目标包含 2D 信息。3D 状态则被建模为多元高斯函数，使用线性卡尔曼滤波更新。这一步与 AB3DMOT 框架^[11]完全一致。但当匹配的目标不包含 3D 信息时，轨迹只通过卡尔曼滤波进行预测步骤以延拓状态。

3.1.3 轨迹生命周期管理

这一步与 AB3DMOT 类似，当一个轨迹在 Age_{max} 帧没有被匹配后，将被舍弃，未被匹配的新目标会被当作新轨迹加入后续匹配中。但论文考虑到 2D 检测结果往往比 3D 检测结果更可信，只有当一个轨迹连续 Age_{2D} 帧与 2D 信息匹配成功后，才会被认为是一个可信的新轨迹。

3.2 评价指标

衡量一个目标跟踪框架好坏有多种指标,早期常用的是CLEAR MOT^[28]指标。近些年 HOTA^[29]指标成为了各个目标跟踪比赛的主流指标。

3.2.1 CLEAR MOT 指标

理想追踪器有 3 个需求: (1) 在所有时刻找出正确的物体数量, (2) 尽可能精确地估计出每个物体的位置, (3) 指派连续的 ID, 即使有遮挡。

假设在跟踪场景中, 某一帧的目标位置信息为 $\{o_1, o_2, \dots, o_n\}$, 而跟踪算法给出的目标假设信息 (hypothesis) 为 $\{h_1, h_2, \dots, h_m\}$ 。我们需要 (1) 在 h 和 o 之间建立最大可能性连接, 这一点往往是给定阈值 α , 当 h 与 o 之间的交比超过 α 时, 设定为匹配成功。(2) 对于每一个连接, 使用位置估计算误差 (3) 积累所有的连接误差: 没有设定假设 h 的目标 o (Miss), 没有对于目标 o 的假设 h (False Positive); 同一个目标 ID 改变 (Mismatch)。

最终论文导出了两大指标: MOTA(Multiple Object Tracking Accuracy) 与 MOTP(Multiple Object Tracking Precision)

MOTA 用于衡量跟踪器的准确度:

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad (3.4)$$

其中 m_t , fp_t 与 mme_t 分别为丢失 (miss), 错检 (false positive) 与 ID 切换 (mismatch) 的数量。

MOTP 用于衡量跟踪器的精确度:

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad (3.5)$$

其中 d^i 为匹配的 $\{o, h\}$ 对之间的距离, c_t 匹配的 $\{o, h\}$ 对的个数。

由于目前流行的还是“检测-跟踪”范式, 即检测器输出检测结果, 跟踪器只需指出前后两帧间目标的对应关系。因此, 跟踪器的精确度主要取决于检测器的精度, MOTP 不太能反映跟踪器本身的性能。所以 MOTA 被广泛采用为首要指标。

3.2.2 HOTA 指标

MOTA 虽然在长时间里被用作跟踪器性能好坏的指标，但是它过于强调检测器的性能。一种极端情况是，检测的性能非常优秀，但是所有检测到的目标不作跟踪，而是全部分配一个相同的 track id，此时的 MOTA 会非常高，因为 $mme = 0$ 。

HOTA 指标则综合考虑了不同匹配阈值 α 下的检测精度与关联精度

$$\text{HOTA} = \int_0^1 \text{HOTA}_\alpha d\alpha \approx \frac{1}{19} \sum_{\alpha \in \left\{ 0.05, 0.1, \dots, 0.9, 0.95 \right\}} \text{HOTA}_\alpha \quad (3.6)$$

其中 HOTA_α 为阈值 α 下的 HOTA 指标

$$\text{HOTA}_\alpha = \sqrt{\frac{\sum_{c \in TP} \mathcal{A}(c)}{|TP| + |FN| + |FP|}} \quad (3.7)$$

$$= \sqrt{\text{DetA}_\alpha \cdot \text{AssA}_\alpha} \quad (3.8)$$

其中 DetA_α 与 AssA_α 分别衡量跟踪器的检测精度与匹配准度

$$\text{DetA}_\alpha = \frac{|TP|}{|TP| + |FN| + |FP|} \quad (3.9)$$

$$\text{AssA}_\alpha = \frac{1}{|TP|} \sum_{c \in TP} \mathcal{A}(c) \quad (3.10)$$

$$\mathcal{A}(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|} \quad (3.11)$$

TP, FP, FN 分别为正确的检测 (True Positive), 错检 (False Positive) 和漏检 (False Negative) 的数量。对于每一个正确的检测 c , 类似有正确关联 TPAs (True Positive Associations), 错误关联 FNAs (False Negative Associations) and 漏关联 FPAs (False Positive Associations) 等指标。

3.3 实验结果

我们下载了论文中提到的检测器与数据集，如表3-3所示。

表 3-3 EagerMOT 数据集与检测器设置

框架	框架名称
数据集	KITTI, NuScenes
2D 检测器	Track-RCNN, MOTSFusion, RCC
3D 检测器	Point-RCNN, Point-GNN

注意到，论文中提到的检测器，其实并非训练好的模型或者权重文件，而是检测器对 KITTI 数据集评测产生的检测结果。这意味着如果我们想要使用这些检测器用于其他数据集，需要下载模型重新进行训练。

论文采用了 CLEAR-MOT, HOTA, AB3DMOT 等指标作为评测标准，在如表3-4所示的基准测试 (benchmark) 中进行了评测：

表 3-4 EagerMOT 论文评测的数据集

数据集	划分 (split)
KITTI 2D MOT	测试集 & 评测集
KITTI 3D MOT	评测集
NuScences 3D MOT	测试集 & 评测集

我们用 MOTSFusion+RCC 作为 2D 检测器，Point-GNN 作为 3D 检测器，HOTA 作为评价指标在 KITTI 数据集的评测集上对 car 类型目标进行了评测。评测结果如图3-5

表 3-5 EagerMOT 评测结果

KITTI Sequence	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr
0002	53.711	49.066	59.081	50.789	86.084	61.606	88.639
0006	80.969	84.762	77.656	88.168	89.059	79.912	91.349
0007	83.244	81.566	85.05	90.391	83.947	91.531	87.421
0008	76.27	81.024	71.936	84.769	85.533	74.265	87.491
0010	81.261	78.375	84.384	83.929	87.868	86.242	92.934
0013	39.585	17.872	87.752	89.895	17.979	89.895	89.895

续表 3-5

KITTI Sequence	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr
0014	77.485	75.13	80.428	81.38	83.41	85.91	87.112
0016	81.565	84.502	79.052	87.761	88.077	80.403	91.35
0018	86.987	85.657	88.485	90.081	88.63	91.454	91.406
COMBINED	78.529	76.839	80.506	82.917	85.081	84.239	89.777

由于 test set 的真值 (ground truth) 储存在 KITTI 官网的 server 中，不对外开放，我们无法直接与论文中的结果进行比较，但在 val set 中得到的评测结果与论文在 test set 中得到的结果 (HOTA 74.39, DetA 75.27, AssA 74.16) 差距在合理的范围内，甚至比论文中的结果更高。我们用论文提供的可视化代码生成了视频文件，大部分轨迹连续，且目标切换频率较少。验证了该算法在 KITTI 数据集上的有效性。

3.4 本章小节

本章首先介绍了 EagerMOT 目标跟踪框架的原理，再介绍了目标跟踪常用的评测方法，最后在 KITTI 数据集上复现了该框架，并与论文结果进行了对比。

第四章 目标检测器训练

如第三章所言，目标跟踪器的精度很大程度上取决于目标检测器的精度。而目前的目标检测器往往针对某一个比赛如 KITTI Object Detection, VOC (The PASCAL Visual Object Classes Challenge), COCO (Common Objects in Context) 等等。它们虽然能在某一个数据集上表现优异，但是迁移到其他数据集时性能位置。因此，将这些目标检测框架在合适的路侧数据集上重新进行训练是一件必要的事情。我们在 DAIR-V2X 数据集上分别训练了 2D 和 3D 的目标检测器。

4.1 2D 目标检测器

2D 目标检测框架在 2010 年前主要是基于一些人工定义的图像特征，例如 HOG(histogram of gradient), DPM(Deformable Parts Models)^[30] 等。在 2012 年 AlexNet^[31] 出现后，基于深度学习的图像检测方案成为主流。其中又可以分为两大分支。一支为以 YOLO 系列 (2016-2022)^{[32] [33]}, SSD(2016)^[34] 等为代表的一阶段检测框架。它们是端到端 (end to end) 的算法，将整张图片作为输入，直接识别目标的类别与边界框。其特点是检测速度快，精确度一般，训练输入一般为 2D 边框。另一只是以 Fast RCNN (2015)^[35], Faster RCNN(2015)^[36], MaskRCNN(2017)^[37] 等两阶段的检测框架。首先使用区域提议网络 (RPN) 生成稀疏的候选框，然后由检测网络 RCNN 识别候选目标类别与 3D 边界框。其特点是精确度高，但由于区域提议网络计算量大，检测速度较慢。从 MaskRCNN 后往往依赖于像素级标注，或掩码 (Mask) 信息，如图 4-8 所示。

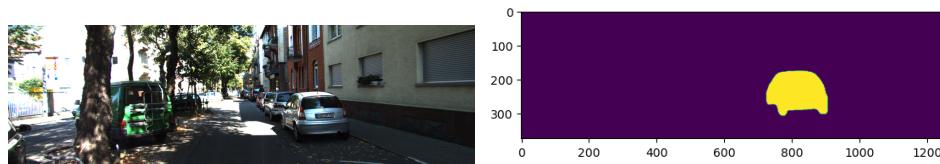


图 4-8 KITTI 数据集的掩码标注

由于我们希望较高的检测速度，以应对在线跟踪场景，且 DAIR 数据集不含掩码标注，不适合现有主流二阶段框架，我们选择了 YOLO 系列框架。

为方便起见，我们采用了应用生态较为成熟的 YOLOv4 框架，没有使用最新的检测器。

4.1.1 YOLOv4 网络模型

YOLOv4 在 YOLOv3 的基础上，在数据增强、网络结构、网络训练方式、损失函数计算方式等方面进行改进。YOLOv4 网络模型的主要贡献为：

1. 提出了一种高效有力的目标检测模型，它可以使用一块 1080Ti 显卡或 2080Ti 显卡训练出快速且精确的目标检测器。
2. 论文确认了检测器训练阶段，最先进的 Bag of Specials 和 Bag of Freebies 方法的效果。
3. 论文修改了最先进的方法并且使它们在单 GPU 训练时更加高效合适，包括了 CBN, PAN, SAM 等

其中，Bag of Freebies 指不增加模型复杂度，也不增加推理的计算量的提高模型的准确度的训练方法技巧，Bag of Specials 指增加少许模型复杂度或计算量，但可以显著提高模型的准确度的训练技巧。YOLOv4 使用大量此类的技巧，提高了模型的性能。

目标检测器无论是一阶段还是二阶段，往往都可以划分为如图4-9结构，由输入 (Input)，主干网络 (Backbone)，颈部网络 (Neck)，检测头 (Head) 与预测层 (Prediction) 组成。

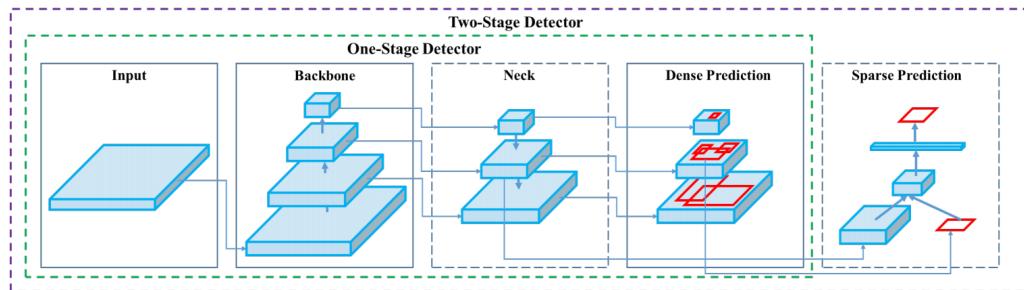


图 4-9 目标检测框架流程图^[33]

YOLOv4 的网络结构如表4-6所示

表 4-6 YOLOv4 网络模型

Module	Method
Backbone	CSPDarknet53
Neck	SPP, PAN
Head	YOLOv3

输入层 Input YOLOv4 网络模型在输入时进行了许多创新改进，例如 Mosaic 数据增强、cmBN、SAT 自对抗训练等。

主干网络 Backbone YOLOv4 采用的主干网络为 CSPDarknet53，CSP-Darknet53 是在 Yolov3 主干网络 Darknet53 的基础上，借鉴 2019 年 CSP-Net(Cross Stage Partial Network) 的经验，产生的 Backbone 结构，其中包含了 1 个 CBM 和 5 个 CSP 模块。它解决了网络优化的梯度信息重复问题，既保证了推理速度和准确度，又减小了模型尺寸。YOLOv4 还将 DarknetConv2D 的激活函数由 LeakyReLU 修改成了 Mish。

颈部网络 Neck 空间金字塔池化层 SPP 全称 Spatial Pyramid Pooling，可以用来解决不同尺寸的特征图如何进入全连接层，增加主干网络的感受野，显著的分离最重要的上下文特征，补充语义信息。PAN(Path Aggregation Network) 替代了 YOLOv3 的 FPN(Feature Pyramid Networks)，加强了信息在特征金字塔中，自下而上的路径，提高了特征提取的能力。

检测头 Head 主体还是 YOLOv3 算法的检测头，但是把损失函数修改为了 CIOU(Complete IoU)。CIOU 将目标与 anchor 之间的距离，重叠率、尺度以及惩罚项都考虑进去，使得目标框回归变得更加稳定，不会像 IoU 和 GIoU(Generalized IoU) 一样出现训练过程中发散等问题。而惩罚因子把预测框长宽比拟合目标框的长宽比考虑进去。

4.1.2 YOLOv4 训练及结果

YOLOv4 有诸多复现版本，我们采用了 AlexeyAB 的版本在如表4-7所示的环境中对 DAIR-V2X 进行了训练。训练集由 5042 张图片构成，评测集由 2016 张图片构成。类别只有一种，即 car。

训练时损失函数下降图如图4-10所示。可以看出下降曲线相对平缓，在大约 2000 步迭代后已经趋于收敛。IOU 阈值为 50% 时，平均精度 mAP(mean

表 4-7 YOLOv4 训练环境

配置	OS	GPU	CPU	CUDA	CUDNN	OpenCV
版本	Windows 10	Nvidia Geforce RTX 2060m	R7 4800H	12.1	8.x	4.7

of Average Precision) 达到了 96.25% 以上。

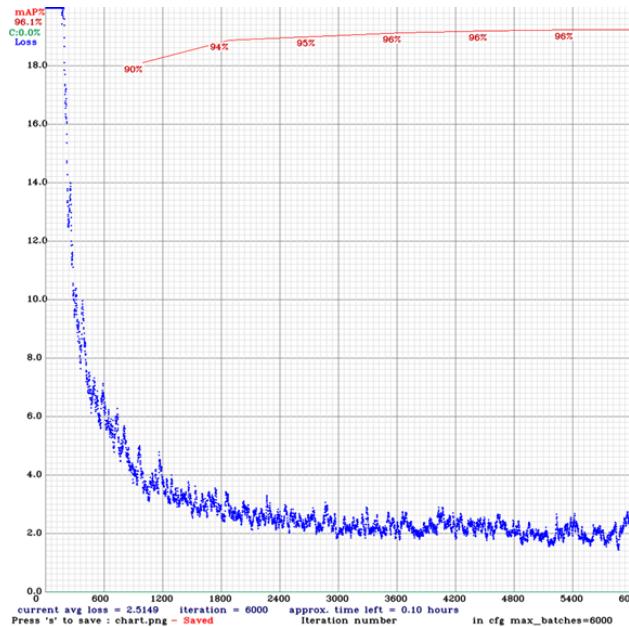


图 4-10 YOLOv4 训练 loss 曲线

评测结果如表4-8所示。其中，精确率定义为被正确预测的正样本数量占预测结果为正样本数量的比例，召回率定义为：被正确预测的正样本数量占真实结果为正样本数量的比例。目标检测过程中会生成大量预测框，每个预测框会有一个置信度属性。设置不同的置信度阈值，可以得到不同的输出。置信度越低，输出的预测框数量越大，召回率越高，但精确度越低，反之相反。如果设置不同的置信度阈值，就可以得到不同的召回率和准确率，将它们作为纵坐标与横坐标可以绘制一条曲线。曲线下方的区域面积定义为平均准确率 AP(Average Precision)。所有类别的 AP 值的平均值定义为 mAP。

如图4-11为随机选取的几张图片的检测结果。在制作标注文件时，我们

表 4-8 YOLOv4 测试结果

检测器	精确率 precision	召回率 Recall	mAP@0.5	平均 IoU
YOLOv4	0.87	0.94	0.9625	0.77

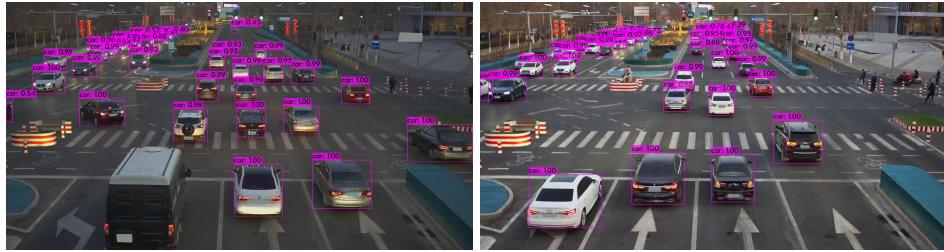


图 4-11 YOLOv4 在评测集上的检测结果

没有将类型 Van 和 Bus 合并至类型 Car 中，因此部分大车型漏检严重，但常见的小轿车准确度较高。

4.2 3D 目标检测器

4.2.1 PointRCNN 网络模型

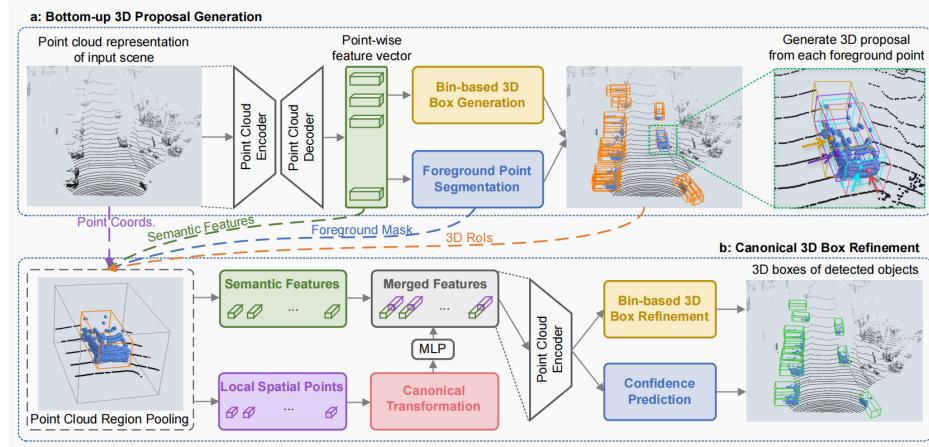
PointRCNN^[7]是一个两阶段的点云目标检测算法，其可以直接对原始点云数据进行目标检测，无需对其投影至某一视角或对其进行体素化。该算法分为两大阶段：（1）基于点云分割的提议区域生成（2）标准坐标系下的边界框优化。整体示意图如图所示

第一阶段：基于点云分割的提议区域生成

首先，为了获得原始点云的逐点判别特征，算法采用了 PointNet++^[6]作为骨干网络。其中，其他的网络也可以替代这个位置，例如 VoxelNet^[9]等。

其次，算法对点云进行了前景点云分割处理。前景点提供了预测目标位置和方向的丰富信息。通过学习分割前景点，点云网络捕获了上下文信息以进行准确的逐点预测，这也有利于 3D 边界框生成。

最后，算法还添加了一个边界框回归头，用来同步地生成 3D 提议区域。为了约束生成的 3D 边界框，相对精准地预测目标的位置信息，每个前景点的周围区域，被沿着 X 轴与 Y 轴划分为了一系列离散的”bins”。算法

图 4-12 PointRCNN 网络模型示意图^[7]

对前景点的 X 轴与 Y 轴设置了一个搜索范围 S , 每个范围都被划分为了长宽均为 δ 的 bins, 从而表示不同的目标中心坐标 (x, z) 。

点云池化处理

在获得第一阶段生成的候选框后, 算法对候选框进行了逐点池化处理, 通过分割掩码, 进行候选框内部前景点与背景点的分类。其中不包含点的候选框会被清除。

第二阶段: 标准坐标系下的边界框优化

首先算法将池化后的点的坐标通过正规坐标变换, 变换到标准坐标系, 使得 (1) 坐标原点位于边界框中心, (2) 坐标系的 $X - Y$ 平面与地面平行, 且 X 轴与边界框朝向一致, (3) Y 轴与激光雷达坐标系的 Y 轴一致。

然后, 算法学习了候选框优化特征。并增加了深度信息

$$d^{(p)} = \sqrt{(x^{(p)})^2 + (y^{(p)})^2 + (z^{(p)})^2}$$

以补偿坐标变化下的位置信息损失。在对候选框优化时, 算法采用了如下的损失函数, 完成回归。

$$\begin{aligned} \mathcal{L}_{\text{refine}} &= \frac{1}{\|\mathcal{B}\|} \sum_{i \in \mathcal{B}} \mathcal{F}_{\text{cls}}(\text{prob}_i, \text{label}_i) \\ &+ \frac{1}{\|\mathcal{B}_{\text{pos}}\|} \sum_{i \in \mathcal{B}_{\text{pos}}} \left(\tilde{\mathcal{L}}_{\text{bin}}^{(i)} + \tilde{\mathcal{L}}_{\text{res}}^{(i)} \right) \end{aligned} \quad (4.1)$$

4.2.2 PointRCNN 训练及结果

我们在 DAIR-V2X 数据集上应用了 PointRCNN 算法。设置 batchsize 为 4, epoch 为 200, 训练了 RPN 网络, 设置 batchsize 为 2, epoch 为 70, 训练了 RCNN 网络。然后在评测集上进行了测试, 实验结果如表4-9所示。

表 4-9 PointRCNN 测试结果

难度	bbox	bev	3d	aos
简单 (Easy)	81.5738	80.1970	69.2464	77.56
中等 (Moderate)	80.4673	78.2905	65.6314	76.28
困难 (Hard)	79.4982	78.2205	65.5195	77.14

其中难度是数据集根据车辆被遮挡, 截断的程度设置的检测难易程度。bbox 是指 2D 检测框的准确度, bev 指鸟瞰图下的检测框准确度, 3d 为 3D 检测框的准确度, aos 为检测目标的旋转角的准确度。值得注意的是, 表4-9中所有结果均是 IoU 为 0.7 的设置下得到的。

如图4-13所示为序列 000018 的检测结果。其中左图为点云可视化结果, 红色边界框为标注数据, 即真值, 蓝色边界框为预测结果。右图为标注信息投影至图像中的结果, 不包含目标检测结果。检测框架仅对图像中出现的车辆进行检测, 由于第二章所述坐标系变换的原因, 部分距离较近的车辆没有被识别。远处的车辆, 被遮挡严重的没有被识别。在中等距离下的识别效果较好。值得注意的是, 数据集本身的标注质量较为一般, 例如图中最靠近观察点的三辆车, 即图片最下方的 3 个车辆, 标注的车辆高度明显比车辆的实际高度要小, 车辆的车顶甚至比标注的边界框高出了不少。但我们识别出的车辆边界框还是比较符合现实的, 比标注的要大一些。



图 4-13 DAIR-V2X 序列 000018 的检测结果

4.3 本章小结

本章首先介绍了 2D 检测器 YOLOv4 的原理，训练环境与网络配置，训练结果，结果分析。再介绍了 3D 检测器 PointRCNN 的原理，训练环境与网络配置，训练结果，结果分析。

第五章 基于优化的路侧导航增强原理

为了提高对车辆的导航定位精度，有许多基于车载传感器的方案，例如视觉 SLAM 或激光 SLAM。它们通过车辆对周边环境的观测，结合自身 IMU/GNSS 传感器的定位，用状态估计理论预测车辆位置。路侧单元对车辆的观测，本质与 IMU, GNSS 等传感器无异，也是观测量的一种。为了实现路端导航增强，我们需要将路侧观测考虑在传统的状态估计模型中。

5.1 车辆状态估计的数学描述

假设车辆装备着 IMU, GNSS, 相机等传感器在一个未知的环境里运动。由于传感器采样的时间是离散的，假设采样时刻为 $t = 1, 2, \dots, k$ 。 x_t 为车辆在 t 时刻的位置¹， $X = \{x_1, x_2, \dots, x_k\}$ 为车辆的轨迹。车载传感器 GNSS 对自身速度的观测量为 $G = \{g_1, g_2, \dots, g_k\}$ 。车载传感器 IMU 对自身速度的观测量为 $U = \{u_1, u_2, \dots, u_k\}$ 。路侧单元在 t 时刻对车辆的观测为 o_t ，轨迹为 $O = o_1, o_2, \dots, o_t$ 。我们的状态估计问题为，已知时刻 $t = 1, 2, \dots, k$ 对应的车辆对自身的观测 U 和 G ，路侧单元对车辆的观测 O ，预测车辆位置 X 。这是一个最大后验估计问题 MAP(Maximum a Posteriori Inference)，其数学描述为

$$X^{MAP} = \arg \max_X p(X|Z) \quad (5.1)$$

其中 $Z = \{G, U, O\}$ 是观测量的集合。

根据贝叶斯概率公式

$$\begin{aligned} X^{MAP} &= \arg \max_X \frac{p(Z|X)p(X)}{p(Z)} \\ &= \arg \max_X p(Z|X)p(X) \end{aligned} \quad (5.2)$$

第三行是由于 Z 是已知的观测量，所以上式中的分母 $p(Z)$ 不会对最大后验概率对应的状态值产生影响。

¹这里为了简化模型，不同传感器的采样时刻假设为一致，现实中各传感器的时刻虽不一致，但不影响理论推导。

5.2 基于因子图的状态估计理论

利用条件概率公式, (5.2)可以继续分解为条件概率的乘积

$$\begin{aligned}\phi(X) &= p(X|Z) \sim p(Z|X)p(X) \\ &= p(x_1) \prod_i p(x_i|x_{i-1}, u_i) \prod_i p(g_i|x_i) \prod_i p(o_i|x_i) \\ &= \phi_0(x_1) \prod_i \phi_i(x_{i-1}, x_i) \prod_i g_i(x_i) \prod_i o_i(x_i)\end{aligned}\quad (5.3)$$

其中, 第二个等号用到了各个传感器间相互独立的假设, 第三个等号使用了如下记号:

$$\phi(X) = p(X|Z) \quad (5.4)$$

$$\phi_0(x_1) = p(x_1) \quad (5.5)$$

$$\phi_i(x_i, x_{i+1}) = p(x_i|x_{i-1}, u_i) \quad (5.6)$$

$$g_i(x_i) = p(x_i|g_i) \quad (5.7)$$

$$o_i(x_i) = p(x_i|o_i) \quad (5.8)$$

这种将函数分解为乘积形式的写法可以进一步用因子图来描述, 如图5-14所示。图中包含了4种因子节点, 先验节点 ϕ_0 , 运动模型节点 ϕ_i , 和来自 GNSS/RSU 的观测节点¹ g_i , o_i 。

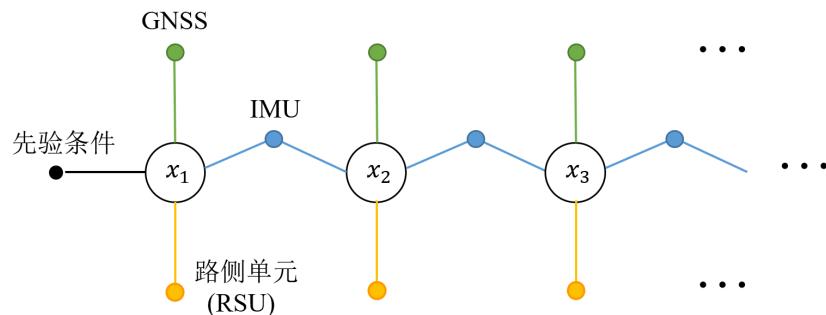


图 5-14 车辆估计的因子图描述

¹本文定义的观测节点和 SLAM 里的观测节点概念不同, 前者是环境 (GNSS/RSU) 对车辆的观测, 后者是车辆对环境的观测。

假设每个因子对应的误差模型都满足零均值的多元高斯分布

$$\begin{aligned} N(\theta; \mu, \Omega) &= \frac{1}{\sqrt{\|2\pi\Omega\|}} \exp \left\{ -\frac{1}{2} \|\theta - \mu\|_{\Omega}^2 \right\} \\ &= \frac{1}{\sqrt{\|2\pi\Omega\|}} \exp \left\{ -\frac{1}{2} \|r\|_{\Omega}^2 \right\} \end{aligned} \quad (5.9)$$

其中 $\mu \in \mathbb{R}^n$ 是均值, Ω 是 $n \times n$ 的协方差矩阵, θ 为分布的自变量, $r = \|\theta - \mu\|$ 为残差,

$$\|\theta - \mu\|_{\Omega}^2 = (\theta - \mu)^T \Omega^{-1} (\theta - \mu) \quad (5.10)$$

是马氏距离 (Mahalanobis Distance) 的平方。

求解原问题的最大后验概率分布等价于求该概率负对数的最小值。

$$\begin{aligned} X^{MAP} &= \arg \max_X p(X) \\ &= \arg \max_X \exp \left\{ \|r_{x_0}\|_{\Omega_0}^2 \right\} \prod_i \exp \left\{ \|r_{x_i}\|_{\Omega_i}^2 + \|r_{o_i}\|_{R_i}^2 + \|r_{g_i}\|_{Q_i}^2 \right\} \\ &= \arg \min_X \left(\|r_{x_i}\|_{\Omega_i}^2 + \sum_i (\|r_{x_i}\|_{\Omega_i}^2 + \|r_{o_i}\|_{R_i}^2 + \|r_{g_i}\|_{Q_i}^2) \right) \end{aligned} \quad (5.11)$$

其中 r_{x_0} , r_{x_i} , r_{g_i} 和 r_{o_i} 分别为先验模型, 运动模型和来自 GNSS/RSU 的观测模型的残差项。

IMU 因子节点论文^[38]给出了一种 IMU 预积分算法。该算法可以计算因子图中两个节点间 IMU 测量值的预积分, 为优化步骤提供高频率的状态更新。时刻 $i-1$ 到时刻 i 间的残差定义如下:

$$\mathbf{r}_{x_i} = [\mathbf{r}_{\Delta \mathbf{R}_i}^T, \mathbf{r}_{\Delta \mathbf{p}_i}^T, \mathbf{r}_{\Delta \mathbf{v}_i}^T]^T \in \mathbb{R}^9 \quad (5.12)$$

其中, $\mathbf{r}_{\Delta \mathbf{R}_i}^T, \mathbf{r}_{\Delta \mathbf{p}_i}^T, \mathbf{r}_{\Delta \mathbf{v}_i}^T$ 分别为来自姿态角 \mathbf{R} (Rotation), 位置 \mathbf{p} (Pose) 和速度

v(velocity) 的残差项。

$$\begin{aligned}\mathbf{r}_{\Delta \mathbf{R}_i} &= \log(\Delta \tilde{\mathbf{R}}_i(\mathbf{b}_{i-1}^g)) \mathbf{R}_{i-1}^T \mathbf{R}_i \\ \mathbf{r}_{\Delta \mathbf{p}_i} &= \mathbf{R}_{i-1}^T \left(\mathbf{p}_i - \mathbf{p}_{i-1} - \mathbf{v}_{i-1} \Delta t_{\Delta i} - \frac{1}{2} \mathbf{g} \Delta t_i^2 \right) - \Delta \tilde{\mathbf{p}}_{\Delta i}(\mathbf{b}_{i-1}^g, \mathbf{b}_{i-1}^a) \\ \mathbf{r}_{\Delta \mathbf{v}_i} &= \mathbf{R}_{i-1}^T (\mathbf{v}_i - \mathbf{v}_{i-1} - \mathbf{g} \Delta t_{\Delta i}) - \Delta \tilde{\mathbf{v}}_i(\mathbf{b}_{i-1}^g, \mathbf{b}_{i-1}^a) \\ \mathbf{r}_{\Delta \mathbf{b}_i} &= \mathbf{b}_i - \mathbf{b}_{i-1}\end{aligned}\quad (5.13)$$

具体的推导可参见[附录 A](#)

GNSS 因子节点可以表示为

$$g_i(x_i) = \frac{1}{\sqrt{|2\pi R|}} \exp \left\{ -\frac{1}{2} \|h_g(x_i) - g_i\|_R^2 \right\} \quad (5.14)$$

其中 $h_g(X)$ 是高斯分布的均值，其现实意义为，当物体轨迹为 X 时，观测量 G 的理论值， h_g 为测量函数， R 为高斯分布的协方差矩阵。

RSU 因子节点可以表示为

$$o_i(x_i) = \frac{1}{\sqrt{|2\pi Q|}} \exp \left\{ -\frac{1}{2} \|h_o(x_i) - o_i\|_Q^2 \right\} \quad (5.15)$$

其中 $h_o(X)$ 是高斯分布的均值，其现实意义为，当物体轨迹为 X 时，观测量 O 的理论值， h_o 为测量函数， Q 为高斯分布的协方差矩阵。

5.3 基于因子图的非线性最优化算法

在获得(5.11)后，问题被转化为了一个非线性优化问题。非线性问题的复杂度和困难度远大于线性问题。非线性函数含有高阶无穷小项，因此难以获得函数的所有解析解。但我们可以用线性化的方法，获得函数的局部最优解。以(5.15)为例，将其在 x_0 线性化可以将问题转化为

$$\xi^* = \arg \min_{\xi} \|h_o(x_0) + \mathbf{J}_{x_0} \xi - o_i\|_Q^2 \quad (5.16)$$

传统的方法有梯度下降法，高斯牛顿法或列文伯格-马夸尔特 (L-M) 法等。

梯度下降法沿着梯度下降的方向进行优化计算，步长 ξ 计算如下

$$\mathbf{J}_{x_0} \xi = o_i - h_o(x_0) \Rightarrow \xi = \lambda \mathbf{J}_{x_0}^T (o_i - h_o(x_0)) \quad (5.17)$$

其中 λ 是迭代系数，控制下降的速度。

高斯牛顿法计算了函数的二阶导数来拟合非线性函数。对(5.16)进行展开，获得误差函数的二阶翻书，求导可以得到极值点，从而获得扰动变量 ξ 的求解方程

$$\mathbf{J}_{x_0}^T \mathbf{J}_{x_0} \xi = \mathbf{J}_{x_0}^T (o_i - h_o(x_0)) \quad (5.18)$$

L-M 算法结合了梯度下降法和高斯牛顿算法，把迭代步长控制在一个范围内

$$(\mathbf{J}_{x_0}^T \mathbf{J}_{x_0} + \lambda \mathbf{I}) \xi = \mathbf{J}_{x_0}^T (o_i - h_o(x_0)) \quad (5.19)$$

I 是单位矩阵， λ 是调控系数，当 λ 比较大时，方法等价于梯度下降法，反之，当 λ 趋近于 0 时等价于高斯牛顿法。

此外，GTSAM^[39]也被广泛用于非线性最优化问题。GTSAM 是一种用于视觉 SLAM 和非线性优化的开源 C++ 库。它是一种高效、通用的图优化框架，可以在多个领域中应用，包括机器人、无人机、自动驾驶汽车等。GTSAM 支持多种传感器类型，例如相机、IMU、GPS 等，并提供了基于因子图的概率推理框架，能够精确地估计状态变量，并根据测量数据进行更新优化。此外，GTSAM 还提供了一组易于使用的 API，使得使用者可以方便地构建 SLAM 系统，并加入自定义的因子模型。

5.4 本章小结

本章首先介绍了路侧增强的数学模型，再介绍了基于因子图的状态估计理论，各个因子节点的残差计算，最后介绍了因子图理论中的常用的优化方法。

第六章 全文总结

6.1 全文工作总结

在自动驾驶的落地过程中，车路一体化系统是尤为重要的一环。目前，车辆终端导航定位主要依赖于全球卫星导航系统（GNSS），但该系统受限于卫星相关误差、传播途径相关误差、接收机相关误差等限制，对车辆的定位、测速精度有限。这给车辆自身的定位能力带来了挑战。为了在 GNSS 脆弱环境下增强车辆的导航能力，本文对基于多传感器的路侧导航增强原理进行了研究。针对如何提高车辆对自身位置的估计精度，展开了以下研究。

首先调研了车端和路端框架下的公开数据集，总结了近年来适用于自动驾驶与车路协同的公开数据集。再调研了 KITTI 数据集和 DAIR-V2X 数据集的传感器配置，标签与标定格式，其中包括了激光雷达与摄像头的标定原理及其坐标转换关系。其次，在 KITTI 数据集上验证了 EagerMOT 跟踪算法，HOTA 到达了 78%。然后调研了 YOLOv4 和 PointRCNN 的原理，在 DAIR-V2X 数据集上训练了 2D 与 3D 检测器。2D 检测器的精度达到了 95%，3D 检测器包围框精度达到了 60%。最后提出了车辆导航增强原理，先给出了问题的数学阐述，然后基于因子图进行建模，转化为了非线性最优化问题，最后介绍了常用的非线性最优化算法。

6.2 创新点

在文献调研中，我们发现虽然基于车端视角或路端视角的车辆跟踪框架有很多。但是鲜有将跟踪结果发挥车端，和车端自身定位结合估计车辆位置的工作。我们基于前人比较成熟的目标检测，目标跟踪工作，搭建好路侧跟踪框架后，提出了基于路侧单元的导航增强原理。

6.3 不足与展望

我们采用的路侧单元数据集只适合目标检测任务。但就在最近 1 个月，该团队发布了适合目标跟踪任务的数据集。我们最后也和另一数据集 IPS+

的作者取得了联系，获得了数据集使用授权。但此时工作进度已接近尾声，我们没有在更加合适的数据集上验证目标跟踪算法的性能。

此外，受到计算资源的限制，我们使用的检测器也比较落后了，YOLO 系列已经迭代到了第 7 代，点云检测模型最近也涌现了许多更好的工作。

受到实验环境的限制，硬件平台尚未搭建完成，我们没有完成最后的路侧导航增强算法的实验部分，仅给出了理论推导。

未来进一步的工作是在最新的数据集上，尝试用不同检测模型训练，最后验证跟踪框架的性能。我们还计划在硬件平台完成后进行实车实验，收集数据验证导航增强原理的效果。

参 考 文 献

- [1] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.
- [2] 邢亚男. 车路协同感知融合研究[D/OL]. 吉林大学, 2022. DOI: [10.27162/d.cnki.gjlin.2022.006243](https://doi.cnki.gjlin.2022.006243).
- [3] 国家发展改革委. 智能汽车创新发展战略[EB/OL]. 2020. <https://www.ndrc.gov.cn/xxgk/zcfb/tz/202002/P020200224573058971435.pdf>.
- [4] MAO J, SHI S, WANG X, et al. 3d object detection for autonomous driving: a review and new outlooks[A]. 2022.
- [5] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.
- [6] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in neural information processing systems, 2017, 30.
- [7] SHI S, WANG X, LI H. Pointrcnn: 3d object proposal generation and detection from point cloud[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 770-779.
- [8] 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.
- [9] 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.

- [10] WENG X, WANG J, HELD D, et al. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics[A]. 2020.
- [11] WENG X, WANG J, HELD D, et al. 3d multi-object tracking: A baseline and new evaluation metrics[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 10359-10366.
- [12] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters[C]//2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010: 2544-2550.
- [13] HENRIQUES J F, CASEIRO R, MARTINS P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12. Springer, 2012: 702-715.
- [14] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(3): 583-596.
- [15] DANELLJAN M, HAGER G, SHAHBAZ KHAN F, et al. Convolutional features for correlation filter based visual tracking[C]//Proceedings of the IEEE international conference on computer vision workshops. 2015: 58-66.
- [16] BROMLEY J, GUYON I, LECUN Y, et al. Signature verification using a "siamese" time delay neural network[J]. Advances in neural information processing systems, 1993, 6.
- [17] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]//Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14. Springer, 2016: 850-865.
- [18] 王思信. 基于三维激光雷达与视觉融合的车辆跟踪与驾驶行为研究[D]. 武汉理工大学, 2020.

- [19] KIM A, OŠEP A, LEAL-TAIXÉ L. Eagermot: 3d multi-object tracking via sensor fusion[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 11315-11321.
- [20] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3354-3361.
- [21] HUANG X, CHENG X, GENG Q, et al. The apolloscape dataset for autonomous driving[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 954-960.
- [22] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2446-2454.
- [23] CAESAR H, BANKITI V, LANG A H, et al. nuscenes: A multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11621-11631.
- [24] SUN P, SUN C, WANG R, et al. Object detection based on roadside lidar for cooperative driving automation: a review[J]. Sensors, 2022, 22(23): 9316.
- [25] WANG H, ZHANG X, LI Z, et al. Ips300+: a challenging multi-modal data sets for intersection perception system[C]//2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 2539-2545.
- [26] YU H, LUO Y, SHU M, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 21361-21370.
- [27] CRESS C, ZIMMER W, STRAND L, et al. A9-dataset: Multi-sensor infrastructure-based dataset for mobility research[C]//2022 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2022: 965-970.

- [28] BERNARDIN K, STIEFELHAGEN R. Evaluating multiple object tracking performance: the clear mot metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008: 1-10.
- [29] LUITEN J, OSEP A, DENDORFER P, et al. Hota: A higher order metric for evaluating multi-object tracking[J]. International journal of computer vision, 2021, 129: 548-578.
- [30] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008: 1-8.
- [31] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [32] REDMON J, FARHADI A. Yolov3: An incremental improvement[A]. 2018.
- [33] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[A]. 2020.
- [34] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016: 21-37.
- [35] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- [36] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
- [37] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

- [38] FORSTER C, CARLONE L, DELLAERT F, et al. On-manifold preintegration for real-time visual–inertial odometry[J]. IEEE Transactions on Robotics, 2016, 33(1): 1-21.
- [39] DELLAERT F. Factor graphs and gtsam: A hands-on introduction[R]. Georgia Institute of Technology, 2012.

IMU 因子节点的残差推导（附录 A）

论文^[38]给出了一种 IMU 预积分算法。该算法可以计算因子图中两个节点间 IMU 测量值的预积分，为优化步骤提供高频率的状态更新。时刻 $i-1$ 到时刻 i 间的残差定义如下：

$$\mathbf{r}_{x_i} = [\mathbf{r}_{\Delta \mathbf{R}_i}^T, \mathbf{r}_{\Delta \mathbf{p}_i}^T, \mathbf{r}_{\Delta \mathbf{v}_i}^T]^T \in \mathbb{R}^9 \quad (\text{A.1})$$

其中， $\mathbf{r}_{\Delta \mathbf{R}_i}^T, \mathbf{r}_{\Delta \mathbf{p}_i}^T, \mathbf{r}_{\Delta \mathbf{v}_i}^T$ 分别为来自姿态角 \mathbf{R} (Rotation)，位置 \mathbf{p} (Pose) 和速度 \mathbf{v} (velocity) 的残差项。

$$\begin{aligned} \mathbf{r}_{\Delta \mathbf{R}_i} &= \log(\Delta \tilde{\mathbf{R}}_i(\mathbf{b}_{i-1}^g)) \mathbf{R}_{i-1}^T \mathbf{R}_i \\ \mathbf{r}_{\Delta \mathbf{p}_i} &= \mathbf{R}_{i-1}^T \left(\mathbf{p}_i - \mathbf{p}_{i-1} - \mathbf{v}_{i-1} \Delta t_{\Delta i} - \frac{1}{2} \mathbf{g} \Delta t_i^2 \right) - \Delta \tilde{\mathbf{p}}_{\Delta i}(\mathbf{b}_{i-1}^g, \mathbf{b}_{i-1}^a) \\ \mathbf{r}_{\Delta \mathbf{v}_i} &= \mathbf{R}_{i-1}^T (\mathbf{v}_i - \mathbf{v}_{i-1} - \mathbf{g} \Delta t_{\Delta i}) - \Delta \tilde{\mathbf{v}}_i(\mathbf{b}_{i-1}^g, \mathbf{b}_{i-1}^a) \\ \mathbf{r}_{\Delta \mathbf{b}_i} &= \mathbf{b}_i - \mathbf{b}_{i-1} \end{aligned} \quad (\text{A.2})$$

其中 \mathbf{R}, \mathbf{p} 和 \mathbf{v} 在连续的两个时刻 $i-1$ 和 i 有如下关系

$$\begin{aligned} \mathbf{R}_i &= \mathbf{R}_{i-1} \exp \left(\left(\tilde{\omega}_{i-1} - \mathbf{b}_{i-1}^g - \eta_{i-1}^{gd} \right) \Delta t_i \right) \\ \mathbf{v}_i &= \mathbf{v}_{i-1} + \mathbf{g} \Delta t_i + \Delta \mathbf{R}_{i-1} (\tilde{\mathbf{a}}_{i-1} - \mathbf{b}_{i-1}^a - \eta_{i-1}^{ad}) \Delta t_i \\ \mathbf{p}_i &= \mathbf{p}_{i-1} + \mathbf{v}_{i-1} \Delta t_i + \frac{1}{2} \mathbf{g} \Delta t_i^2 + \mathbf{R}_{i-1} (\tilde{\mathbf{a}}_{i-1} - \mathbf{b}_{i-1}^a - \eta_{i-1}^{ad}) \Delta t_i^2 \end{aligned} \quad (\text{A.3})$$

$\Delta \mathbf{R}_i, \Delta \mathbf{p}_i, \Delta \mathbf{v}_i$ 是姿态角在时刻 t 相对于时刻 $t-1$ 人为定义的增量，只有 $\Delta \mathbf{R}_i$ 符合实际的“增量”意义，其他两个量只是为了让“增量”与 $i-1$ 时刻的

状态和重力影响无关

$$\begin{aligned}
 \Delta \mathbf{R}_i &= \mathbf{R}_{i-1}^T \mathbf{R}_i = \exp \left(\left(\tilde{\boldsymbol{\omega}}_{i-1} - \mathbf{b}_{i-1}^g - \boldsymbol{\eta}_{i-1}^{gd} \right) \Delta t_i \right) \\
 \Delta \mathbf{v}_i &= \mathbf{R}_{i-1}^T (\mathbf{v}_i - \mathbf{v}_{i-1} - \mathbf{g} \Delta t_i) \\
 &= \Delta \mathbf{R}_{i-1} (\tilde{\mathbf{a}}_{i-1} - \mathbf{b}_{i-1}^a - \boldsymbol{\eta}_{i-1}^{ad}) \Delta t_i \\
 \Delta \mathbf{p}_i &= \mathbf{R}_{i-1}^T \left(\mathbf{p}_i - \mathbf{p}_{i-1} - \mathbf{v}_{i-1} \Delta t_i - \frac{1}{2} \mathbf{g} \Delta t_i^2 \right) \\
 &= \frac{1}{2} (\tilde{\mathbf{a}}_{i-1} - \mathbf{b}_{i-1}^a - \boldsymbol{\eta}_{i-1}^{ad}) \Delta t_i^2
 \end{aligned} \tag{A.4}$$

$\Delta \tilde{\mathbf{R}}_i$, $\Delta \tilde{\mathbf{v}}_i$ 和 $\Delta \tilde{\mathbf{p}}_i$ 是预积分测量量 (preintegrated measurement), 代表了对应“增量”的主要部分, 被如下定义

$$\begin{aligned}
 \Delta \tilde{\mathbf{R}}_i &= \exp ((\tilde{\boldsymbol{\omega}}_{i-1} - \mathbf{b}_{i-1}^g) \Delta t_i) \\
 \Delta \tilde{\mathbf{v}}_i &= (\tilde{\mathbf{a}}_{i-1} - \mathbf{b}_{i-1}^a) \Delta t_i \\
 \Delta \tilde{\mathbf{p}}_i &= 0
 \end{aligned} \tag{A.5}$$

论文假设了 IMU 采样时刻比相机更加密集, 因此推导了从时刻 i 到时刻 j 之间的残差表达。如本文 5.1 所假设的所有传感器采样时间一致, 问题简化为相邻两时刻产生的残差。代入论文中的公式可知, 本问题中的预积分测量量全部为 0。

攻读学位期间学术论文和科研成果目录

致 谢

感谢我的指导教师张欣副研究员，在课题开展的过程中提供了细致、耐心的指导，也带我快速学习导航领域的基础知识。课题组的战兴群副院长也在百忙中为课题思路提供了指导。同课题组的袁文翰师兄，刘佳辉师兄，王士壮师兄，池澄师兄都给予过我帮助，帮助我了解研究方向的最新进展，理清研究思路。

感谢印子斐老师，带领我初入科研道路。在参加 PRP 项目时，印老师从原理到代码都提供了大量指导与帮助。老师治学严谨，也时常鞭策我们进步。

感谢给我带来过优秀课堂的众多教师。陈克应老师的数学分析课让我感受到了世界一流大学该有的授课水平与教学氛围，老师幽默风趣又不失数学严谨性的授课方式至今让我记忆犹新。徐海光老师的大学物理课别具一格，用看似出乎意料实则在情理之中的授课方式，让我们不拘泥于题目，而是关注物理学本身的哲学。刘玉琴老师和邵鹏洁老师的大学俄语课也对我帮助很大。俄语是一门很难的学科，两位老师对我们的提问总是不吝啬自己的学识。如果没有两位老师负责认真的态度，我来到莫斯科不会有和当地人对话的基础。张峰老师的电路理论和电路实验课更是国家精品课程，我第一次认识到了“电路之美”。四年里还遇到了许多优秀老师，他们或者授课严谨，或者风格独特，一时难以写完。没有他们精彩的课堂，过去的四年将失去许多颜色。

致远工科的老师和同学是非常可爱的群体。师生间的融洽氛围让我第一次对交大有了归属感。在这里老师组织了许多有意思的活动，即使在 20 年疫情居家隔离期间，也组织了线上的音乐会。在这里，我认识很多有趣，聪明的小伙伴，他们勤勉的学习态度让我受益良多。回忆起来，在致远学习的时间仍然是四年里最充实最怀念的时间。

感谢航空航天学院对我们的关怀。大学是一个人走向社会的缓冲期，学院积极组织暑期实习，实地参观，但我们认识行业一线，逐渐认清自己未来的道路。学院还提供了来莫斯科交流的机会。第一次独立生活让我走出象牙塔，感受到了生活的烟火气。在留学生宿舍里，我感受到了世界的参差。

感谢我的舍友，在我失意的时候帮助我走出阴影，逐渐认识自我。

感谢我的家人，在我求学期间提供了温暖和关怀。