

博士学位论文

机器人视觉定位

姓      名：张会  
学      号：1510456  
所在院系：电子与信息工程学院  
学科门类：工学  
学科专业：控制科学与工程系  
指导教师：陈启军教授

二〇二一年四月

A dissertation submitted to  
Tongji University in conformity with the requirements for  
the degree of Doctor of Philosophy

## **Robots Monocular Visual Localization**

Candidate : Hui Zhang  
Student Number : 1510456  
School/Department : School of Electronic and  
Information Engineering  
Discipline : Engineering  
Major : Control Science and Engi-  
neering  
Supervisor : Prof. Qijun Chen

April, 2021

# 学位论文版权使用授权书

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构递交论文的复印件和电子版；在不以盈利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年      月      日



# 同济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，是本人在导师指导下，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，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年      月      日



## 摘要

本文主要贡献和创新点为：

**关键词：**单目相机，视觉定位，绝对定位，增量定位，卷积神经网络，几何约束

## ABSTRACT

The main contributions and innovations of this article are:

**Key Words:** Visual Localization, State Estimation, Graph Optimization, Probabilistic Graphical Model, Convolutional Nerual Network, Homomorphism , Patch Agreement

# 目录

主要符号对照表 .....	IV
第1章 引言 .....	1
1.1 机器人定位研究背景和意义 .....	1
1.2 国内外研究现状介绍 .....	3
1.2.1 单目视觉绝对定位研究现状 .....	3
1.2.2 增量式定位--单目视觉里程计研究现状 .....	6
1.3 本文内容与贡献 .....	10
1.4 全文架构 .....	10
第2章 数学原理及研究热点数学建模 .....	11
2.1 单目视觉绝对定位数学原理 .....	11
2.2 单目视觉增量式定位数学原理 .....	11
第3章 动态环境中绝对定位之单目深度特征提取与压缩 .....	14
3.1 单目视觉绝对定位深度特征提取方法 .....	15
3.1.1 基于 AlexNet 网络的深度学习特征提取 .....	17
3.1.2 基于 AlexNet 网络的单目视觉绝对定位深度特征降维 .....	18
3.1.3 单目视觉绝对定位匹配矩阵核化处理与归一化 .....	19
3.2 基于 AlexNet 的单目视觉绝对定位实验 .....	21
3.3 本章小结 .....	22
第4章 增量式定位--单目视觉里程计几何尺度估计 .....	25
4.1 道路几何模型计算 .....	26
4.1.1 基于深度一致性的路面特征点筛选 .....	27
4.1.2 基于路面模型一致性的特征点筛选 .....	29
4.1.3 路面模型与绝对尺度计算 .....	31
4.2 KITTI 数据集单目视觉里程计实验 .....	33
4.2.1 对现有单目视觉里程计开源算法的改进 .....	33
4.2.2 算法不同模块效果分析 .....	40
4.3 本章小结 .....	44
第5章 单目视觉里程计尺度计算: 从手动建模到自主学习 .....	47
5.1 引言 .....	47
5.2 基于场景建模的尺度计算方法 .....	47
5.2.1 场景建模 .....	47
5.2.2 尺度计算 .....	48
5.2.3 基于场景建模的尺度计算验证实验 .....	49
5.3 本章小结 .....	49

第 6 章 传统单目位姿估计与深度学习尺度恢复.....	50
6.1 引言 .....	50
第 7 章 路面驾驶机器人单目视觉里程计简化.....	51
7.1 方法 .....	54
7.1.1 运动聚焦和运动解耦.....	54
7.1.2 模型与训练.....	58
7.2 实验 .....	59
7.2.1 数据集和实验平台.....	60
7.2.2 实验结果 .....	60
7.2.3 Discussion .....	66
7.3 本章小结 .....	67
致谢 .....	68
个人简历、在学期间发表的学术论文与研究成果 .....	69

## 主要符号对照表

$\mathbb{R}$	实数集合
$\mathbb{I}$	0-255 闭区间内所有整数集合
$\text{SO}(3)$	3 维空间特殊正交群
$\text{SE}(3)$	3 维空间特殊欧式群
$\mathbf{R}$	旋转矩阵 $\mathbf{R} \in \text{SO}(3)$
$\mathbf{R}_{ij}$	矩阵 $\mathbf{R}$ 的第 $i$ 行第 $j$ 列的元素
$\mathbf{r}$	旋转向量
$\mathbf{t}$	平移向量
$\mathbf{t}_i$	向量 $\mathbf{t}$ 的第 $i$ 个元素
$\bar{\mathbf{t}}$	相对尺度下的平移向量
$[\mathbf{t}]_{\times}$	斜对称矩阵 $[\mathbf{t}]_{\times} = \begin{pmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ t_1 & t_2 & 0 \end{pmatrix}$
$\mathbf{T}$	运动矩阵 $\mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \in \text{SE3}$
$\mathbf{T}_t^{t+1}$	$t$ 时刻到 $t + 1$ 时刻机器人的运动矩阵
$\tau$	运动向量 $\tau = \log(\mathbf{T}) \in \text{se3}$
$\underline{\tau}$	训练集中 $\tau$ 的真值
$\log(\mathbf{R})$	旋转矩阵的对数映射 (结果为旋转运动的角轴表示)
$\exp(\mathbf{r})$	旋转向量反对称矩阵的指数映射
$\mathbf{I}^t$	第 $t$ 帧图像
$\mathbf{D}^t$	第 $t$ 帧图像对应的深度图
$\Omega_{\mathbf{I}}$	图像集合
$\mathbf{I}_m^n$	图像和图像组成的图像对
$\Omega_{\mathbf{I}_m^n}$	图像对集合
$\otimes$	定义在图像对集合上的二元运算, $\mathbf{I}_m^n \otimes \mathbf{I}_t^n = \mathbf{I}_m^n$
${}^p \mathbf{I}^t$	第 $t$ 帧图像的第 $p$ 块子区域
${}^p \mathbf{I}_m^n$	图像对的第 $p$ 块子区域
$\Omega_f$	特征点集合 $\Omega_f = \{f_0, f_1, \dots\}$
$\Omega_{\mathbf{I}^t}$	图像 $\mathbf{I}^t$ 中的像素点集合
$f_i^t$	第 $t$ 帧图像的第 $i$ 个特征点
$\check{f}$	特征点 $f$ 属于路面区域
$\hat{f}$	特征点 $f$ 不属于路面区域

---

$\mathbf{n}_r$	路面所在平面在相机坐标系下的法向
$\mathbf{u}_i^t$	第 $t$ 帧第 $i$ 个像素点的 2 维像素坐标 $\mathbf{u}_i = (u_i, v_i)$
$\mathbf{u}_i$	第 $i$ 个像素点的 2 维像素坐标 $\mathbf{u}_i = (u_i, v_i)$
$\vec{\mathbf{u}}_i$	第 $i$ 个像素点的扩充像素坐标 $\vec{\mathbf{u}}_i = (u_i, v_i, 1)$
$\mathbf{x}_i$	第 $i$ 个像素点的 3 维空间坐标 $\mathbf{x}_i = (x_i, y_i, z_i)$
$d_i$	第 $i$ 个像素点的深度
$I_p$	像素点 $p$ 的像素值
$\mathbf{u}_p$	像素点 $p$ 的像素位置
$d_p$	像素点 $p$ 的深度值
$\underline{d}_p$	训练集中像素点 $p$ 深度值的真值
$\mathbf{x}_p$	像素点 $p$ 的空间坐标
$I_p^t$	第 $t$ 帧图像中像素点 $p$ 的像素值
$a \propto b$	$a$ 与 $b$ 正相关
$\mathbf{v}^T$	向量的转置
$\mathbf{M}^T$	矩阵的转置
$\mu(\mathbf{x})$	$\mathbf{x}$ 的均值
$\sigma(\mathbf{x})$	$\mathbf{x}$ 的标准差
$\sigma^2(\mathbf{x})$	$\mathbf{x}$ 的方差
$Q_{\frac{1}{2}}(\mathbf{x})$	$\mathbf{x}$ 的中位数

# 第1章 引言

## 1.1 机器人定位研究背景和意义

本论文以移动机器人在已知和未知环境中基于单目相机的自我定位为研究背景，主要研究绝对式定位的深度特征提取与压缩、增量式定位之单目视觉里程计（VO）的绝对尺度运算及 VO 问题的求解框架简化。本章节将从国际人工智能发展背景、国内机器人行业发展需求、移动机器人的发展需求及研究热点等多个角度逐层剖析。

国内背景：中国制造 2025<sup>[1]</sup>，从第四代工业革命介绍人工智能的发展潮流，其中机器人尤其是无人驾驶的地位，简述机器人的研究难点，VO 在其中的重要性，VO 的难点，VO 的关键技术人工智能技术的出现，为人们生活带来翻天覆地的转变，机器人正在以多种存在形式逐渐融入人类的日常生活，大大提高了人类的生活质量与幸福感。人工智能技术的发展历经曲折却又在一代又一代人的努力下，不断创新与颠覆。1956 年达特茅斯会议，标志着 AI 的诞生，第一个神经网络由 Rosenblatt 在 1957 年提出；但是由于计算能力的限制，没能使机器完成大规模数据训练和复杂任务，AI 进入第一个低谷时期；1986 年，BP 算法的出现使得大规模神经网络的训练成为可能，带来 AI 的第二个黄金时代；但是，由于 AI 计算机 DARPA 没能实现，政府支持度降低，资金投入缩减，使 AI 进入了第二个低谷；2006 年，Hinton 提出“深度学习神经网络”，使 AI 出现了突破性转折，并先后在机器视觉识别、智能语音等领域取得成功，宣告了 AI 的爆发式发展，表明人类正式进入智能感知时代，带入第四代工业革命。机器人正以不同的身份出现在人们生活的方方面面，逐渐更新我们的生活方式。机器人出现在林林总总的服务行业中，改变了传统产品形态与服务模式，比如送餐机器人、青少儿教育机器人、老人陪护与家庭助理型机器人、虚拟/增强现实等交互手段等。同时，智能机器人具有人力劳动无法比拟的优点，比如精度更高、安全性更高、连续工作，既高效完成任务又减少了人力劳动的成本与危险。智能机器人在工业、农业、交通等方面具有绝佳的研究价值与商业价值。比如智能机器人能够完成工业行业的高强度与高难度的各种作业——核电作业、能源矿山作业、矿山勘探开采、水下考察等任务；在农业方面常见的有播种机器人、收割机器人等；在安全监测、智慧巡检、特殊运输等方面也出现了很多研究热点，例如深圳一清科技研发的“夸父-I”无人车在重灾区、一线城市郊野和产业园区等场所往返运输生活用品、医疗器械等；在医疗保健行业也存在广大的应用场景，比如打造智

慧医疗、疫情防控、救援抢险、助老助残等智能体，本质上解决医生的长时间工作疲劳问题，同时提高医生工作操作精准度，降低患者痛苦加速伤口愈合。外骨骼机器人——模仿人体骨骼结构设计的一种机电一体化装置，通过感知穿戴者的意图，帮助残障人士更自主的生活，同时帮助负重等工作人士的身体损伤，提升人体机能。医药物资搬运机器人等，投入多家医院使用。在军事防御、外太空空间探索、极地科考、深海遨游等极限环境中降低对人类的挑战，拓展人类能力，利用多模态的感知手段为复杂地形下的自主探测提供了新的解决方案，例如2016年10月26日，美国军方用3架F/A-18“超级大黄蜂”战斗机搭载并释放了103架“山鹑”（Perdix）微型无人机组成的机群，进行了无人机集群飞行试验。2017年以色列在第二届国际陆战会议上展示了“卡梅尔”自动驾驶装甲车。

智能机器人的发展趋势呈现自主化、协作化、灵巧化。

从以上不同应用领域我们不难发现，移动机器人是第四代工业革命中不可或缺的核心技术，是保证人工智能服务于人类的最常见的机器人外在形态之一。移动机器人的代表之一就是无人驾驶车辆，其是指在没有人工干预的情况下，能够感知环境并进行路径规划、导航避障自主到达目的地的汽车。被科学家普遍认可的无人驾驶汽车出现在1921年，是美国军队在俄亥俄州空军基地展示的一种三轮拖车，因为这辆拖车上配备了无线电控制系统，所以拖车上无需坐人。通用汽车在1956年的时候，推出的一辆Firebird II概念车，成为了无人驾驶概念车的先驱。1957年，在内布拉斯加州一条高速公路上，美国一些研究人员，通过埋在地下的探测器检测道路上的障碍物。1960年，英国研究人员利用埋在道路中的信号电缆网络为提供给汽车道路信息，控制汽车自动转向、加速或制动等。要实现自主导航、运动跟踪、障碍物检测和规避等一系列功能，要求移动机器人必须保证能够获得随着时间的推移获取自身位置的改变。

近年来为了保障机器人能够获得随着时间的推移其位置的改变，研究人员和工程师们开发了各种用于移动机器人定位的传感器、技术和系统，如轮式里程计IMU、惯性导航系统(INS)、激光或超声波测距法、全球定位系统(GPS)和视觉里程计(VO)等。然而，每种方案都有自己的不可避免的弱点：IMU是最简单的位置估计技术，但由于车轮滑移，它存在位置漂移问题<sup>[2]</sup>；INS也极易产生漂移，而高精度的INS价格昂贵，对于商业用途来说是不可行的解决方案；激光测距仪以激光器作为光源，由光电元件以一定的工作频率向目标物体发射并接收目标反射的激光束，由计时器来测定激光束从发射出去到接收的时间差，从而计算出观测者到目标物体的距离，由于不受主动光源的影响，激光测距仪可以在黑暗环境中正常工作。超声波测距原理与激光测距原理相似，不过发射装置发出的是超声波而不是激光；GPS是最常见的定位解决方案，因为它可以提供绝对的位置而不会积累误差，但它只在天空视野清晰的地方有效，它不能

在室内、深海、密闭空间等环境中使用<sup>[3]</sup>。商用 GPS 所估计位置的误差也较大，通常其误差是在米级。这种误差被认为对于要求精度以厘米为单位的精确应用来说太大，例如室内服务机器人小面积精准定位和自主停车等。差分式全球定位系统和实时运动式全球定位系统可以提供厘米级精度的位置，但这些技术成本较高价格昂贵。GPS 是一项已经深深的融入到了我们老百姓的日常生活中的定位方式，包括车载型、通讯型、便携型、船载型、指挥型等多种用户终端，具体如每日必不可少的车辆导航、手机定位等。GPS 接收器的可以保证以较低的价格就可以立即获取我们在地球上所处的 10 米级的位置，包括纬度，经度和海拔。根据美国政府有关全球定位系统 (GPS)<sup>①</sup>的官方信息，截至 2020 年 5 月，有 29 颗可运行卫星。卫星导航系统虽然最早是美国发明并投入使用的，但其他国家也开发了自己的卫星导航系统，比如我们国家自主研发的北斗卫星导航系统、俄罗斯的格洛纳斯系统等。

本文研究重点之一是增量式单目视觉里程计，其是解决机器人定位的最快捷方便且低成本的方式之一。

## 1.2 国内外研究现状介绍

### 1.2.1 单目视觉绝对定位研究现状

在绝对定位方中，机器人对世界的认知须以地图的形式存储，进而与当前的观察结果进行对比，通过 GPS 等 tag 完成定位。通过对图像进行特征提取并检索以完成匹配。然而用一个稳定鲁棒的特征来表示一个处于变化的动态场景是一个很大的挑战，如图??所示。

#### 1.2.1.1 图像特征提取相关工作

根据运算过程中采用的图片帧数，单目视觉定位的方法可分为两类：基于单帧图像的定位方法，基于两帧或多帧的定位方法。基于单帧图像的定位方法包括基于特征点的定位 (Perspective-n-Point)、基于直线或平面特征的定位，其关键点在于快速准确地实现投影图像与模板之间的特征匹配。基于两帧或多帧图像的定位方法的关键在于实现多帧投影图像之间的对应特征元素匹配，如 SLAM。

从相机采集的视觉数据中提取特征，即图像特征提取是机器人位置识别的一个基本问题，其中图像特征提取的方式是影响定位性能的关键。局部特征 (local features)，仍是近年来研究的一个热点。局部特征指一些能够稳定出现并且具有

<sup>①</sup> <https://www.gps.gov/chinese.php>

良好的可区分性的稳定特征点。局部特征数量丰富，特征间相关度小，不容易受到部分遮挡、光照等噪声的干扰，因为不会因为部分特征的消失而影响其他特征的检测和匹配，这样如果我们用这些稳定出现的点来代替整幅图像，可以大大降低图像原有携带的大量信息，起到减少计算量的作用，且在物体不完全受到遮挡的情况下，一些局部特征依然稳定存在，以代表这个物体（甚至这幅图像），方便进一步分析与运算。

因此，大量的计算机视觉研究集中在如何手动提取特征上，以发现和描述从图像中提取的特征，如 SURF<sup>[4]</sup>、ORB<sup>[5]</sup>、BRIEF<sup>[6]</sup>、SIFT<sup>[7]</sup>、Harris<sup>[8]</sup>、SIFT<sup>[7]</sup> 和 HOG<sup>[9]</sup>。

如果用户对整个图像的整体感兴趣，而不是前景本身感兴趣的话，全局特征用来描述总是比较合适的。但是无法分辨出前景和背景却是全局特征本身就有 的劣势，特别是在我们关注的对象受到遮挡等影响的时候，全局特征很有可能就 被破坏掉了。

非手工特征采用三种方法：基于卷积神经网络 (CNN) 的深度转移学习特征、主成分分析网络 (PCAN) 和紧凑的二进制描述符 (CBD)<sup>nanni2017handcraft</sup>。在我们以前的工作中，我们还尝试用 IPCA 来减少特征维数<sup>zhang2017Dynamic</sup>，我们的结果证明了 33 维深度特征可以在匹配矩阵中以高精度识别。最近，非手工制作的特征能够通过深度卷积神经网络 (DCNN) 从数百万标记图像中自动学习鉴别特征，这在计算机视觉和机器学习社区的几乎所有重要任务中都取得了最先进的性能<sup>Radford2016UnsupervisedChen20143D[10]Simonyan2014Very[11][12]</sup>。

最近的文献提出了多种方法来解决这个领域的挑战<sup>[13][14][15][16][17][18][19]</sup>。众所周知，在 2012 年的 AlexNet 大规模视觉识别挑战赛 (ILSVRC) 上一个新的网络模型 CNN 获得了令人难以置信的准确率<sup>[20]</sup>。

文章<sup>[21][22][20][23]</sup>研究表明，来自神经网络的网络特征优于传统的手动特征<sup>Citesharif2014cnn,ORB2011orb,surf2006surf,lowe2004distinctive</sup>。该网络由 5 个卷积层，以及 3 个全连接层和 soft-max 层组成，它在 120 万张有标签的图像上进行了预训练。根据从 AlexNet 中提取的特征对图像进行分类。每个单独层的输出可以作为一个全局的图像描述符。我们还可以根据这些特征对图像进行匹配，然后定位机器人。<sup>[21]</sup> 表示，来自 CNN 中层的特征可以更有效地消除数据集的偏差。<sup>[24]</sup> 比较了不同层特征的性能。他们的结果表明，来自 ConvNet 层次结构中的中间层的特征对一天中的时间、季节或天气条件引起的外观变化表现出鲁棒性。Conv3 层的特征在面对极端的外观变化时表现得相当好。表 1 列出了 AlexNet 网络中不同层的向量尺寸。<sup>[24]</sup> 证明了 Conv3 层的特征在外貌变化方面的表现相当好。此外，fc6 和 fc7 在视角变化方面优于其余层。但是，当外观变化时，fc6 和 fc7 完全失效。Conv3 的维度为 64896，即一幅图像显示为 64896 维度的矢量。在

线定位将持续接收来自摄像头的图像。毋庸置疑，大量高维度向量数学运算增加了运算时间。它在 120 万张有标签的图像上进行了预训练。根据从 AlexNet 中提取的特征对图像进行分类。每个单独层的输出可以作为一个全局的图像描述符。我们还可以根据这些特征对图像进行匹配，然后定位机器人。

为了方便后续的二值化，<sup>[25]</sup> 将这些特征投向一个规范化的 8 位整数格式。然后利用汉明距离对所有二进制特征进行匹配，计算出一个匹配矩阵。他们的研究结果表明，对特征进行压缩可以极大程度上降低其描述符的冗余度，而精度只降低了约 2%。此外，他们对特征的二值化允许使用汉明距离，这也代表了位置匹配的加速。在减少特征集的情况下，改进了地点识别。

### 1.2.1.2 图像特征匹配

机器人视觉图像匹配是指机器人定位领域的场所识别，是继特征提取之后的另一个挑战。毋庸置疑，在绝对定位方中，机器人对世界的认知须以地图的形式存储，进而与当前的观察结果进行对比，通过 GPS 等 tag 完成定位。文章<sup>[26]</sup> 指出，根据视觉传感器的不同，以及识别场景种类的不同，地图框架也有所不同。可分为纯图像检索、拓扑地图和拓扑-度量地图。纯图像检索只存储环境中每个地方的外观信息，没有相关的位置信息，例如 FAB-MAP 中使用的 Chow-Liu 树结构<sup>[27]</sup>。FAB-MAP<sup>[27]</sup> 描述了一种概率方法来解决匹配图像和地图增强的问题。他们使用了基于向量的描述符，如与 bag-of-words 联合的 SURF 特征。FAB-MAP<sup>[27]</sup> 描述了一种概率方法来解决匹配图像和地图增强的问题。他们使用了基于向量的描述符，如 SURF 与 Bag-of-Words 联合使用。他们通过构建一个 Chow-Liu 树结构<sup>[28]</sup> 来捕捉视觉词的共现统计，学习了一个图像深度网络特征的生成模型。Chow-Liu 树由节点和边组成。变量之间的相互信息关联度由节点之间边的粗细来显示，图中的每一个节点对应一个由传感器数据转换而来的词袋，图中的每一个节点对应一个由传感器数据转换而来的词袋，变量之间的相互信息通过树的边的粗细来显示。图中的每一个节点对应一个由输入感官数据转换而来的词袋表示。在具有挑战性的户外环境中，FAB-MAP 能够成功地检测到了大部分的闭环场景。但<sup>[17]</sup> 的结果显示，在跨季节的数据集中，OpenFABMAP2 只找到了少数正确的匹配，原因是，传统手工特征描述符是不可重复的。论文 citenaseer2014robust 将图像匹配制定为数据关联图中的最小成本流问题，以有效利用序列信息。他们通过最小成本流定位车辆。他们的方法即使在高度变化的动态场景也表现良好。SeqSLAM<sup>[13]</sup> 将图像识别问题构思为在局部邻域内寻找所有与当前图像最佳匹配的模板。这很容易实现。然而，<sup>[13]</sup> 很容易受到机器人速度的影响。这种影响限制了机器人进行长时间自我定位。

### 1.2.2 增量式定位--单目视觉里程计研究现状

里程计，英文是"odometry"，该词源于两个希腊词 *hodos*（意为"旅程"或者"旅行"）和 *metron*（意为"测量"）<sup>[2]</sup>。这一推导与估计机器人姿势（平移和方向）随时间的变化有关。移动机器人使用来自运动传感器的数据来估计它们相对于初始位置的位置；这个过程被称为里程测量。里程计是使用来自运动传感器的数据来估算其位置随着时间变化的一种定位技术，一些腿式或轮式机器人在机器人定位技术中使用它来估计其相对于起始点的位置。由于对速度测量值进行了时间积分，因此该方法对误差很敏感，可以给出位置估计值。在大多数情况下，需要快速而准确的数据收集，仪器标定校准和快速处理才能有效使用里程计。

视觉里程计，即 VO (visual odometry)，是一种视觉定位技术，视觉里程计 (VO) 是一种仅仅通过从连接到机器人的单个或多个摄像头获取的图像序列来实现机器人定位<sup>[29]</sup> 移动机器人的增量式定位过程。这些图像包含足够多的有意义的信息（颜色、纹理、形状等），以估计相机在静态环境中的运动<sup>[30]</sup>。增量定位是连续观察机器人姿态的变化，通过累积运动计算机器人当前姿态。VO 是未探索环境中机器人自定位和自主导航的基本模块，因为它不依赖于预先构建的地图<sup>[31][32]</sup>。单目视觉里程计 (MVO) 由于配备了最便宜、使用最广泛的传感器，也是最方便校准的传感器，因此引起了机器人界的广泛研究兴趣。同时，它不受固定基线长度的限制，可以在不同的场景中广泛使用。然而，当单目相机将三维 (3D) 世界投影到二维 (2D) 平面空间时，它失去了物体的深度信息和绝对尺度。因此，MVO 只能获得相对的，而不是机器人运动的绝对距离。这种尺度模糊可以积累尺度误差，称为尺度漂移。尺度模糊和尺度漂移统称为尺度问题。VO 已经研究了 30 多年，最初是由 NASA 的火星探索项目推动的<sup>[33]</sup>。与 IMU 和雷达等其他增量定位系统相比，单目视觉里程法 (MVO) 的优点是显而易见的。它配备了最便宜和使用最广泛的传感器，也是最方便的校准。因此，MVO 是机器人界的一个活跃的研究领域。

尺度问题严重限制了 MVO<sup>[34]</sup> 的精度，相应的解可分为相对尺度修正和绝对尺度估计。前者主要包括光束法平差 (BA)<sup>[35]</sup> 和环路闭合 (LC) 检测。虽然它们确实在限制尺度问题上起作用，但它们无法检索机器人的度量信息。对于 MVO 系统，只有借助先验知识，引用绝对量度信息才能解决尺度问题。流行的绝对量度信息参考包括基线距离（双目摄像机）、摄像机高度（道路模型）和从其他传感器或离线训练中获得的像素深度等<sup>[36]</sup>。其中，挂载相机绝对高度是常用的，因为它既不需要其他传感器的帮助，也不需要离线训练，是最方便测量和校准的。此外，在车辆运行过程中，摄像机绝对高度保持稳定。

在已安装摄像机高度的先验知识下，MVO 的精度取决于相对尺度下的道路

几何估计。当使用相机高度作为尺度恢复的绝对参考时，有必要获得道路的几何模型。其中包含估计的摄像机高度。许多方法<sup>[37][38][39]</sup>根据先验知识选择一个感兴趣的区域 (ROI)，或自动检测的确定固定区域<sup>[40]</sup> 作为道路区域。然而，基于 ROI 的方法有两个缺点。首先，不能保证所选区域始终为路面。此外，图像信息不能得到充分利用。道路检测解决方案更合理，因为它从整个图像中提取特征。此外，由于深度学习方法在多个领域取得超越性性能，<sup>[41]</sup> 提出的分割方法和训练分类器也用于道路检测。但这种方法的计算成本较高且对不熟悉的情况不够鲁棒。此外，所有以前基于分类器的方法都集中在道路的颜色信息上，虽然在深度学习的帮助下，基于道路颜色信息的视觉里程法可以得到很大的改进，此类方法对光照、阴影和材料等因素依然较敏感。因此，我们将颜色信息替换为结合道路几何约束的道路点选择。通过在线更新道路分类器使<sup>[42]</sup> 中的框架更加鲁棒。

虽然融合传感器测量系统目前在精度、鲁棒性和可靠性方面处于领先地位<sup>[43][44]</sup>。单目视觉里程计 (MVO) 却有可能取代它们。MVO 面临的许多挑战<sup>[45]</sup> 主要存在于大规模、动态或无特征的环境中)。

### 1.2.2.1 相对尺度校正

光束法平差<sup>[35]</sup> 和环路闭合检测相对尺度校正的两种重要方法。尺度校正被光束法平差描述为一个非线性最小二乘问题，以产生联合最优的三维结构和摄像机姿态估计。Mouragon *et al.*<sup>[46]</sup> 第一次在实时 VO 中利用光束法平差，其次是并行跟踪和映射 (PTAM)，这是定向 FAST 和旋转 BRIEF 同时定位和映射 (ORB-SLAM) 的主要动机。然而，局部和全局光束法平差优化都存在严重的累积尺度误差。环路闭合检测是一种在先前访问过的地点进行比对的技术，以修正产生的位移偏差；

<sup>[47]</sup> 提出了一种词袋 (BoW) 方法来表示关键帧。基于快速外观的映射 (FAB-MAP)<sup>[27]</sup> 是一种经典的位置识别方法，它与 Chow-Liu 树<sup>[28]</sup> 一起构造了 BoW 模型的视觉词汇表，以表达其特征相似性。然而，在实际的交通场景中，环路很少出现，关键帧的选择严重影响了环路闭合检测的准确性<sup>[48]</sup>。此外，在长距离驾驶中，光束法平差下的尺度漂移也变得严重<sup>[46][49]</sup>。

FAB-MAP<sup>[27]</sup> 是一种经典的位置识别方法。它不仅限于定位任务，而且可以判断一个新的观察是否来自地图上已经存在的地方。FAB-MAP 与 Chow-Liu 树<sup>[28]</sup> 一起构造了 BoW 模型的视觉词汇表，以表达其特征相似度。该方法也应用于另一个优秀的工作 ORB-SLAM2<sup>[50]</sup>。该方法基于面向 FAST 和旋转的关键帧的 BRIEF (ORB) 描述符离线训练大量的 BoW。当摄像机返回到以前的场景时，它将获得类似的 BoW 描述符，从而检测环路闭合。

### 1.2.2.2 绝对尺度恢复

绝对尺度恢复方法可以补偿以已知的度量信息作为参考的相对尺度校正的局限性，例如从深度学习中学到的安装相机高度和图像深度。

**相机高度-固定方法** 摄像机高度约束方法的区别主要在于道路平面的检测和建模方法。许多方法<sup>[37], [38], [39]</sup> 假定 ROI 为道路。来自运动的单目大规模多核结构 (MLM-SFM) 方法<sup>[38]</sup> 中，扩展了<sup>[51]</sup> 和<sup>[52]</sup>，假设图像的下三分之一的中五分之一为 ROI。MLM-SFM 提出了一种数据驱动机制，将多个线索组合在一个框架中，该框架反映了它们的每帧相对机密性，这显示了很有前途的性能。然而，当所选区域被汽车或其他东西遮挡时，基于 RIO 的方法无法工作，如 KITTI 数据集的序列 07 中所发生的那样，因此 MLM-SFM 不能像我们所预期的那样在它该环境运行。同时图像信息可能无法充分利用，因为 ROI 只是图像的一小部分。第二个缺点可以用双向或全向相机<sup>[53], [54], [55]</sup> 弥补。但是考虑到传感器的成本及使用便捷性，普通单目相机更具有研究价值。

道路平面估计方法将判断哪些点属于道路进行帧对帧运动估计，会进行特征点筛选并充分利用整张图像信息，因此更合理。这些方法可根据其特征点大小分为稀疏<sup>[56]</sup>，半稠密<sup>[57][58]</sup>，稠密<sup>[59]</sup>。描述符也可能由一些视觉过程<sup>[60]</sup> 增强或从卷积神经网络 (CNNs) 中提取<sup>[61][62]</sup>。除了特定特征点的方法外，多种描述符相结合的方法也很流行。该方法将来自稠密和稀疏匹配点的线索结合起来，并使用分类器基于各种特征检测尺度异常值，这确实提高了对各种地面结构的鲁棒性。然而，它依赖于稠密的特性，如果没有 GPU 的帮助，它就不能很容易地在移动嵌入式系统上实现。我们的方法在没有稠密特征的任何帮助下取得了有竞争力的结果。

在道路点检测后，一些方法<sup>[63]</sup> 选择利用三角稀疏地面点计算高度，然后估计绝对尺度。传统上，采用三点 RANSAC<sup>[64]</sup> 来实现鲁棒平面拟合。在非传统方法中，<sup>[65]</sup> 用逆向选择代替 RANSAC，<sup>[66]</sup> 用一种快速匹配方法对一组选定的点进行三角剖分，该方法称为有效的大规模立体声<sup>[67]</sup>。

**基于图像深度的方法** 最近，MVO 有一个与深度学习相结合的流行趋势，其中包括从图像与 CNN 估计的深度。对于 MVO 的训练数据，<sup>[68]</sup> 中的结果表明，从合成训练输入中获得的尺度估计精度与从实际数据中获得的估计精度相似。在<sup>[69]</sup> 中，使用深度 CNN 与 CRF 相结合的方法来估计在小规模和大规模环境中拍摄的单目图像的深度。Luo *et al.* 将在线自适应深度与直接单目 SLAM 相结合<sup>[70]</sup>，提高了不同场景的深度预测精度。它有望解决两个核心挑战：地图完整性低和尺度

模糊。然而，单帧图像的深度估计<sup>[69, 71-74]</sup>比来自连续帧的深度估计更复杂<sup>[73]</sup>。<sup>[75]</sup>提出了一种新颖的监督系统从估计的深度图计算平移的尺度，将条件随机场与 CNN 网络相结合以优化深度图，这是通过考虑两个连续图像和运动约束来改进的。

深度预测的准确性对单目 SLAM 中的特征跟踪误差有巨大影响。CNN-SLAM<sup>[76]</sup> 扩展了大规模直接 SLAM(LSD-SLAM)<sup>[77]</sup>，通过部署深度神经网络的预测深度图来产生密集的 3D 地图。它在室内数据集<sup>2014A, [78]</sup>中取得了很好的效果，但在多个关键帧重叠时，预测的深度图无法优化，从而使得重建和映射的精度降低。DVS0<sup>[79]</sup> 使用与<sup>[80]</sup>类似的虚拟立体视图，将深度预测纳入几何单目测绘流水线。Luo 等人<sup>[70]</sup> 将在线适应深度与直接单眼 SLAM 相结合，提高不同场景的深度预测精度。这些方法都有希望解决地图完整性低和比例尺模糊性这两个核心难题。

从连续帧中进行深度估计比从单幅图像中进行深度估计更容易<sup>[73, 81]</sup>。<sup>[36]</sup>的方法从连续图像中提取密集的光流，并训练一个基于深度 CNN 的估计器来进行自运动估计。本研究中的新型监督系统通过考虑两幅连续图像和运动约束，从估计的深度图中计算出转换的尺度，并对其进行改进。他们的网络是通过并发 CNN 和条件随机场来构建的，以完善深度图。除了估计单视角深度，<sup>[82]</sup>还尝试估计双视角光流作为另一个中间输出。最近，Xue 等人<sup>[83]</sup>提出了一种利用密集法线进行道路检测的新方法，在几何约束方面与我们的方法类似。这些基于端到端深度学习的 SLAM 系统已经取得了令人印象深刻的性能，然而，CNN 的深度预测不准确会严重导致单目 SLAM 中的特征跟踪误差，且它们都需要进行离线训练，增加了时间成本与计算成本。此外，也不能保证它们能泛化到新的环境中。我们的系统不仅可以在新的环境中工作，而且可以用低成本的硬件达到较好的性能。

视觉里程计的实现方法？按照传感器数目的不同可以分为：多目相机、双目相机、单目相机，其中相机的类型包括普通相机、鱼眼相机和全景相机等。双目里程计与单目里程计是视觉里程计的研究热点，双目 VO 的优势在于，对机器人的运动轨迹估计更加精确，且具备明确的距离单位。而在单目 VO 中，如果没有已知量度大小作为参考我们只能知道物体在 x/y 方向上移动了 1 个或多个单位，却不知道具体的单位大小。但是，单目视觉里程计比双目里程计更有研究价值，原因如下：单目相机比双目相机更加便于标定校准、当物体距离机器人很远的，双目系统又退化为单目系统。而且当机器人形体很小时，对单目相机被占据更小空间，即使安装了双目相机，但是因为两个相机距离及其近，又可被近似为单目系统。

视觉里程计的主要工作就是计算从图像  $I_t$  到图像  $I_{t+1}$  位置变换  $T_k$ ，然后集

成所有的姿态变换恢复出相机相对于初始位置坐标  $\mathbf{P}_0$  的位姿  $\mathbf{P}_t$ ，这意味着 VO 是一种增量式轨迹重建方法。

### 1.3 本文内容与贡献

本文提出了哪几种方案，主要贡献是：绝对定位之单目深度特征提取与压缩  
增量定位之单目视觉里程计的绝对尺度运算增量定位之单目视觉里程计求解新  
框架简化

1. 基于固定相机高度，基于路面几何约束的单目视觉里程计尺度恢复。
2. 基于单目模型的单目尺度计算：从手动建模到自主学习
3. 传统视觉位姿估计与深度学习尺度恢复的结合
4. 视觉定位及其与视觉里程计的结合

### 1.4 全文架构

## 第2章 数学原理及研究热点数学建模

### 2.1 单目视觉绝对定位数学原理

大规模绝对视觉定位的最先进技术包括基于 2D 图像的图像检索法和基于 3D 结构的 2D-3D 匹配方法。基于 2D 图像的图像检索法根据当前图像特征，从具有地理标记图像的数据库中调取出与当前图像最相似的模版图像，并以该模版图像已知的地理标记（常用全球定位系统，即 GPS 信息）作为自身的地理位置，并返回最相关的数据库图像的姿势。基于 3D 结构的方法采用场景的 3D 模型，使用 2D-3D 匹配与 3D 模型进行相机姿势估计来非常精确地估计相机的 6-DOF 姿势。然而，构建大规模的 3D 模型仍然是一个巨大的挑战。在<sup>[84]</sup> 中通过大量实验证明，当检索到足够多的相关数据库图像时，基于二维图像检索的单目视觉绝对定位可以恢复出精确的相机姿势，大规模的 3D 模型并不是准确的视觉定位所必需的。相比之下，基于二维图像检索的方法只需要一个地理标记图像的数据库，节省了大量的地图构建和维护成本。

### 2.2 单目视觉增量式定位数学原理

在本章我们首先介绍了该方法的背景和表示法，然后详细介绍了我们在道路模型计算中的道路点选择算法。最后，我们使用 RANSAC 计算相机的初始高度，并采用中值滤波器来减少噪声干扰。

在我们提出的方法中，在摄像机高度保持不变，地面局部平面的假设下，摄像机到路面平面的绝对高度  $h_0$  被认为是一个参考。所提出的 MVO 尺度恢复算法的结构如图4.1所示。初始自我运动 ( $R$  和  $t$  在相对尺度上) 和匹配特征由初始 VO 过程给出。道路模型估计模块计算安装摄像机的相对高度，即摄像机的光学中心到地面的距离，并有经过验证的道路点。详细来讲，图像首先被 Delaunay 三角剖分进行分割，每个三角形首先通过考虑深度一致性来确定是否属于道路区域。其余点再通过 Delaunay 三角剖分，并通过考虑道路模型一致性进行选择。通过筛选的路面特征点被用于计算路面模型  $\mathbf{n}_i^T \mathbf{x}_i - \bar{h}_i = 0$ ，以求得摄像机相对高度  $\bar{h}_i$ ，通过与给定相机绝对高度比较恢复相机运动绝对尺度  $s$ ，以求得绝对运动估计  $R$  和  $t$ 。

MVO 旨在求解相机相对于初始位置坐标  $\mathbf{P}_0$  的位姿  $\mathbf{P}_t$ 。两帧图像之间  $I_t$  和  $I_{t-1}$  之间的相机运动  $\mathbf{R}$  与  $\mathbf{t}$  可通过累积求取： $\mathbf{P}_t = \mathbf{P}_{t-1} \mathbf{T}$ ，这里估计运动

$\mathbf{T} = [\mathbf{R}, \mathbf{t}; \mathbf{0}, \mathbf{1}]$ 。两帧图像  $I_{t-1}$  和  $I_t$  之间的匹配特征可以分别表示为  $\mathbf{M}_{t-1}$  和  $\mathbf{M}_t$ 。对于最初的两个帧，最常用的解决方案是求解基本矩阵，因为特征点的三维坐标是未知的<sup>[85]</sup>：

$$\mathbf{M}_{t-1}^T \mathbf{F} \mathbf{M}_t = 0 \quad (2.1)$$

这里  $\mathbf{F} = \mathbf{K}^{-1} [\mathbf{t}]_\times \mathbf{R} \mathbf{K}^{-1}$  是基础矩阵。 $\mathbf{K} = [f_x, 0, c_x; 0, f_y, c_y; 0, 0, 1]$  是标定得到的相机内参，其中  $c_x$  和  $c_y$  是光心像素坐标（主光轴在物理成像平面上的角点）， $f_x, f_y$  是焦距，即左右投影中心（光心）到物理成像平面的距离。 $[\mathbf{t}]_\times = [0, -t_3, t_2; t_3, 0, -t_1; -t_2, t_1, 0]$  是位移矩阵的斜对称矩阵。在公式(2.1)中，基础矩阵  $\mathbf{F}$  首先被解决，然后通过  $\mathbf{F}$  的矩阵分解得到旋转矩阵  $\mathbf{R}$  和位移矩阵  $\mathbf{t}$ <sup>[86]</sup>。我们通过观察得知如果对位移矩阵  $\mathbf{t}$  乘以一个系数时  $s \in \mathbb{R}$ ，公式(2.1)仍然成立。

$$\mathbf{M}_{t-1}^T \mathbf{K}^{-1} s [\mathbf{t}]_\times \mathbf{R} \mathbf{K}^{-1} \mathbf{M}_t = 0. \quad (2.2)$$

我们可以看到，图像  $I_i$  中的度量信息是在同一尺度上的，但 MVO 在没有先验知识的帮助下，不能直接计算出位移向量  $\mathbf{t}$  在帧  $I_i$  中的绝对尺度  $s_i$ ，单纯 MVO 无法获得绝对尺度  $\bar{\mathbf{t}}$ 。这意味着 MVO 可以保证在不同时间计算的相对转换向量  $\bar{\mathbf{t}}$  在同一尺度上，但转换向量  $\mathbf{t}$  的绝对尺度无法通过分解基本矩阵  $\mathbf{F}$  实现。所以强调了绝对先验知识的重要性。

对于接下来的帧，在获取初始运动后，三角测量法计算出特征点  $\bar{\mathbf{x}}_i$  的三维坐标，该坐标与  $\bar{\mathbf{t}}$  的比例相同。下一个相机姿势是由 3D 地图和当前帧通过透视-n-point (PnP) 方法计算出来的<sup>[87]</sup>，通过求解

$$\mathbf{R}, \bar{\mathbf{t}} = \underset{\mathbf{R}, \bar{\mathbf{t}}}{\operatorname{argmin}} \sum_{\bar{\mathbf{x}}_i, \bar{\mathbf{u}}_i} \left| \frac{\mathbf{K}(\mathbf{R}\bar{\mathbf{x}}_i + \bar{\mathbf{t}})}{\bar{\mathbf{x}}_{i3}} - \bar{\mathbf{u}}_i \right| \quad (2.3)$$

其中  $\bar{\mathbf{x}}_{i3}$  是向量  $\bar{\mathbf{x}}_i$  的第三个元素。特征点  $i$  在帧  $I_t$  中的 2D 像素坐标表示为  $\mathbf{u}_i = (u_i, v_i)$ 。这种方法可以保持尺度，但误差会累积。大多数方法，如直接稀疏 odometry (DSO)<sup>[56]</sup>，大规模直接单目 SLAM (LSD-SLAM)<sup>[77]</sup>，ORB-SLAM<sup>[88]</sup>，以及半间接视觉 odometry(SVO)<sup>[58]</sup>，试图通过光束法平差和环路闭合检测技术来对抗尺度漂移，而不是考虑绝对尺度计算。在不与 IMU 和 GPS 等其他传感器融合的情况下，利用周围环境中已知的绝对比例尺来还原比例尺是一种便捷的方法。借助环境中的度量信息  $l$ ，我们根据其相对尺度计算出尺寸  $\bar{l}$ ，并通过  $\frac{l}{\bar{l}}$  计算出尺度系数。 $s = l/\bar{l}$ 。位移矩阵是根据公式  $\mathbf{t} = s\bar{\mathbf{t}}$  计算恢复的绝对位移  $\mathbf{t}$ 。

在本文中，所有的标量、向量和矩阵分别用纯字母（如  $s$ ）、粗体小写（如  $\mathbf{t}$ ）和粗体大写（如  $\mathbf{R}$ ）表示。默认情况下，向量是列式的。矩阵  $\mathbf{R}$  的  $i_{th}$  行和  $j_{th}$  列中的元素用  $R_{ij}$  表示。上面带有条形的变量为相对尺度（例如， $\bar{\mathbf{t}}$ ）。特别是，我

们把一个向量的斜对称矩阵表示为  $[*]_x$  (例如,  $[\mathbf{t}]_x$ )。数学集用希腊大写字母表示。例如,  $\nabla$  表示通过 Delaunay 三角测量分割的三角形集,  $\Theta$  表示属于道路的验证三角形。 $\Omega$  表示初始特征点集, 算法4中验证的道路点集用  $\Gamma$  表示。这些三角形  $\nabla$  的内部区域和顶点分别表示为  $\tilde{\nabla}$  和  $\hat{\nabla}$ 。

## 第3章 动态环境中绝对定位之单目深度特征提取与压缩

如何在动态环境中快速准确地自主定位机器人是机器人路径规划、导航、避障等一些问题的保障。与深度学习相结合的单目视觉定位已经获得了令人难以置信的结果。然而，从深度学习中提取的特征维度巨大，匹配算法也很复杂。如果自动驾驶汽车只能在单一场景中进行训练，将很难满足复杂多变的现实场景。如何减少尺寸与精确的定位是困难之一。本章提出了一种新的方法，通过对动态环境中的大尺度图像训练，来探索满足一定精度要求的深度特征维度。我们从 AlexNet 网络中提取特征并通过 IPCA 减少了特征的维度，更重要的是，我们用核化方法、归一化和形态学等方法处理得到匹配矩阵，消除了图像的冗余匹配与歧义。最后，我们在跨季节的动态环境数据集 Norland 中在线检测最佳匹配序列，证明了经过特征降维后该深度特征仍然能够表达图像信息，仍可以快速定位机器人。

在过去的几年里，为了找到不受大幅度变化影响又能表达场景信息的特征，人们已经研究了各种类型的特征用于本地化<sup>[13, 27, 89, 90]</sup>。图像描述符可分为基于特征的局部特征和整体的图像全局特征。基于特征的描述符在计算机视觉中起着重要的作用。到目前为止，一些手工制作的特征已经获得了一定的成功<sup>[4, 91-93]</sup>。然而，机器人在动态环境中往往无法通过这些手工制作的特征来进行定位。

图像全局描述符根据图像不变特征表达整幅图像信息。最近的结果表明，从卷积神经网络中提取的通用描述符非常强大<sup>[94]</sup>。2012 年，CNN 在 AlexNet 大规模视觉识别挑战赛（ILSVRC）上获得了令人难以置信的准确性<sup>[20]</sup>。这表明，从 CNN 中提取的特征在分类上明显优于手工制作的特征。他们在 120 万张标注的图像上训练了一个名为 AlexNet 的大型 CNN。对于图像根据 AlexNet 提取的特征进行分类，我们也可以根据这些特征来定位机器人。<sup>[21]</sup>研究表明来自 CNN 中层的特征可以更有效地消除数据集偏差。<sup>[24]</sup>比较了来自不同层的特征的性能。他们的结果表明，来自 ConvNet 层次结构中的中间层的特征对一天中的时间、季节或天气条件引起的外观变化表现出鲁棒性。来自 Conv3 层的特征在面对极端外观变化时表现得相当好。然而，CNNs 特征的主要障碍是昂贵的计算成本和内存资源，这对实时性能是一个很大的挑战。如果我们将巨大维度的图像特征与记录的数据集逐一在线比较，将耗费大量的时间，因此有必要降低 CNNs 特征的计算成本和内存资源。所以有必要降低这些向量的维度。<sup>[25]</sup>将 CNN 特征的冗余数据压缩成一个可控的比特数。通过应用简单的压缩和二值化技术来减少最终的描述符，以便使用汉明距离进行快速匹配。压缩意味着丢失一定量的信息。但

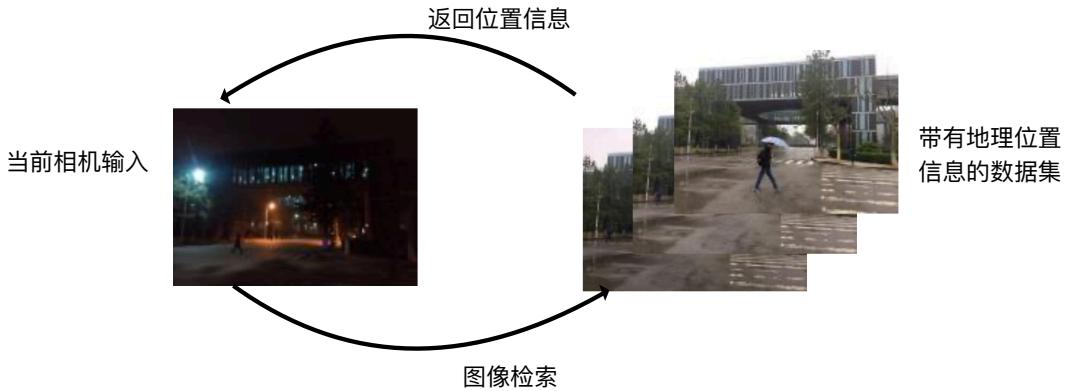


图 3.1 基于图像检索的单目绝对定位示意图

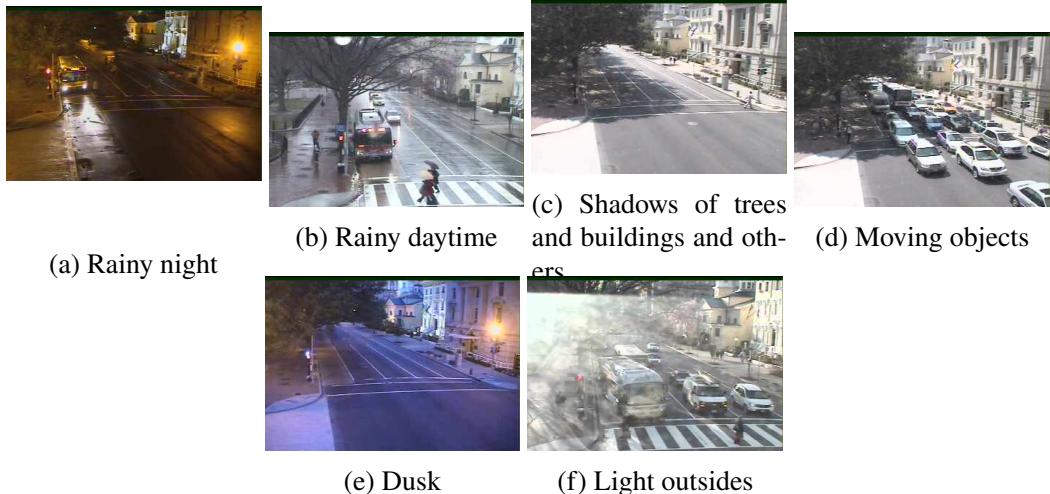


图 3.2 复杂多变的动态环境。

是，我们可以尽可能地保留数据之间的重要关系。我们通过在数据分析中广泛使用的增量 PCA(Principal Component Analysis) 来实现这一目的。在本章中，我们提出了一种新型的机器人跨季节动态环境定位算法。

本章的主要贡献有：1) 我们通过深度学习特征的维度减少，提出了一种新型的动态环境下的定位系统。2) 我们减少了从 AlexNet 中提取的特征的维度。它不仅可以加快计算速度，而且可以减少从数据集引起的图像与大多数在线图像匹配的混乱匹配。3) 代替复杂的数据关联图，通过对匹配矩阵进行形态学处理，在线找到最佳匹配序列。

### 3.1 单目视觉绝对定位深度特征提取方法

在本文中，我们提出一个新的视觉定位图像特征提取方法，它结合 CNNs 网络特征表示的优势，在不同季节等环境条件下执行基于单目视觉的鲁棒定位，正如方法框架图（图3.1.1）所示，我们的工作过程如下：1) 从 AlexNet 的 Conv3 中提取特征，并通过 IPCA 进行图像特征降维。2) 将在线图像的向量与已存数据集的向量通过余弦距离逐一匹配。通过核化方法对匹配矩阵进行归一化，以减

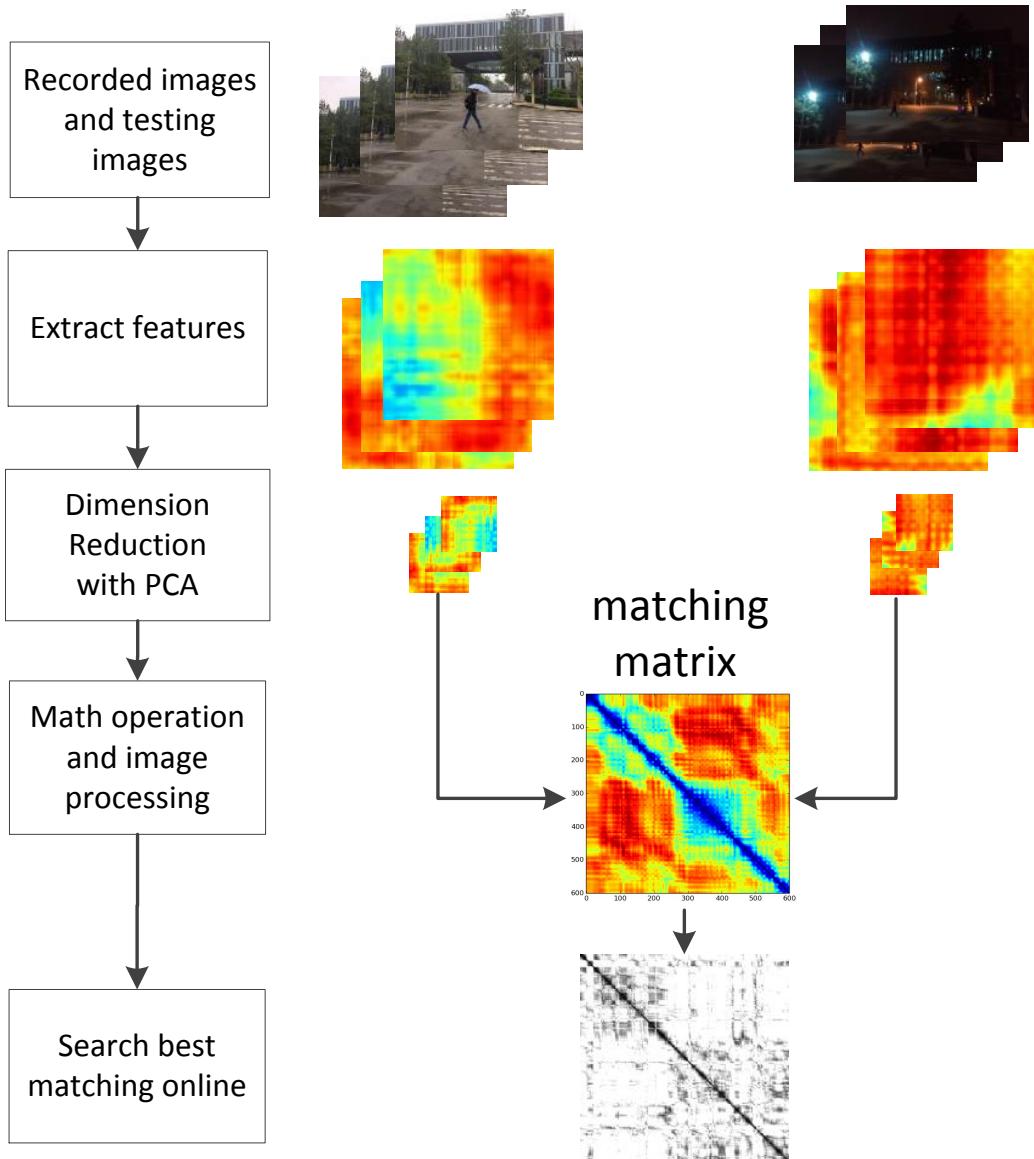


图 3.3 基于 AlexNet 的深度特征单目图像匹配定位框架。

少因大部分在线图像匹配的数据集混乱造成的歧义。3) 对匹配图像进行图像处理, 包括图像二值化、图像侵蚀等。4) 设置参数, 通过 RANSAC (随机样本共识) 在线寻找最佳匹配序列。本文的研究过程如下。在第 3 节中, 我们描述了我们的方法的细节。在第 4 节中, 我们在 Norland 数据集上做了一个动态环境下的在线定位实验。在第 5 节中, 我们对结果和未来的工作进行了讨论。

### 3.1.1 基于 AlexNet 网络的深度学习特征提取

本文算法框架如算法所示。关于地图框架, 我们采用的是纯图像检索, 但数据集是按照图像传入时间的先后顺序存储的。不仅可以保证精度, 同时保证了计算效率。我们选择 AlexNet 的 Conv3 中的特征作为我们的整体图像描述符。Conv3 的维度是 64896, 也就是说, 一张图像显示为 64896 维的向量  $\mathbf{f}$ 。我们用每

**Algorithm 1:** 视觉定位算法

---

**Input:** 视觉地图  $\{[\mathbf{f}, \mathbf{l}]\}_{i=1}^n$ , where  $\mathbf{f}$  is the feature vector of image on location extracted from AlexNet;  $\mathbf{l}$  and  $n$  is the size of visual map; Current image sequences  $\{\mathbf{I}\}_{j=t-m+1}^t$ , where  $m$  is the sequence size; last robot location  $\hat{l}_{t-1}$

**Initialize:**  $\hat{l}_t = \hat{l}_{t-1}$  **Output:** 机器人当前位置  $l_t$

**for**  $t = 2$  to  $n$  **do**

- 计算图像  $\{\mathbf{I}\}_{j=t-m+1}^t$  的特征表示  $\{\hat{\mathbf{f}}\}_{j=t-m+1}^t$
- 计算匹配矩阵  $\mathbf{M}$ , 其中每个元素  $\mathbf{M}_{ij} = \mathcal{F}(\mathbf{f}_i, \hat{\mathbf{f}}_j)$
- 矩阵归一化  $\mathbf{M}_{ij} = \frac{255(\mathbf{M}_{ij} - \mathbf{M}_{min})}{\mathbf{M}_{max} - \mathbf{M}_{min}}$  take  $\mathbf{M}$  as a gray image  $\mathbf{I}_g$
- 用合适的阈值对矩阵  $\mathbf{I}_g$  进行二值化, 得到  $\mathbf{I}_b$
- 对匹配图像  $\mathbf{I}_b$  进行处理得到  $\mathbf{I}_m$
- 使用 RANSAC 法得到最佳匹配曲线  $y = kx + b$  on  $\mathbf{I}_m$
- 在视觉地图中当前图像的最佳匹配特征是  $\mathbf{f}_{km+b}$
- 所以当前位置是  $\hat{l}_t = l_{km+b}$

**end**

return  $\hat{l}_t$

---

张图像的位置建立可视化地图  $\{[\mathbf{f}, \mathbf{l}]\}_{i=1}^n$ 。所以当前的图像序列表示为  $\{\mathbf{I}\}_{j=t-m+1}^t$ 。为了减少高维度向量运算耗时, 我们通过 PCA (主成分分析) 来减少深度特征维度。虽然图像特征描述在一定程度上丢失了信息, 但同时减少了因天空、地面和树木等数据集中因背景信息而导致的模糊匹配。在线图像的向量将通过余弦距离与数据集向量进行逐一比较, 然后得到匹配矩阵  $\mathbf{S}$ , 其组成元素是位于(0,1)之间的浮点数。通过核化方法对匹配矩阵进行归一化处理, 以减少因与大多数在线图像匹配的数据集混淆而造成的歧义。将匹配矩阵保存为灰色图像, 然后通过合适的阈值将其转换为二进制图像。此外, 我们对匹配的灰度图像进行了核化与侵蚀, 以消除接近最佳匹配的相似匹配的影响。我们尝试调整参数, 然后通过 RANSAC 在线寻找最佳匹配序列。匹配矩阵中当前图像的最佳匹配特征为  $\mathbf{f}_{km+b}$ 。那么当前图像在视觉图中的最佳匹配图像为  $\mathbf{l}_{km+b}$ 。

我们采用 Tensorflow 深度学习框架, 从 AlexNet 的 Conv3 中提取特征, 作为每张图像的全局特征描述符。Conv3 层的向量维度为 64896, 也就是说每一张图像的深度特征由 64896 维的向量来表示。AlexNet 的不同层特征适用于不同的机器视觉任务。AlexNet ConvNets 中不同层的向量尺寸见表3.1<sup>[20]</sup>。层次较高的层在语义上更有意义<sup>[24]</sup>, 但同时失去了过多的语义信息, 而对同类型场景无法区分。所以使用网络的那一层作为图像特征表示具有重要意义。来自 Conv3 层的特征在剧烈的外观变化条件下仍表现较好。此外, fc6 和 fc7 在视角变化方面优

于其余层。然而，当外观变化时，fc6 和 fc7 完全失败。基于以上研究，我们用 AlexNet 的 Conv3 的特征来表达图像。

表 3.1 AlexNet 不同层特征维度

Layer	dimensions	Layer	dimensions
Conv1	96×55×55	Conv4	384×13×13
pool1	96×27×27	Conv5	256×13×13
Conv2	256×27×27	fc6	4096×1×1
pool2	256×13×13	fc7	4096×1×1
Conv3	384×13×13	fc8	1000×1×1

### 3.1.2 基于 AlexNet 网络的单目视觉绝对定位深度特征降维

表 3.2 信息保有率与参数 n\_components 的关系。

n_components	Ratio	n_components	Ratio
316	99%	51	93%
187	98%	44	92%
136	97%	38	91%
99	96%	33	90%
76	95%	29	89%
62	94%	25	88%

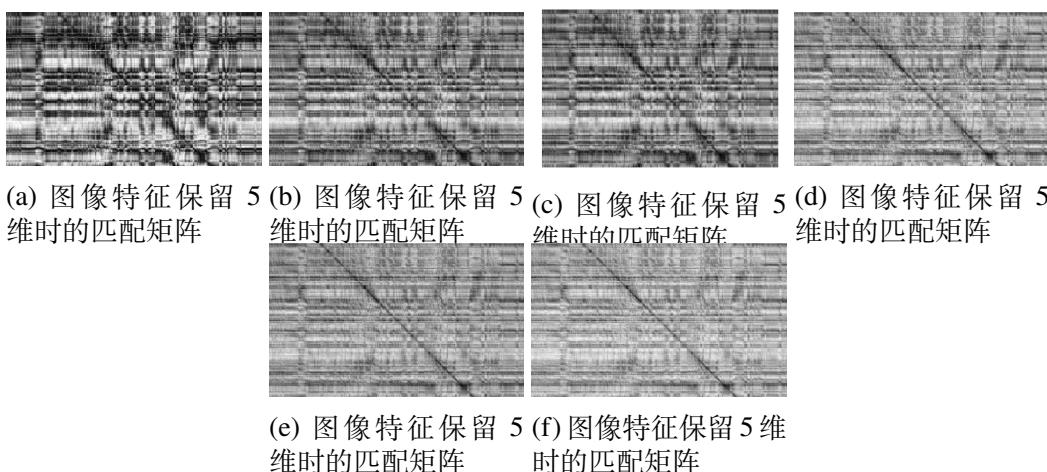


图 3.4 深度特征分别保留 5, 10, 20, 33, 51, 99 维时的匹配矩阵

我们在 Norland 数据集上进行测试，以选择平衡计算消耗和准确性的图像特征维度。Norland 数据集是：\*\*\*\*\* 我们选择 300 张春季的图像序列进行深度

特征训练，500张秋季的图像序列作为测试。我们在 scikit-learn 中使用 IPCA 进行大量的图像匹配。PCA 是高维数据分析的重要手段之一。PCA 通过线性变换将高维数据转化为低维数据。AlexNet 各层的维度如表 1\*\*\*\*\* 所示。从表中可以看出，我们保留的维度越多，获得的信息就越多，但也很耗时，所以首要任务是由参数 `n_components` 为参照以确定每个向量保留多少维度，该参数与主信息 Ratio 之间的关系列在表3.2中。一般来说，在保持一定精度的情况下，我们最好保持至少 90% 的主信息比。我们还对不同维度的匹配结果进行了比较，比较结果如图 ?? 所示。随着特征维度的降低，匹配曲线变得模糊，当图像特征维度降低至 20 以下时无法检测出最佳匹配线。而保留 33 个维度时匹配矩阵足够清晰，同时也节省了计算量。综上所述，我们选择 33 个维度的向量作为图像描述符。

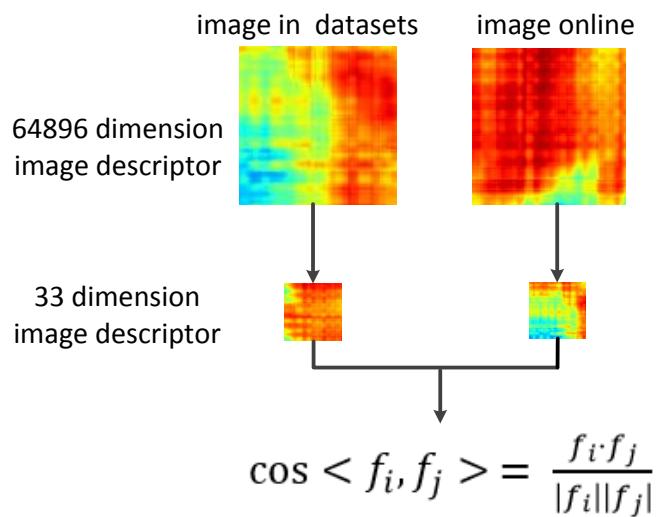


图 3.5 图片特征描述子降维匹配矩阵示意图

### 3.1.3 单目视觉绝对定位矩阵核化处理与归一化

我们的任务是准确地找到匹配矩阵中的最佳匹配线。我们利用数学变换使这条线更加清晰比如核化方法，包括对匹配矩阵的元素进行反演和指数化。选择这种方法的原因如下。1) 2幅图像之间的余弦距离，即匹配矩阵元素与图像相似度不是完全正相关关系。2) 核化方法会扩大假阴性与真阳性之间的距离。图3.6是由余弦距离计算出的函数曲线比较，如公式 (3.1) 所示，核化距离如公式 (3.2) 所示。红色圈线代表两个图像向量的余弦距离。绿色星线代表的是核法距离右图是两个图像向量的核法距离。我们可以看到，核化法可以增强两个不同地点之间的差异。最佳匹配图像的颜色会显示为黑色，不同的地方会显示为白色，如图3.7所示。更重要的是，通过内核法对匹配矩阵进行归一化处理，可以减少大部分在线图片匹配数据集混乱造成的歧义。将匹配矩阵保存为灰度图像，以便进行后续的处理，包括形态变换和二值化。我们对匹配矩阵进行了归一化处理，范围为 0 ~

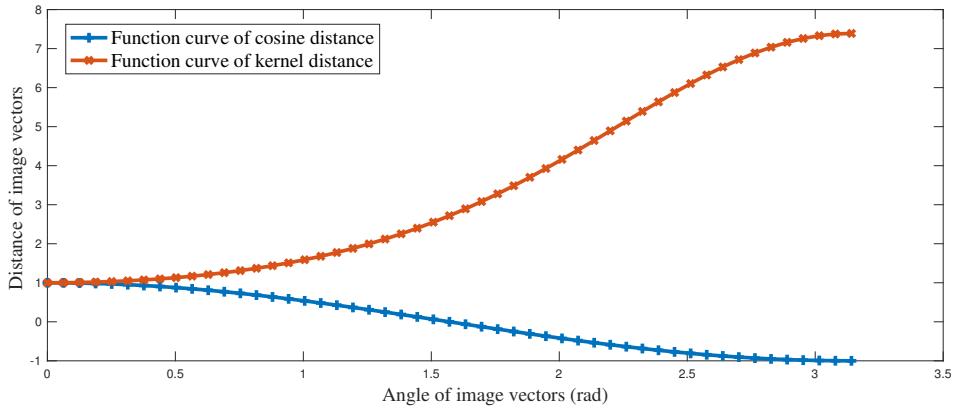


图 3.6 矩阵核化运算前后函数曲线对比。

255，公式为 (3.2)，经过核方法后，匹配矩阵就变得更加明显了。

我们在 Norland 数据集中选取了 3000 张在同一个地方拍摄的春季和冬季图像，测试了核化方法。由于两个图像序列起点是同一地点，因此，在对角线上出现的匹配线即最佳匹配序列。匹配结果如图3.7所示，我们通过余弦距离  $\cos \langle \mathbf{f}_i, \mathbf{f}_j \rangle$  将在线图像与记录的数据集图像逐一进行匹配。因此，一条线出现在对角线上，为最佳匹配序列。匹配结果如图3.7(b)所示。我们通过余弦距离  $\cos \langle \mathbf{f}_i, \mathbf{f}_j \rangle$  将在线图像与地图数据集图像逐一匹配。然而，匹配的图像显示为图??，这意味着错误匹配和最佳匹配之间产生了混淆。然而，通过核方法和归一化，对角线变得很明显，如图3.7(b)所示。最后，将匹配矩阵保存为灰度图像，通过适当的阈值将其转换为二进制图像。

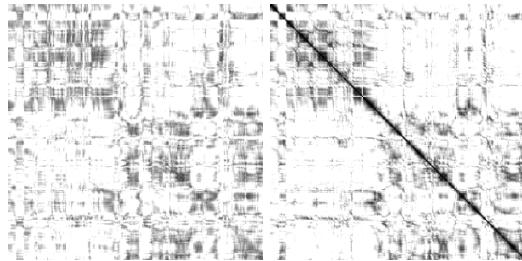
$$\cos \langle \mathbf{f}_i, \mathbf{f}_j \rangle = \frac{\sum_{i=1}^{33} a_i b_i}{\sqrt{\sum_{j=1}^{33} a_j^2} \sqrt{\sum_{k=1}^{33} b_k^2}} \quad (3.1)$$

$\mathbf{f}_i = \{\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_{33}\}$ ,  $i \in D$ ,  $D$  is set of datasets images  $\mathbf{f}_j = \{\mathbf{b}_1 \mathbf{b}_2 \dots \mathbf{b}_{33}\}$ ,  $j \in O$ ,  $O$  is set of online images

$$\mathbf{M}_{ij} = \frac{255 (\mathbf{M}_{ij} - \mathbf{M}_{min})}{\mathbf{M}_{max} - \mathbf{M}_{min}} \quad (3.2)$$

## 3.2 基于 AlexNet 的单目视觉绝对定位实验

我们的实验旨在展示我们的方法对图像进行深度特征降维以及匹配矩阵核化处理后对定位效果的影响。我们的方法能够(i)在跨季节的场景中进行定位，忽



(a) 欧式距离匹配矩阵 (b) 核化处理后的匹配矩阵

图 3.7 Norland 数据集春-冬跨季节匹配矩阵

略动态物体、不同天气和季节变化。(ii) 节省时间和计算成本。我们在 SeqSLAM 中使用的公开数据集 Norland 上进行了评估<sup>[13]</sup>。采用该数据集中 64x32 的灰色图像图像频率为 1 帧/秒，如果我们的方法在这种不清晰和微小的低信息含量图像中仍然有效，那么它可以节省大量的时间和计算消耗。我们可以看到在图3.7(b)中，最佳匹配线已经很明显了。我们通过经典的 RANSAC 算法找到它的数学模型，在数据集中找到相应的索引。我们选择了 300 张秋季图片作为训练地图，对春



(a) 核化及标准化后的匹配矩阵

(c) 匹配矩阵中的最佳匹配查找

图 3.8 Norland 数据集秋-春跨季节匹配矩阵

季进行视觉定位，如图??所示，我们选择了一个 300 张秋季图片序列作为地图，由春季图像进行在线定位。我们可以看到，从 AlexNet 的 Conv3 中提取的特征并没有影响匹配结果。相反，如图??所示，减少了背景信息的影响。图??是匹配图像的二值化结果。你可以看到，大部分干扰信息已经被擦掉了。我们可以看到，在图??中，绿色的线只是这段时间的最佳匹配。当前图像在匹配矩阵中的最佳匹配特征为  $\mathbf{f}_{km+b}$ 。那么当前图像在视觉图中的最佳匹配图像是  $\mathbf{l}_{km+b}$ 。在图中3.11，我们绘制了 3 条线来评估我们方法的在 3000 张匹配图像上的实验结果。其中蓝线是图像索引 Ground Truth，红线是匹配图像索引，黄线是匹配索引误差。

### 3.3 本章小结

我们的论文提出了一种在动态环境中对机器人进行跨季节定位的新型算法。我们从 AlexNet 的 Conv3 中提取特征，该深度特征的匹配精度优于传统手工提取特征，通过 PCA 减少维度是一个新的尝试。

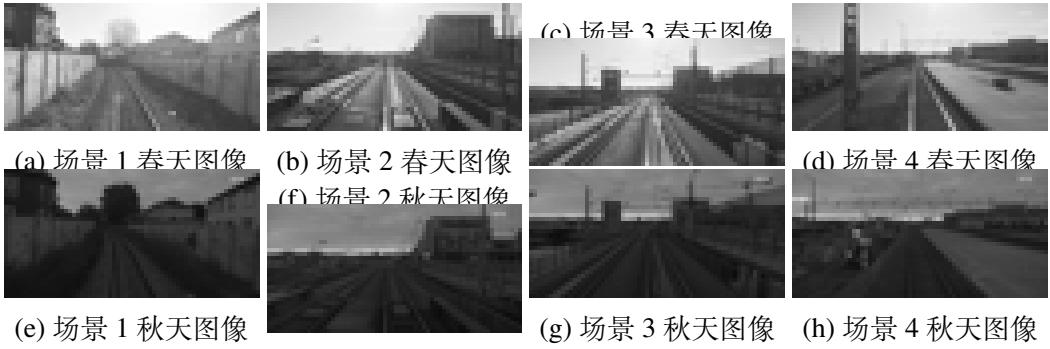


图 3.9 匹配易失败场景春秋对比图

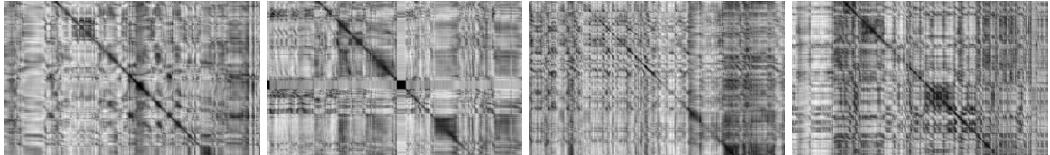


图 3.10 部分序列匹配矩阵

事实证明，Conv3 是机器人本地化的最佳选择。相对于更底层的深度特征它具有更快的计算速度，并保留了一定语义信息，减少了图像混乱匹配。我们通过核化法将在线图像向量与地图向量逐一进行比较。我们从 AlexNet 的 Conv3 中提取特征，通过 PCA 减少维度特征的匹配精度仍优于传统手工提取特征。

这个过程扩大了正确匹配和错误匹配之间的差异。此外，通过适当阈值核化处理、图像放大和侵蚀，将复杂的数据关联图转化为简单的图像处理。在序列匹配方面，我们采用经典的 RANSAC 算法，在短时间内找到最佳匹配线。我们的实验结果表明，深度特征降维是加快计算速度和减少混乱匹配的好主意，且该算法对季节变换、动态环境、天气变化等都有很强的适应性。本章算法的局限性主要受限制于图像采集设备。在完全黑暗的环境下，由于没有主动光源，图像信息很难表现出来，如图??所示。实际上在第 1872 至 2016 张图像中没有对应的匹配线。在 AlexNet 的 Conv3 中，对于难以表达深度特征的图像，匹配图像显示为黑色区域，我们将考虑加入激光的辅助。此外，我们还将研究特征维度与定位精度之间的更具体的影响关系。我们希望通过训练一个针对性的网络结构，得到不受季节变化、天气变化、动态环境等因素的影响的全局图像特征。

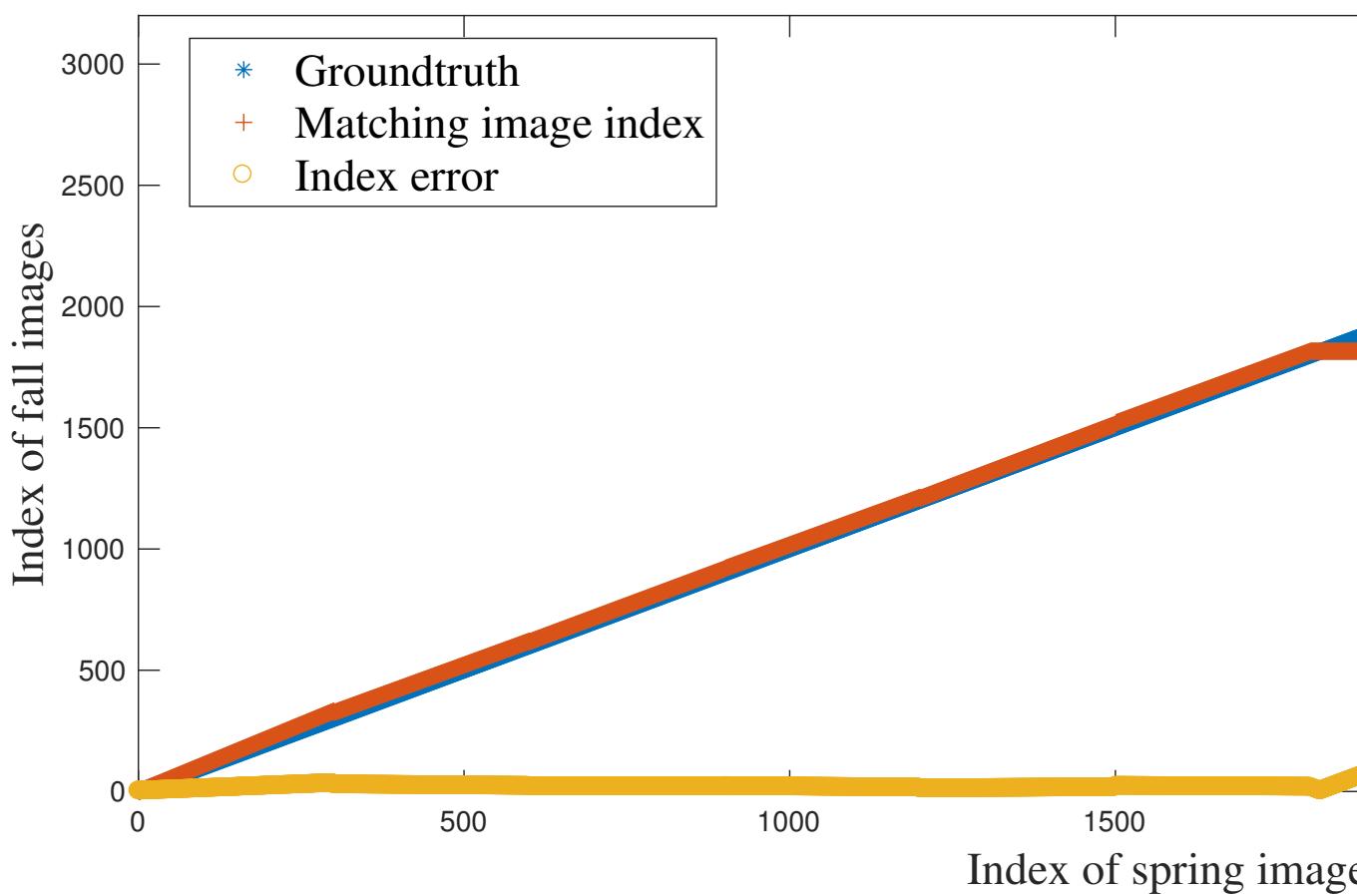


图 3.11 3000 张图片匹配结果



图 3.12 Norland 数据集中隧道图像采集样例

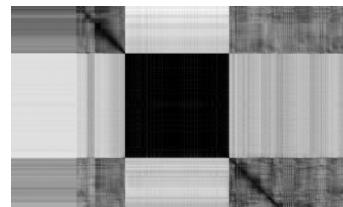


图 3.13 春季数据集中包含隧道片段的匹配矩阵

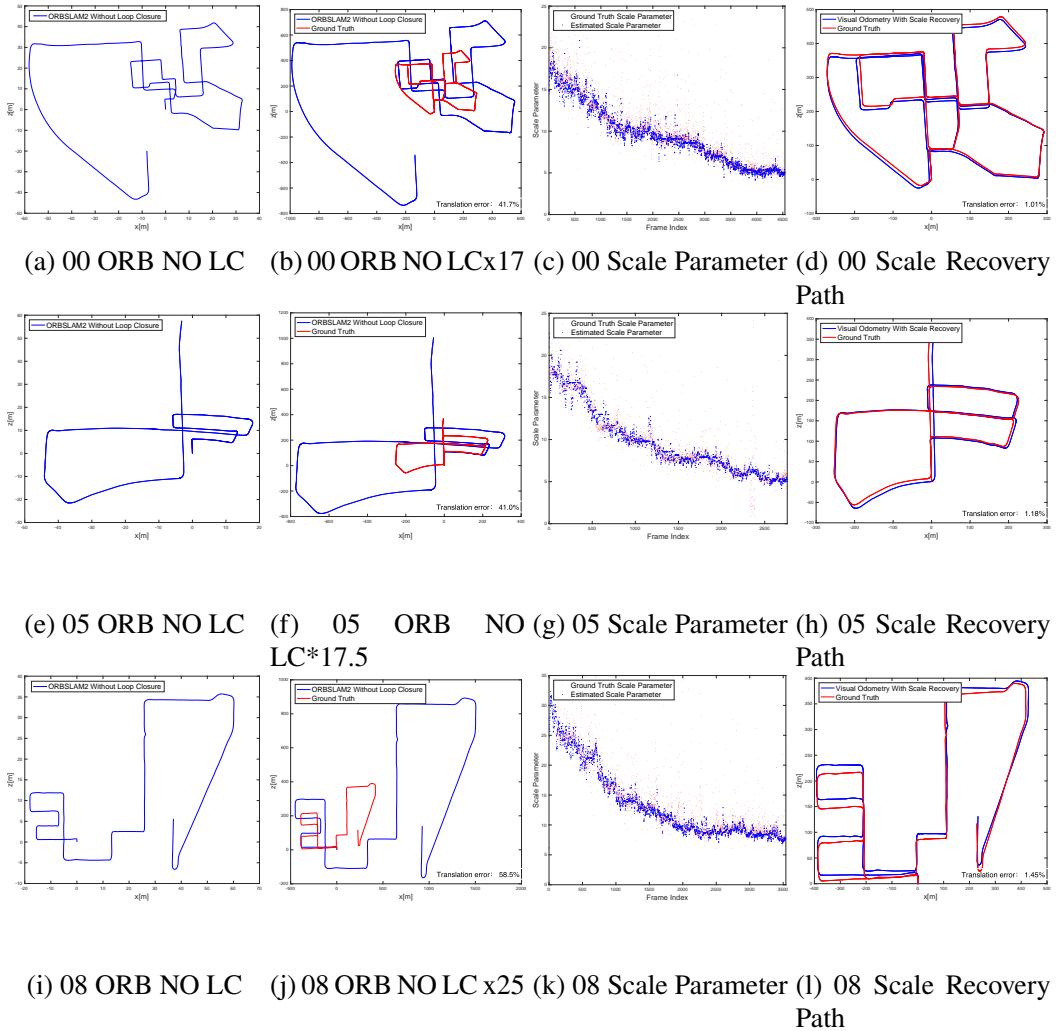


图 3.14 在 KITTI 数据集序列 00、05、08 上与无 LC 的 ORB-SLAM2 的尺度恢复性能比较。第一列中的三个数字是没有环路闭合的单目 ORB-SLAM2 轨迹，显然尺度发生了明显的错误。第二列为通过对应序列的前 100 帧尺度校正乘以 17.0、17.5、25.0 三个固定尺度参数得到的轨迹。第四列是第三列中乘以我们估计的尺度参数得到的轨迹。

## 第4章 增量式定位--单目视觉里程计几何尺度估计

近年来，人们提出了多种视觉里程表恢复方法。从尺度 **scale** 概念来看，解决思路大致可以分为两类：相对尺度校正和绝对尺度恢复。前者致力于将机器人自我运动保持在同一尺度下，以保持全局一致性；后者借助给定的绝对度量参考，计算每一帧的真实尺度。我们将固定在机器人上的相机至路面高度作为绝对尺度参考，提出了一种基于路面几何模型的单目视觉里程计尺度恢复算法。在从未探索过的环境种，单目视觉里程计是机器人实现自我定位和自主导航的核心模块，而尺度恢复是弥补单目视觉里程计不可或缺的功能，因为它弥补了相机所丢失的度量信息造成的尺度漂移等问题。当将相机高度视为绝对尺度参考系时，尺度恢复的精度取决于路面特征点的筛选和路面模型的建立。大多已有方法将这两个问题独立解决：他们的路面特征点筛选是基于路面颜色信息或者已知的图像固定区域，并没有利用两个方法的优势，将其进行有利结合。

如图4.1所示：单目图像序列、图像相对运动位姿（ $\mathbf{R}$  和  $\bar{\mathbf{t}}$ ）、匹配特征点作为输入，由 VO 初始化操作提供。每一帧都会被 Delaunay 三角剖分进行分割，相邻图像的匹配特征点即三角形的顶点，每个三角形都会通过深度一致性约束进行筛选以判断其是否属于路面。最终筛选得到的路面特征点会用来帮助恢复相机运动  $s$ 。最终，相机绝对尺度下的运动位姿估计  $\mathbf{R}$  和  $\mathbf{s}\bar{\mathbf{t}}$  得到了解决。在第一次 Delaunay 三角剖分后的结果图中（左中），所有的点都是通过初始视觉里程测量过程得到的匹配特征点，蓝色的点是选择的满足深度约束的特征点（如 III-B1 所述），红色的点是不满足深度约束的特征点；在第二次 Delaunay 三角剖分后的结果图中（左下），蓝色的点是满足道路模型约束的特征点（如 III-B1 所述），初步通过筛选的路面特征点被用于计算路面模型  $\mathbf{n}_i^T \mathbf{x}_i - \bar{h}_i = 0$ ，该模型又反过来结合路面模型约束进行再一次的路面特征点筛选。

我们提出迭代求解道路点选择和道路几何模型计算：我们考虑估计的道路几何模型来检测道路点；考虑检测到的道路点在线更新道路几何模型。筛选出的路面特征点会用于估计路面几何模型，路面几何模型又是路面特征点筛选的一个几何约束；同时，我们采用路面几何信息代替路面颜色信息进行特征点筛选，使得系统更加鲁棒，这两个问题可以互相受益。此外，对于道路点的选择，也采用了新的解决方案处理这个关键任务。详细地，我们利用 Delaunay 三角剖分将图像分割成一组以匹配特征点为顶点的三角形。每个三角形通过考虑深度和道路模型一致性两个约束条件来确定是否属于道路区域。此外，我们通过随机样本共识 (RANSAC)<sup>[95]</sup> 估计具有验证道路点的几何道路模型，并通过中值滤波器去

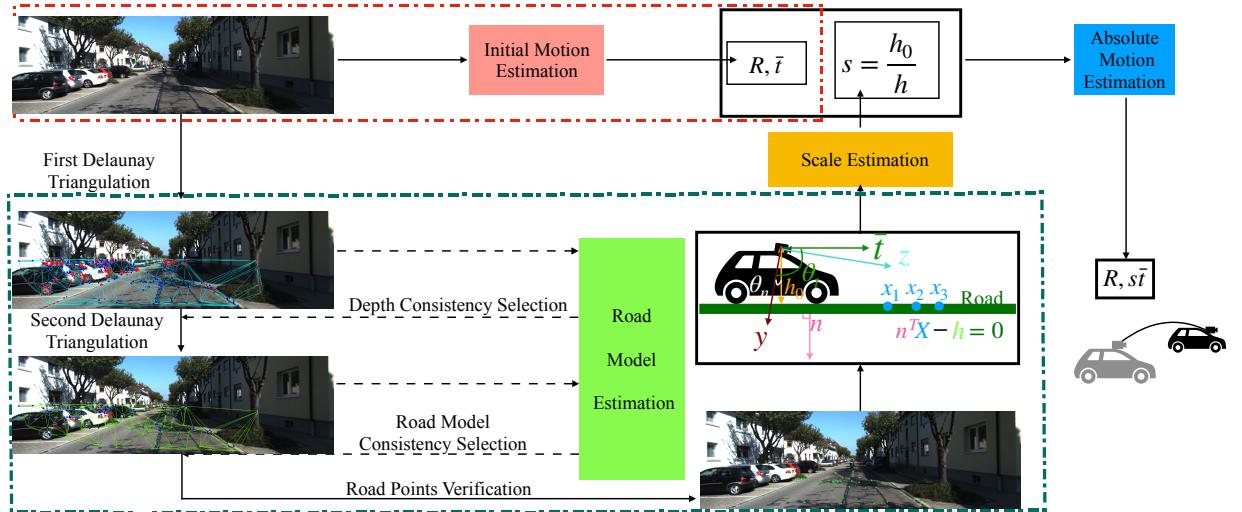


图 4.1 基于路面几何模型的单目尺度恢复算法框架

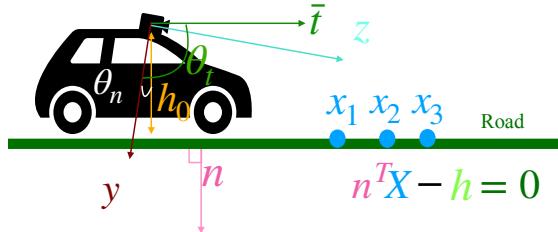


图 4.2 路面模型法向量  $n$  与车辆平移矩阵  $t$  方向垂直。道路俯仰角是道路几何模型的一个中心单元，可以从自我运动来估计。通过考虑道路几何模型。 $\Theta_t, \Theta_n, h_0, n, t$

除尺度噪声。对于道路建模，我们用不同的 RANSAC 参数检验了翻译误差。研究了中值滤波器和均值滤波器的比较。此外，在 VO 中，我们根据深度一致性和道路模型约束两个规则来选择道路点。综上所述，我们工作的主要贡献如下：

1. 本章我们提出了一种基于路面几何信息的鲁棒单目视觉里程计尺度估计方法，将道路点选择和道路几何模型计算结合为一个问题——基于道路几何模型检测道路点，并根据检测到的道路点更新道路模型。
2. 本章提出了一种新的道路点选择策略，该策略受深度一致性和道路模型一致性的约束，并结合 Delaunay 三角剖分方法提出了该策略。该方法提高了 MVO 的精度，实现起来很简单。我们公开了可用的源代码<sup>①</sup>。

## 4.1 道路几何模型计算

毫无疑问，图像特征点集中只有一部分属于道路区域，所以需要对属于路面的特征点进行筛选。道路区域检测的目的是计算包含计算高度摄像头的道路模型公式。本论文以道路几何结构的约束，而非颜色信息来侦测道路区域，因为几何结构比颜色信息更为鲁棒且几何模型在每一帧都会更新。

<sup>①</sup> <https://github.com/TimingSpace/MVOScaleRecovery>

道路模型估计模块如图所示，4.2。我们的道路点选择和道路几何模型计算是迭代进行的。它们可以相互受益。通过考虑道路几何模型来检测道路点，然后通过选择的三维道路特征点来更新道路几何模型，经过验证的道路点用绿色标记。详细来说，我们提出利用初始运动来粗略估计道路特征的初始选择中的摄像机俯仰角。利用 Delaunay 三角剖分<sup>[96]</sup> 将已知三维坐标的特征点划分为多个三角形区域。然后，我们根据深度一致性（算法2和3）剔除道路异常点，被验证的道路点用蓝色标记，如图4.2所示。然后对剩余的点再次使用 Delaunay 三角测量，我们根据道路模型的一致性继续剔除道路异常点（算法4）。

### 4.1.1 基于深度一致性的路面特征点筛选

#### 4.1.1.1 直接剔除法

---

##### **Algorithm 2:** 基于深度一致性的特征点筛选 (直接剔除法)

---

**Input:** 准路面特征点集  $\Omega = \{f_0, f_1, \dots, f_n\}$ ，每个点像素坐标  $(u_i, v_i)$  及其深度  $\bar{d}_i$

**Output:** 认证路面特征点集  $\Lambda \subset \Omega$

依据其像素位置，对特征点  $\Omega$  进行三角剖分，得到三角形集合

$$\nabla = \{\nabla_0, \nabla_1, \dots, \nabla_m\}, \nabla_i = \{f_i \in \Omega, f_j \in \Omega, f_k \in \Omega\}$$

设置  $\Lambda = \Omega$

**for**  $t_i = 0$  to  $m$  **do**

```

for  $\forall \{f_i, f_j\} \subset \nabla_i$  do
    if  $(v_i - v_j)(d_i - d_j) > 0$  then
        |  $\Lambda = \Lambda - \{f_i, f_j\}$ 
    end
end

```

**end**

---

匹配的特征点的三维坐标在初始 VO 处理后就可以得到。它们与实测尺度的坐标保持相同的几何结构。第  $I_t$  帧图像的初始特征点表示为  $\Omega = \{f_0, f_1, \dots, f_n\}$ 。每个点的二维像素坐标为  $\mathbf{u}_i = (u_i, v_i)$ ，相对尺度下的深度和三维坐标分别表示为  $\bar{d}_i$  和  $\bar{\mathbf{x}}_i$ 。路点选择方法以  $\mathbf{u}_i$  和  $\bar{d}_i$  为基础。首先，根据特征点在帧  $I_t$  中的二维投影坐标  $(u_i, v_i)$ ，用 Delaunay 三角测量法将特征点划分成一系列三角形区域  $\nabla$ ，其顶点就是特征点。如果一个特征点  $f_i$  满足路面几何模型，那么它的深度为

$$\bar{d}_i = \frac{\bar{h}_i f_y}{v_i - c_y}, \quad (4.1)$$

所以我们可以得出结论:  $\bar{d}_i \propto \frac{1}{v_i}$ 。此外, 对于路面上的任意两个特征点  $f_i = (u_i, v_i, \bar{d}_i)$  和  $f_j = (u_j, v_j, \bar{d}_j)$ , 有以下关系必然成立:

$$\sigma(i, j) = (v_i - v_j)(\bar{d}_i - \bar{d}_j) \leq 0. \quad (4.2)$$

如果  $\sigma > 0$ , 则至少有一个特征不属于道路, 或者其深度  $\bar{d}_i$  不正确。对于这两种情况中的任何一种, 我们选择将其排除。然而  $f_i$  和  $f_j$  中哪一个点应该被删除是不确定的。我们提出了两种选择机制: 一种是算法2中所示的直接删除法, 另一种是算法3中所示的最大团和集成学习法。

#### 4.1.1.2 最大团和集成学习剔除法

---

##### **Algorithm 3:** 基于深度一致性的路面特征点筛选 (最大团筛选)

---

**Input:** 准路面特征点集  $\Omega = \{f_0, f_1, \dots, f_n\}$ , 每个点的像素坐标  $(u_i, v_i)$  及

其深度  $\bar{d}_i$

**Output:** 认证路面特征点集  $\Lambda \subset \Omega$

设置准路面特征点属于路面的阈值  $p_s$ , 根据像素坐标对准路面特征点集  $\Omega$  进行三角剖分得到三角形集合

$\nabla = \{\nabla_0, \nabla_1, \dots, \nabla_m\}, \nabla_i = \{f_i \in \Omega, f_j \in \Omega, f_k \in \Omega\}$  以  $\text{Omega}$  为节点  $V$ , 即  $V = \Omega$ , 根据  $\nabla$  生成无向  $G = V, E$ , 图中的最大团为  $\nabla$  **for**  $\forall \nabla_i \in \nabla$

**do**

设置每个特征点的投票初始值为 0:  $\text{vote}(f_i) = 0, \forall f_i \in \Omega$  **for**

$\forall \{f_i, f_j\} \subset \nabla_i$  **do**

根据(4.3)计算特征点属于最大团的概率 **if**  $p(f_i|\nabla_i) \geq p_s$  **then**

|  $\text{vote}(f_i) = \text{vote}(f_i) + 1$

**end**

**else**

|  $\text{vote}(f_i) = \text{vote}(f_i) - 1$

**end**

**end**

$\Lambda = \{f_i\}, \forall f_i \in \Omega \text{ and } p(\tilde{f}_i) > 0.5$

**end**

---

算法2是一种最简单直接的算法, 它将不满足方程(4.2)的点  $f_i$  和  $f_j$  都删除。此外, 一个特征点可能存在于多个三角形中, 这可能会导致  $\sigma$  被重复计算, 如图4.3所示。为了避免冲突, 我们提出了另一种基于最大团和集成学习法的特征点选择方法。此外, 一个特征点可能存在于多个三角形中, 这可能导致  $\sigma$  被重

|c|c|c|c|

表4.1 两点之间（四种情形）的势函数定义。在第1-2列中，0表示该点被移除，1表示该点被保留。

A	B	$P_e$ ( $\sigma \geq 0$ )	$P_e$ ( $\sigma \leq 0$ )
0	0	3	1
0	1	2	2
1	0	2	2
1	1	0	4

复计算。为了避免冲突，我们提出了另一种基于最大团和集成学习法的特征点选择方法。如图4.3所示，将每个三角形视为一个最大团，一个点可能存在于三个最大团中。这个点是否从路面特征点集合中删除由所有最大团的投票决定。每个最大团内的计算细节在算法3中描述，基于非定向图  $G$ ，最大团  $\nabla_i$  中的特征点  $f_i$  属于道路区域的概率  $p(f_i|\nabla_i)$  可以估计为条件概率分布。

$$p(f_i|\nabla_i) = \frac{\sum_{\check{f}_i, f_j, f_k} P_m(f_i, f_j, f_k | \sigma_{ij}, \sigma_{jk}, \sigma_{ik})}{\sum_{f_i, f_j, f_k} P_m(f_i, f_j, f_k | \sigma_{ij}, \sigma_{jk}, \sigma_{ik})} \quad (4.3)$$

其中  $\check{f}_i$  表示某点属于道路。 $P_m$  为各最大组的势函数，其计算方法为：

$$\begin{aligned} P_m(f_i, f_j, f_k | \sigma_{ij}, \sigma_{jk}, \sigma_{ik}) &= P_e(f_i, f_j | \sigma_{ij}) \\ &\cdot P_e(f_j, f_k | \sigma_{jk}) \cdot P_e(f_i, f_k | \sigma_{ik}) \end{aligned} \quad (4.4)$$

其中  $P_e(f_i, f_j | \sigma_{ij})$  表示当观测值为  $\sigma_{ij}$  时， $P_e(f_i, f_j | \sigma_{ij})$  的有效性概率，当观测值是  $\sigma_{ij}$  时，由公式(4.2)计算出来的。我们为图中的每一条边定义势函数  $P_e(f_i, f_j | \sigma_{ij})$ ，如图4.3中右侧表格所示。两点之间有四个删除或保留的决策，1表示倾向于保持，反之为0。在集成学习的思想下，每个最大组根据概率  $p(f_i|\nabla_i)$  对特征点进行投票。只有在  $\text{vote}(f_i)$  中有一半以上的赞同率， $f_i$  才会被验证为道路特征点。

#### 4.1.2 基于路面模型一致性的特征点筛选

我们根据道路模型的一致性，包括角度和距离的一致性，继续选择属于道路的三角形。由于道路模型约束选择是深度一致性选择之后的步骤，我们选择算法3的输出作为算法4的输入。我们用 Delaunay 三角测量法再次对剩余的点进行三角测量，即 Delaunay 三角测量法<sup>[96]</sup>。每个三角形的平面三维几何模型可以通过以下方式求解：

$$\mathbf{n}_i^T [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}] - [\bar{h}_i, \bar{h}_i, \bar{h}_i]^T = [0, 0, 0]^T, \quad (4.5)$$

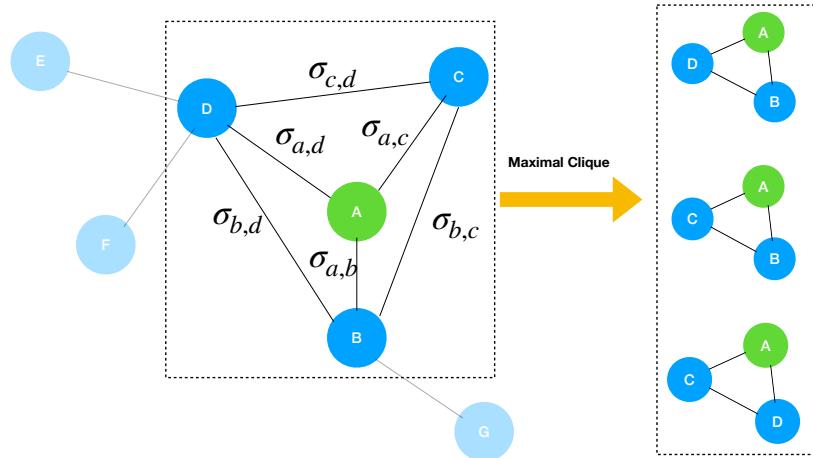


图 4.3 Illustration of maximum clique.

上述方程的解是：

$$\bar{\mathbf{n}}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}]^{-1} [1, 1, 1]^T \quad (4.6)$$

$$\begin{cases} \mathbf{n}_i = \frac{\bar{\mathbf{n}}_i}{\|\bar{\mathbf{n}}_i\|}, \\ \bar{h}_i = \frac{1}{\|\bar{\mathbf{n}}_i\|}. \end{cases} \quad (4.7)$$

我们在方程(4.5)中再增加两个约束条件， $\|\mathbf{n}_i\| = 1$  和  $n_{iy} > 0$ ，得到唯一解。在得到每个三角形区域的几何结构  $\mathbf{n}_i^T \mathbf{x} - h_i = 0$  后，我们先根据角度约束选择三角形，再根据距离约束选择。初始，我们假设摄像头安装时是向前看的，三角形应该位于摄像头下方，所以直接删除  $barh_i < 0$  的点。

首先，考虑到移动的连续性，连续帧的法向量应相似，所以只保留法向量与前一帧路面法向量接近的三角形。我们将三角形区域的几何模型与前一帧的道路模型进行比较，由摄像机运动  $\mathbf{R}$  和  $\mathbf{bart}$  计算出的法线应该接近估计的道路法线， $\mathbf{n}_t$ 。每个三角形区域法线的俯仰角可以通过以下方式计算出来：

$$\theta_i = \arcsin \left( -\frac{n_{i2}}{|\mathbf{n}_i|} \right) \quad (4.8)$$

其中  $n_{i2}$  表示  $\mathbf{n}_i$  的第二个元素。然后将  $\theta_i$  与道路正常向量的角度  $\theta_r$  进行比较。一般来说， $\theta_r$  可以直接从最后一帧的道路正常向量中估算出来。但是，对于最初的几帧，道路的法向量是未知的，我们必须从运动的  $\bar{\mathbf{t}}$  来计算。严格来说， $\theta_{\bar{\mathbf{t}}}$  只有在摄像机俯角  $\theta_{\mathbf{R}}$  小于某个阈值时才有效，所以这两个变量都需要计算。

$$\theta_{\bar{\mathbf{t}}} = \begin{cases} \arcsin \left( -\frac{t_2}{|\bar{\mathbf{t}}|} \right) & \text{if } |\bar{\mathbf{t}}| \neq 0, \\ \text{NaN} & \text{if } |\bar{\mathbf{t}}| = 0. \end{cases} \quad (4.9)$$

当  $|t| = 0$  时, 说明机器人无运动, 无需恢复尺度。 $\theta_R$  绝对值是

$$|\theta_R| = \begin{cases} |\arctan(-\frac{\mathbf{R}_{32}}{\mathbf{R}_{33}})| & \text{if } \mathbf{R}_{33} \neq 0, \\ \frac{\pi}{2} & \text{if } \mathbf{R}_{33} = 0. \end{cases} \quad (4.10)$$

如果  $\theta_R$  足够小, 那么道路法线的俯仰角被估计为  $\theta_r = \theta_t - \frac{\pi}{2}$ , 因为当车辆在道路上行驶时, 运动矢量  $\bar{\mathbf{t}}$  与路面相切, 与道路法线正交, 如图所示。只有满足以下条件:

$$||\theta_r - \theta_i|| < \theta_0 \quad (4.11)$$

的点将被保留。这里  $\theta_0$  是角隙阈值, 我们在实验中设置为 10 度。

---

#### Algorithm 4: 基于路面模型的特征点选择

---

**Input:** 准路面特征点集  $\Omega = \{f_0, f_1, \dots, f_n\}$ , 每个点的像素坐标  $(u_i, v_i)$ ,  
相机坐标系下的点坐标  $(\bar{x}_i, \bar{y}_i, \bar{z}_i)$ , 路面模型  $\mathbf{n}_i^T \mathbf{x}_i - \bar{h}_i = 0$ , 运动  
向量  $\bar{\mathbf{t}}$

**Output:** 认证路面特征点集  $\Gamma \subset \Omega$

根据像素坐标对准路面特征点集  $\Omega$  进行三角剖分, 得到三角形集合

$$\nabla = \{\nabla_0, \nabla_1, \dots, \nabla_m\}, \nabla_i = \{f_i \in \Omega, f_j \in \Omega, f_k \in \Omega\}$$

**if** 上一帧的路面模型未知 **then**

$$|\theta_r = \theta_t - \frac{\pi}{2}, \theta_t \text{ 是 } \bar{\mathbf{t}} \text{ 的仰角初始值}$$

**end**

将认证路面三角形集初始化为空集  $\Theta = \emptyset$

**for**  $\forall \nabla_i \in \nabla$  **do**

计算相机高度  $\bar{h}_i$ , 三角形  $\nabla_i$  的法向量  $n_i$  和仰角  $\theta_i$

**if**  $||\theta_r - \theta_i|| < 10$  **then**

$$|\Theta = \Theta + \nabla_i$$

**end**

**end**

Calculate the median of the distance between effective triangles and camera

optical center  $h_t$

**for**  $\forall \nabla_i \in \Theta$  **do**

**if**  $\bar{h}_i < h_t$  **then**

$$|\Theta = \Theta - \nabla_i$$

**end**

**end**

$$\Gamma = \{f_k\}, \forall f_k \in \nabla_i, \forall \nabla_i \in \Theta$$


---

其次，我们继续以摄像机光学中心与三角形区域的距离为约束进行选择。根据道路点相对较低的约束条件，选择三角形有效距离的中值作为阈值，将高于该阈值的点视为道路点。利用所选三角形区域  $\Gamma$  的顶点更新路面几何模型。具体过程如算法4所示。

#### 4.1.3 路面模型与绝对尺度计算

在本节中，只对第4.1节中两次选取后存活下来的特征点进行验证，估计道路几何模型和尺度。我们主要利用平面拟合的方法来确定相对尺度的相机高度，并利用中值滤波来降低尺度噪声。我们假设道路是一个平面，在帧  $I_t$  中的几何模型可以表示为：

$$\mathbf{n}_i^T \mathbf{x}_i - \bar{h}_i = 0, \quad (4.12)$$

其中  $\mathbf{n}_i$  为道路平面法线， $\bar{h}_i$  为计算高度。比例尺可以用以下方法恢复：

$$s_i = h_0 / \bar{h}_i, \quad (4.13)$$

其中  $h_0$  是给定的摄像机安装高度。应用 RANSAC 方法<sup>[97]</sup> 对经过验证的道路点进行道路平面估计。如果选择的特征点数量小于一个阈值，我们跳过 RANSAC 步骤，并保持道路几何模型与上次验证的帧相同。我们在实验中设置这个阈值为 12。根据车辆速度不会发生剧烈变化的假设，我们在时间维度上对尺度进行过滤，以削弱尺度噪声的影响。直接采用独立于噪声模型的中值滤波器来消除噪声，即用前  $q$  帧估计的尺度系数的中值作为尺度系数。RANSAC 不同参数下所获得的性能以及不同滤波器大小的影响将在第4.2.2.2节中进行分析。

$$s_i = median(s_{i-q+1}, s_{i-q+2}, \dots, s_q). \quad (4.14)$$

由此得到绝对尺度下的位移矩阵是：

$$\mathbf{t} = s_i \bar{\mathbf{t}}. \quad (4.15)$$

最后，计算出机器人的绝对运动估计得到解决  $\mathbf{R}$  和  $\mathbf{t}$ 。

#### 4.2 KITTI 数据集单目视觉里程计实验

我们在常用的公开测评数据集 KITTI<sup>[98]</sup> 上评估了我们的视觉里程计尺度恢复方法，可用于评估视觉里程计算方法的准确性，是目前应用最广泛的测试环境

之一。它由 22 个序列组成，覆盖了城市、村庄、高速公路等环境，运行长度从数百米到数公里不等。其中，前 11 个序列提供了真实的运动轨迹。但我们忽略序列 01，因为大多数 VO 方法在这种高速场景下无法提供满意的初始结果。此外，我们的主要评估标准是相对姿态误差 (RPE)<sup>[98]</sup> 和绝对轨迹误差 (ATE)<sup>[99]</sup>。RPE 测量每个序列中每个固定距离段的  $\mathbf{R}$  和  $\mathbf{t}$  的平均相对误差。ATE 计算  $\mathbf{t}$  的绝对误差，这些指标可以通过预处理相似性转换<sup>[88]</sup> 来评估尺度漂移消除性能。我们用 Python 实现了我们的算法，源代码是公开的。所有的实验都是在英特尔酷睿 i5，2.7GHz，使用单线程进行的。

我们的实验由三部分组成。首先，我们将我们的单目尺度恢复方法与其他简单开源的 VO/SLAM 方法结合起来，定量和定性地测试对它们性能改进。其次，设置 MonoVO<sup>①</sup> 和 ORB-SLAM2<sup>[88]</sup> 作为我们的初始自我运动估计，将我们的方法与四个最先进的 VO 算法进行比较。最后，给出了每个模块的性能分析和参数探索。

#### 4.2.1 对现有单目视觉里程计开源算法的改进

为了展示比例尺恢复方法的性能，我们将我们的比例尺校正方法移植到 ORB-SLAM2、LibVISO<sup>[66]</sup> 和 MonoVO 等三种基于特征的开源定位算法上，并将它们的性能分别与原方法进行比较。

##### 4.2.1.1 对 ORB-SLAM2 视觉里程计性能的提升

我们将我们的方法与 ORB-SLAM2(无 LC) 进行定性比较，如图4.4和图4.5所示，并与 ORB-SLAM2(有 LC) 进行定量比较，如表4.2所示。

首先，我们将我们的算法添加到单目 ORB-SLAM2 (无 LC) 中，以比较固定和计算尺度的轨迹。在计算 ORB-SLAM2 中摄像机的初始运动之前，我们关闭全局优化和环路闭合检测，以避免额外的尺度漂移优化。

考虑到较长的车辆运行面临较明显的尺度漂移问题，选择了 KITTI 数据集中的 00 (3.7 公里)、02 (5.1 公里)、05 (2.2 公里) 和 08 (2.8 公里) 四个长距离序列。从图4.4b, 4.4f, 4.4j 和4.4n 中可以看出，虽然我们对第一列的初始路径进行了对齐，但仍然存在严重的尺度漂移问题，尤其是在后期。因此，用固定比例尺恢复轨迹是不可行的。图中第三列4.4显示了我们估计的尺度参数和地面真实尺度参数的对比。地面真值尺度参数的计算方法为

$$s_g = \frac{\|\mathbf{t}_g\|}{\|\bar{\mathbf{t}}_i\|} \quad (4.16)$$

<sup>①</sup> <https://github.com/uoiip/monoVO-python>

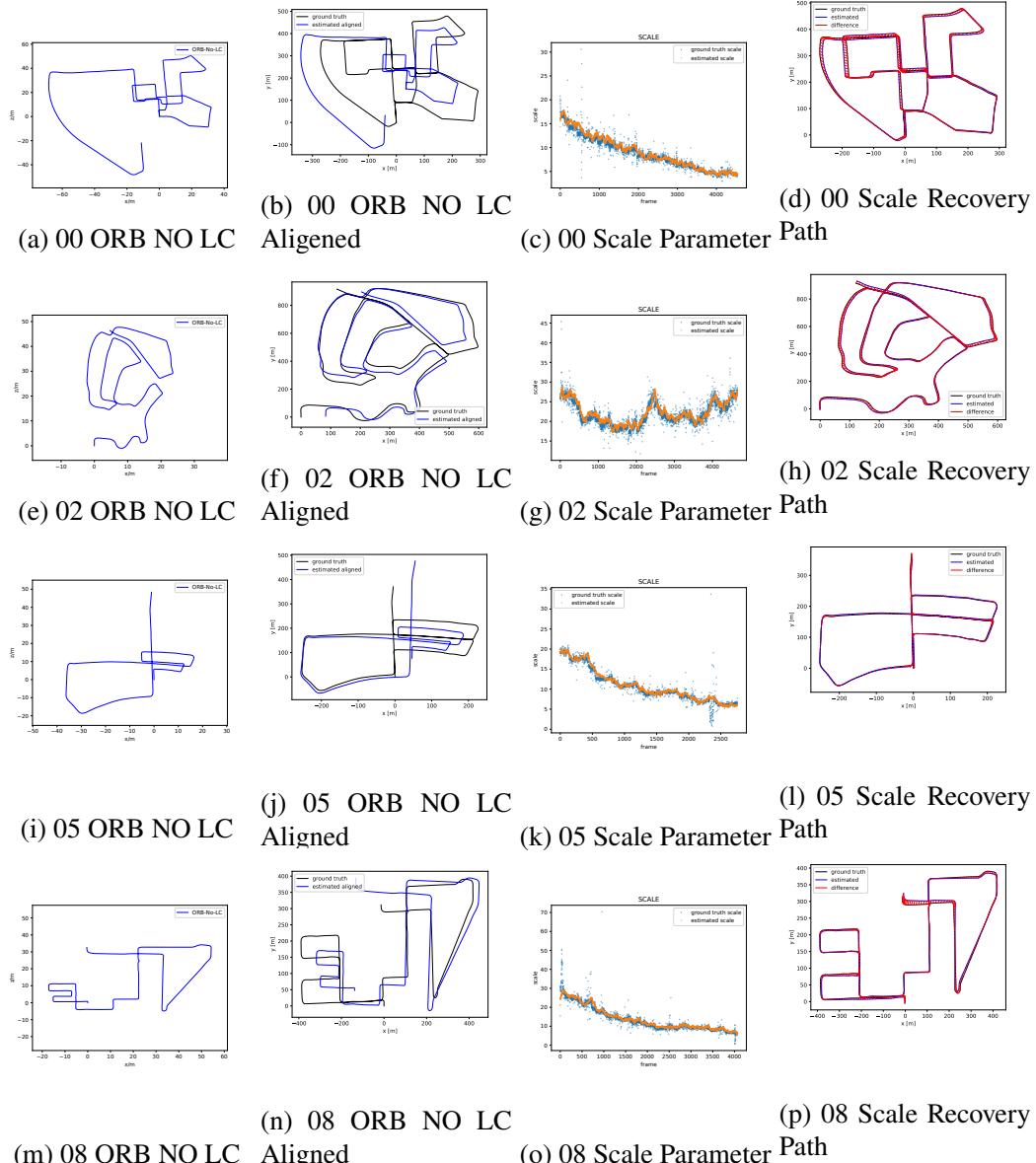


图4.4 在KITTI数据集序列00、02、05、08上与无LC的ORB-SLAM2的尺度恢复性能比较。第一列中三个图像是没有环路闭合检测的单目ORB-SLAM2轨迹，显然尺度发生了明显的错误。第二列为对应序列的前100帧通过7-dof尺度校正得到的轨迹。第四列是第一列图像与我们估计的尺度参数（第三列）相运算得到的轨迹。

其中  $\mathbf{t}_g$  是以自我运动估计位移的 ground-truth。估计的尺度参数由中值滤波器过滤。通过将我们估计的尺度参数与初始的相对位移矩阵相乘，我们得到了绝对尺度下的自我运动估计结果，如图4.4d, 4.4h, 4.4l和4.4所示。加入我们的尺度漂移消除模块后，定性比较来看机器人移动轨迹与 ground-truth 更接近。

以ORB-SLAM2(无LC)的初始运动作为基准，在KITTI数据集序列00和02-10上得到的视觉里程计轨迹。每个子图中的红线是轨迹ground-truth，蓝线是经过我们的绝对尺度恢复方法校准后的轨迹。

其次，我们对ORB-SLAM2(带LC)和我们的方法进行了性能量化比较，说明我们的尺度恢复算法在消除尺度漂移方面效果更好。评价标准是ATE，如表4.2所

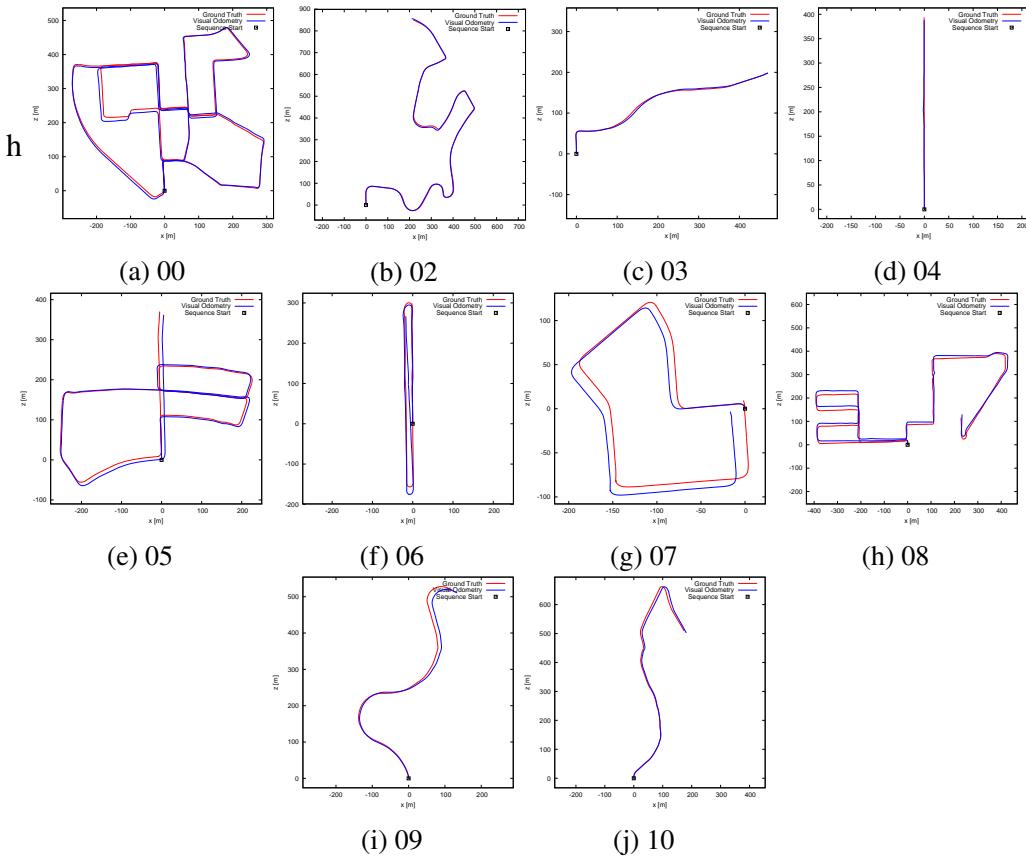


图 4.5

示。使用 ATE 的原因是 ORB-SLAM2 的单目版不提供尺度计算，不能直接计算 RPE。具有 7-自由度对齐（平移、旋转和尺度）的 ATE 误差可以在评估前在绝对尺度上对齐轨迹，以评估尺度模糊性的影响。

在大多数序列中，我们的 VO 效果优于 ORB-SLAM2（有 LC 模块）的效果。对于小尺度的轨迹，如 04、06 和 07，我们的方法和 ORB-SLAM2 与 LC 之间的误差差距很小。但是，由于我们的算法可以恢复绝对尺度，消除尺度漂移，所以在所有长轨迹上都可以达到满意的误差。

#### 4.2.1.2 对 LIBVISO2 性能的提升

LibVISO2 是一个开源的 C++ 库，用于测量视觉里程计。它通过双核和非最大抑制的响应提取特征。根据这三个点，计算出它们的矩阵旋转矩阵  $R$  和位移矩阵  $t$  来表示两帧之间的姿态变换。它提供了一种和我们一样基于局部平面地面和摄像机高度的简单比例计算方法，我们用本文提出的比例恢复算法来代替它，以测试对 LibVISO2 单目版本算法的改进。为方便起见，本文将 LibVISO2 使用 pybind11 封装成 Python 库<sup>①</sup>。如表4.3所示，通过与我们的尺度恢复算法相结合，LibVISO2 的性能得到了极大的提升。以序列 00 为例，加入我们的尺度恢复算

<sup>①</sup> <https://github.com/SummerHuiZhang/Libviso-Python>

表 4.2 ORB-SLAM2 中，采用尺度漂移与环路闭合检测两种方法时对绝对平移误差（ATE）的影响。

Seq	Running distance	ORB+LC	ORBnoLC+Our SR
	[88] (m×m)	ATE(m)	ATE(m)
00	3724	6.68	<b>5.56</b>
02	5067	21.75	<b>4.36</b>
03	561	1.59	<b>1.36</b>
04	394	<b>1.79</b>	2.36
05	2206	8.23	<b>8.03</b>
06	1233	<b>14.68</b>	18.36
07	695	<b>3.36</b>	5.16
08	3223	46.58	<b>5.64</b>
09	1705	7.62	<b>2.53</b>
10	920	8.68	<b>2.33</b>
Average	1972.80	12.10	<b>5.56</b>

表 4.3 Improvement on LibVISO2.

Seq	Running distance	Rotation error	LibVISO2	LibVISO2+Our SR
	[88] (m×m)	[88] RPE(deg/m)	[66] RPE(%)	RPE(%)
00	3724	0.0267	9.77	<b>6.51</b>
02	5067	0.0136	16.40	<b>3.89</b>
03	561	0.0200	22.85	<b>4.81</b>
04	394	0.0292	18.79	<b>6.70</b>
05	2206	0.0382	12.22	<b>12.0</b>
06	1233	0.0255	9.42	<b>9.19</b>
08	3223	0.0223	9.60	<b>7.83</b>
09	1705	0.0143	9.82	<b>5.20</b>
10	920	0.0378	18.70	<b>7.86</b>
Average	1903.30	0.0253	14.18	<b>7.11</b>

法后，RPE 从 9.77% 下降到 6.51%。在序列 02、03、04 和 10 中，这种改进尤为明显。平均 RPE 从 14.18% 下降到 7.11%。且当车头通常被其他车辆挡住时，LibVISO2 很难运行，如序列 07 中出现的情况。

#### 4.2.1.3 对 MonoVO 性能的提升

MonoVO 是一个简单的基于 OpenCV 的开源 MVO 项目，它用 FAST 描述符<sup>[100]</sup> 提取特征，并用光流跟踪它们。MonoVO 确实提供了一个方便的五点运动

估计，但它缺乏尺度计算。因此，我们将我们的尺度恢复算法与原始 MonoVO 和带有特征稀疏化的 MonoVO 结合起来，测试我们的方法在这种简单 VO 方法上的性能。

表 4.4 Improvement on MonoVO

Seq	Rotation error	MonoVO	MonoVO-ROI	MonoVO-SR
	[88] RPE(deg/m)	RPE(%)	RPE(%)	RPE(%)
00	0.0046	36.44	20.32	<b>2.51</b>
02	0.0059	46.19	7.28	<b>2.33</b>
03	0.0048	46.19	12.06	<b>5.65</b>
04	0.0036	55.92	14.12	<b>2.40</b>
05	0.0316	35.43	27.93	<b>8.48</b>
06	0.0057	22.94	17.14	<b>2.32</b>
08	0.0072	32.47	16.27	<b>3.05</b>
09	0.0062	33.50	12.26	<b>2.15</b>
10	0.0162	28.32	20.00	<b>4.92</b>
Average	0.0096	37.50	16.38	<b>3.76</b>

结合我们的尺度恢复算法 (称为 MonoVO-SR)，对原始 MonoVO 的改进如表4.4??所示。MonoVO-ROI 和 MonoVO-SR 都是 MonoVO 和本文提出的方法的组合，不同的是 MonoVO-SR 的性能比原来的 MonoVO 和 MonoVO-ROI 都要好，MonoVO-ROI 假设 ROI 是固定的道路。以序列 00 为例，MonoVO 和 MonoVO-ROI 的 RPE 分别为 36.44% 和 20.32%。加入我们的规模恢复算法后，降低到 2.51%。改善明显，平均 RPE 从 37.75% 急剧下降到 3.76%。

此外，我们观察到，分散的特征点可以进一步帮助提高 MonoVO 的性能。我们将原始 MonoVO 的特征点分散开来，并在每个切割的  $30 \times 30$  像素区域随机选择两个点。然后我们在分散的 MonoVO 上加入我们的尺度恢复方法 (称为 ST-MonoVO-SR)，结果如表??所示。我们可以看到，ST-MonoVO-SR 在所有测试序列上都优于特征稀疏化的 MonoVO，平均 RPE 从 10.30% 提高到 2.13%。此外，从表4.4和表??的比较中，我们也可以得出结论，分散的特征点避免了很多模糊性，确实有助于提高 MonoVO 的性能。ST-MonoVO-SR 将是我们与其他机器人自我运动估计算法比较时的系统之一。

我们将我们的比例尺恢复方法与散射的 MonoVO (ST-MonoVO-SR)，以及没有 LC 的 ORB-SLAM2 (ORB-SLAM2-SR) 结合起来，并与四种最先进的视觉里程计方法，在 KITTI 数据集的序列 00 和 02-10 上进行比较，他们分别来

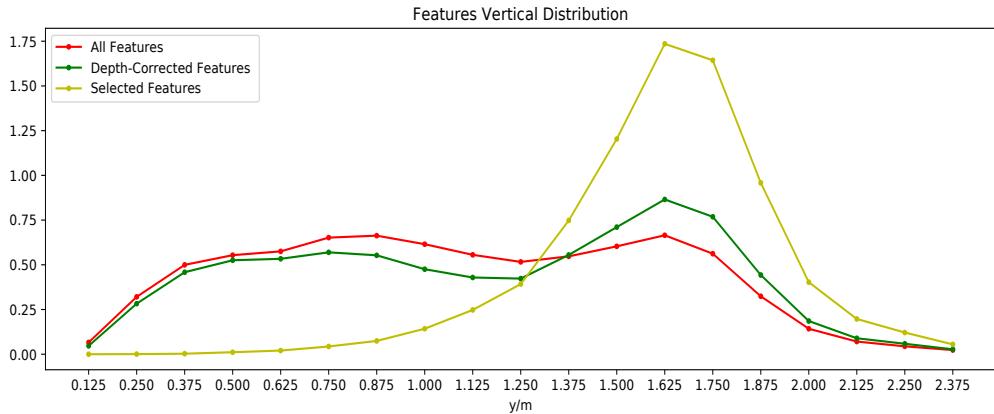


图 4.6 算法 1-3 对 KITTI 序列 00 的前 500 帧特征点分布的初步测试。绿线表示深度一致性选择（算法 1 和 2）后的点分布。黄线表示深度和道路模型一致性选择后的点分布（算法 1-3）。

自<sup>[38]</sup>、<sup>[66]</sup>、<sup>[42]</sup> 和<sup>[39]</sup>。如表7.3所示，我们算法的平均误差低于其他单目尺度恢复算法，与 LibVISO2-stereo 算法相当。我们的方法也优于在大多数序列上失败的<sup>[39]</sup>的方法。虽然 Lee<sup>[42]</sup> 提出的方法在序列 01（一个高速场景）上有效，但它在其他序列上的表现比我们的差。<sup>[38]</sup> 的算法假设一个 ROI 作为路面，所以当车前被其他物体挡住时，它很难计算出路面的几何模型，就像序列 07 中发生的那样。但是，当使用 ORB-SLAM2 作为我们的前端时，我们仍然可以在序列 07 中通过在线识别路面面积获得相对稳定的结果 1.73%。

值得强调的是，虽然 ORB-SLAM2-SR (ORB-SLAM2 和本方法的组合) 的性能更好，但 ST-MonoVO-SR (MonoVO 和本方法的组合) 更能证明本方法的效果。MonoVO 原本方法比 ORB-SLAM2 简单得多且精度较差，我们的尺度恢复方法大大提高了它的精度。

## 4.2.2 算法不同模块效果分析

开源代码包括离线版和在线版。离线版需要保存初始运动和特征点，然后运行比例恢复代码；在线版的代码与初始 VO 代码融合，同时运行。本章给出了各模块的消融研究，包括特征选择模块（基于深度和道路模型的一致性）和参数探索模块（中值滤波和 RANSAC）。使用经过特征稀疏化的 MonoVO 和 ORB-SLAM2 提供初始运动估计。

### 4.2.2.1 路面特征点选择测试实验

选取的道路点受到两个规则的约束：深度一致性和道路模型一致性。我们算法的性能如图4.6和4.8所示，图4.6中表示在 KITTI 数据集中序列 00 上第 10 帧与第 20 帧两帧图像上测试算法 1-3 的结果。红色折线表示经过两帧图像匹配后可

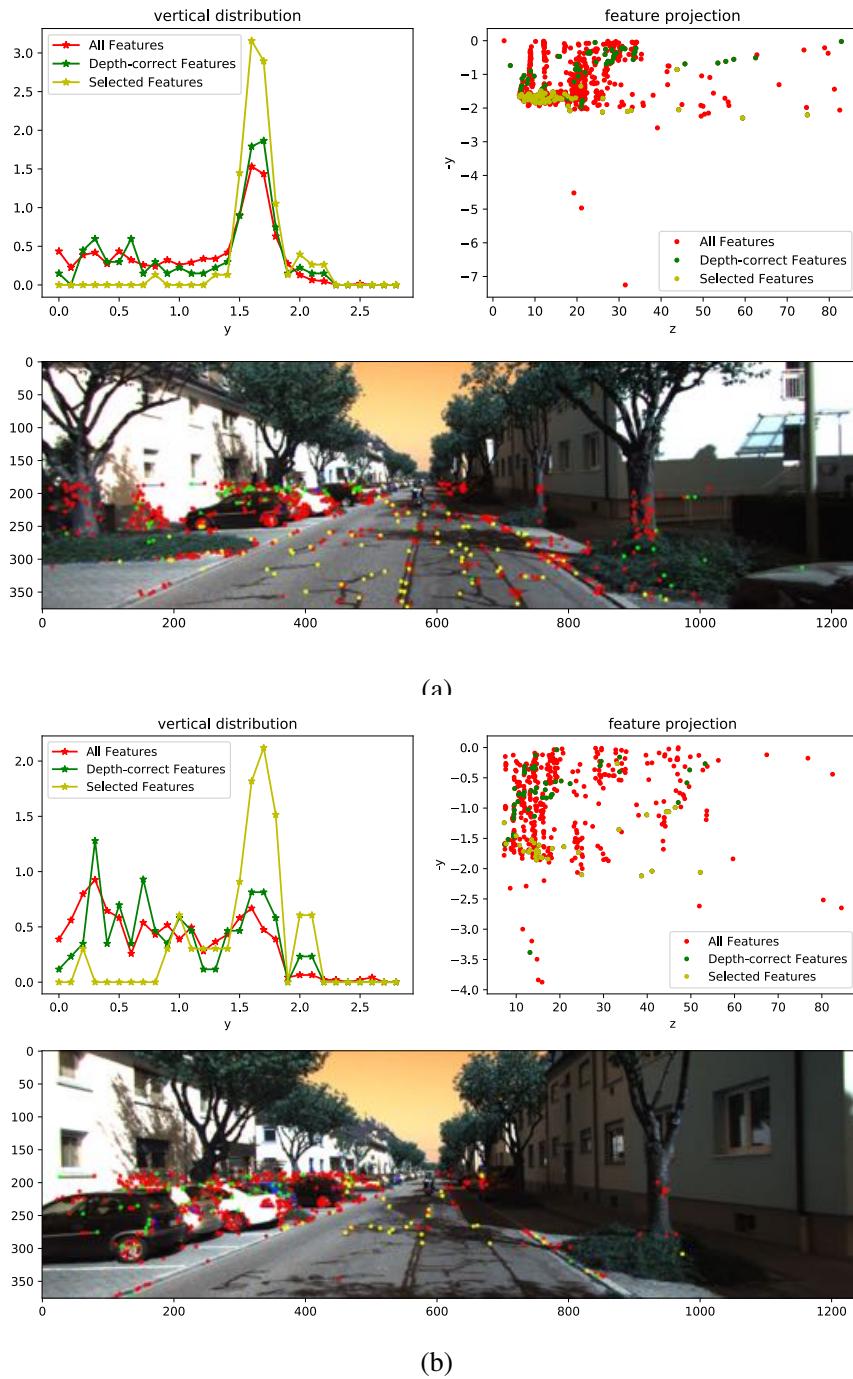


图 4.7 在 KITTI 序列 00 的两帧上对算法 1 至 3 进行初步测试。标签"Depth-Correct Features" 代表算法 1 和 2 之后的分布，标签"Selected Features" 代表算法 1 至 3 的分布。(a) 第 10 帧。(b) 第 20 帧。

能的路面特征点，绿色折线"Depth-Correct Features" 表示经过算法 1 和算法 2 后的特征点分布，黄色折线"Selected Features" 表示经过算法 1-3 后的特征点分布。

表4.6-4.8所示，证明了本文提出的特征点选择算法确实可以有效地拒绝道路异常值，提高轨迹估计的精度。我们先在一些帧上进行定性测试，再在 KITTI 数据集的 10 个序列上进行定量测试。在此基础上，本文提出的轨迹估计算法确实能够有效地剔除道路异常值，以提高轨迹估计的精度。

Seq	Zhou <i>et al.</i> (from <sup>[39]</sup> )		LibVISO2-stereo (from <sup>[66]</sup> )		Lee <i>et al.</i> (from <sup>[42]</sup> )		Song <i>et al.</i> (from <sup>[38]</sup> )		ST-MonoVO-SR		ORB-SLAM2-SR	
	Trans	Rot	Trans	Rot	Trans	Rot	Trans	Rot	Trans	Rot	Trans	Rot
	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)
00	2.17	0.0039	2.32	0.0109	4.42	0.0150	2.04	0.0048	2.17	0.0053	<b>1.01</b>	<b>0.0014</b>
01	-	-	-	-	6.91	0.0140	-	-	-	-	-	-
02	-	-	2.01	0.0074	4.77	0.0168	1.50	0.0035	1.81	0.0041	<b>0.93</b>	<b>0.0018</b>
03	2.70	0.0044	2.32	0.0107	8.49	0.0192	3.37	0.0021	1.45	0.0035	<b>0.52</b>	<b>0.0010</b>
04	-	-	<b>0.99</b>	0.0081	6.21	0.0069	1.43	<b>0.0023</b>	2.21	0.0049	1.16	<b>0.0023</b>
05	-	-	1.78	0.0098	5.44	0.0248	2.19	0.0038	1.51	0.0041	<b>1.45</b>	<b>0.0014</b>
06	-	-	<b>1.17</b>	0.0072	6.51	0.0222	2.09	0.0081	2.91	0.0060	2.92	<b>0.0027</b>
07	-	-	-	-	6.23	0.0292	-	-	-	-	<b>1.73</b>	<b>0.0023</b>
08	-	-	2.35	0.0104	8.23	0.0243	2.37	0.0044	2.34	0.0035	<b>1.18</b>	<b>0.0017</b>
09	-	-	2.36	0.0094	9.08	0.0286	1.76	0.0047	1.85	0.0032	<b>1.17</b>	<b>0.0020</b>
10	2.09	0.0054	1.37	0.0086	9.11	0.0322	2.12	0.0085	1.83	0.0048	<b>0.93</b>	<b>0.0029</b>
AVG	2.32	0.045	2.02	0.0095	6.86	0.0212	2.03	0.0045	2.01	0.0049	<b>1.25</b>	<b>0.0020</b>

表 4.5 与其他最先进的 VO/SLAM 算法的比较

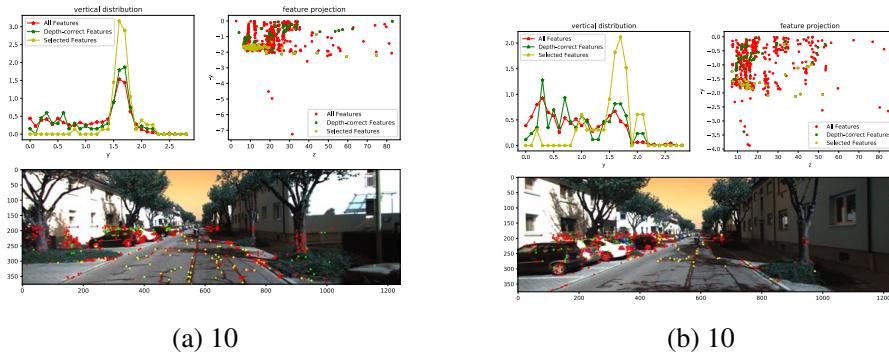


图 4.8 算法 1-3 在两帧图像上测试结果

首先，我们用我们的特征选择方法对 KITTI 数据集序列 00 上的前 500 帧进行定性分析，并计算其在纵轴上的分布，如图4.6所示，特征过滤后，靠近道路的点的比例增加。然后，选取个别帧来观察道路点选择的效果。我们比较 KITTI 序列 00 中第 10 帧（图4.8a）和第 20 帧（图4.8b）在道路点选择前后的特征分布。红色曲线是所有特征点的分布。特征点属于路面的概率在特征选择后变得突出。

然后，选取 KITTI 序列 00 中的个别帧第 10 帧（图4.8a）和第 20 帧（图4.8b），查看道路点选择前后的特征分布。可以看出，第 20 帧上的特征点数量较少，不足以计算道路模型。特征点选择后，道路点的百分比变得很突出。此外，我们将剩余的点用不同的颜色可视化，我们可以看到大部分被选择的点（用黄色标记）位于道路上。深度一致性选择后，在随机选择的帧上进行两次 Delaunay 三角测量的结果。图??直观地显示了深度一致性选择（图??a）和深度与道路模型一致性

表4.6 采用同一种路面模型(RM)特征点筛选的情形下，加入直接踢除法(DD)或最大团模型(MC)特征点筛选对旋转矩阵 $R$ 的影响。

Seq	Rotation error (deg/m)	RM (%)	RM+DD (%)	RM+MC (%)
00	0.0053	2.87	2.25	<b>2.17</b>
02	0.0041	2.03	1.87	<b>1.81</b>
03	0.0035	1.20	<b>1.19</b>	1.45
04	0.0049	2.08	<b>1.77</b>	2.21
05	0.0041	2.97	1.81	<b>1.51</b>
06	0.0060	4.51	3.10	<b>2.91</b>
07	0.0092	4.07	3.72	<b>3.21</b>
08	0.0035	3.28	2.47	<b>2.34</b>
09	0.0032	1.86	1.87	<b>1.85</b>
10	0.0048	3.04	2.45	<b>1.83</b>
Average	0.0049	2.80	2.25	<b>2.13</b>

选择(图??b)后，在随机选择的帧上进行两次Delaunay三角测量的结果。二维特征点集 $(u_i, v_i)$ 被分割成一组三角形。一些由移动障碍物或不良特征跟踪产生的错误匹配特征被移除，如图??所示。

其次，我们在KITTI数据集的序列00、02-10上量化了我们的路面特征点选择策略的性能。我们分别测试基于深度一致性(4.6)和模型一致性(4.7)的路面点选择，然后在表4.8中进行比较，实验重复十次，记录相对位移误差的平均百分比。旋转误差是由MonoVO算法获得的经过稀疏化的特征的RPE<sup>[98]</sup>。%表示相对位移误差的百分比。

为了评估基于深度一致性的路点选择的性能，我们用相同的道路模型一致性进行测试，如表4.6所示。我们可以得出结论，直接删除算法2和最大小团筛选算法3在道路异常值剔除中都能发挥作用，算法3在8个序列上的表现略优于算法2。

为了评估基于道路模型一致性的道路点选择的性能(在4.1.2中的算法4)，我们设计了另一个类似的对比实验，结果显示在4.7中。使用相同的最大团深度一致性，MC-ROI显示了基于ROI的最大团深度一致性选择的性能。对于基于ROI的方法将一个固定的区域视为道路而不进行选择，其性能是最差的。第5列中的结果是将最大团和道路模型方法结合起来得到的，其性能优于仅有最大团一致性约束的性能(第4列)。我们还可以看到，使用直接删除后，道路模型一致性可以进一步拒绝更多的道路离群值，相对位移误差的比例从3.77%下降到2.13%。

在对深度和道路模型一致性性能单独分析后，我们收集的结果如表4.8所示。

表 4.7 采用同一种最大团模型 (MC) 特征点筛选的情形下, 加入路面模型 (RM) 特征点筛选对旋转矩阵  $R$  的影响。MC-ROI 是采用固定区域作为路面而未进行特种筛选的情形。

Seq	Rotation error (deg/m)	MC-ROI (%)	MC (%)	MC+RM (%)
00	0.0053	4.05	4.09	<b>2.17</b>
02	0.0041	4.08	2.75	<b>1.81</b>
03	0.0035	5.05	2.73	<b>1.45</b>
04	0.0049	6.32	<b>1.93</b>	2.21
05	0.0041	3.72	3.73	<b>1.51</b>
06	0.0060	2.99	3.71	<b>2.91</b>
07	0.0092	3.67	5.61	<b>3.21</b>
08	0.0035	3.59	3.55	<b>2.34</b>
09	0.0032	4.25	4.54	<b>1.85</b>
10	0.0048	1.97	5.02	<b>1.83</b>
Average	0.0049	3.97	3.77	<b>2.13</b>

表 4.8 基于最大团模型 (MC) 特征点筛选和路面模型一致性 (RM) 特征点筛选下的不同旋转矩阵  $R$  的变化。

Seq	Rotation error (deg/m)	No selection (%)	MC (%)	RM (%)	MC+RM (%)
00	0.0053	13.28	4.09	2.87	<b>2.17</b>
02	0.0041	11.27	2.75	2.03	<b>1.81</b>
03	0.0035	7.070	2.73	<b>1.20</b>	1.45
04	0.0049	9.200	1.93	<b>2.08</b>	2.21
05	0.0041	9.860	3.73	2.97	<b>1.51</b>
06	0.0060	3.000	3.71	4.51	<b>2.91</b>
07	0.0092	12.30	5.61	4.07	<b>3.21</b>
08	0.0035	10.82	3.55	3.28	<b>2.34</b>
09	0.0032	16.77	4.54	1.86	<b>1.85</b>
10	0.0048	10.50	5.02	3.04	<b>1.83</b>
Average	0.0049	10.41	3.77	2.80	<b>2.13</b>

我们可以看到, 它们的组合性能 (第 6 列) 确实优于单个算法。算法3和4的组合比单独使用效果更好。我们还可以推断, 道路模型一致性选择的效果比深度一致性选择的效果好, 道路模型的约束性更强。这是有道理的, 因为深度一致性约束的主要作用只是消除一些匹配错误的点, 但道路模型一致性约束可以拒绝所有角度或高度与道路模型矛盾的分割三角形。

#### 4.2.2.2 RANSAC 与 Median 滤波参数性能比较

我们使用 RANSAC 方法计算几何道路模型，并使用中值滤波法去除尺度噪声，如4.1.3节所述。我们评估了不同的 RANSAC（最大迭代）和过滤器参数（过滤器大小）下视觉里程测量的位移误差。我们用不同的 RANSAC 和过滤器参数评估视觉里程测量的位移误，以决定最合适的选择。初始运动由 ORBSLAM2 提供。

如果选择的特征点数量小于 12 个，我们跳过 RANSAC 步骤，保持道路几何模型与最后一帧验证的模型相同。为了确定最合适的选择 RANSAC 参数，我们设计了三个实验，结果如图所示4.9。

首先，我们在 KITTI 数据集的序列 00 上用 10 次不同的最大迭代和 17 个滤波器大小进行测试。我们对每一种情况运行 10 次，平均位移误差如图4.9a 所示。我们可以看到，虽然这十条线代表着不同的最大迭代次数，但在过滤器大小为 6 之前，它们都会急剧减少，然后随着过滤器大小的增加而变得稳定。

所以我们暂时将过滤器大小设置为 6，记录不同最大迭代下翻译误差的变化趋势，如图4.9b 所示。随着最大迭代次数的增加，性能不断提高，在 15 次之后变得稳定。考虑到计算成本和系统的鲁棒性，我们在实验中留有余地，将 RANSAC 的最大迭代次数设置为 20 次。

滤波器的大小决定了考虑多少张之前的图像来获得尺度  $s$  的中值。同样，我们将 RANSAC 的最大迭代次数设置为 20，并记录不同滤波器大小下的翻译误差方差，如图4.9c 所示。我们可以观察到，在一定的阈值下，随着滤波器大小的增加，更多的路面特征点异常值被去除。当滤波器大小达到 6 左右时，我们的方法实现了比较小的误差。这意味着所选取的特征点大多在道路平面上。但是，当超过一定的阈值后，过滤器尺寸过大，无法跟随尺度漂移，从而降低了尺度恢复方法的性能。最后，我们在实验中采用 6 作为最佳滤波大小。位移误差的平均值和中值分别约为 1.0207% 和 1.0008%。此外，位移误差的标准误差约为 0.0712%。

在确认了中值滤波的最佳滤波尺寸后，我们对 KITTI 数据集中带中值滤波的序列 00 上的尺度和轨迹进行了展示，如图所示，4.10。我们可以观察到，在这个序列中，没有尺度滤波器的原始轨迹发生了严重的尺度漂移。然而，在中值滤波器的帮助下，大部分尺度噪声被去除，我们的尺度恢复方法估计的 VO 轨迹非常接近真实路线。

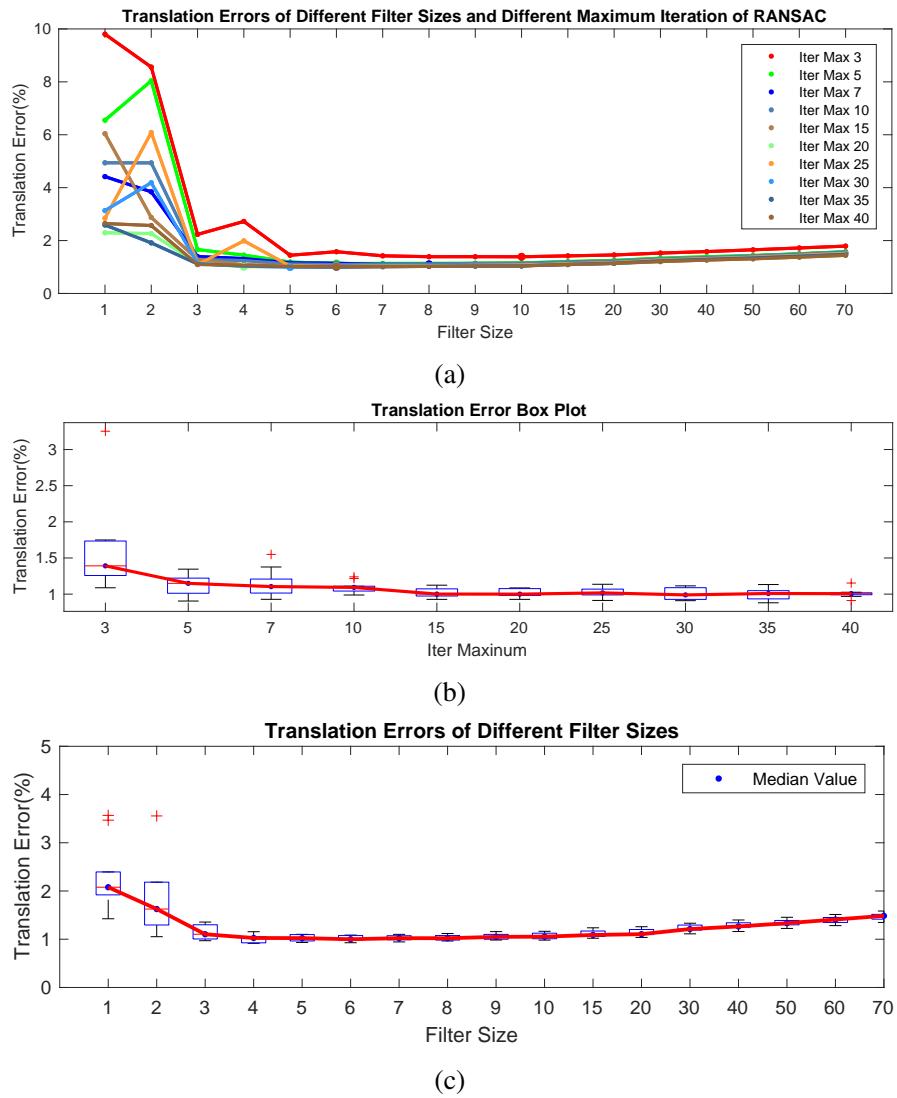


图 4.9 RANSAC 和中指滤波不同参数下的旋转误差。(a)RANSAC 最大迭代次数的影响。(b) 中值滤波不同滤波尺寸的影响。

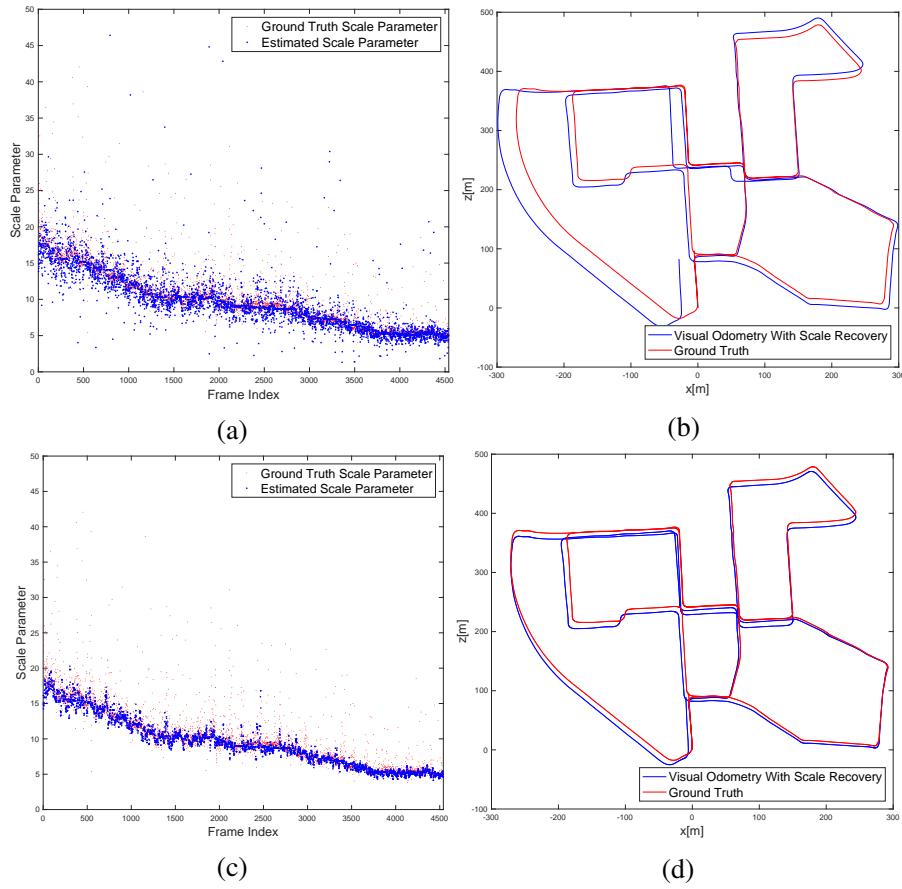


图 4.10 值滤波加入前后尺度系数和轨迹比较。(a) 加中值滤波前的尺度系数。(b) 加中值滤波前的轨迹。(c) 加中值滤波后的尺度系数。(d) 加中值滤波后的轨迹。

### 4.3 本章小结

。我们提出了一种以摄像机高度为绝对参考，基于道路几何约束的实时单目视觉里程计尺度恢复方法。我们的方法的新颖之处在于，道路特征点的选择和道路模型估计是迭代计算的——的估计道路几何模型被认为是选择道路点的反馈。考虑检测到的道路点，对几何道路模型进行在线更新。选取的路点用于估计道路模型，进而限制路点的选择。通过 Delaunay 三角测量法<sup>[101]</sup> 将每一帧图像分割成一组三角形。通过考虑深度一致性和道路模型一致性，检查每个三角形是否属于道路区域。在道路点的选择上，我们用 Delaunay 三角测量法对点位进行分割，根据深度一致性和道路模型一致性选择道路点。因此，对于经过深度一致筛选的剩余三角形，我们将三角形区域的几何模型与前一帧的道路模型进行比较，只有在后续帧中法线  $n_i$  相似的三角形以及高度符合要求的才会被认证为路面特征点。摄像机运动  $\mathbf{R}$  和  $\mathbf{bart}$  计算出的法线应该接近估计的道路法线  $n_t$ 。

对于经过验证的道路点，我们采用 RANSAC 来估计道路模型，并采用中值滤波器来去除尺度噪声。实验结果表明，我们的路面特征点选取策略能够有效地剔除道路异常值，RANSAC 和中值滤波器的参数探索有助于提高系统的精度和鲁棒性。实验结果还表明，现有的开源 VO 或 SLAM 方法，包括 ORB-SLAM2(有

LC 和无 LC)、LibVISO 和 MonoVO，通过与我们的尺度恢复方法相结合，得到了显著的改进。将 ORB-SLAM2(无 LC) 和 MonoVO 作为我们的初始自我运动估计，我们的方法在四种最先进的 VO 方法中取得了最好的性能。因此，通过简单的单相机校准和安装的相机高度测量，我们的方法可以帮助机器人在未探索的环境中进行精准的自我定位。

在未来，我们计划探索与 IMU 等廉价传感器融合的绝对尺度估计，因为所有基于地面平面的视觉方法在地面平面被严重遮挡或不能作为平面消耗时都会失败。此外，考虑到我们基于点的算法在低纹理环境下可能会因为缺乏足够的特征而失效，我们也会关注更丰富的线或点线结合的特征。

## 第5章 单目视觉里程计尺度计算: 从手动建模到自主学习

### 5.1 引言

前一章中介绍了路面几何模型的单目视觉里程计绝对尺度计算方法, 该方法将相对稳定的路面几何模型作为先验信息, 本章将介绍一种场景中稳定区域的自主建模方法。

### 5.2 基于场景建模的尺度计算方法

本文提出了一种基于场景建模的尺度计算方法, 本节中将依次介绍场景建模方法以及基于场景模型的尺度计算方法。

#### 5.2.1 场景建模

##### 5.2.1.1 场景模型表征

绝对尺度计算直接依赖于场景中特征点的绝对深度, 故本文使用像素深度概率模型来表征机器人所在的场景。该方法将机器人视野纵横分为多个栅格区域  $G_{ij}$ , 并使用高斯分布建模每个栅格区域的深度模型。

$$D(G_{ij} | S^k)N(\mu_{ij}^k - \sigma_{ij}^k) > 0 \quad (5.1)$$

其中  $\mu_{ij}^k$  和  $\sigma_{ij}^k$  分别为场景  $S_k$  中栅格  $G_{ij}$  所服从的高斯分布中的均值和标准差。

##### 5.2.1.2 场景建模方法

场景建模的训练集需要包扩在特定场景中拍摄的连续图像序列  $\mathbf{I}_t$  和相邻图片拍摄位置的绝对距离  $l_{t-1}^t$ 。首先依据相邻帧图像  $\mathbf{I}_{t-1}$  和  $\mathbf{I}_t$  进行特征检测匹配和相对运动估计以及相对深度估计, 具体操作如下在图像  $\mathbf{I}_{t-1}$  过量提取特征点, 得到初始特征点集合  $\Omega_f$  和  $u_f$ , 可计算出机器人的相对运动  $\mathbf{R}\bar{\mathbf{t}}$ , 由于单目的尺度歧义性, 无法准确获取  $\bar{\mathbf{t}}$  的绝对大小, 但可以根据此运动通过三角测量的方式获取每个像素点的像素深度  $\bar{d}_f$  在获取特征点的相对深度后, 本文对特征点不同区域的深度分布进行统计学建模。首先根据相邻图片的绝对距离  $l_{t-1}^t$ , 确定绝对尺

度与相对尺度之前的系数

$$s = \frac{l_{t-1}^t}{\|\bar{\mathbf{t}}_{t-1}^t\|} \quad (5.2)$$

进而可以获取，每个像素点在真实世界中的绝对深度  $d = s\bar{d}$ 。同时根据像素位置，将特征点归属于不同的栅格  $G_{ij}$ 。在执行完场景中全部帧之后，使用高斯分布建模每个栅格内的特征点深度均值  $\mu_{ij}$  和标准差  $\sigma_{ij}$  考虑到训练数据中可能存在多种分布不同的场景，简单的将所有场景用统一中模型表征会是偏差较大。本文提出使用聚类方法，将场景按照深度分为  $K$  个场景，对于每一个单独进行建模。场景聚类方法简述如下：首先我们将每一帧图像  $\mathbf{I}^t$  按照特征点的分布位置和深度进行编码表示为  $C_t \in \mathbb{R}^{hw2}$

$$C_t[i_w, i_h] = (N(f), \mu(d)) \quad (5.3)$$

其中  $N(f)$  为特征点数量， $(d)$  为特征点深度的均值。即我们使用各个栅格的特征点数量和深度均值作为编码基本单元来表征每帧图像的结构信息（查查参考文献）。我们定义图像结构的距离为

$$\|C^{t1} - C^{t2}\| = \sum(|N(f)^{t1} - N(f)^{t2}| + N(f)^{t1}N(f)^{t2}(\mu(d)^{t1} - \mu(d)^{t2})) \quad (5.4)$$

依据如上编码方式和编码距离的定义，可以完成聚类。

### 5.2.2 尺度计算

获取场景绝对深度模型之后，使用绝对深度与相对深度的比值即可计算尺度系数，在计算过程中，首先根据相对深度计算场景结构编码，然后选取与当前结构最近的场景用于计算尺度

$$S_k = \operatorname{argmin} S_k \quad (5.5)$$

在计算过程中，以栅格深度的均值作为该位置区域的绝对深度，通过栅格深度的标准差评价绝对深度的可靠性，并根据如下公式计算尺度的最优值

$$s = \frac{\sum f_i \left( \frac{1}{\sigma_{f_i}^2} \right) \frac{u_i}{d_{f_i}}}{\sum f_i \left( \frac{1}{\sigma_{f_i}^2} \right)} \quad (5.6)$$

获取尺度系数之后，即可恢复机器人相对运动的绝对尺度。

### 5.2.3 基于场景建模的尺度计算验证实验

#### 5.2.3.1 模型参数测试实验

#### 5.2.3.2 算法有效性消冗实验

#### 5.2.3.3 与其它方法性能对比

### 5.3 本章小结

## 第 6 章 传统单目位姿估计与深度学习尺度恢复

### 6.1 引言

## 第7章 路面驾驶机器人单目视觉里程计简化

人类的日常生活越来越多地涉及到移动机器人，包括自主地面车辆（AGV）、无人驾驶飞行器（UAV）和服务机器人。移动机器人在复杂的环境中进行导航和执行任务时，必须对自己进行定位。在环境未开发的情况下，既没有全球定位系统（GPS），也没有环境地图，无法进行绝对状态估计，只能采取相对状态估计，又称为增量式定位方法，如视觉测距。该方法是通过累积式增量计算机器人相对于到它的起始坐标系中的增量自我运动（包括平移和旋转运动），与构建的地图无关。大多数 VO 利用基于几何学的方法来估计自我运动（ $\mathbf{Rt}$ ），通过最小化重投影误差<sup>[88]</sup>（基于特征的方法）或最小化光学误差<sup>[102]</sup>（直接法）。然而，这些方法需要准确的传感器校准和手动参数调整，以便在不同的环境中良好地工作<sup>[103]</sup>。但是，参数调整过程需要工程师根据状态反馈不断进行调整，是一项消耗精神与时间的繁琐工作。

为了减少人为调整参数的努力，很多研究工作都投入到基于学习的端到端方法中。用基于学习的方法进行自我运动估计是由 Roberts 等人开始的<sup>[103]</sup>，他们尝试用 K-Nearest Neighbors 模型学习从光流到二维运动的映射。其他许多开创性的方法也在探索建立从光流到自我运动的映射模型<sup>[104-107]</sup>。Wang 等人<sup>[108, 109]</sup>首先提出了从原始图像到自我运动的端到端模型。除此之外，Wang 等人还通过使用递归神经网络来处理图像序列信息。为了减少标签数据的依赖性，Zhou 等人<sup>[110]</sup>提出了两个无监督的网络结构，与图像深度预测与自我运动估计相结合，将重投影图像的残差作为损失函数。在此基础上，许多研究人员设计了不同损失函数进行网络改进：增加额外的 3D 几何损失函数<sup>[111]</sup>，双目损失函数<sup>[112]</sup>，深度特征重建损失函数<sup>[113]</sup>，动态和光流损失函数<sup>[114]</sup>或相对损失函数<sup>[115]</sup>等。为了实现系统的鲁棒性，Klodt 等人<sup>[116]</sup>和 Yang 等人<sup>[117]</sup>进一步提出测评自我运动和深度的不确定性。而 Clement 等人<sup>[118]</sup>则试图用图像转换模型来增加算法对光照的鲁棒性。

H 然而，基于端到端深度学习方法的表现仍然不容乐观。我们发现其中一个基本问题是基于学习的方法总是依赖于一个庞大而多样化的数据集来训练一个性能良好的模型。如 Imagenet 数据集<sup>[119]</sup>用于物体检测，Cityscape 数据集<sup>[120]</sup>用于语义分割。然而，对于地面车辆的自我运动估计，最显著的数据集如 KITTI<sup>[121]</sup>和 Robotcar<sup>[122]</sup>，而用于移动机器人视觉定位的开源训练数据集在图片数量和多样性上仍十分有限。Zhou 等人的自监督方法<sup>[110]</sup>可以减弱对地面车辆运动 ground truth 的依赖性，降低了训练数据采集难度，但他们并没有解决对训练数据数量

的要求，因为他们仍然依赖于图像序列。为了应对训练数据集的局限性，Slinko 等人<sup>[123]</sup> 提出基于 RGB-D 图像的随机重投影来生成训练集，增加的训练数据的可利用性。此外，Wang 等人<sup>[124]</sup> 通过模拟不同的环境和具有挑战性的光照条件，收集了一个更大的具有复杂运动模式的数据集。

为了降低网络对数据集的依赖，我们提出通过简化学习目标，在有限的 KITTI 数据集上学习视觉里程计深度模型。

我们观察发现，以往的路面机器人在实现自我定位时，学习的是六自由度的运动  $\mathbf{R}$  和  $\mathbf{t}$ 。然而，在车辆的行驶过程中，其主要运动（如图7.1所示）。由于地面车辆的运动受到其机械结构和动力学的限制，所以如果我们的深度网络结构只专注于主要运动时并不会导致过多的姿势偏移。另外，我们观察发现由于噪声的存在，观察到的次要运动总是具有较低的信噪比。基于以上观察我们得出结论，当深度学习网络聚焦于学习地面车辆的主要运动时，可以减少对训练数据量的依赖，简化学习问题的同时可以降低其余自由度带来的冗余误差，是一个可观的深度学习视觉定位方案。

地面车辆的约束运动模型已被广泛采用在基于几何学的视觉里程测量方法中。Scaramuzza 等人<sup>[125]</sup> 基于运动约束和阿克曼转向定律<sup>[126]</sup> 提出 1-point-RANSAC 用于自我动作估计，以提高实时性能。Choi 等人<sup>[127]</sup> 考虑了突然的颠簸或相机振动，并放宽了平面假设。Scaramuzza 等人<sup>[128]</sup> 利用全向相机提供的地面特征点根据单应性矩阵法计算机器人自我运动位姿。

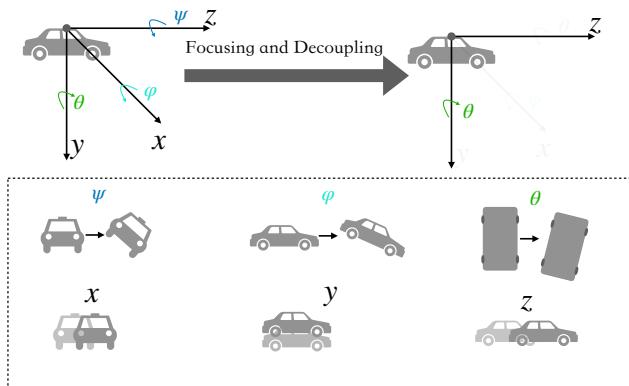


图 7.1 动机：将网络学习目标集中在路面车辆的主要运动上以简化深度学习目标并降低冗余误差。

本文尝试了一种全新的路面车辆运动模型，我们将神经网络的学习目标集中在主要运动的自由度上（命名为运动聚焦），并定量评估了当忽略某个次要运动自由度时所引起的姿态位移，并探讨了通过考虑地面车辆运动模型（命名为运动解耦）来最小化姿态位移。

此外，我们构建了仅有 4 个卷积层的轻型卷积神经网络来学习地面车辆的

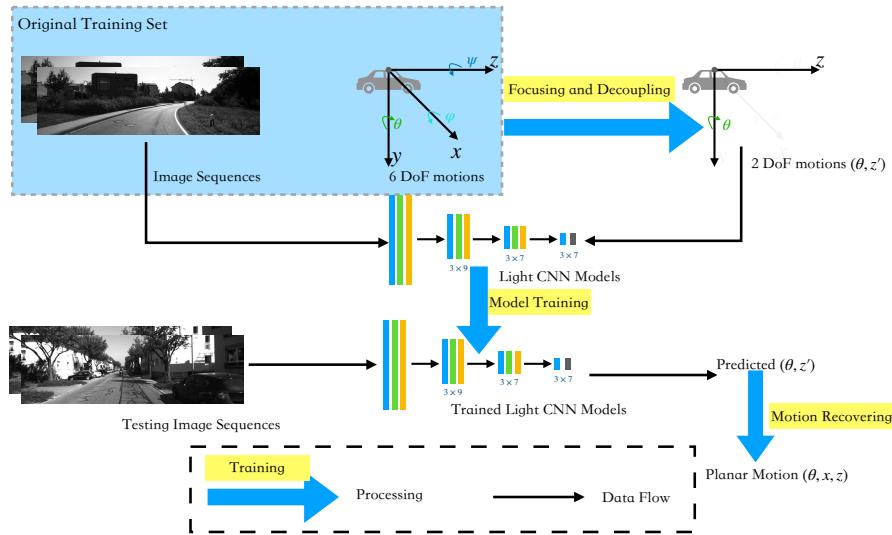


图 7.2 运动聚焦与解耦算法系统结构。

显著运动，并进行了实验，表明运动聚焦与运动解耦可以提高自我运动估计性能。整个系统的结构如图所示7.2。本章的主要贡献是：

1. 通过对运动聚焦引起的地面真实姿势位移进行定量评估，发现其位移相对较小，可以通过实验证明运动聚焦的可行性；
2. 分析了X轴意外平移的原因，建立了X轴平移和Y轴旋转的关系模型，并利用该模型来减少运动解耦引起的姿势位移；
3. 我们在KITTI数据集上进行了对比实验，表明所提出的运动聚焦和解耦可以减少训练时间，提高学习性能；
4. 构建了轻型卷积神经网络来模拟地面车辆的主要运动，模型足够轻，可以在GPU上用2G左右的内存进行训练，并在CPU上实时运行（每秒超过200帧）。为了促进视觉里程计的发展，我们公开了源代码<sup>①</sup>。

本文的其余部分组织如下。第一节7.1描述了我们的方法，包括运动聚焦和训练细节。在第??节中，我们的方法在KITTI数据集<sup>[121]</sup>上进行了评估。我们在??中对本文进行了总结并讨论了未来的工作。

## 7.1 方法

本节在7.1.1首先介绍了运动聚焦和运动解耦的数据处理方法；然后在章节7.1.2中介绍了关于地面车辆视觉里程模型的网络结构和训练细节。

<sup>①</sup> <https://github.com/TimingSpace/DMVOGV>

### 7.1.1 运动聚焦和运动解耦

运动聚焦是通过将注意力集中在主要运动上来简化学习目标的一种思想，这在7.1.1.1中有详细描述，运动聚焦引起的姿势位移在节7.2.2.1中进行了评估。运动解耦减少了运动聚焦引起的姿势位移，这在7.1.1.2中进行了描述，在7.2.2.2中进行了定量评价。运动聚焦和运动解耦引起的性能提升在7.2.2.3节中进行评估。

#### 7.1.1.1 运动聚焦

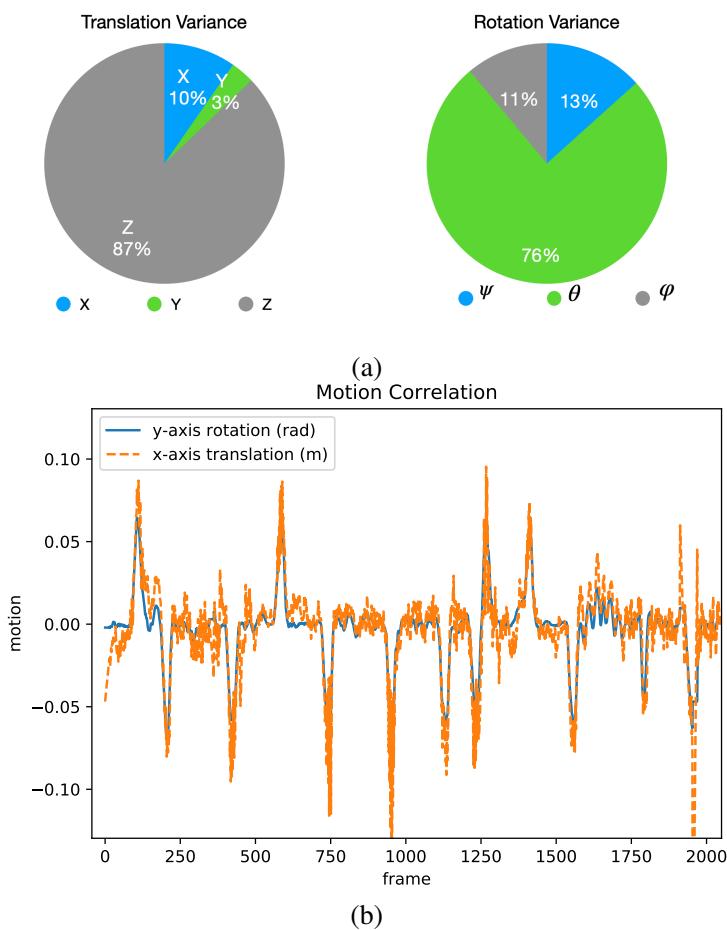


图 7.3 运动模式分析。(a) 沿着或围绕不同轴的运动比较；(b)X 轴平移和 Y 轴旋转的相关性。

运动聚焦就是忽略地面车辆的微不足道的运动，集中精力对大部分运动进行建模。我们在分解运动时采用常规的相机坐标系作为参考系（如图7.1所示），该坐标系为右手系，定义如下：原点为相机的光学中心，z 轴定义为前进光轴，x 轴水平向右，y 轴垂直向下。围绕 x 轴、y 轴和 z 轴的旋转运动分别用 Euler 角  $\psi$ 、 $\varphi$  和  $\theta$  表示。沿不同轴的平移运动分别用  $x$ 、 $y$  和  $z$  表示。

我们在一个典型的地面车辆运动估计数据集——KITTI 视觉里程数据集<sup>[121]</sup>上对地面车辆的运动模式进行了定量评估。我们计算了 KITTI 序列 00 中关于

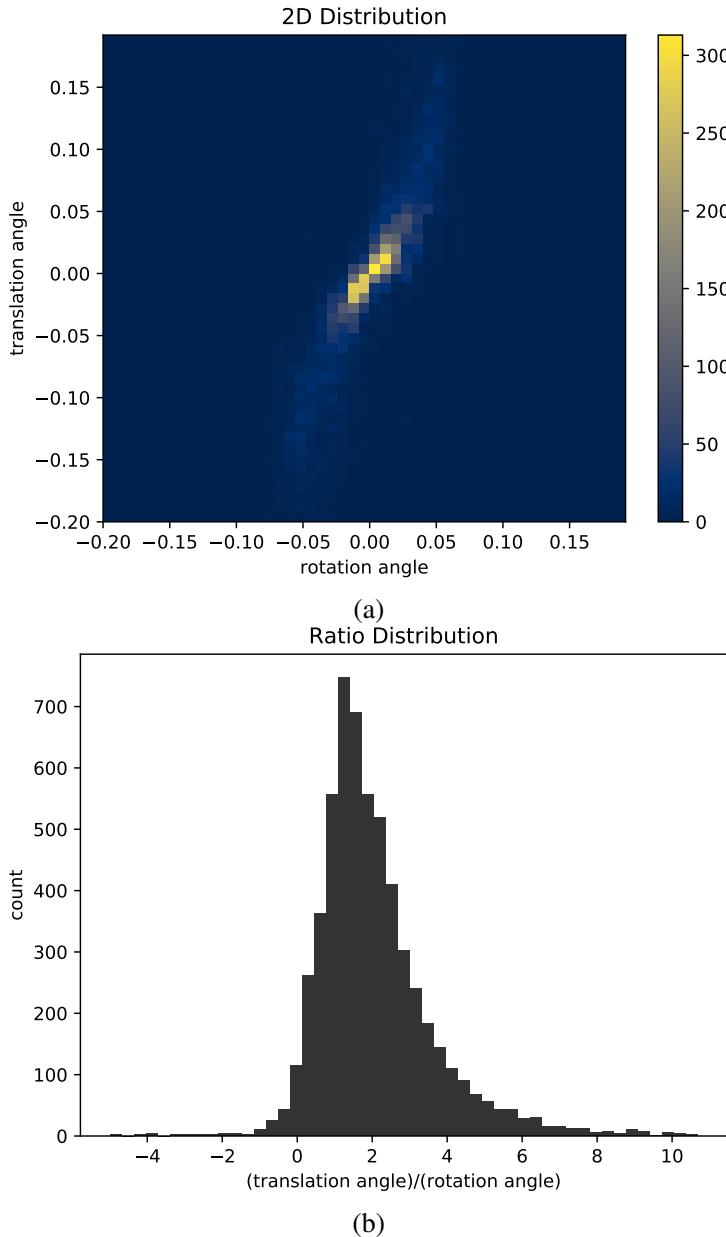


图 7.4 旋转角和平移角的关系。(a) 平移角和旋转角的二维直方图; (b) 平移角和旋转角的一维直方图。

不同轴的平移运动和旋转运动的方差, 如图7.3(a)所示, 更多的方差可视化与图7.3(a)类似, 详见附录。图7.3(a)显示了地面车辆的大部分运动是 z 轴方向的平移运动和 y 轴方向的旋转运动。所以我们建议简化运动估计目标, 只关注大多数运动, 我们称这个建议为运动聚焦。由运动聚焦引起的姿势位移在第7.2.2.1节中进行了评估。

### 7.1.1.2 运动解耦

然而, 如图7.3(a)所示, 沿 x 轴仍有不可忽略的平移运动, 具体分析见表7.1。由此可知, 如果直接全部忽略 X 轴的运动, 会造成较大的漂移姿势 (10%)。然

而，考虑到动力学约束，地面车辆不能沿 X 轴移动太多，为什么存在高达 10% 的 X 轴平移呢？当深入研究地面车辆的运动模式时，我们发现 X 轴运动是由运动表示方法产生的。

$$\begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ 0 & 1 \end{pmatrix} \quad (7.1)$$

这里  $\mathbf{I}$  是一个  $3 \times 3$  的单位矩阵。在这种表示方式中，平移运动  $\mathbf{t}$  先于旋转运动  $\mathbf{R}$ 。因此，当车辆有旋转运动时，参考坐标系统已经发生了变化，前向运动  $z'$  被映射成较小的前向运动  $z$  与侧向运动 ( $x$  轴平移  $x$ )，如图7.5(a)所示。它引起的平移角  $\alpha$ ，定义为：

$$\alpha = \arctan\left(\frac{x}{z}\right) \quad (7.2)$$

这里  $x$  和  $z$  分别表示  $x$ -轴与  $z$ -轴的平移。当我们把  $x$  轴平移和  $y$  轴旋转可视化后，就可以得到验证，如图7.3(b)所示，因为这两个运动是高度相关的。图7.3(b)只能可视化一个子序列的局部相关性。我们利用图7.4中的两个直方图来可视化 KITTI 数据集序列 00-10 中，所有  $Y$  轴旋转角  $\theta$  和平移角  $\alpha$  的全局关系。我们利用图中的两个直方图 (7.4) 来可视化所有 KITTI 序列 00-10 中所有  $Y$  轴旋转角  $\theta$  和平移角  $\alpha$  的全局关系。图7.4(b)中的 1d 直方图显示了  $\alpha/\theta$  的分布，图7.4(a)中的 2d 直方图可视化了  $\alpha$  和  $\theta$  的联合分布。这两个分布都表明  $y$  轴的旋转角  $\theta$  和平移角  $\alpha$  是相关的。那么如何重新制定运动表示法来减少运动的相关性呢？一个简单的方法是将平移重写为：

$$\begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}' & \mathbf{0} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{t}' \\ 0 & 1 \end{pmatrix} \quad (7.3)$$

在这个公式中，车辆的旋转是先于平移的，所以平移运动是相对于旋转运动后的参考系而言的，不会被重新映射。可以得出的关系是： $\mathbf{R}' = \mathbf{R}$ ,  $\mathbf{t}' = \mathbf{R}^{-1}\mathbf{t}$ 。然而，如图??所示，绕  $y$  轴的旋转角度  $\theta$  不等于由旋转产生的平移角  $\alpha$ 。我们需要找出  $\alpha$  和  $\theta$  之间的关系  $\alpha = f(\theta)$ 。然后，我们就只需保持  $y$  轴的旋转  $\theta$  和汽车的前移  $z$ ，使用平移角  $\alpha$ ，用公式(7.4)来恢复车辆的运动。

$$(x, z) = z'(\sin(f(\theta)), \cos(f(\theta))) \quad (7.4)$$

在图??中，A 点是车辆后轴的中心，标记 B 是安装摄像头的位置， $l$  表示 A 和 B 之间的距离。通过视觉里程测量法估算出的车辆平移距离等于  $B$  和  $B'$  之间的距离，用  $z'$  表示。我们将图??简化为图??。根据阿克曼转向定律<sup>[126]</sup>， $OA \perp AB$  且

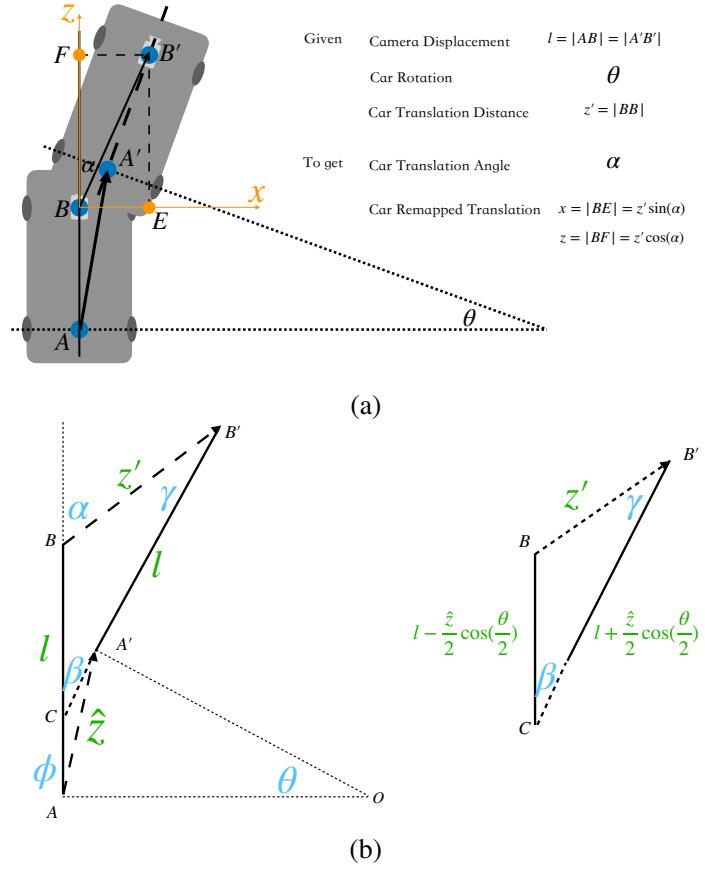


图 7.5 地面车辆旋转模型。(a) 旋转模型; (b) 简化的旋转模型。

$OA' \perp A'B'$ , 所以  $\phi = 0.5\beta = 0.5\theta$ 。在三角形  $CBB'$  中, 根据正弦定律可知:

$$\frac{\sin(\gamma)}{\sin(\beta)} = \frac{l - \frac{\hat{z}}{2} / \cos(\frac{\theta}{2})}{z'} \quad (7.5)$$

因为  $\theta$  趋近于 0, 所以  $\cos(\frac{\theta}{2}) \approx 1$ , 且  $\frac{\gamma}{\beta} \approx \frac{\sin(\gamma)}{\sin(\beta)}$ , 因此

$$\frac{\gamma}{\beta} \approx \frac{l - \frac{\hat{z}}{2}}{z'} \quad (7.6)$$

又根据三角形  $CBB'$  中的余弦定律,  $d = |AC| \approx 0.5|AA'| = 0.5\hat{z}$

$$\begin{aligned} z'^2 &= (l + d)^2 + (l - d)^2 - 2(l + d)(l - d) \cos(\beta) \\ &= 2l^2 + 2d^2 - 2(l^2 - d^2) \cos(\beta) \approx 4d^2 \end{aligned} \quad (7.7)$$

因此  $z' \approx \hat{z}$ , 可知平移角度  $\alpha$  和旋转角度  $\theta$

$$\alpha = \beta + \gamma \approx (\frac{l}{z'} + 0.5)\beta = (\frac{l}{z'} + 0.5)\theta \quad (7.8)$$

我们用位移角度  $a$  构建旋转矩阵  $\mathbf{R}_\alpha$ ,

$$\mathbf{R}_\alpha = \begin{pmatrix} \cos(a) & 0 & \sin(a) \\ 0 & 1 & 0 \\ -\sin(a) & 0 & \cos(a) \end{pmatrix} \quad (7.9)$$

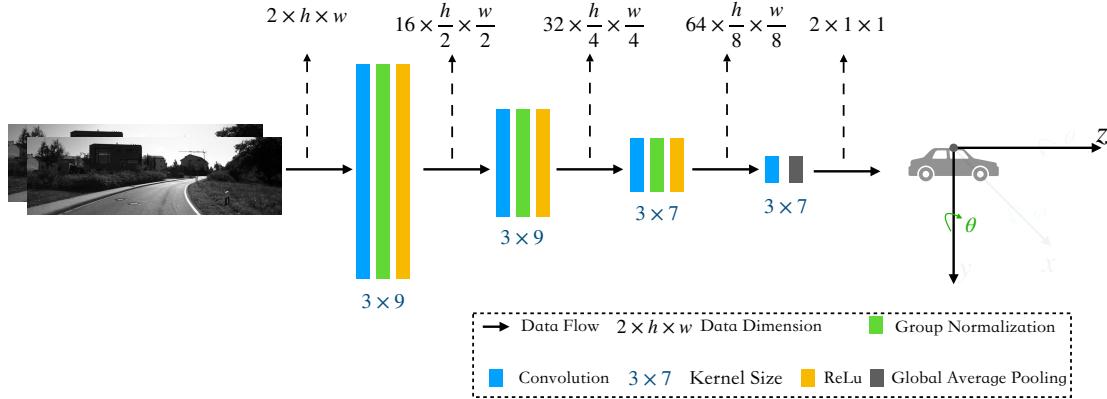


图 7.6 自我运动估计的轻型卷积神经网络

然后得到  $\mathbf{t}'$  变形为

$$\mathbf{t}' = \mathbf{R}_\alpha^{-1} \mathbf{t} \quad (7.10)$$

所需的车辆前行运动  $z'$  是  $\mathbf{t}'$  的第三个元素。到目前为止，地面车辆的规划者运动可以由两个变量来表示：旋转角  $\theta$  和重映射前向运动  $z'$ 。我们专注于学习二维运动，以简化学习目标。模型和学习细节将在下一节介绍，运动聚焦和解耦引起的性能提升将在 7.2.2.3 中进行评估。

### 7.1.2 模型与训练

我们构建一个轻型网络结构来学习地面车辆的主要运动。如图 7.6 所示。模型主要由卷积层组成，除了最后一层外，每个卷积层后面都有一个组归一化层<sup>[129]</sup>和 ReLU 层。与 Zhou 等人相同的是<sup>[110]</sup>，我们使用全局平均池化层<sup>[130]</sup>而不是全连接层作为最后一层，以减少过拟合。我们观察到，地面车辆拍摄的图像的光流主要是水平的，特别是当车辆转弯时，如图 7.7 所示，所以我们利用卷积层与非正方形核来实现更大的水平感受野。此外，我们还采用了扩张卷积层<sup>[131]</sup>，以较少的参数获得更大的感知场。

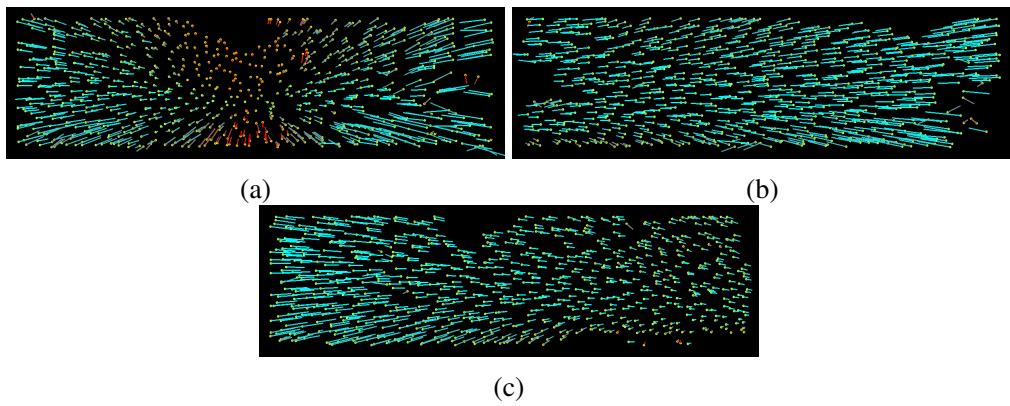


图 7.7 地面车辆的光流。(a) 前进时的光流；(b) 左转时的光流；(c) 右转时的光流。

模型的输入是堆叠的灰色图像对，我们不仅使用序列图像来构造图像对(帧

间隔为 0)，而且在 [-4,4] 中为每个样本随机选择一个帧间隔，作为一个数据增量过程。

输出是由 y 轴旋转  $\theta$  和重映射前向平移  $z'$  所代表的相应主要相机运动。 $z'$  是根据(7.10)得到的。L2 损失被用作监督函数：

$$L_2 = \|\theta_g - \theta_p\|_2 + \|z'_g - z'_p\|_2 \quad (7.11)$$

下标为  $p$  的变量  $\theta_p$  和  $z'_p$  代表预测的结果，下标为  $g$  的变量  $\theta_g$  和  $z'_g$  代表 the 地面真值。我们使用 ADAM<sup>[132]</sup> 来优化模型参数，学习率设置为 0.001，50 个 epoch 后线性衰减。

模型训练完成后，输入一个新图像序列后，从模型输出中可以得到旋转角  $\theta$  和前向运动  $z'$ 。我们首先计算平移角  $\alpha$ ，并假设关于其他轴的旋转为零，并构造旋转矩阵  $\mathbf{R}_\theta$  和  $\mathbf{R}_\alpha$ ，则车辆转换向量  $\mathbf{t}_\alpha = \mathbf{R}_\alpha(0, 0, z')^T$ ，该方程等同于(7.4)，称为运动恢复。路面车辆整体运动矩阵可以表示为：

$$\mathbf{T}_i = \begin{pmatrix} \mathbf{R}_\theta & \mathbf{t}_\alpha \\ 0 & 1 \end{pmatrix} \quad (7.12)$$

然后，通过累积可以得到车辆位姿为：

$$\mathbf{P}_i = \mathbf{P}_{i-1}\mathbf{T}_i \quad (7.13)$$

## 7.2 实验

我们在 KITTI 数据集<sup>[121]</sup> 上进行了四个实验来评估所提出的方法的性能。首先，我们对 KITTI 数据集和实验平台进行介绍说明。其次，我们详细介绍了四个实验：姿势位移评估、运动解耦性能、自我运动估计改进以及与其他方法的比较。最后，我们对实验结果进行讨论和分析。

### 7.2.1 数据集和实验平台

KITTI 基准提供了 22 个测试序列，其中前 11 个序列为地面真实姿态评估。每个测试序列中都提供了 RGB 图像、灰色图像和激光雷达点云。

训练和测试我们的模型时，我们只利用单眼灰色图像与地面真实姿势 (KITTI 数据集序列 00-10)。

我们把训练数据集分成了四个不同的训练-评估模型，用于定量评估所提出的运动聚焦和解耦对自我运动估计的改进，详细内容见章节7.2.2.3；为了与其他

相对方法进行比较，在章节7.2.2.4中，我们使用 KITTI 00-08 进行训练，09-10 进行评估，这与其他基于学习的方法是一样的，以便进行公平比较。

我们使用 RPE（相对姿势误差的简称）的平均值包括相对旋转误差和相对位移误差<sup>[121]</sup> 作为评估指标。

我们的算法是基于 PyTorch 用 Python 实现的，PyTorch 是一个成熟的深度学习框架，具有方便的 Python 接口。

我们的代码已在 github 网站公开，该算法在个人笔记本电脑上进行测试，其配置内存为 16GB，CPU 为 Intel Core i7-7700@2.80GHz，主频为 2.80GHz，Nvidia 1060 GPU，6GB 图形内存。测试环境为 Ubuntu 18.04，使用 CUDA 10.0 和 Python 3.6.9。模型训练只需要 2.0G 的 GPU 内存当批量大小设置为 30 时，且即使只用 CPU 进行测试的情况下，测试频率也可达到 200FPS(帧/秒) 以上。

## 7.2.2 实验结果

首先，我们在第7.2.2.1节中通过 RPE 评估运动聚焦所造成的姿势位移。第二，在第7.2.2.2节中，我们评估了运动解耦后姿势位移的缓解 (mitigation)。第三，在第7.2.2.3节中，我们评估了所提出的运动聚焦和解耦对自我运动估计的改进。最后，我们在章节7.2.2.4中比较了我们与其他基于学习和基于几何的方法的结果。

### 7.2.2.1 Motion Displacement by Motion Focusing

%由于地面飞行器的运动受其动力学和机械结构的限制，其大部分运动是沿 z 轴和绕 y 轴的。

为了说明运动聚焦的可行性，我们定量地评估了当忽略部分或所有其他无关紧要的运动维度时，运动聚焦的程度对姿态漂移的影响。

我们重建了运动减少后的姿势，并利用 RPE<sup>[121]</sup> 来评估姿势位移。KITTI 数据集序列 00-10 的平均 RPE 记录在表7.1中。在表7.1中，列和行的名称分别代表

表 7.1 Average RPE When Only Keeping Part of Vehicle Motion

R / t	z RPE(%) /NID	cxz RPE(%) /NID	yz RPE(%) /NID	zyz RPE(%) /NID
y	2.20 /4	2.06 /3	2.45 /3	2.34 /2
xy	1.92 /3	1.77 /2	1.76 /2	1.56 /1
zy	2.05 /3	1.91 /2	1.47 /2	1.27 /1
xyz	1.92 /2	1.81 /1	0.49 /1	0 /0

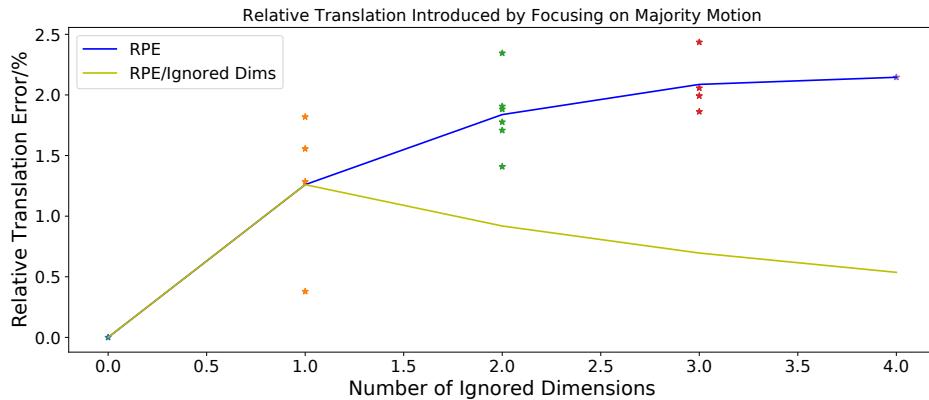


图 7.8 当忽略不同数量的维度时 RPE 平均值比较。

保留的旋转轴和位移轴, NID 表示忽略的维度数量。当我们只保留 z 轴平移和 y 轴旋转时, 忽略四个维度 (NID=4) 时, 重建路径的 RPE 为 2.20/位移主要积累在 z 轴上, z 轴上的位移也受运行环境的影响, 因为当路面有高低起伏时, 位移会比较大 (如图??中表示的序列 10), 而当路面几乎平坦, 位移就比较小 (如图??中的序列 07)。更多可视化的重构路径详见附录。

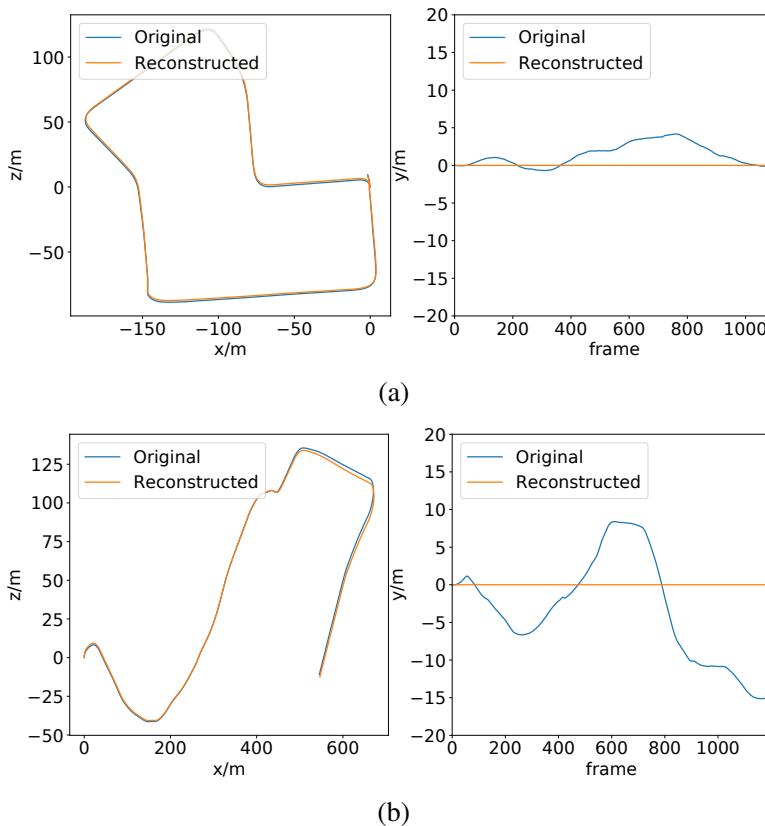


图 7.9 重建路径的可视化。(a) 重建的 KITTI 07; (b) 重建的 KITTI 10。

为了更好的理解, 我们还在图 RPEs7.8 中可视化了 RPE。蓝色的线表示忽略不同维度数时的平均 RPEs 误差, 我们将平均的 RPE (成本) 除以所进行的维度数 (收益), 如图7.8中的黄线所示, 该比率可视为成本收益指标。只保留 z 轴平

移和 y 轴旋转的成本收益比相对较小。

### 7.2.2.2 Pose Displacement Improvement by Motion Decoupling

运动解耦的目标是减少忽略 x 轴平移时的姿态位移，其方法将在章节7.1.1.2中介绍。为了显示所提出的运动解耦的效率，我们定量评估了运动解耦所减少的姿态位移，如表7.1所示。表中第一行表示当忽略 X 轴的平移时将导致更多的姿势偏移（从 2.06 根据公式(7.8)，平移角  $\alpha$  和旋转角  $\theta$  之间是线性关系。然而，线性映射的斜率不是固定的，而是相对于不固定的前向运动  $z$  而言的。

为了简化问题，我们首先使用固定斜率参数来变换所有的前向运动，并使用不同比例测试 RPE。结果如图所示7.10(a)。当比值设为 1.7 时，我们得到的 RPE 最小。可以被解释为当车辆处于旋转状态时，车辆的平均前移距离约为  $\frac{1.7-0.5}{l}$  米。为了更好地理解图7.10(a)中的条形图，我们使用不同的颜色来表示不同的情形。运动解耦的目标是减少忽略 x 轴平移时的姿态位移，其方法将在章节7.1.1.2中介绍。为了显示所提出的运动解耦的效率，我们定量评估了运动解耦所减少的姿态位移，如表7.1所示。表中第一行表示当忽略 X 轴的平移时将导致更多的姿势偏移（从 2.06 根据公式(7.8)，平移角  $\alpha$  和旋转角  $\theta$  之间是线性关系。然而，线性映射的斜率不是固定的，而是相对于不固定的前向运动  $z$  而言的。

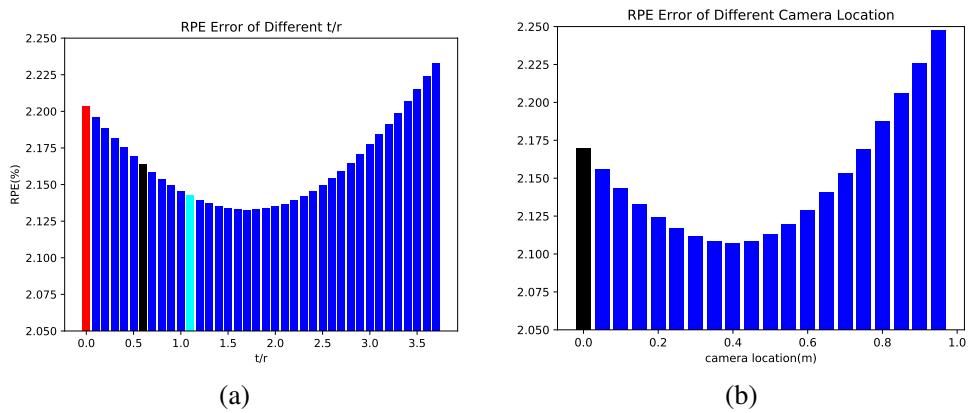


图 7.10 Decoupling performance visualization. (a) Different fixed ratio; (b) Different dynamic ratio

固定的比例会忽略车辆前行距离的影响，所以我们尝试使用动态比例。我们使用不同的摄像机位置  $l$ ，用公式(7.8)计算比率，然后评估重建路径的 RPE，如图7.10(b)所示。当摄像头位置设置为距离后轴 0.4m 时，RPE 误差最小。图7.10(b)中的黑色条形图代表着平移角  $\alpha = 0.5\theta$ ，这与图7.10(a)中的黑色条形图相同。计算比率时，用(7.8)计算，则评估重建路径 RPE，如图7.10(b)所示。当摄像头位置设置为距离后轴 0.4m 时，RPE 误差最小。图7.10(b)中的黑条也代表着平移角  $\alpha = 0.5\theta$ ，这与图7.10(a)中的黑条相同。图??中可以直观地看到动态

解耦和静态解耦的比较。在图??中，蓝色条形代表没有运动解耦的 RPE，黄色和绿色条形分别代表静态解耦（比例设为 1.7）和动态解耦（摄像机位置  $l$  设为 0.4m）的 RPE，红色条形代表我们同时保持 x 轴和 z 轴平移运动时的 RPE。可以发现，动态解耦和静态解耦都可以降低 RPE。当我们同时保持 x 轴平移和 z 轴平移时，动态解耦的 RPE 比静态解耦的 RPE 要低，而且更接近 RPE。

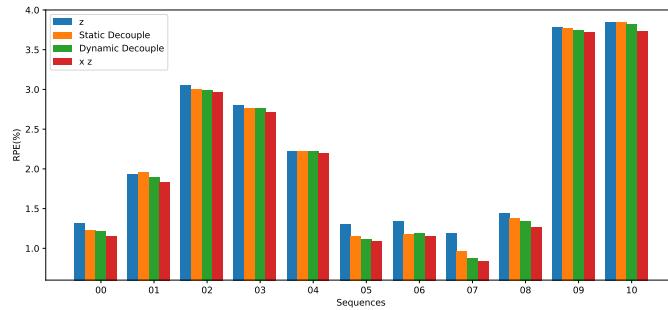


图 7.11 采用固定的解耦比率与动态比率时解耦性能比较。

### 7.2.2.3 Performance Improvement by Motion Focusing and Decoupling

表 7.2 The improvement of motion focusing

Train	Test	Learn All Motion		Learn $R_y, t_z$	
		Trans (%)	Rot (deg/m)	Trans (%)	Rot (deg/m)
00	02 04 06 08 10	26.8	0.137	23.9	0.110
00 02	04 06 08 10	18.3	0.095	16.7	0.070
00 02 04	06 08 10	17.6	0.091	16.9	0.076
00 02 04 06	08 10	15.3	0.082	13.2	0.065

我们通过实验研究运动聚焦和解耦的影响。我们使用相同的训练数据来训练两种模型：1) MFM（运动聚焦模型），只学习 Y 轴的旋转和 Z 轴的平移；2) AMM（全运动模型），学习六个自由度的运动。在同一测试集上的 RPE 可以作为显示改进的指标。实验是在不同的训练-测试数据分集上进行的，以避免偶然性。我们记录了训练集的损耗变化曲线和测试集的 RPE。如图7.12所示。对于所有的训练分割，MFM 模型收敛比 AMM 模型快。不同训练模型的测试 RPE 记录在表7.2中，并直观地显示在图7.13中。我们可以发现，在不同的训练数据模式下，运动聚焦模型的结果都比学习所有 6 自由度的运动模型结果好，位移误差提高了约 2%，旋转提高了 0.2degree/m。从测试结果中还可以观察到，随着训练数据的增加，运动聚焦模型和所有运动模型的测试效果都越来越好。同时我们注意

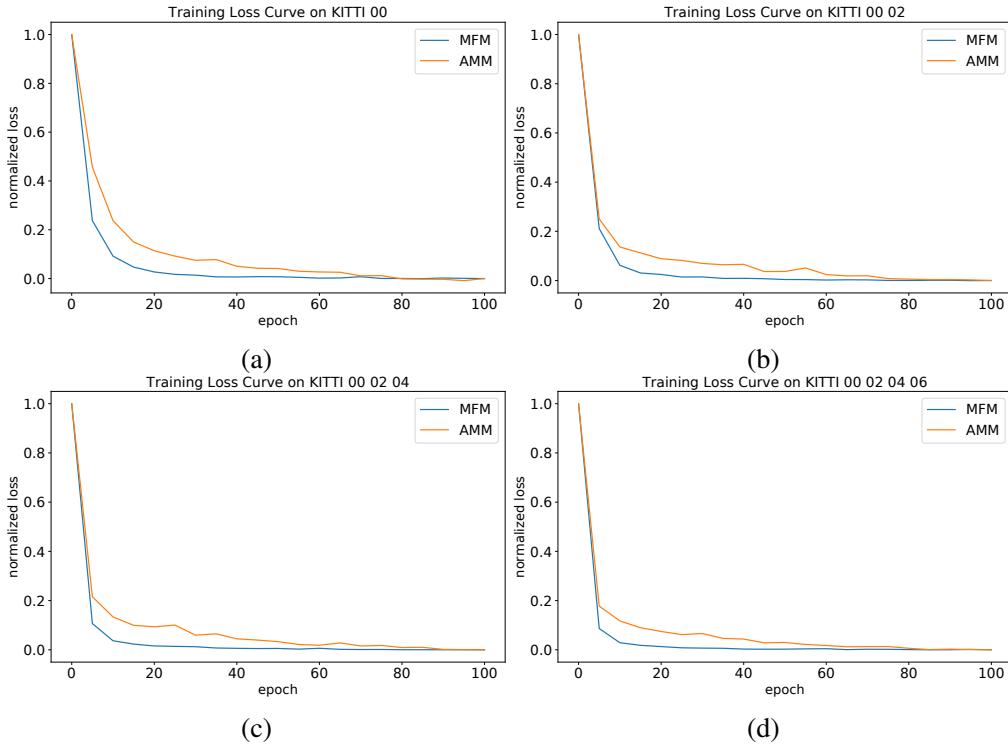


图 7.12 训练损失曲线比较。(a) 在 KITTI 00 上的训练; (b) 在 KITTI 00 02 上的训练; (c) 在 KITTI 00 02 04 上的训练; (d) 在 KITTI 00 02 04 06 上的训练。

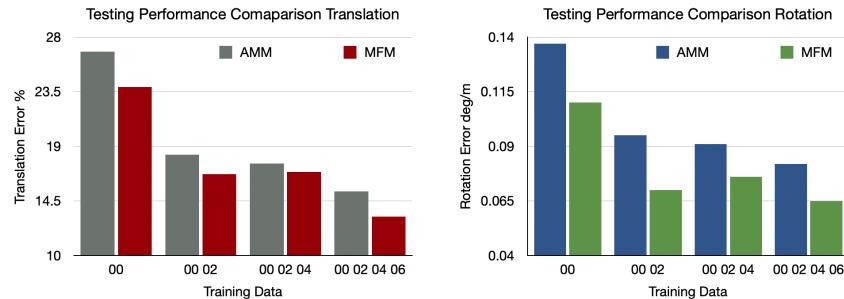


图 7.13 聚焦训练后性能的提升。

到，运动聚焦模型是由带有漂移姿势的地面真值运动聚焦和解耦后进行训练的，但测试性能仍然较好。

#### 7.2.2.4 与其他方法的比较

我们将我们的算法与其他基于深度学习和基于传统几何的方法的算法进行比较。我们的模型在 KITTI 序列 00-08 上进行训练，在 KITTI 09 和 10 上进行测试，数据分割与其他基于卷积神经网络（CNN）方法相同<sup>[110, 113, 114]</sup>。测试的 RPE 记录在表 7.3 和 7.4 中。由于 SfM-Learner<sup>[110]</sup> 和 GeoNet<sup>[114]</sup> 的模型都是以自监督的方式进行无绝对尺度的训练，因此在评价前，它们的路径与地面路径真值是一致的。Zhan 等人的模型<sup>[113]</sup>、DeepVO<sup>[108]</sup> 和我们的方法都是用绝对尺度训练的，所

表 7.3 Comparison with other Learning-based Methods

Seq	Zhan et al. (from <sup>[113]</sup> )		DeepVO (from <sup>[108]</sup> )		SfM-Learner. (from <sup>[110]</sup> )		GeoNet (from <sup>[114]</sup> )		Our Method	
	Trans	Rot	Trans	Rot	Trans	Rot	Trans	Rot	Trans	Rot
	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)
09	11.92	0.0360	-	-	17.84	0.0678	26.93	0.0954	9.26	0.0229
10	12.62	0.0343	8.11	0.0883	37.91	0.1778	24.69	0.0843	9.10	0.0221
Avg	12.27	0.0351	8.11	0.0883	28.88	0.1228	25.81	0.0899	9.18	<b>0.0225</b>

表 7.4 Comparison with Popular Geometry-based Methods

Seq	LIBVISO2 (from <sup>[38]</sup> )		ORB-SLAM (from <sup>[88]</sup> )		Our Method	
	Trans	Rot	Trans	Rot	Trans	Rot
	(%)	(deg/m)	(%)	(deg/m)	(%)	(deg/m)
09	4.04	0.0143	15.30	0.0026	9.26	0.0229
10	25.20	0.0388	3.68	0.0048	9.10	0.0221
Avg	14.62	0.0266	9.49	0.0037	<b>9.18</b>	0.0225

以不需要对齐。单目模式的 ORB-SLAM<sup>[88]</sup> 和 LIBVISO<sup>[66]</sup> 的尺度也是与地面真实路径对齐的。

如表7.3所示，我们的方法优于其他基于(自我运动模型主要由卷积层构建的)CNN的方法<sup>[110, 113, 114]</sup>，并与基于 CNN-RNN(RNN 是 Recurrent Neural Network 的缩写)的方法<sup>[108]</sup> 竞争，后者可以利用时间信息优化姿势。

与两种效果较好的主流传统方法 LIBVISO 单目法 (LIBVISO monocular<sup>[66]</sup>) 和 ORB-SLAM 单目法 (ORB-SLAM monocular<sup>[88]</sup>) 比较，我们得到了较好的平均位移性能。

### 7.2.3 Discussion

在本节中，我们将对结果进行总结，对性能进行分析，并说明所提出的方法的局限性。

### 7.2.3.1 The Efficiency of the Proposed Method

从上面的实验结果来看，可以得出四个方面的结论。1) 运动聚焦不会带来太大额外的位姿偏移。平均 RPE 只有 2% 左右，也就是说车辆姿态跑了 100 米后会有 2 米左右的漂移。路径可视化显示，重建后的路径是可以接受的。2) 运动解耦可以减少姿势位移。运动解耦利用 y 轴旋转和 x 轴平移的相关性来降低重构的姿势位移偏移。动态解耦的性能优于静态解耦。对于动态解耦，需要对摄像机的位置有所要求。在上述实验中，摄像头位置是根据地面真实数据计算出来的。在实际操作中，如果训练数据是自己采集的，也可以直接测量得到。3) 运动聚焦和解耦可以从两个方面提高自我运动估计性能。首先，它缩短了训练时间，所有的训练实验都在 20 个 epoch 内收敛，但所有运动的模型在大约 60 个 epoch 后才收敛，所以运动聚焦模型与所有运动模型相比，可以减少大约 2/3 的训练时间。另外，尽管运动聚焦和运动解耦后训练的地面真值姿势有一定的漂移，但 MFM 的测试性能比 AMM 更好。4) 在与基于几何的方法比较的过程中，发现基于几何的方法并不稳健，在不同序列上获得的性能各异。我们的方法更加稳定和稳健，平均性能更好。我们的方法也优于其他基于 CNN 的方法，在与利用 RNN 提高性能的 DeepVO<sup>[108]</sup> 比较时，我们也获得了更好的相对旋转性能，但位移性能比 DeepVO 差。

### 7.2.3.2 Why Motion Focusing and Decoupling Works

运动聚焦和解耦性能较好的原因有三点：首先，地面车辆的运动是受约束的，且分布不均衡，忽略不明显的运动不可能造成太大的姿态位移误差，这一点已在第7.2.2.1节中得到证明。而这也是运动聚焦的根本基础；第二，不重要的运动太有限，没有足够的信噪比，所以模型在瞄准它们建模时，很容易被噪声干扰；第三，当我们只聚焦于二维运动时，训练任务变得简单很多。当采用轻型模型时，训练数据相对丰富。实验证明，增加训练数据量确实能提高测试性能，如表7.2所示。

### 7.2.3.3 The Limitation

当汽车有近似平面运动时，所提出的方法可以获得更好的性能。当有明显的 x 轴旋转时，性能将下降。如图??所示，序列 09 和 10 的 RPE 误差相对较高，因为在这两个序列中，地面车辆的运动不是平面的，如图7.9所示。为了解决这个局限性，一个可行的方法是采用其他传感器，如 IMU（惯性测量单元）来估计 X 轴的旋转，这可以作为所提出的模型所估计的平面运动的补充。

此外，另一个限制是因为我们的算法是基于摄像机是平放且是向前看的假设，所以如果摄像机有俯仰角，那么车辆的向前运动将被映射成 Z 轴运动和 Y 轴运动。如果摄像机有一个俯仰角，那么车辆的向前运动将被映射成 Z 轴运动和 Y 轴运动。在这种情况下，摄像机俯仰角  $\sigma$  应在训练前进行校准。然后，平移运动应该被重新定义为

$$t_\sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\sigma) & -\sin(\sigma) \\ 0 & \sin(\sigma) & \cos(\sigma) \end{pmatrix} t \quad (7.14)$$

这些局限性也可以通过利用视觉里程测量的多模型结构来解决，每个子模型只关注地面车辆的一个维度运动，六自由度可以用六个分离的模型来学习，在这种情况下，应该分析模型权重分摊的影响。

### 7.3 本章小结

本文通过提出运动聚焦和运动解耦，将地面车辆的运动聚焦为两个自由度上的运动。实验证明了运动聚焦的可行性，并通过定量的姿态位移评估，进一步降低了姿态位移。基于地面车辆的旋转模型，我们提出了运动解耦，进一步降低了姿势位移。我们构建了一个轻型 CNNs 网络模型来模拟二自由度运动，它可以在 CPU 上实时运行。实验证明，运动聚焦和解耦可以提高小我运动估计性能，缩短收敛时间。在 KITTI 数据集上与其他方法的比较表明，所提出的方法的性能与其他端到端的视觉里程测量方法相当，甚至更好，并且比基于几何的方法更稳健。

## 致谢

## 参考文献

- [1] TANG J. Made in China 2025[J]. Integración & comercio, 2017: págs. 204-215.
- [2] FERNANDEZ D, PRICE A. Visual odometry for an outdoor mobile robot[C]//IEEE Conference on Robotics, Automation & Mechatronics. 2005.
- [3] GONZALEZ R, RODRIGUEZ F, GUZMAN J L, et al. Combined visual odometry and visual compass for off-road mobile robots localization[J]. Robotica, 2011, 30(06): 865-878.
- [4] BAY H, TUYTELAARS T, VAN GOOL L. Surf: Speeded up robust features[J]. Computer vision-ECCV 2006, 2006: 404-417.
- [5] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF[C]//Computer Vision (ICCV), 2011 IEEE international conference on. 2011: 2564-2571.
- [6] CALONDER M, LEPETIT V, STRECHA C, et al. Brief: Binary robust independent elementary features[C]//European conference on computer vision. 2010: 778-792.
- [7] LOWE D G. Object recognition from local scale-invariant features[C]//Computer vision, 1999. The proceedings of the seventh IEEE international conference on: vol. 2. 1999: 1150-1157.
- [8] HARRIS C, STEPHENS M. A combined corner and edge detector.[C]//Alvey vision conference: vol. 15: 50. 1988: 10-5244.
- [9] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005: 886-893.
- [10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. 2012: 1097-1105.
- [11] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
- [12] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[J], 2015: 770-778.
- [13] MILFORD M J, WYETH G F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights[C]//Robotics and Automation (ICRA), 2012 IEEE International Conference on. 2012: 1643-1649.
- [14] CORKE P, PAUL R, CHURCHILL W, et al. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation[C]//Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on. 2013: 2085-2092.
- [15] NEUBERT P, SÜNDERHAUF N, PROTZEL P. Superpixel-based appearance change prediction for long-term navigation across seasons[J]. Robotics and Autonomous Systems, 2015, 69: 15-27.
- [16] MCMANUS C, UPCROFT B, NEWMAN P. Learning place-dependant features for long-term vision-based localisation[J]. Autonomous Robots, 2015, 39(3): 363-387.
- [17] NASEER T, SPINELLO L, BURGARD W, et al. Robust Visual Robot Localization Across Seasons Using Network Flows.[C]//AAAI. 2014: 2564-2570.
- [18] CHURCHILL W, NEWMAN P. Practice makes perfect? managing and leveraging visual experiences for lifelong navigation[C]//Robotics and Automation (ICRA), 2012 IEEE International Conference on. 2012: 4525-4532.

- [19] LOWRY S M, MILFORD M J, WYETH G F. Transforming morning to afternoon using linear regression techniques[C]//Robotics and Automation (ICRA), 2014 IEEE International Conference on. 2014: 3950-3955.
- [20] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
- [21] DONAHUE J, JIA Y, VINYALS O, et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.[C]//Icml: vol. 32. 2014: 647-655.
- [22] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
- [23] SERMANET P, EIGEN D, ZHANG X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. ArXiv preprint arXiv:1312.6229, 2013.
- [24] SÜNDERHAUF N, SHIRAZI S, DAYOUB F, et al. On the performance of convnet features for place recognition[C]//Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. 2015: 4297-4304.
- [25] ARROYO R, ALCANTARILLA P F, BERGASA L M, et al. Fusion and binarization of CNN features for robust topological localization across seasons[C]//Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on. 2016: 4656-4663.
- [26] LOWRY S, SÜNDERHAUF N, NEWMAN P, et al. Visual place recognition: A survey[J]. IEEE Transactions on Robotics, 2016, 32(1): 1-19.
- [27] CUMMINS M, NEWMAN P. FAB-MAP: Probabilistic localization and mapping in the space of appearance[J]. The International Journal of Robotics Research, 2008, 27(6): 647-665.
- [28] CHOW C, LIU C. Approximating discrete probability distributions with dependence trees[J]. IEEE transactions on Information Theory, 1968, 14(3): 462-467.
- [29] SCARAMUZZA D, SIEGWART R Y. Monocular omnidirectional visual odometry for outdoor ground vehicles[C]//International Conference on Computer Vision Systems. 2008.
- [30] RONE W, BEN-TZVI P. Mapping, localization and motion planning in mobile multi-robotic systems[J]. Robotica, 2013, 31(PT.1): 1-23.
- [31] LIU H, MEI T, LUO J, et al. Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing[C]//Proceedings of the 20th ACM international conference on Multimedia. 2012: 9-18.
- [32] SALARIAN M, ILIEV N, CETIN A E, et al. Improved Image-Based Localization Using SFM and Modified Coordinate System Transfer[J]. IEEE Transactions on Multimedia, 2018, 20(12): 3298-3310.
- [33] SCARAMUZZA D, FRAUNDORFER F. Visual Odometry Part II: Matching, robustness, optimization and applications[J]. IEEE Robotics & Automation Magazine, 2012, 19(2): 78-90.
- [34] SCARAMUZZA D, FRAUNDORFER F. Visual Odometry: Part I - The First 30 Years and Fundamentals[J]. IEEE Robotics and Automation Magazine, 2011, 4.
- [35] TRIGGS B, MCLAUCHLAN P F, HARTLEY R I, et al. Bundle adjustment—a modern synthesis[C]//International workshop on vision algorithms. 1999: 298-372.
- [36] COSTANTE G, MANCINI M, VALIGI P, et al. Exploring Representation Learning With CNNs for Frame-to-Frame Ego-Motion Estimation[J]. IEEE Robotics and Automation Letters, 2015, 1(1): 18-25.
- [37] KITT B M, REHDER J, CHAMBERS A D, et al. Monocular visual odometry using a planar road model to solve scale ambiguity[J]., 2011.

- [38] SONG S, CHANDRAKER M, GUEST C. High Accuracy Monocular SFM and Scale Correction for Autonomous Driving[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015: 1-1.
- [39] ZHOU D, DAI Y, LI H. Reliable scale estimation and correction for monocular Visual Odometry[C]// Intelligent Vehicles Symposium (IV), 2016 IEEE. IEEE, 2016: 490-495.
- [40] CHEN O T C, CHEN C C. Automatically-determined region of interest in JPEG 2000[J]. IEEE Transactions on Multimedia, 2007, 9(7): 1333-1345.
- [41] HOIEM D, EFROS A A, HEBERT M. Recovering surface layout from an image[J]. International Journal of Computer Vision, 2007, 75(1): 151.
- [42] LEE B, DANIILIDIS K, LEE D D. Online self-supervised monocular visual odometry for ground vehicles[Z]. Conference Paper. 2015.
- [43] YE H, CHEN Y, LIU M. Tightly coupled 3d lidar inertial odometry and mapping[C]// 2019 International Conference on Robotics and Automation (ICRA). 2019: 3144-3150.
- [44] ZHANG J, SINGH S. LOAM: Lidar Odometry and Mapping in Real-time.[C]// Robotics: Science and Systems: vol. 2: 9. 2014.
- [45] SCARAMUZZA D, FRAUNDORFER F. Visual odometry [tutorial][J]. IEEE robotics & automation magazine, 2011, 18(4): 80-92.
- [46] MOURAGNON E, LHUIILLIER M, DHOME M, et al. Real time localization and 3d reconstruction[C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06): vol. 1. 2006: 363-370.
- [47] NISTER D, STEWENIUS H. Scalable recognition with a vocabulary tree[C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06): vol. 2. 2006: 2161-2168.
- [48] PIAO J C, KIM S D. Real-Time Visual-Inertial SLAM Based on Adaptive Keyframe Selection for Mobile AR Applications[J]. IEEE Transactions on Multimedia, 2019, 21(11): 2827-2836.
- [49] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces[C]// 2007 6th IEEE and ACM international symposium on mixed and augmented reality. 2007: 225-234.
- [50] MUR-ARTAL R, TARDOS J D. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras[J]. IEEE Transactions on Robotics, 2016, 33(5): 1255-1262.
- [51] SONG S, CHANDRAKER M. Robust scale estimation in real-time monocular SFM for autonomous driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1566-1573.
- [52] SONG S, CHANDRAKER M, GUEST C C. Parallel, Real-Time Monocular Visual Odometry[C]// IEEE International Conference on Robotics and Automation. 2013.
- [53] GUTIÉRREZ-GÓMEZ D, PUIG L, GUERRERO J J. Full scaled 3d visual odometry from a single wearable omnidirectional camera[C]// 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012: 4276-4281.
- [54] SCARAMUZZA D, SIEGWART R. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles[J]. IEEE transactions on robotics, 2008, 24(5): 1015-1026.
- [55] SCARAMUZZA D, MARTINELLI A, SIEGWART R. A flexible technique for accurate omnidirectional camera calibration and structure from motion[C]// Fourth IEEE International Conference on Computer Vision Systems (ICVS'06). 2006: 45-45.
- [56] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(3): 611-625.

- [57] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-Scale Direct Monocular SLAM[M]. Springer International Publishing, 2014: 834-849.
- [58] FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: Fast semi-direct monocular visual odometry[C]//2014 IEEE International Conference on Robotics and Automation (ICRA). 2014: 15-22.
- [59] NEWCOMBE R A, LOVEGROVE S J, DAVISON A J. DTAM: Dense tracking and mapping in real-time[C]//2011 International Conference on Computer Vision. 2011: 2320-2327.
- [60] CHEN L, XU D, TSANG I W, et al. Tag-based image retrieval improved by augmented features and group-based refinement[J]. IEEE Transactions on Multimedia, 2012, 14(4): 1057-1067.
- [61] LIN J, DUAN L Y, WANG S, et al. Hnip: Compact deep invariant representations for video matching, localization, and retrieval[J]. IEEE Transactions on Multimedia, 2017, 19(9): 1968-1983.
- [62] CHADHA A, ANDREOPoulos Y. Voronoi-based compact image descriptors: Efficient region-of-interest retrieval with VLAD and deep-learning-based descriptors[J]. IEEE Transactions on Multimedia, 2017, 19(7): 1596-1608.
- [63] GALL J, GEHLER P, LEIBE B. [Lecture Notes in Computer Science] Pattern Recognition Volume 9358 || Fast Techniques for Monocular Visual Odometry[J]., 2015, 10.1007/978-3-319-24947-6(Chapter 24): 297-307.
- [64] CHOI S, KIM T, YU W. Performance evaluation of RANSAC family[J]. Journal of Computer Vision, 1997, 24(3): 271-300.
- [65] PEREIRA F I, LUFT J A, ILHA G, et al. A Novel Resection–Intersection Algorithm With Fast Triangulation Applied to Monocular Visual Odometry[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(11): 3584-3593.
- [66] GEIGER A, ZIEGLER J, STILLER C. StereoScan: Dense 3D Reconstruction in Real-time[C]//Intelligent Vehicles Symposium (IV). 2011.
- [67] GEIGER A, ROSER M, URTASUN R. Efficient large-scale stereo matching[C]//Asian conference on computer vision. 2010: 25-38.
- [68] RUKHOVICH D, MOURITZEN D, KAESTNER R, et al. Estimation of Absolute Scale in Monocular SLAM Using Synthetic Data[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019: 0-0.
- [69] SAXENA A, CHUNG S H, NG A Y. Learning depth from single monocular images[C]//Advances in neural information processing systems. 2006: 1161-1168.
- [70] LUO H, GAO Y, WU Y, et al. Real-time dense monocular SLAM with online adapted depth prediction network[J]. IEEE Transactions on Multimedia, 2018, 21(2): 470-483.
- [71] KARSCH K, LIU C, KANG S B. Depth transfer: Depth extraction from video using non-parametric sampling[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(11): 2144-2158.
- [72] RANFTL R, VINEET V, CHEN Q, et al. Dense monocular depth estimation in complex dynamic scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4058-4066.
- [73] YANG X, GAO Y, LUO H, et al. Bayesian DeNet: monocular depth prediction and frame-wise fusion with synchronized uncertainty[J]. IEEE Transactions on Multimedia, 2019, 21(11): 2701-2713.
- [74] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in neural information processing systems. 2014: 2366-2374.

- [75] YIN X, WANG X, DU X, et al. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields[C] // Proceedings of the IEEE International Conference on Computer Vision. 2017: 5870-5878.
- [76] TATENO K, TOMBARI F, LAINA I, et al. Cnn-slam: Real-time dense monocular slam with learned depth prediction[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6243-6252.
- [77] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM[C] // European Conference on Computer Vision. 2014: 834-849.
- [78] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C] // IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012.
- [79] YANG N, WANG R, STUCKLER J, et al. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 817-833.
- [80] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 270-279.
- [81] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C] // Proceedings of the IEEE international conference on computer vision. 2015: 2650-2658.
- [82] ZHAN H, WEERASEKERA C S, BIAN J W, et al. Visual odometry revisited: What should be learnt?[C] // 2020 IEEE International Conference on Robotics and Automation (ICRA). 2020: 4203-4210.
- [83] XUE F, ZHUO G, HUANG Z, et al. Toward Hierarchical Self-Supervised Monocular Absolute Depth Estimation for Autonomous Driving Applications[J]. ArXiv preprint arXiv:2004.05560, 2020.
- [84] SATTLER T, TORII A, SIVIC J, et al. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?[C] // IEEE Conference on Computer Vision & Pattern Recognition. 2017.
- [85] LUONG Q T, FAUGERAS O D. The fundamental matrix: Theory, algorithms, and stability analysis[J]. International journal of computer vision, 1996, 17(1): 43-75.
- [86] Nister, D. An Efficient Solution to the Five-Point Relative Pose Problem[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(6): 756-770.
- [87] LEPETIT V, MORENO-NOGUER F, FUA P. Epnp: An accurate o (n) solution to the pnp problem[J]. International journal of computer vision, 2009, 81(2): 155.
- [88] MUR-ARTAL R, MONTIEL J, TARDÓS J D. Orb-slam: a versatile and accurate monocular slam system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.
- [89] SÜNDERHAUF N, PROTZEL P. BRIEF-Gist-Closing the loop by simple means[C] // Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. 2011: 1234-1241.
- [90] ARROYO R, ALCANTARILLA P F, BERGASA L M, et al. Fast and effective visual place recognition using binary codes and disparity information[C] // Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on. 2014: 3089-3094.
- [91] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.
- [92] TOLA E, LEPETIT V, FUA P. A fast local descriptor for dense matching[C] // Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. 2008: 1-8.

- [93] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF[C]//Computer Vision (ICCV), 2011 IEEE International Conference on. 2011: 2564-2571.
- [94] SHARIF RAZAVIAN A, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014: 806-813.
- [95] FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395.
- [96] SHEWCHUK J R. Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator[C]//Workshop on Applied Computational Geometry. 1996.
- [97] FISCHLER M A. Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography[J]. Comm of Acm, 1981, 24.
- [98] 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. 2012: 3354-3361.
- [99] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012: 573-580.
- [100] ROSTEN E. Machine learning for very high-speed corner detection[J]. Eccv06 May, 2006.
- [101] SHEWCHUK J R. Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator[G]//Applied computational geometry towards geometric engineering. Springer, 1996: 203-222.
- [102] ENGEL J, KOLTUN V, CREMERS D. Direct Sparse Odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.
- [103] ROBERTS R, NGUYEN H, KRISHNAMURTHI N, et al. Memory-based learning for visual odometry[C]//Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on. 2008: 47-52.
- [104] GUIZILINI V, RAMOS F. Semi-parametric models for visual odometry[C]//Robotics and Automation (ICRA), 2012 IEEE International Conference on. 2012: 3482-3489.
- [105] COSTANTE G, MANCINI M, VALIGI P, et al. Exploring representation learning with cnns for frame-to-frame ego-motion estimation[J]. IEEE Robotics and Automation Letters, 2016, 1(1): 18-25.
- [106] PILLAI S, LEONARD J J. Towards visual ego-motion learning in robots[J]. ArXiv preprint arXiv:1705.10279, 2017.
- [107] COSTANTE G, CIARFUGLIA T A. LS-VO: Learning dense optical subspace for robust visual odometry estimation[J]. IEEE Robotics and Automation Letters, 2018, 3(3): 1735-1742.
- [108] WANG S, CLARK R, WEN H, et al. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]//Robotics and Automation (ICRA), 2017 IEEE International Conference on. 2017: 2043-2050.
- [109] WANG S, CLARK R, WEN H, et al. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks[J]. The International Journal of Robotics Research, 2018, 37(4-5): 513-542.
- [110] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//CVPR: vol. 2: 6. 2017: 7.
- [111] MAHJOURIAN R, WICKE M, ANGELOVA A. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5667-5675.

- [112] LI R, WANG S, LONG Z, et al. Undeepvo: Monocular visual odometry through unsupervised deep learning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). 2018: 7286-7291.
- [113] ZHAN H, GARG R, WEERASEKERA C S, et al. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 340-349.
- [114] YIN Z, SHI J. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): vol. 2. 2018.
- [115] ALMALIOGLU Y, SAPUTRA M R U, de GUSMAO P P, et al. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks[C]// 2019 International Conference on Robotics and Automation (ICRA). 2019: 5474-5480.
- [116] KLODT M, VEDALDI A. Supervising the new with the old: learning SFM from SFM[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 698-713.
- [117] YANG N, STUMBERG L V, WANG R, et al. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1281-1292.
- [118] CLEMENT L, KELLY J. How to train a cat: learning canonical appearance transformations for direct visual localization under illumination change[J]. IEEE Robotics and Automation Letters, 2018, 3(3): 2447-2454.
- [119] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, Florida, USA, 20-25 Jun 2009: 248-255.
- [120] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Caesars Palace, Las Vegas, Nevada, United States, Jun 26- Jul 1, 2016.
- [121] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 3354-3361.
- [122] MADDERN W, PASCOE G, LINEGAR C, et al. 1 Year, 1000km: The Oxford RobotCar Dataset[J]. The International Journal of Robotics Research (IJRR), 2017, 36(1): 3-15. eprint: <http://ijr.sagepub.com/content/early/2016/11/28/0278364916679498.full.pdf+html>.
- [123] SLINKO I, VORONSOVA A, ZHUKOV D, et al. Training Deep SLAM on Single Frames[J]. ArXiv preprint arXiv:1912.05405, 2019.
- [124] WANG W, ZHU D, WANG X, et al. TartanAir: A Dataset to Push the Limits of Visual SLAM[J]. ArXiv preprint arXiv:2003.14338, 2020.
- [125] SCARAMUZZA D, FRAUNDORFER F, SIEGWART R. Real-time monocular visual odometry for on-road vehicles with 1-point ransac[C]//2009 IEEE International Conference on Robotics and Automation. 2009: 4293-4299.
- [126] SIEGWART R, NOURBAKHSH I R, SCARAMUZZA D. Introduction to autonomous mobile robots[M]. MIT press, 2011.
- [127] CHOI S, PARK J, YU W. Simplified epipolar geometry for real-time monocular visual odometry on roads[J]. International Journal of Control, Automation and Systems, 2015, 13(6): 1454-1464.
- [128] SCARAMUZZA D, SIEGWART R. Appearance-Guided Monocular Omnidirectional Visual Odometry for Outdoor Ground Vehicles[J]. IEEE Transactions on Robotics, 2008, 24(5): 1015-1026.

- [129] WU Y, HE K. Group normalization[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.
- [130] LIN M, CHEN Q, YAN S. Network in network[C]//International Conference on Learning Representations (ICLR). Banff, Canada, Apr 14 - 16, 2014.
- [131] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[C]//International Conference on Learning Representations (ICLR). San Juan, Puerto Rico, May 2-4, 2016.
- [132] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. ArXiv preprint arXiv:1412.6980, 2014.

# 个人简历、在学期间发表的学术论文与研究成果

## 个人简历

1992年01月9日出生于安徽省宿州市砀山县。

2011年9月考入同安徽济大学电气工程及其自动化学院自动化专业，2015年7月本科毕业并获得工学学士学位，并获得安徽大学优秀毕业生称号

2015年9月免试进入同济大学控制科学与工程系攻读博士学位至今。

2017年9月 - 2018年3月受香港科技大学刘明教授和陈启军导师共同资助去香港科技大学工学院机器人与多感知实验室（RAM-LAB）做访问学者。

## 发表论文：

- [1] Wang, X., Maturana, D., Yang, S., Wang, W., Chen, Q., and Scherer, S. (2019, November). Improving learning-based ego-motion estimation with homomorphism-based losses and drift correction. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 970-976). IEEE. (机器人领域顶级会议 CCF C 类)
- [2] Wang, X., Zhang, H., Yin, X., Du, M., and Chen, Q. (2018, May). Monocular visual odometry scale recovery using geometrical constraint. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 988-995). IEEE. (机器人领域顶级会议 CCF B 类)
- [3] Yin, X\*. , Wang, X\*. , Du, X., and Chen, Q. Scale Correction for Monocular Visual Odometry Using Depth Estimated with Deep Convolutional Neural Fields, *International Conference on Computer Vision 2017* (\*Both authors contributed equally to this paper.) (计算机视觉顶级会议 CCF A 类)
- [4] Wang, X., and Chen, Q. (2015, August). Vision-based entity Chinese chess playing robot design and realization. In International Conference on Intelligent Robotics and Applications (pp. 341-351). Springer, Cham. (EI)
- [5] Zhang, H., Wang, X., Du, X., Liu, M., and Chen, Q. (2017, July). Dynamic environments localization via dimensions reduction of deep learning features. In International Conference on Computer Vision Systems (pp. 239-253). Springer, Cham. (EI)
- [6] Mingxiao, D., Xiaofeng, M., Zhe, Z., Xiangwei, W., and Qijun, C. (2017, October). A review on consensus algorithm of blockchain. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2567-2572). IEEE. (EI)

## 已投稿论文：

- [1] Zhang, H\*, Wang, X\*, Yin, X., Du, M., Liu, C., and Chen, Q., Geometric Constrained Scale Estimation for Monocular Visual Odometry. Submited to IEEE Transaction on Multimedia. 2020 (\*Both authors contributed equally to this paper.)

## 已授权专利：

- [1] 一种基于三角剖分的单目视觉里程计尺度恢复方法 - 201710346708.6 发明专利，导师外

第一发明人，已授权

**已公开专利:**

1. 一种基于图像特征降维的无人车单目视觉定位方法 - 201710333483.0 , 发明专利, 导师外第三发明人, 公开实质审查中, 2017 年
2. 一种无人车单目视觉定位中对匹配矩阵的图像匹配方法 - 201710333485.X, 发明专利, 导师外第三发明人, 公开实质审查中, 2017 年