涉密论文 □　公开论文 □

浙江大学

本科生毕业设计

总结报告

题目　基于深度学习的环境语义地图构建

姓名与学号　许学成 / 3150105319

指导教师　熊蓉

年级与专业　自动化 2015 级

所在学院　控制科学与工程学院

提交日期　2019 年 6 月

浙江大学本科生毕业论文（设计）承诺书

1. 本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2. 本人在毕业论文（设计）中除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 浙江大学 或其他教育机构的学位或证书而使用过的材料。

3. 与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

4. 本人承诺在毕业论文（设计）选题和研究内容过程中没有伪造相关数据等行为。

5. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

6. 本人完全了解 浙江大学 有权保留并向有关部门或机构送交本论文（设计）的复印件和磁盘，允许本论文（设计）被查阅和借阅。本人授权 浙江大学 可以将本论文（设计）的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编本论文（设计）。

作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

致　谢

时光荏苒，在本科阶段的最后，要感谢的人太多，谨以此对那些在大学生活中直接或间接帮助过我的人表示感谢。

首先要感谢我的导师熊蓉老师。熊老师对我的学习和生活都提供了很多的帮助。老师渊博的知识和缜密的逻辑让我受益匪浅。在实验室学习的时间里，老师每周都与我们交流想法，为我们开辟科研的思路，排除遇到的障碍，引导我们自己去寻找解决的方法。这样的交流和督促使得我逐渐形成了科研的习惯。我同样要感谢在本科阶段遇到的其他两位老师。陈积明老师在我大学的前两年里，给我提供了很多课程安排和未来发展方向上的指导,许超老师鼓励我参加比赛,让我积累了实践经验。感谢老师们的指导，使我顺利地度过了大学的四年。

其次感谢王越老师和实验室的师兄师姐们。王越老师丰富的理论知识和实践经验给我提供了很多细节上的指导，也帮助我大致确定了未来研究的方向。焦艳梅师姐带我理清了语义地图的发展过程，让我在本次设计中少走了很多弯路。罗倩慧师姐帮助我在服务器上部署我自己的项目，使得我的语义分割网络能够成功地在 GPU 上运行。周忠祥师兄在物体点云特征方面和我讨论出了合理的方案。丁夏清师姐和尹欢师兄在 SLAM 方法的改进上为我提供了他们的研究经验。傅博师兄给我提供了他的答辩经验，帮助我在开题答辩时取得了好成绩。潘一源师兄对我的论文提出了很多宝贵的建议。有你们的帮助才能有我的这篇论文。

同时，感谢一同奋战毕业设计的同学们，崔瑜翔、李威杰、刘立陆、殷隆基等同学在我做实验时提供了很多帮助，刘立陆同学在点云处理的方法上与我讨论并给出了合理的方法。李威杰同学在深度学习网络的搭建和调参方面给我了很多参考。崔瑜翔同学在图像处理方面给我提供了很多建议。难忘和你们一起讨论分析问题，一起约饭刷夜的日子。

最后要感谢家人对我的支持和帮助，是你们成就了现在的我。

摘　要

地图是移动机器人运动的依据，是其定位导航的基础。语义地图则在此基础上为智能移动机器人提供周围环境中物体的类别信息，使得机器人能够更直观地理解场景。研究语义地图的构建方法，在机器人定位、导航以及人机交互等方面都有重要的价值。本设计研究如何提升语义分割精度、语义标注融合优化以及语义地图表示形式问题，取得了以下成果：

1. 针对一般采用深度学习方法对单帧图像进行语义分割精度有待提高问题，本文采用循环神经网络将多视角图像用于语义分割，发现并解决了其中数据关联模块无法正常工作的问题，提高了语义标注的准确性，从而可以构建更为精准的语义地图。
2. 针对因错误语义标注而导致在语义点云地图中引入错误标注点的问题，本文提出了一种基于语义概率的优化方法，通过引入语义分割网络输出的概率分布以及融合后点云的空间关系，优化了部分语义标注，提高了语义点云地图的准确性。
3. 针对现有语义地图采用点云形式，对定位、导航和人机交互几乎没有直接应用价值，本文设计了物体层面的语义地图构建优化方法，通过去除噪声、表面重建来优化语义物体提取质量，通过相似性判断和频次存储策略降低存储空间要求，从而提高语义地图的可利用性。

关键词：语义地图，深度学习，物体提取

Abstract

The map is the basis of mobile robot movement and its localization and navigation. On this basis, the semantic map provides the category information of objects in the surrounding environment for the intelligent mobile robot, so that the robot can intuitively understand the scene. It is of great value to study the method of semantic mapping in robot localization, navigation, and human-robot interaction. This project studied how to improve the semantic segmentation accuracy, semantic annotation fusion, and semantic map representation, and achieved the following results:

1. In view of the problem that the deep learning method for single frame image semantic segmentation accuracy needs to be improved, this paper adopts the recurrent neural network that uses multiple perspective image in the training process. This study found and solved the problem that the data association module in the neural network cannot work normally, and finally improved the accuracy of the semantic annotation, which is crucial in building a more accurate semantic map.

2. Aiming at the problem of importing wrong annotation points in the semantic map caused by wrong image annotation, this study proposes an optimization method based on semantic probability. By introducing the probability distribution from the semantic segmentation network and the spatial relationship of point cloud after fusion, most semantic annotations are optimized, thus improving the accuracy of the map.

3. The existing semantic map which in the form of point cloud has almost no direct application value for localization, navigation, and human-robot interaction. This study designed a new method for the optimization of the semantic map. By removing noise and reconstructing surface, the quality of the semantic object extraction is improved. By judging the similarity of the objects and adopting frequency storage strategy, the storage requirements can be loosed, so as to improve the availability of semantic map.

Keywords: Semantic map, Deep learning, Object extraction,

目　　录

第一部分　毕业设计

第二部分　文献综述和开题报告

文献综述和开题报告封面

指导教师对文献综述和开题报告具体内容要求

目录

《浙江大学本科生文献综述和开题报告考核表》

第一部分

毕业设计

1 引言

1.1 研究背景及意义

地图是机器人自主移动必不可少的知识，它不仅可以为移动机器人提供自身位置的信息，还能够描述周围环境的各种属性。早期，Thrun 在机器人建图问题的综述[1]中写道，机器人建图领域大致可以分为两个部分：一是采用度量的方法建图，二是采用拓扑的方法建图。前者构建的度量地图包含了环境中的几何属性，后者则表现了连通性的特征。

上述两种方法构建的地图都只保存了环境的空间信息而没有体现环境的属性信息，这不仅不符合人类的认知范式，也不利于移动机器人和人之间的交互。智能的移动机器人需要更加自主地执行各种任务以及更好地与人进行交互，因此还需要其他更高层面的信息，如场景中物体、结构等语义信息。近年来，随着智能决策、语义导航等移动机器人发展需求的增长，语义地图成为新的发展趋势和研究热点。

语义地图是在度量地图的基础上带有物体类别信息的一类地图。近年来，语义地图在移动机器人领域取得了广泛的关注。语义地图的可理解性使其成为未来具有高智能机器人必不可少的工具。然而目前对于语义地图的各种研究才刚刚起步，研究过程中也反映出了很多问题。

现有语义地图构建方法通常先对机器人移动过程中获取的二维图像进行语义分割，再通过估计每帧图像的相机位姿，将各帧语义分割的结果融合，得到一个三维带语义标注的点云地图。这类方法主要依赖于计算机视觉领域的语义分割技术，精确的语义分割是构建准确语义地图的关键因素。

早先语义地图的精确度不高，主要是因为很多对图像进行语义分割或是直接对点云地图进行分割的算法准确率不高。近期，深度学习的相关算法在图像分割任务中表现出了优越的性能，它能够较为准确的从大量原始数据中自主学习特征，且能够获得比传统方法更好的效果，因此被广泛应用于语义建图的过程中。但是，如何有针对性地在语义建图问题上使用深度学习语义分割方法，是需要研究的关键。最简单的语义建图框架是，进行单帧语义分割和多视角语义信息融合的两阶段操作，

这样的方法没有将机器人移动过程中的其他信息引入语义分割网络，实际上只是在做三维重建。在语义分割中利用多视角图像提高语义标注准确性是一个值得研究的问题。

在机器人系统中，地图构建只是机器人的一个环节，移动机器人还需要考虑后续的定位和导航应用，故语义地图的表示形式应该满足定位和导航的需求。目前大多方法构建的语义地图都是以带标注的三维点云地图的形式存在，这样的语义地图与度量地图的差别不大，较难被用于之后语义层面的定位和导航中。

本论文针对上述问题研究语义地图构建准确性和可利用性问题，对推动语义地图领域的发展、提高 SLAM 和定位导航的准确性、提升移动机器人的智能性和人机交互性能具有重要意义。

1.2 国内外研究现状

在语义地图发展的过程中，逐渐涌现出两种主流的方法，一是建立物体模型数据库，将传感器数据经过简单处理后与事先建好的数据库进行匹配，再转换到地图上进行语义标注；二是借助图像语义分割或者点云分割的方法对传感器数据进行处理，并通过相机位姿估计或同时定位及地图构建（Simultaneous Localization And Mapping, SLAM）的框架对其进行优化。

1.2.1 基于数据库模型的语义建图

受限于早期的算法以及计算能力，早期语义建图的研究工作大多在高度可控的情况下进行，采用事先建立好模型数据库的方式。

早期语义分割的准确率不高，通过识别模型特征并进行数据关联的方法比较主流，一种是通过物体轮廓来识别物体。如 Nüchter[2]提出了通过训练物体轮廓分类器的方法对场景中物体进行语义标注。该方法利用事先建好的物体点云数据库来训练得到轮廓分类器识别物体，识别后将模型与数据库中对应的物体进行最近点迭代（Iterative Closest Point, ICP）评估。另一种是通过物体的其他特征，如 Salas-Moreno[3]根据先验知识（室内场景中有很多重复、特定的物体和结构）构建了重复物体的数据库，来进行物体层面上的语义建图。通过点对特征（Point Pair Feature,

PPF）将由 RGB-D 相机构建的场景点云与数据库中的物体模型进行匹配，一旦发现匹配，就将预先存储好的物体模型放入地图，来实现地图的语义标注。如图 1.1，最上方为建好的物体地图，由数据库中的物体模型构成，中间为相机采集的 RGB 图像以及深度图像，最下方对应的由数据库模型组成的 RGB 图像以及深度图像。文章最后还利用物体信息进行了重定位，实现了 SLAM 与语义的结合。

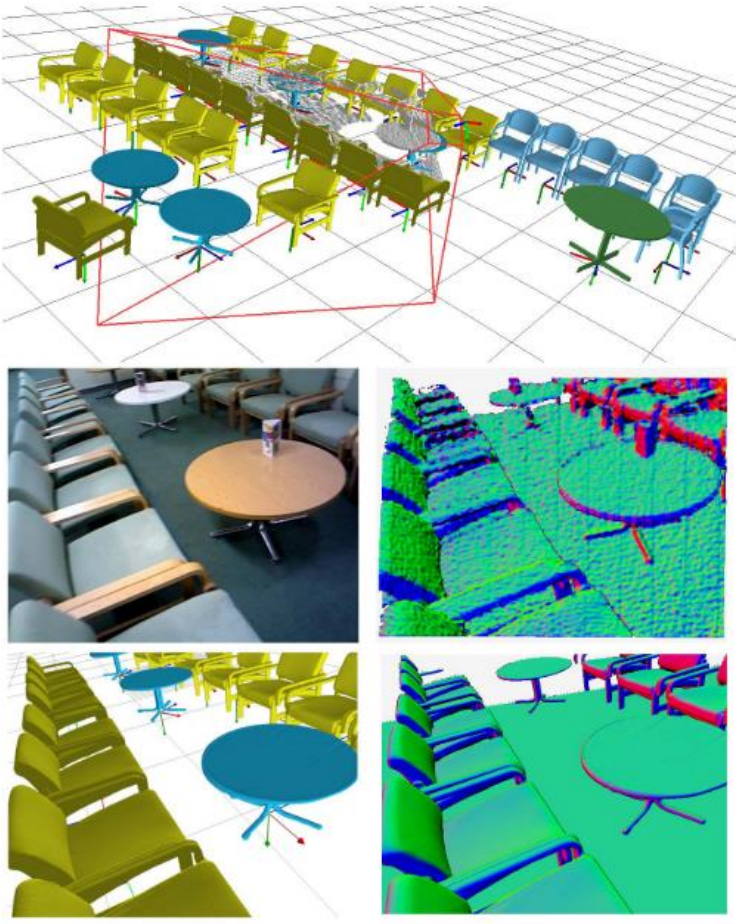

图 1.1[3] SLAM++构建的语义物体地图

上述方法都仅仅是将模型数据库作为一种语义标注的手段，而并没有对建图过程有任何作用。针对这个问题，Fioraio[4]提出了物体和相机位姿的联合预测。区别于 Salas-Moreno 直接在点云层面进行物体匹配，该方法是在视觉层面进行物体匹配，其过程如图 1.2，它将视觉 SLAM 中摄像头获取的图像与事先建好的数据库中模型的 2D 或 3D 特征进行匹配，并将匹配误差（图中菱形部分）加入图模型中与相机姿态节点（图中椭圆形部分）一起优化，实现了语义层面的光束平差方法 (Bundle Adjustment, BA)。

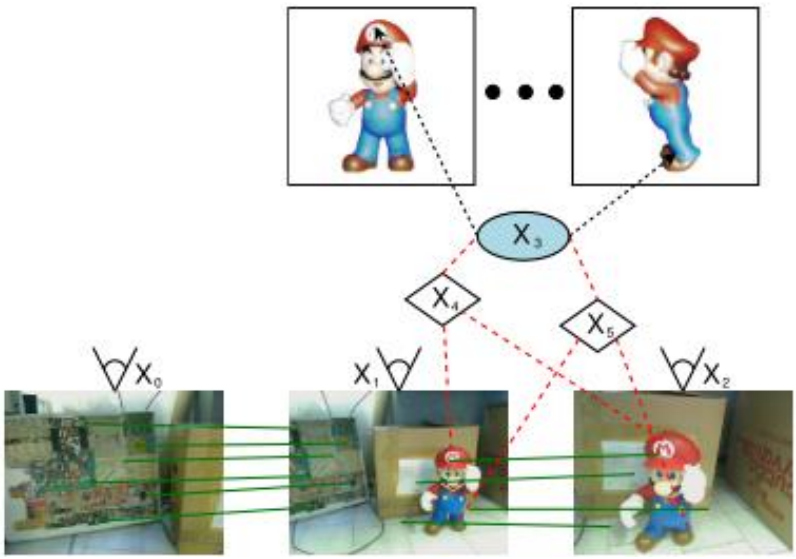

图 1.2[4] 将图片与 3D 模型进行特征匹配并将误差项加入图模型

采用数据库模型的语义建图，语义信息在地图上的表现形式大多为经过渲染的物体模型，并且只能结合数据库中的物体进行语义标注，其认知受数据库的影响和约束，无法识别标注数据库模型之外的物体。

1.2.2 基于分割的语义建图

基于分割的语义建图方法主要借助图像分割算法和点云分割算法，将单帧数据进行语义分割，然后借助 SLAM 的位姿估计，将分割的结果映射到三维空间，实现语义建图，这类方法中数据关联的可靠性是比较关键的问题。

最早的研究主要基于当时的图像分割方法，如基于随机森林分类器的分割：Stuckler[5]提出用随机决策森林算法对二维图像进行像素级的分类。通过在训练过程中引入深度信息，可以得到一个无缝融合形状和纹理特征的尺度不变分类器。经过分类器完成像素级语义标注的图像，根据贝叶斯准则生成了场景的语义 3D 体素模型。Hermans[6]借助类似的图像分割方法，用一种渐进的方式进行 2D 分割至 3D 的语义建图。文章中用随机决策树进行 2D 图像的粗分割，再由全连接条件随机场（Dense Conditional Random Field, DenseCRF）算法进行精细分割，得到的结果结合视觉里程计得到的位姿进行三维重建。

随着目标检测方法的成熟，也有一部分研究从目标检测的方法出发，结合点云的空间信息进行语义地图构建。如 Sunderhauf[7]将无监督的 3D 点云分割方法与单

发多框（Single-shot Multi-box Detector, SSD）目标检测算法结合提出了构建物体语义地图的算法，不仅做到了像素级的语义分割，还将物体的点云储存作为地图的元素。文中观测物体与物体数据库的数据关联是通过类似 ICP 的方法，计算两个物体点云的欧式距离，判定是否为同一个物体。但是 SSD 是基于室外场景训练的，对于室内场景，获取的物体语义信息就会比较少，如图 1.3 所示，最后构建的语义地图仅仅只有很少的类别，漏掉了很多语义信息。

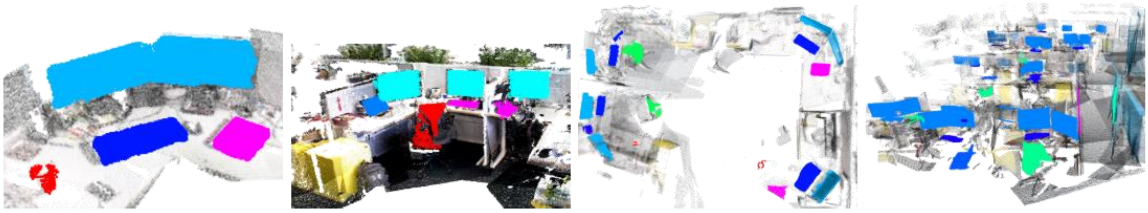

图 1.3[7]　针对具有平面特征的物体语义分割

当前一种将多视角信息融合到语义分割网络的方法被验证能提高语义标注的准确性，如 Xiang[8]提出直接将 RGB-D 图像作为输入，利用带有数据关联的循环神经网络（Recurrent Neural Network, RNN）进行像素级的图像语义标注。通过类似于门控循环单元（Gated Recurrent Unit, GRU）的结构，将 SLAM 框架中得到的位姿加入到神经网络的训练中，即将帧间信息也考虑在内，提高了语义标注的精度，并通过 KinectFusion[9]方法将语义标注融合。

基于图像分割的语义建图方法，大多涉及机器学习，其对计算资源的需求较高。这类方法构建的语义地图大多为带语义标注的三维点云地图，在后续定位阶段，如何利用带语义标注的三维点云地图是需要考虑的一个问题。

1.2.3 国内外研究发展方向概述

从近几年国内外的语义地图研究发展来看，基于深度学习的语义地图构建方法变得越来越普遍也越来越有针对性，除了构建地图，一些研究已经开始讨论语义地图的应用问题。

在语义地图构建方面，有必要研究面向这一特定问题的语义标注，以提升标注的准确性和效率。研究者如 Sunderhauf[7]针对建图问题提出结合点云分割以及目标识别的语义分割算法，而不是单纯地套用计算机视觉领域中的算法。机器人的相关问题中会涉及很多不同类型的感知信息，比如点云信息，而不只是计算机视觉中的

图像信息，涉及的处理方法也相应地会更加丰富，比如点云处理，这就为语义地图构建注入了新的可能性。Xiang[8]提出的数据关联循环单元就是利用了点云处理中的最近点迭代方法来处理循环计算中的关联问题。

如何应用语义地图也是当前的研究问题。构建语义地图是为了更好地服务机器人的后续应用，如定位、导航和人机交互等，因此语义地图的可利用性就显得尤为重要。Schönberger[10]通过编码-解码器来学习局部语义地图与全局语义地图的关系，从而构造出一种高效的 3D 描述子，用于语义定位。Wang[11]假设已有语义地图，提出了一种语义信息与相机姿态共同优化的深度学习优化方法，实现了定位与语义地图的共同优化。

1.3 本文主要工作

本次毕业设计的研究目标为利用深度学习方法实现语义地图构建，包括实现基于深度学习的 RGB-D 图像语义分割、利用 SLAM 技术融合语义标注建图、提取语义物体等。

本文采用三阶段方法来构建语义地图，第一阶段将机器人移动过程中获取的 RGB-D 图像通过改进的循环神经网络进行语义分割；第二阶段通过 ORB-SLAM2 的方法连续估计每帧图像对应的相机位姿，将各帧语义分割结果融合，得到一个以起点为坐标系原点的三维语义点云地图；第三阶段通过提取语义物体点云并进行优化，实现了物体层面的语义地图构建。具体流程见图 1.4。

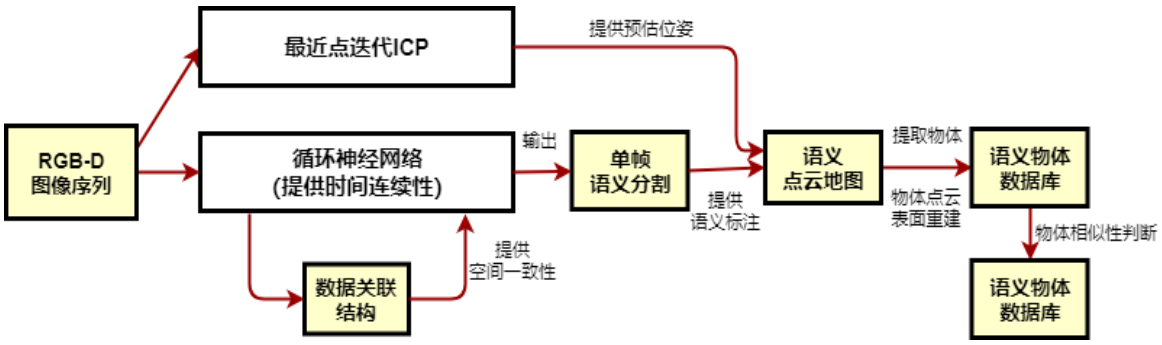

图 1.4 本文语义地图构建的流程图

主要完成工作如下：

（1）实现了结合多视角信息进行图像语义分割的方法。采用 DA-RNN[8]提出的网络结构，修改了其中的数据关联方式，搭建了可以融合多视角信息进行语义分

割的网络，并在公开数据集上进行训练和测试，验证了网络的可行性及语义分割的准确性。

（2）引入语义标注的概率分布实现了语义点云地图的优化。针对融合后语义点云地图出现的一类典型错误，结合图像语义标注的概率分布和点云的空间信息对该类典型的错误标注点进行优化。

（3）实现了语义物体的提取及优化，并提出物体模型存储方案。在上述过程中构建的三维语义点云地图中提取语义物体，形成单个模型，并进行曲面重建等优化工作，最后存储在一个可以在线运行的数据库中。

2 RGB-D 图像的语义分割

2.1 引言

语义分割是计算机视觉中的基础任务，其目标是将视觉输入分为不同的语义类别，语义的类别指分类的类别在真实世界中是有意义的。例如图 2.1 所示，语义分割区分图像中属于汽车的所有像素，并把这些像素涂成蓝色。语义分割让机器视觉对图像的理解比图像分类和目标检测为更详细。这种细节信息在无人驾驶、医疗影像分析以及机器人领域有着非常重要的意义。

图 2.1 语义分割实现的效果

相对于现有一般二阶段方法将单帧语义分割与融合割裂的现象，本文语义分割部分采用 DA-RNN 中的网络及数据关联结构，利用多视角的语义信息训练语义分割网络，以提升语义分割的准确率。但是由于复现 DA-RNN 网络过程中出现数据关联模块无法运行的问题，本文实现由 ORB-SLAM2 框架提供数据关联的方式。

本章首先介绍语义分割网络的基础结构，然后介绍 Xiang[8]提出的数据关联循环单元和网络结构，最后介绍本文对该结构的修改与实验。

2.2 语义分割网络基础

语义分割发展的过程中出现过几种比较经典的算法：基于图分割的 Graph cut[12]、Grab cut[13]，基于条件随机场的图像后处理方法[14]和基于深度学习的全卷积

神经网络[15]等。本文着重于采用基于深度学习的语义分割方法，故在此介绍其主要的三种技术：卷积化、上采样和跳跃结构。

A．卷积化

卷积化操作是将普通分类网络中的全连接层都转换为卷积层，形成全卷积网络FCN[15]，其过程如图 2.2。上半部分的分类网络中三个全连接层转换为卷积层。

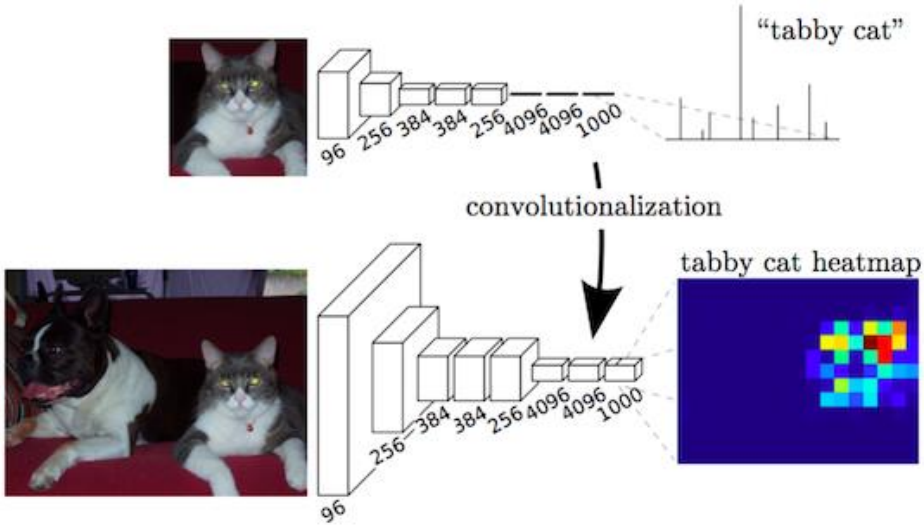

图 2.2 全卷积网络结构

B．反卷积/上采样

普通的池化层会缩小图片的尺寸，在多次卷积池化操作之后，图片会被缩小很多倍，如 VGG16 在经过五次池化之后，图片的尺寸被缩小了 32 倍，故为了在全卷积网络的输出层得到与原图等大的分割图片，需要反卷积操作。反卷积操作和卷积操作类似，相当于卷积的反向操作。其过程如图 2.3 所示，蓝色部分为被反卷积的原图，绿色部分为经过反卷积后得到的图片。

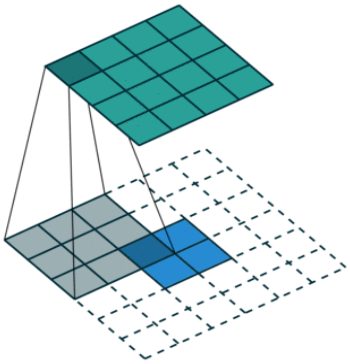

图 2.3 反卷积操作示意图

C．跳跃结构

跳跃结构的作用在于优化分割的粒度，直接通过全卷积网络得到的语义分割是比较粗糙的，这是因为多次的卷积池化操作将图像中信息抽象到无法恢复细节的程度。故需要将不同池化层的结果经过反卷积后来优化输出，具体的结构和效果如图2.4所示，可见多层的跳跃结构可以更好地捕捉语义物体边缘的信息，从而达到更细致的分割。

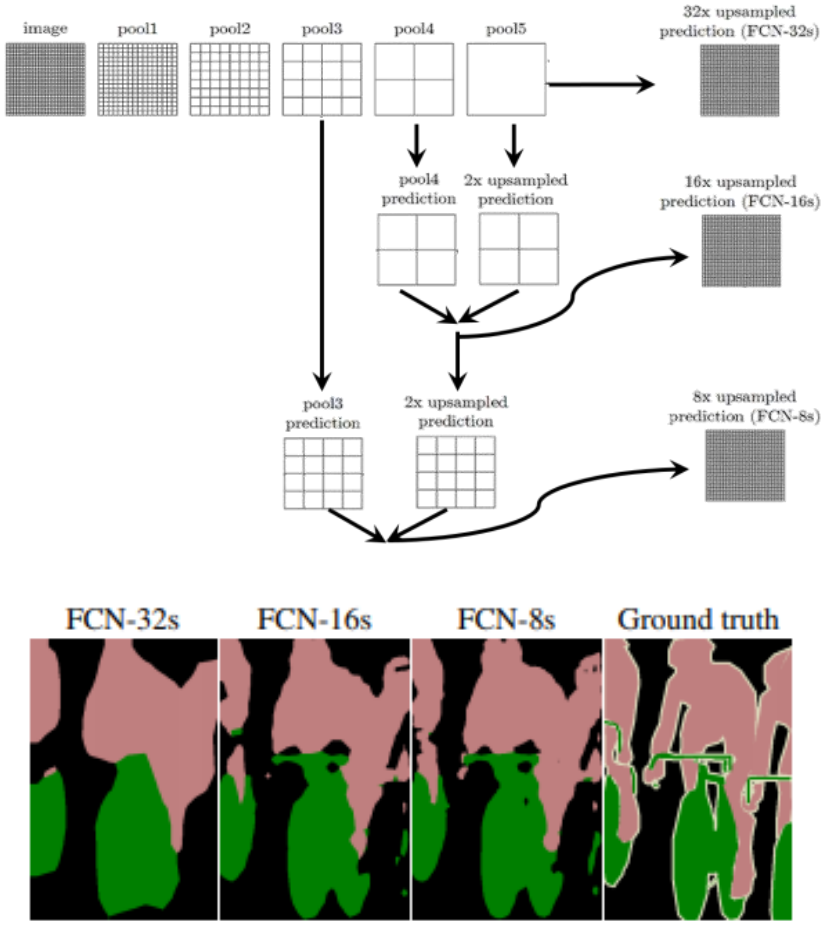

图 2.4[15] 跳跃结构示意图以及其效果

2.3 循环神经网络（RNN）

循环神经网络是一种可以用来分析时间序列数据的网络，它可将任意序列长度的数据作为输入，通过隐藏状态的储存来传递序列中不同时间步的信息，从而达到学习时间序列数据中的相关性，也就是说，循环神经网络不仅将当前的输入作为网络输入，同时还将之前感知到的一并作为输入。由于循环神经网络对于序列数据处

理的有效性，其被广泛用于情感分析、图像文字描述生成、自然语言处理、视频标记等方面。

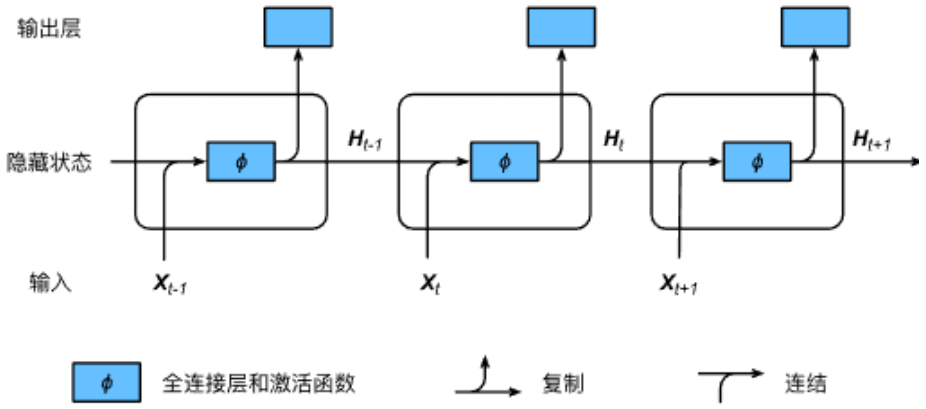

图 2.5 循环神经网络的一般结构

循环神经网络的主要结构如图 2.5 所示。考虑输入数据具有时间相关性，假设$X_t \in \mathbb{R}^{n\times d}$是序列中时间步 t 的输入，$H_t \in \mathbb{R}^{n\times \mathrm{h}}$是时间步 t 的隐藏变量，$W_{\mathrm{hh}} \in \mathbb{R}^{\mathrm{h}\times \mathrm{h}}$是用于衡量上一时间步的隐藏变量如何转移到当前时间步的权重参数，$W_{\mathrm{xh}} \in \mathbb{R}^{\mathrm{d}\times \mathrm{h}}$为隐藏层的权重参数，$b_{\mathrm{h}} \in \mathbb{R}^{1\times \mathrm{h}}$为偏差，激活函数为$\phi$。定义了上述变量后，可以计算当前时间步的隐藏变量：

$$H_t = \phi(X_t W_{\mathrm{xh}} + H_{t-1} W_{\mathrm{hh}} + b_{\mathrm{h}}) \tag{2.1}$$

输出层的计算方式与一般前馈网络相似：

$$O_t = H_t W_{\mathrm{hq}} + b_q \tag{2.2}$$

$O_t \in \mathbb{R}^{d\times \mathrm{h}}$为网络的输出，$W_{\mathrm{hq}} \in \mathbb{R}^{\mathrm{h}\times q}$为输出层的权重，$b_q \in \mathbb{R}^{1\times q}$为偏差。这些模型参数在每个时间步中保持一致，因此模型的参数数量不会随时间而增长。

与一般的前馈神经网络不同，循环神经网络中有一项$H_{t-1}W_{\mathrm{hh}}$，这一项中包含了上一个时间步的信息，也就将时间相关性引入网络中。由于每个时间步的隐藏状态计算时都使用了上一时间步的隐藏状态，因此整个网络的计算是循环的，循环神经网络也因此得名。

2.4 门控循环单元与数据关联循环单元

节 2.3 介绍了计算隐藏状态的方法，但由于引入了循环计算，其在时间步较大或较小时，容易出现梯度衰减或爆炸的情况，这是因为一些过小或过大的梯度信息在很长的时间间隔中无法正常传递。

为了避免梯度衰减和梯度爆炸问题，常用的方法是引入门控循环单元，它通过可以学习的门来控制信息的流动。

A．门控循环单元

门控循环单元包含三个部分：重置门、更新门和候选隐藏状态。其结构如下图，⊙运算为矩阵按元素乘法；σ为 sigmoid 函数运算，它将元素的值变换到[0,1]。

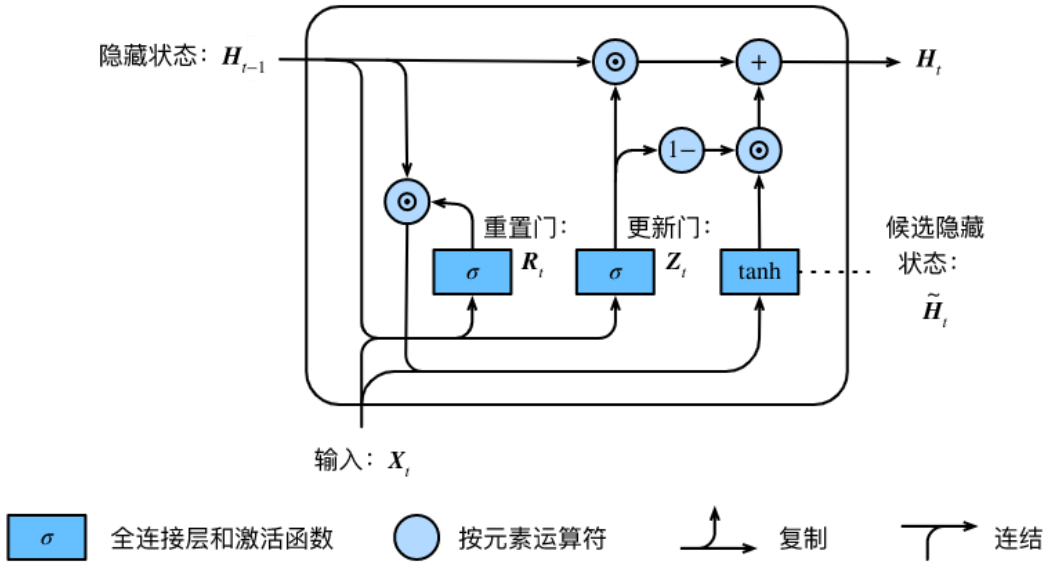

图 2.6 门控循环单元结构示意图

重置门、更新门和候选隐藏状态的计算方法如下：

$$R_t = \sigma(X_t W_{\mathrm{xr}} + H_{t-1} W_{\mathrm{hr}} + b_{\mathrm{r}}) \tag{2.3}$$

$$Z_t = \sigma(X_t W_{\mathrm{xz}} + H_{t-1} W_{\mathrm{hz}} + b_{\mathrm{z}}) \tag{2.4}$$

$$\tilde{H}_t = tanh(X_t W_{\mathrm{xh}} + (R_t \odot H_{t-1}) W_{\mathrm{hh}} + b_{\mathrm{h}}) \tag{2.5}$$

$W_{\mathrm{xr}}, W_{\mathrm{xz}}$和$W_{\mathrm{hr}}, W_{\mathrm{hz}}$为权重参数，$b_{\mathrm{r}}, b_{\mathrm{z}}$为偏差，$\sigma$为 sigmoid 函数。由上述式子可见，重置门控制了上一时间步的隐藏状态是如何流入当前时间步的候选隐藏状态。更新门控制了隐藏状态是如何权衡上一步隐藏状态和当前候选隐藏状态的。简言之，重置门有助于学习时间序列中短期的相关性，更新门有助于学习时间序列中长期的相关性。

B．数据关联循环单元[8]

针对输入为图像数据的网络，门控循环单元需要适当修改。由 Xiang[8]提出的数据关联循环单元是一种将图像序列中的数据关联性作为隐藏状态传递标准的网络层，该单元的结构如图 2.7，在循环神经网络基本结构上加入了数据关联模块。一个单元对应一个像素，*t* 时刻像素通过某种数据映射方式如果在 *t+1* 时刻找到与其对应的像素，则将 *t* 时刻的隐藏状态和权重向量传递到 *t+1* 时刻的该单元，没有找到则置零。

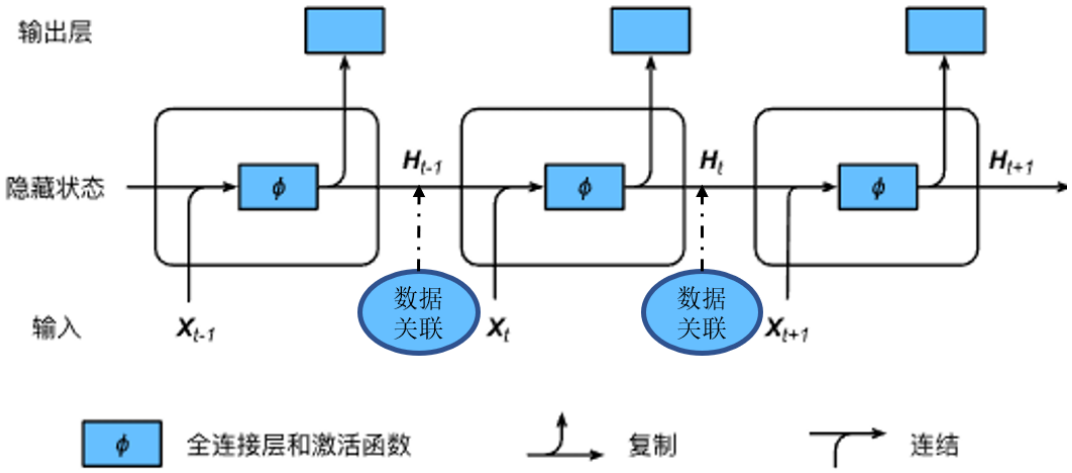

图 2.7 门控循环单元结构示意图

数据关联方法：

$$\langle \tilde{h}_{t+1}^{i}, \tilde{w}_{t+1}^{i} \rangle = \begin{cases} \langle 0,0 \rangle, & \text{若没有关联} \\ \langle h_t^{i'}, w_t^{i'} \rangle, & \text{若} p_{t+1}^{i} \text{与} p_t^{i'} \text{有关联} \end{cases} \tag{2.6}$$

其中p_{t+1}^{i}和$p_t^{i'}$分别表示时刻 *t+1* 的单元 *i* 的像素和时刻 *t* 中单位 *i'*对应的像素。$h_t^{i'}, w_t^{i'}$表示时间步 *t* 内网络的隐藏状态和权重参数。

输入的权重参数计算：

$$\hat{w}_{t+1}^{i} = \sigma\left(W\left[\tilde{h}_{t+1}^{i}, x_{t+1}^{i}\right] + b\right) \tag{2.7}$$

其中$\hat{w}_{t+1}^{i}$是输入x_{t+1}^{i}的权重向量，它是前一帧的隐藏状态和当前帧的输入的函数。*W, b* 是循环层的参数，由层中的所有单元共享，σ为 sigmoid 函数，[• ， •]表示两个向量的串联。$W \in \mathbb{R}^{d \times 2d}$为权重参数，$b \in \mathbb{R}^{1 \times d}$为偏差，其中 *d* 是隐藏状态的维度。

更新权重向量：

$$w_{t+1}^{i} = \hat{w}_{t+1}^{i} + \tilde{w}_{t+1}^{i} \tag{2.8}$$

更新隐藏状态：

$$\mathrm{h}_{t+1}^{i} = f\left(\left(\widetilde{\mathrm{w}}_{\mathrm{t+1}}^{\mathrm{i}} \oslash \mathrm{w}_{t+1}^{i}\right) \otimes \tilde{\mathrm{h}}_{t+1}^{i} + \left(\widehat{\mathrm{w}}_{t+1}^{i} \oslash \mathrm{w}_{t+1}^{i}\right) \otimes x_{t+1}^{i}\right) \quad (2.9)$$

其中$f(x) = max(0, max)$是线性整流单元（ReLU）激活函数，$\oslash$,$\otimes$分别表示逐元素除法和逐元素乘法。该式定义了隐藏状态的计算。

当前时刻神经网络的权重向量是前一帧累加权重向量与当前输入的权重向量的总和，在此引入加法运算，可以消除梯度消失的可能，使得数据关联能够在长时间内保存并计算。这样的结构能够将不同时刻不同视角的特征组合起来，并且其可以在深度学习过程中不断优化，以达到最佳的数据关联效果。

2.5 利用多视角图像数据的语义分割网络结构

在语义分割过程中若能够利用多视角图像数据，语义分割网络就可以从中学习到多视角图像中的时空一致性，其准确性也可以得到提升。在自然语言处理等领域，循环神经网络可以有效地处理具有时间连续性的序列。类似的，机器人移动过程中产生的 RGB-D 数据可以类似地理解为一个时间图像序列，这个序列同样具有时间连续性。时间循环神经网络中常用的门控循环单元（GRU）可以将序列中的时间连续性引入网络的训练中。

2.5.1 语义分割的网络结构

基于上述思想，Xiang[8]提出了 DA-RNN，其网络结构如图 2.8 所示。它通过循环神经网络以及数据关联循环单元将多视角图像引入网络的训练中。针对输入的 RGB-D 数据，语义分割网络的构建则相应地分为 RGB 和 Depth 两部分，RGB 图像通过 VGG16 网络提取特征然后进行反卷积操作，深度图像通过类似的结构提取深度特征并与 RGB 特征提取得到的结果加和，相当于将 RGB-D 信息嵌入到一个特征中。通过这个联合的特征即可恢复得到语义分割的粗结果，粗结果通过数据关联循环单元，将多视角图像信息引入网络训练，可以得到优化的结果。

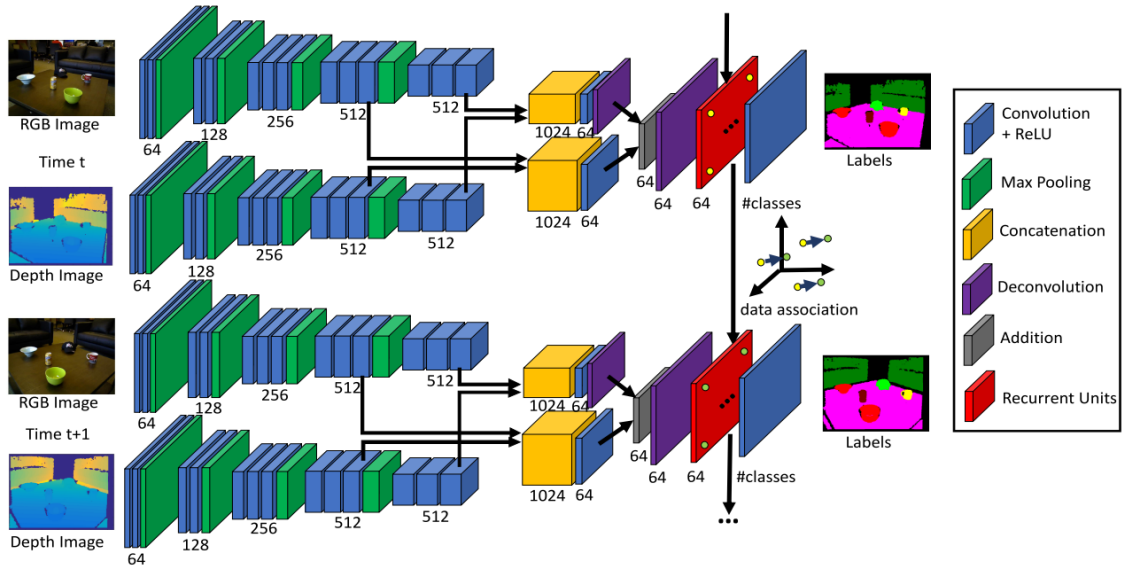

图 2.8[8] DA-RNN 网络结构

数据关联循环单元在节 2.4 中进行了详细的介绍，DA-RNN 网络的训练过程中，数据关联的判断是由 KinectFusion[9]估计的相机位姿决定的，但是在 GPU 上实现的 KinectFusion 无法运行。本文经实验分析发现，KinectFusion 和语义分割网络一起运行会导致 GPU 显存不足，从而出现网络无法训练的结果。

针对 KinectFusion 无法与循环神经网络一同运行提供数据关联的问题，本文采用以下较为轻量的解决方案：在 ORB-SLAM2 的框架下估计两帧 RGB-D 图像的相机位姿，通过估计的位姿将 RGB-D 映射到三维空间的 3D 点云投影到下一帧的二维平面，来判断不同时间步内像素的关联性。

2.5.2 ORB-SLAM2 框架

ORB-SLAM2 是基于 ORB 特征的三维定位与地图构建算法，它常用于度量地图的构建，其框架如图 2.9 所示。ORB-SLAM2 由三个并行的线程组成，它们分别为跟踪、局部建图和回环检测线程，在一次回环检测后，执行另一个 BA (Bundle Adjustment)优化的线程。

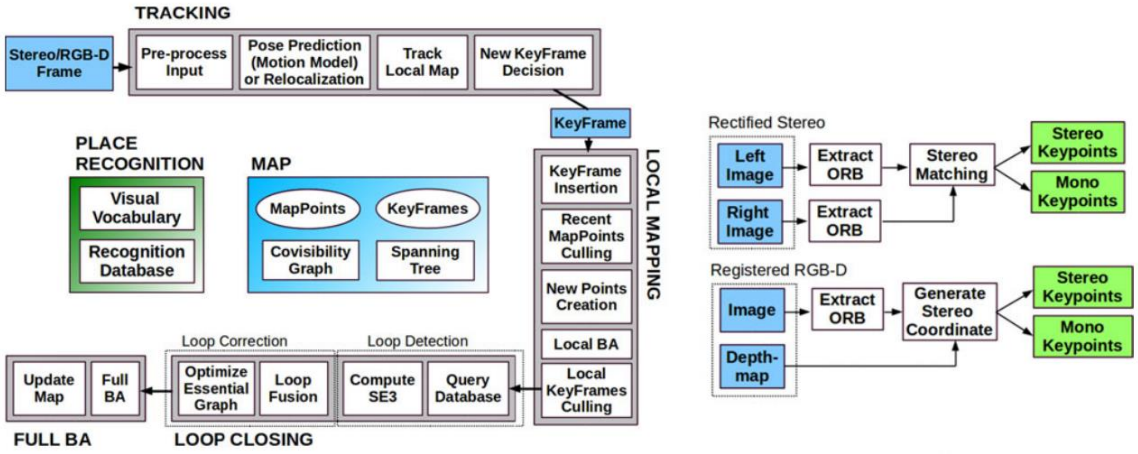

图 2.9[16]　ORB-SLAM2 框架

跟踪线程的工作为寻找每一帧图像与局部地图匹配的特征，以及使用带有运动约束的 BA 算法来最小化重投影误差，跟踪和定位每帧的相机位姿；局部建图线程的工作是建立局部地图并通过执行局部 BA 来进行优化；回环检测线程的工作是检测机器人运动过程中的回环，若检测到回环则用图优化的方法来校正累积误差。

该系统轻量，可以实时地在 CPU 上运行，并不影响在 GPU 上进行训练的语义分割网络，充分利用了计算资源来进行网络的训练。同时，该框架也可以为语义标注融合提供位姿估计。

2.6 语义分割实验

A．实现细节

根据上述描述在主流深度学习框架 Tensorflow[17]中搭建网络，RGB 图像的特征提取采用在 ImageNet[18]上预训练的 VGG16 网络[19]，选取带动量的随机梯度下降方法进行学习，损失函数定义为像素 softmax 交叉熵。考虑到显存及计算资源的限制，设置每个 SGD 小批量为 3 个连续帧的图像序列。

实验平台选择基于 Ubuntu 操作系统的 GPU 服务器进行训练，使用 NVIDIA TITAN X（Pascal 架构）显卡及 CUDA8.0 Toolkit 进行网络训练和测试。TITAN X 提供 12G 显存，使得图像和网络参数能够更快更好的训练。

B．数据集

语义分割数据集采用的是由 Kevin[20]引入的 RGB-D 室内场景数据集，它由 Kinect 在室内场景中采集的 14 个 RGB-D 视频组成。每个场景都被重建为对齐过的

3D 点云。然后这些 3D 点云有 9 个对象类标签和背景标注。原始数据集并不包含图像的语义标注，本设计采用由 Xiang[8]进行像素级标注后的数据集。

C．训练过程

本文在 GPU 服务器上迭代 40000 次后，得到较低的损失，损失函数值的动态变化过程如图 2.10 所示，40000 次迭代后网络基本达到其最优的效果。

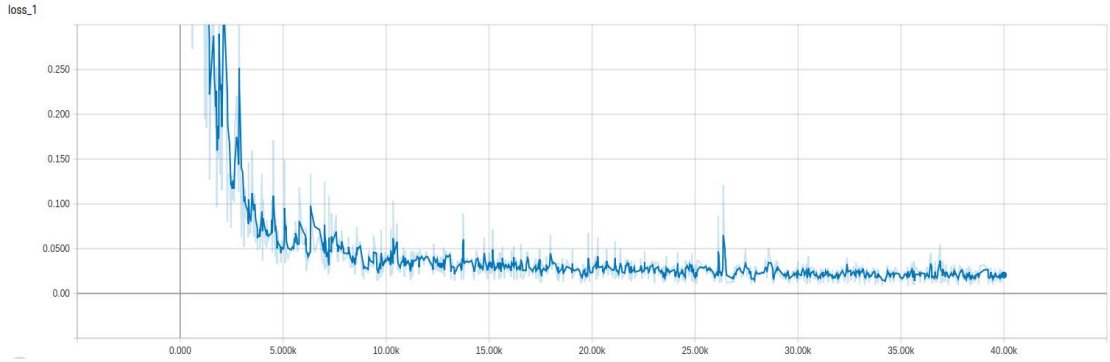

图 2.10 网络训练过程中损失函数的值曲线

D．语义标注结果

RGB-D 数据经过语义分割网络后生成像素级语义标注，部分标注结果如图 2.11 所示。

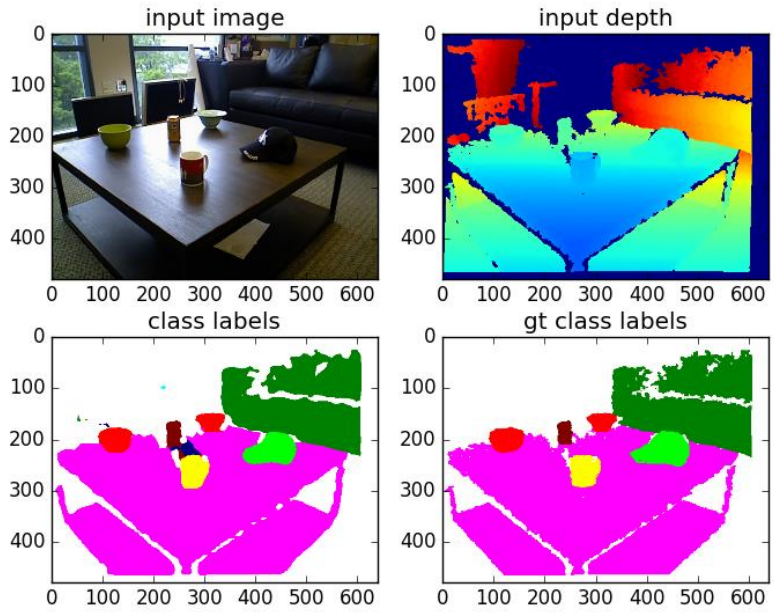

图 2.11 左上为输入的 RGB 图像，右上为输入的深度图像，左下为语义标注结果，右下为真值

可以看到，语义标注基本准确，但是由于物体边界处的标注较为依赖数据关联结构中相机位姿的估计，一旦相机位姿估计存在一定误差，语义标注时，物体边缘就会出现一些不稳定的标注，如图 2.12 所示，这些误标注将会在之后融合的过程中进行优化。

图 2.12 语义标注错误示例

E．评估指标

本设计采用广泛使用的交并比（IoU）作为评估标准，其直观的定义如图 2.13。针对像素级语义标注，此处使用的是像素交并比的评估标准，即，相同语义标注的像素点比该类语义标注像素点真值与标注的总和。该评估指标用于衡量语义标注与真值之间的相似性。IoU 的定义如下：

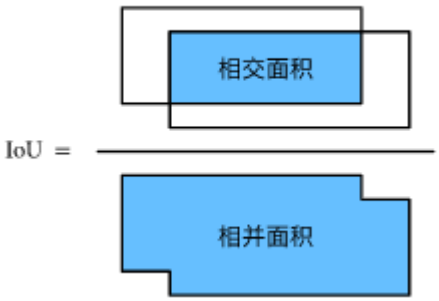

图 2.13 一般 IoU 的定义

F．对比实验及分析

为了验证 ORB-SLAM2 进行数据关联的可行性及有效性，本文比较了以单一视角数据训练、通过 ORB-SLAM2 进行数据关联的多视角数据训练、全卷积网络

三种网络。由于 DA-RNN 中的数据关联模块无法运行，故无法与其做对比实验。从表 2.1 中可以看出，基于 ORB-SLAM2 的数据关联循环单元的确能够通过引入多视角数据从而提高语义分割的准确率。但准确率的提升并不明显。这是因为 ORB-SLAM2 框架的后端优化只有在较为明显的回环处有显著的作用，在一般未闭合的路径中，其相机位姿的估计也存在一定的误差，然而数据集中的采集时相机的运动轨迹都没有回环，这直接导致了 ORB-SLAM2 框架的估计位姿没有优化，使得数据关联存在错误的情况。尽管如此，本文的网络相比 FCN 方法还是有所提升。

表 2.1　图像语义分割准确率统计

方法	FCN	本文网络 – 单帧数据	本文网络 – 多视角数据
背景	96.1	97.4	**97.6**
碗	**97.0**	91.1	89.8
帽子	79.0	77.8	**83.0**
麦片包装盒	87.5	89.3	**89.3**
咖啡杯	75.7	**85.1**	82.5
茶几	95.2	96.1	**96.2**
椅子	71.6	76.6	**81.3**
易拉罐	82.9	**86.9**	85.5
沙发	92.9	**96.4**	95.7
圆桌	89.8	**92.8**	92.3
平均像素 IoU	85.8	89.0	**89.3**

2.7 本章小结

本章介绍了语义分割和循环神经网络的基本概念与结构，分析并实验验证了循环神经网络可通过 ORB-SLAM2 框架进行数据关联的方式融合多视角图像信息，以帮助提高语义标注的准确率。经过足够多迭代的训练，最后得到了一个可以通过输入的 RGB-D 图像序列生成语义标注的网络。

3 采用概率优化的语义点云地图构建

3.1 引言

语义分割网络生成每帧对应的像素级语义标注，结合其深度信息就可以得到单帧的 3D 语义点云，有效地将这些单帧的 3D 语义点云融合起来，最后就可以得到语义点云地图。

多视角点云的融合实际上是一个估计相机位姿的问题，以初始点云相机坐标作为坐标原点，估计后续点云帧的相机位姿并将其转换到坐标原点，就可以逐渐地将带有语义标注的点云融合到同一个坐标系下，成为一张语义点云地图。

在估计相机位姿进行语义标注融合的基础上，本文还通过引入语义标注的概率分布，优化了语义地图中物体边缘的错误标注。

3.2 基于 ORB-SLAM2 框架的语义融合

估计相机位姿的方法有很多，针对利用多帧图像估计运动的有 SLAM 方法，比较经典的有 ORB-SLAM2 框架；对于利用多帧点云估计相机位姿的方法有最近点迭代（ICP）。前者计算快速但精度上略有不足，但其具有回环检测的结构，可以通过回环检测及优化的方式提高精度。回环优化使得其可以在长期运行过程中保持一定的精度。ICP 的方法在位姿变化小的情况下精度高，但计算量大，且由于其没有优化步骤，在对较长序列的融合时会有误差累积的问题。考虑上述两种框架的优缺点，本设计选取 ORB-SLAM2 的框架来估计相机位姿并进行融合。

3.3 语义点云地图的规模控制

由于每帧图像映射到三维空间均有 30 万的点，稠密地将各帧点云数据转换到初始坐标系下将会导致点云规模逐渐变大，最后导致内存泄漏。故在每帧点云融合之后做一次体素网格滤波，将点云数量控制在一定的规模下。

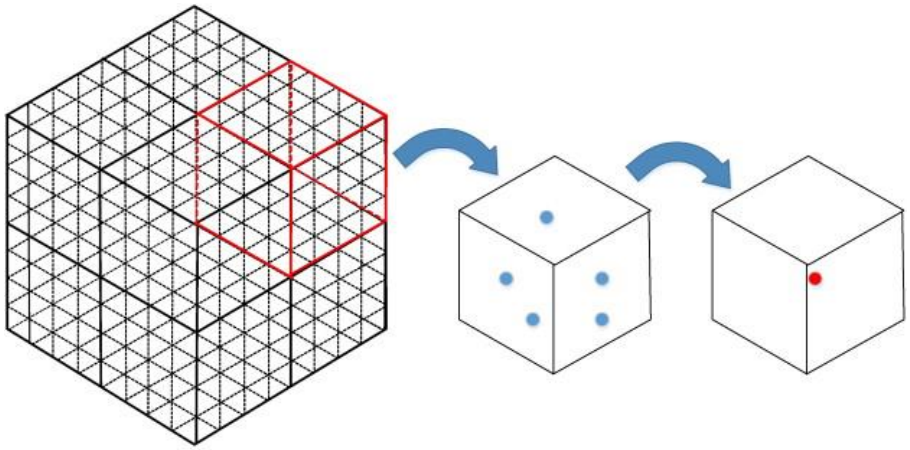

图 3.2 体素网格滤波原理示意图

体素网格滤波的原理如图 3.2 所示，在点云坐标系下创建一个三维体素栅格，相当于创建一个个小的三维立方体，然后在每个体素内，用体素中所有点的重心来近似表示体素中的其他点。这种滤波方法在室内场景下可以通过设置网格的大小来控制点云的规模，同时也可以加速融合的计算。

3.4 基于语义概率的标注优化

语义分割网络在生成语义标注的时候，实际上是计算了一个针对所有类别的概率分布，然后取概率最高的类作为最后结果。本文称生成的标注类别的概率为语义概率。

语义分割网络生成的语义标注存在或多或少的错误标注，这些误标注大致可以分为两类：

1、由于图像语义分割在物体边缘标注的不稳定性而产生的错误标注点，这类点通常在语义分割生成标注时语义概率较低，在点云地图上直观的表现如图 3.3 左；

2、图像语义分割在类似物体上的稳定错误标注，这类错误标注通常语义概率较高，在地图上表现见图 3.3 右，为了清晰地表示，这里选取了地图的稀疏形式。可见椅子的水平面被误识别成了茶几的类别，茶几和椅面都是平面，特征较为相像。

对于这两类错误点，本文主要致力于解决第一类，第二类错误依赖于对语义分割网络的改造，暂且不做处理。

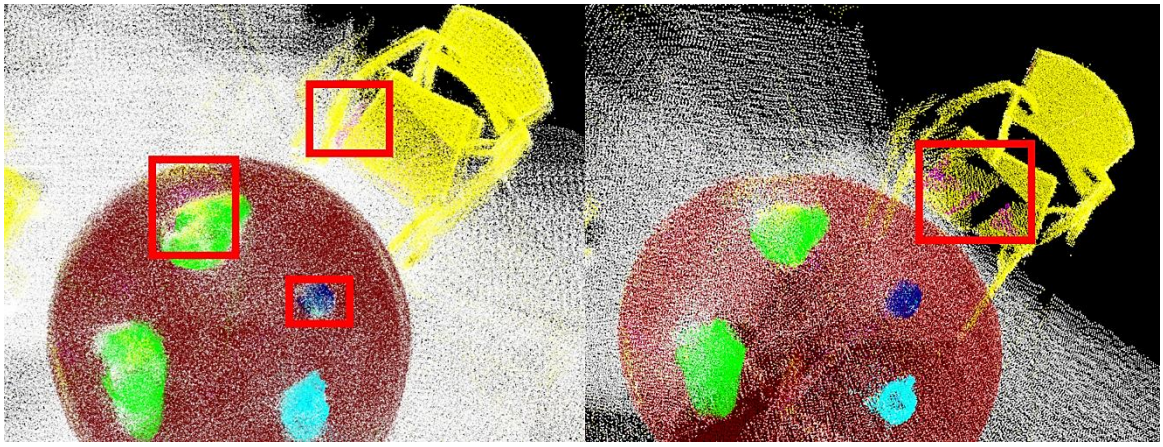

图 3.3 左图为稠密地图，红框标出了第一类错误标注，其多在物体表面；右图为地图的稀疏形式，红框中的紫色点为第二类错误标注。

针对物体边缘标注不稳定而造成的标注错误，其语义概率较低，本文考虑在语义点云地图中添加语义概率这一信息，然后结合点云的几何特征来进行处理。所提出的语义和几何信息联合优化语义点云的算法，不仅可以解决第一类误差，也可以拓展到所有语义概率较低的点上。

算法步骤如下：

- 首先抛弃所有语义概率在 50%以下的点，这类点通常处于多类物体交界处，无法界定类别。
- 然后找到语义概率在一个较低区间的点（本文为 50~80%，大多点的概率在 80%以上），通过 kd-tree 找到它的 k 个语义概率高于 80%的近邻点，统计它们的类别和数目。
- 最后找到语义概率高于 80%的点最多的类别，并将该低概率点标注成这个类别。

该算法可以用于语义标注融合之前，每帧进行一次优化，也可以应用在融合之后，只进行一次优化。

3.5 实验结果

A. 基于 ORB-SLAM2 框架的语义融合结果

经过 ORB-SLAM2 估计相机位姿并进行点云旋转平移变换后得到的融合结果如图 3.4 所示。可以看到语义地图相比于直接重建的地图更为清晰，且不存在光照变

化，可以为后续定位导航应用提供更鲁棒的信息。然而，语义点云地图中出现了很多错误标注点，这些点是语义分割网络误分割的产物，通过引入语义分割中的概率可以优化。

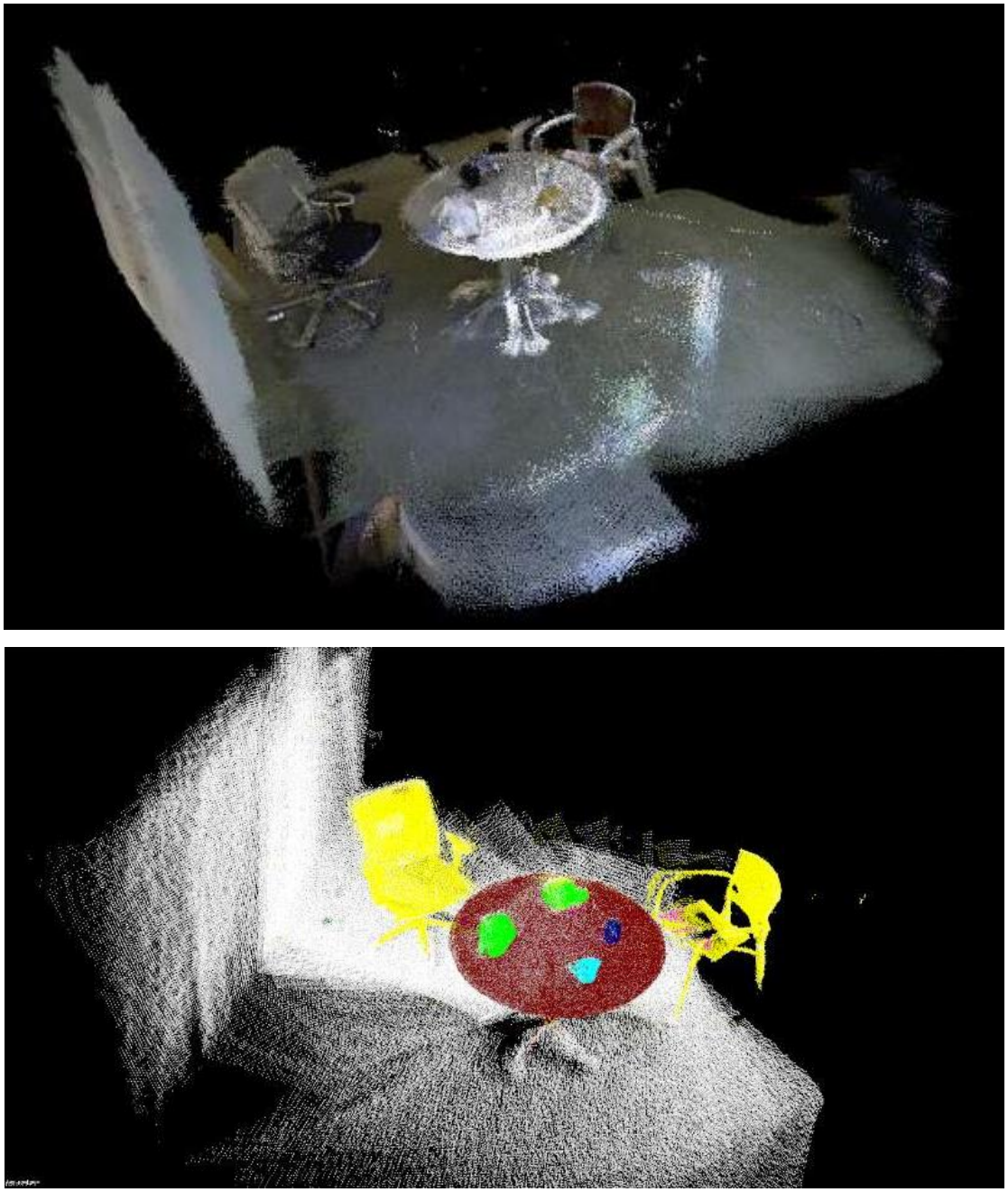

图 3.4 上图为 ORB-SLAM2 重建的度量地图，下图为对应的语义地图

B．语义点云地图的规模控制实验

本文设置体素栅格的大小为$1cm^3$，由于内存限制，需要先对每帧数据进行一次滤波，在语义融合后再进行一次滤波以控制最终语义地图的规模。实验的结果如

图 3.5 所示，上图为仅对每帧数据进行一次体素网格滤波的结果，下图为对每帧点云进行滤波，在融合后再进行一次滤波的结果，可见点云较一次滤波的情况更为稀疏，使得存储的数据量减少。

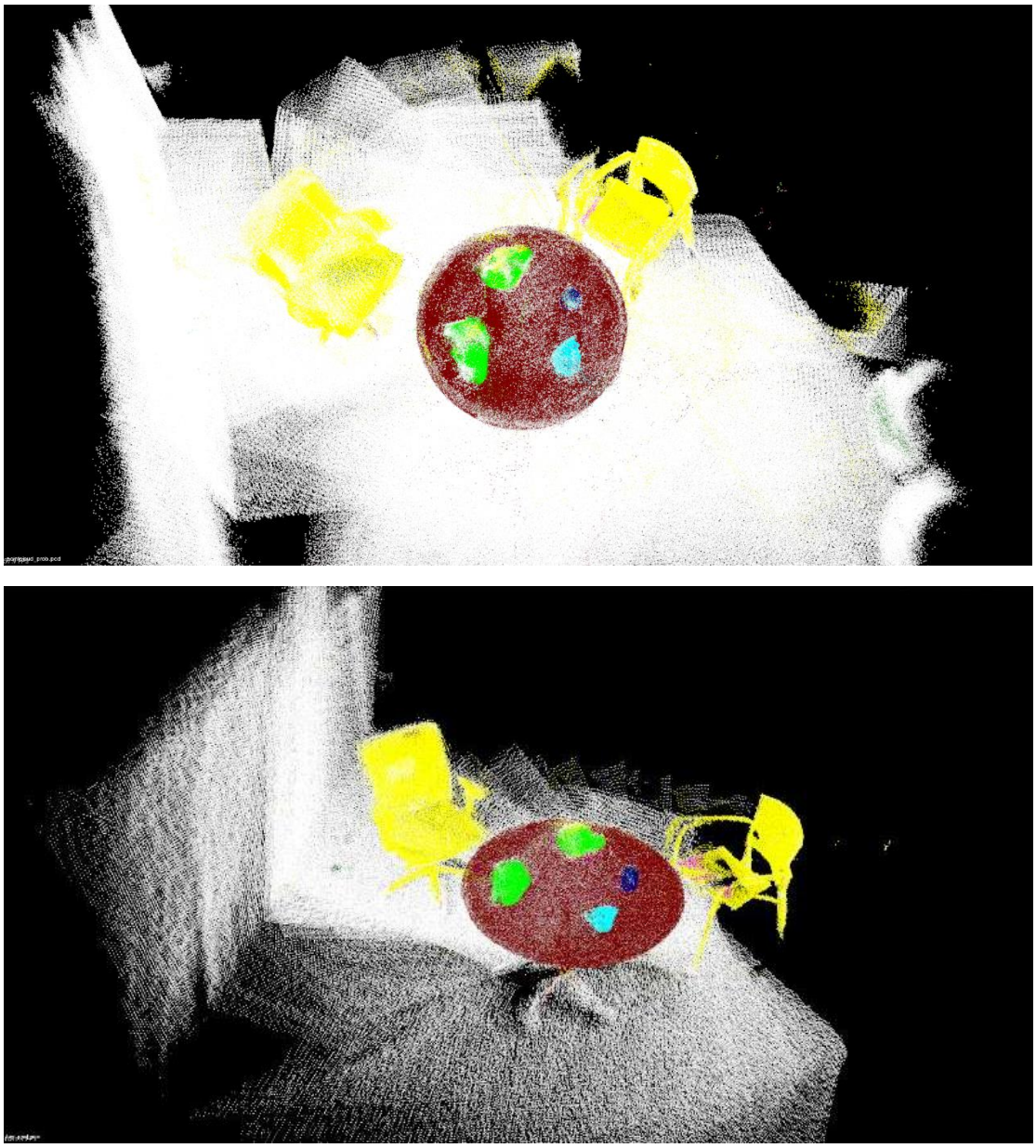

图 3.5 上图为稠密的语义地图，下图为控制规模的语义地图

C．基于概率的标注优化结果

基于引入概率的优化算法可以在两个位置进行优化，一是语义融合后做一次全局优化，二是对每帧进行一次优化，其效果如图 3.6 所示。从图中的红框可见，不论在哪个地方，地图都有不同程度的优化。每帧优化然后融合的次序，物体表面的

误标注点明显变少，说明这些点都是一类错误点。最后只进行一次全局优化的结果，物体表面误标注点相较于没有优化时要少，但是仍有团簇的误标注点，说明多视角融合后，该位置的一些二类错误标注点聚集，影响了优化的结果。

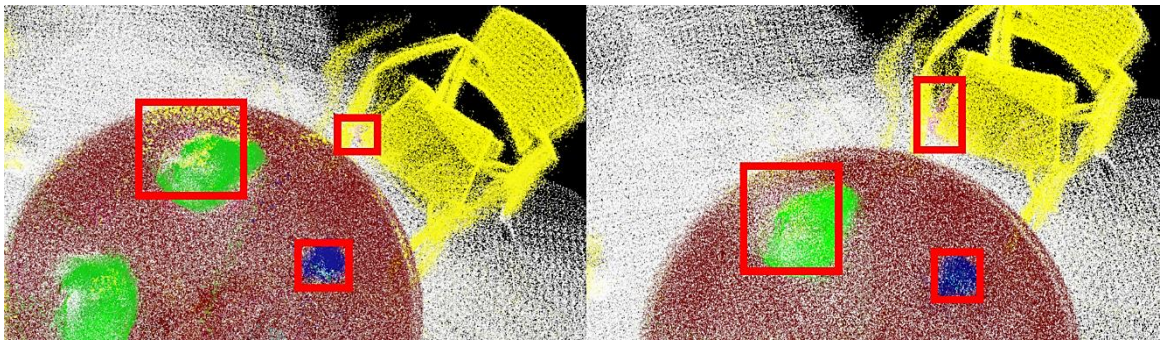

图 3.6 左图为全局一次优化的结果，右图为对每帧分别做一次优化再融合的结果

3.6 本章小结

本章利用语义分割网络生成的语义标注结合 SLAM 方法进行离线的多视角语义信息融合。针对地图的一些标注错误，通过引入语义分割网络输出的语义类别概率进行优化，达到了不错的效果

4 物体层面的语义地图构建与优化

4.1 引言

对于物体层面的语义地图构建问题，在上述研究的基础上，在点云地图中提取相近且具有相同语义标注的 3D 点，优化后形成物体的点云模型，并建立数据库，可以增强语义地图的可理解性。具体步骤包括：语义物体提取以及物体模型关联。

4.2 语义物体提取

4.2.1 基本语义物体提取方法

借鉴 Sunderhauf 提出的语义地图构建方法[7]：结合目标检测结果以及点云分割结果，将物体识别出来并作为地图的物体元素。本设计通过直接的语义分割将物体提取出来作为单独的一个个物体，并建立数据库。这个数据库与物体模型数据库不同，它是将观测过的物体储存下来，用于标记机器人的观测状态，而不是事先存储用于匹配的先验物体模型。

4.2.2 语义物体提取优化

基于建好的语义点云地图，提取相同语义标注的点，即可得到语义点云模型，但这样提取的点云模型往往非常粗糙，会有大量的离群点，以及多个物体的情况。本设计采用基于方差的离群点去除以及聚类分割的方法，以避免出现上述情况。

A．基于方差的离群点去除方法

视觉恢复的 3D 点云存在测量误差，会导致稀疏异常值，即离群点，从而破坏语义地图的精度。这些离群点的出现同时也会使得局部点云特征（例如表面法线或曲率变化）的估计变得复杂，导致错误计算，使得使用地图定位时，无法很好地进行点云帧的配准。不规则离群点中的一部分可以通过对每个点的邻域进行统计分析和修剪来解决。离群点的去除基于点到近邻点距离分布的计算。对于每个点，计算从它到 K 个近邻点的平均距离。假设结果分布是高斯分布，具有均值和标准差，设置阈值，将标准差超出阈值的部分剔除。

B．聚类分割方法

对于场景中有多个物体的情况，在去除离群点后，通过聚类分割的方法提取单个物体。常见的方法有基于欧氏距离分割和基于区域生长分割的方法，它们都是用区分近邻关系来分割的。

欧式聚类分割方法使用近邻之间的欧氏距离作为判定标准，这种方法计算较快，分割精度一般，而区域生长算法利用法线、曲率、颜色等信息来判断点云是否应该聚成一类，计算较慢，对于结构性场景分割精度高。考虑到语义标注信息已经能够使同类物体从场景中分离，其分割难度不高，故使用欧式聚类分割方法。

欧式聚类分割算法实现流程如下：

1、找到空间中的一点 P1，通过 kd-tree 找到它的 k 个近邻，判断这 k 个点到 P1 的距离，将距离小于阈值的点 P2、P3、P4…放入一类 Q 中。

2、在除去 P1 点的类 Q 中找到一点 P2，继续第一步操作,。

3、在除去 P1、P2 的类 Q 中找到一点并重复第一步操作，直到 Q 中没有新的点加入，则算法结束。

对于成片的点云，该方法通常不能奏效，但针对本文，语义标注提供了基本分割后，该方法可以有效地将多物体分为单独的个体。这为之后判断相似性以及优化提供了更好的基础。

C．语义物体表面重建

在经过离群点处理和聚类分割后，物体模型已经初具雏形，但由于数据集采集使用的深度相机测量误差较大，其测试误差如图 4.1。

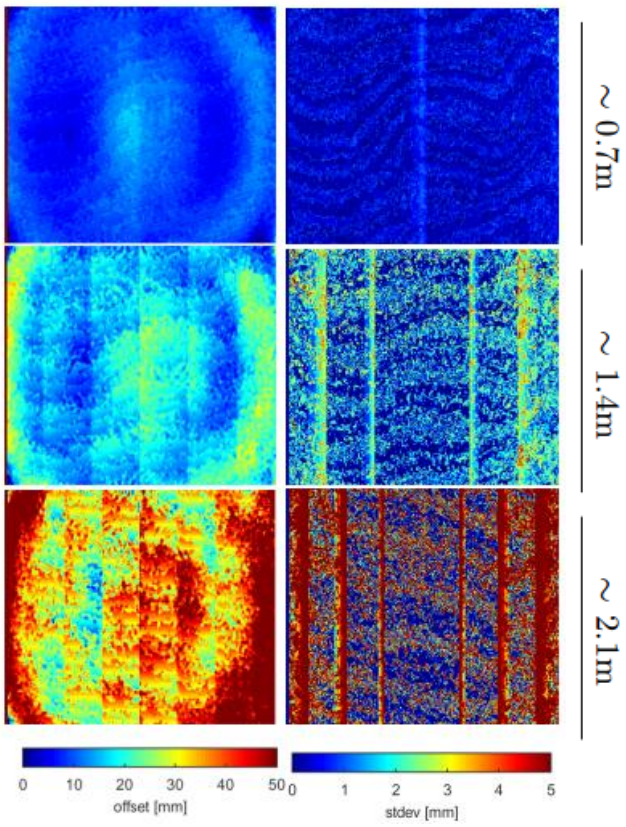

图 4.1[21] KinectV1 的误差分布图

本文中使用的数据集采集的场景中多为距离相机 1m~4m 的物体，由误差分布图可见恢复的点云具有大于 2 厘米的误差，这会导致提取的物体出现一些不规则数据点，这些点直观地表现为物体边缘的重影，如图 4.3 左上所示。

这些不规则数据点无法通过统计分析的方法去除，然而要建立可用完整的模型，必须考虑光泽表面以及数据中的遮挡，在无法获得额外点云数据的情况下，对粗糙模型进行表面重建是一种解决方法。

本文采用基于移动最小二乘法的曲面重建和重采样算法（MLS）[22]来建立可用的完整模型。该算法通过对数据点进行高阶多项式插值来重建表面的缺失部分并通过重采样校正不规则数据点。

4.2.3 语义物体提取与优化实验结果

对于上述算法，本文分别进行了实验，实验结果如下：

A．离群点剔除实验

某点的均值和方差是由它与 50 个近邻点共同计算的，将方差超过 0.5 标准差的点作为离群点去除。该方法的效果如图 4.1 所示，白框中的离群点通过上述算法被剔除了。

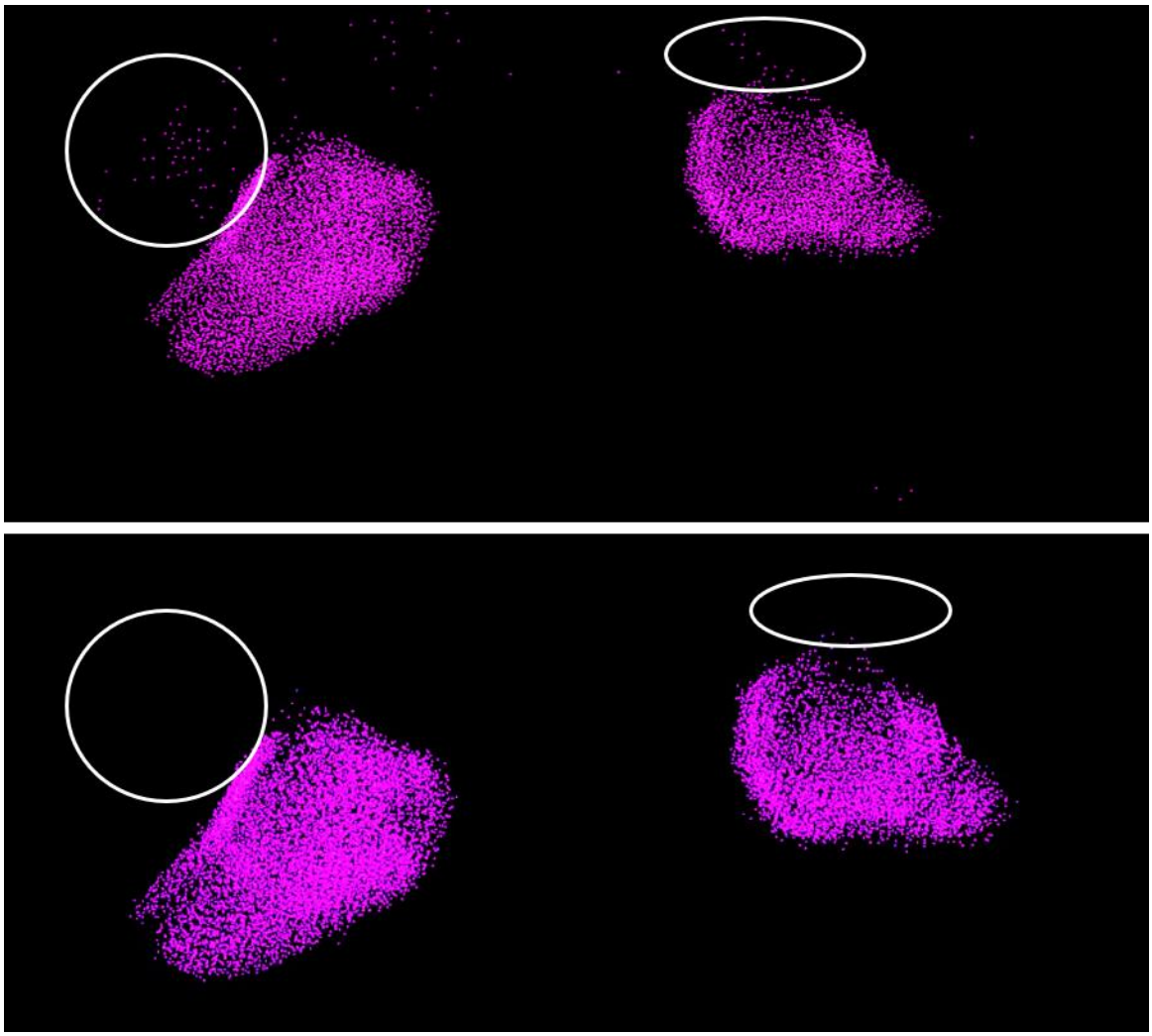

图 4.2 上图为提取相同语义点的结果，下图为剔除离群点后的结果

B．语义物体提取

从得到的语义地图提取物体点云结果如图 4.3 左上，物体点云作为模型太过粗糙，通过 MLS 方法重建得到的结果如图 4.3 右上所示。经过上述提取与优化方法处理后，物体模型就成功提取出来了，其表面更为平滑。该优化方法也可以通过设置重采样算法中的参数，调整物体模型的点云数量，从而控制模型存储的大小。设置点云数量为每$5cm^3$内 300 个点，其点云模型如图 4.3 下所示。

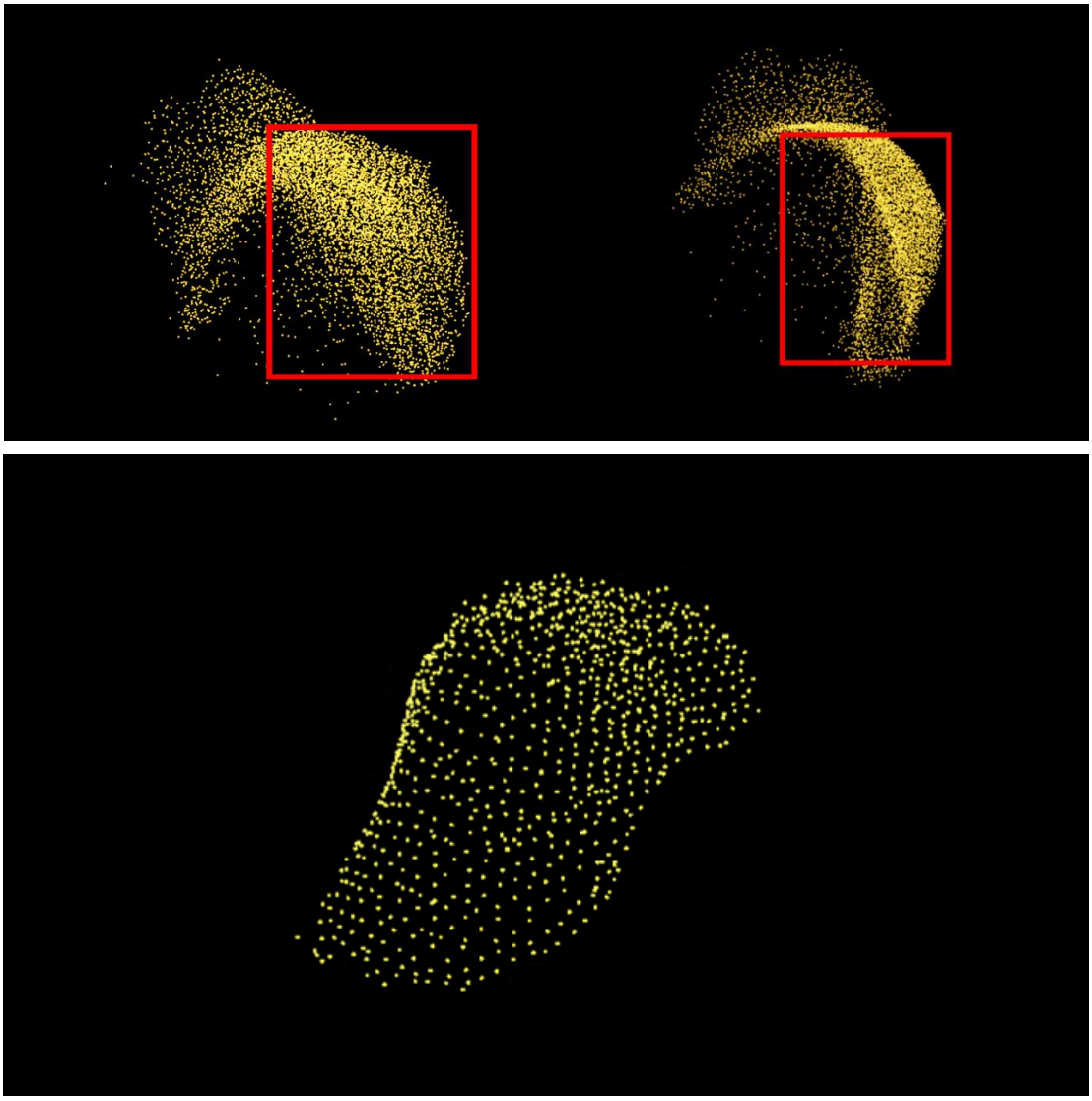

图 4.3　左上为 MLS 重建前，右上为重建后，下方为控制 MLS 参数使点云数量减少的情况

4.3 物体模型存储

本文采用的数据集中，每个场景的物体数量不多，在模型存储方面没有任何困难，但对于本文讨论的语义地图构建问题来说，模型存储的效率与方法就显得至关重要，它直接决定了建图系统是否具有长期稳定的运行效率。考虑在一个很大的室内场景，存在大量可标注语义的物体，机器人在采集完 RGB-D 数据后，离线进行地图构建。在离线构建语义地图的过程中，如何存储提取出来的物体模型，将会影

响到之后机器人在线运行使用该地图时的效率。因此，物体模型存储的原则和方法同样重要，前者决定了点云规模的增长速度，后者则决定了点云利用的效率。

本文采取相似物体只保存一次的原则来解决多个同类相似物体存储导致点云规模剧增的情况。对于存储方法，大数据方法中的数据库管理系统是一种现成有效的点云存储方法。同时，本文也提出了一种针对中小规模点云数据的处理方案，基于语义物体出现频率进行混合点云存储。

4.3.1 基于快速点特征直方图的物体相似性判断

如果将一个大场景中的所有语义物体单独存储，很多类似物体的存在会使存储效率变低，为了提高效率可以采取降低分辨率的方法来减少存储的点云数量，但这就会使得之后以此地图进行定位时，缺少足够的模型点进行配准。

本文采用相似物体只保存一次的原则，在物体模型存储之前进行相似性判断，若判定该物体与已保存的模型相似，则标记该物体，而并不存储它；若判定不相似，则保存该模型。

物体相似性采用快速点特征直方图描述符（FPFH）[23]来衡量。FPFH 是对点特征直方图描述符（PFH）[24]的改进，PFH 通过考虑法线方向之间的相互关系来捕捉表面变化。图 4.4 为点 P_q 的 PFH 影响图，点 P_q 及其 k 个近邻点完全互连。

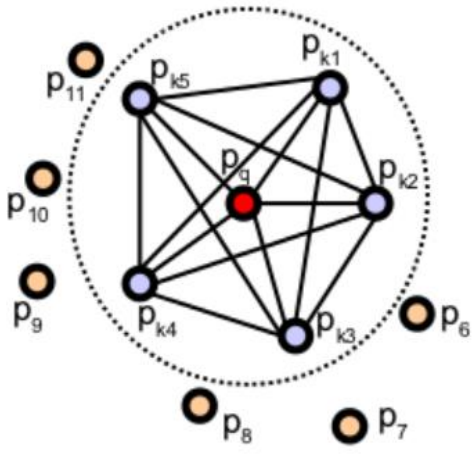

图 4.4　PFH 计算示意图

为计算两个点的关系，定义一个固定的坐标系。如图 4.5 所示，坐标系三轴为 UVW，其中

$$U = n_s \tag{4.1}$$

$$V = U \times (p_t - p_s)/\|p_t - p_s\|_2 \tag{4.2}$$

$$W = U \times V \tag{4.3}$$

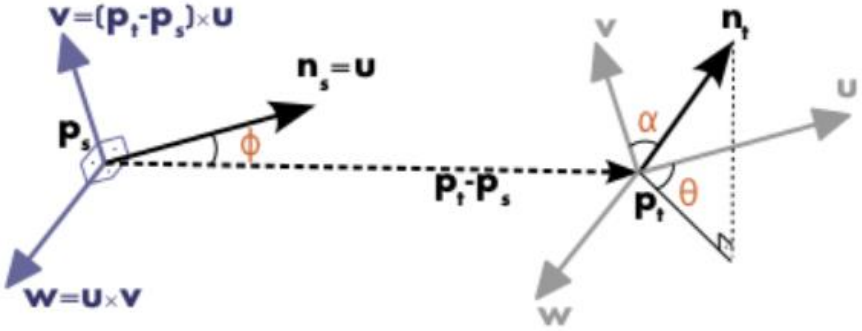

图 4.5　基于 UVW 的坐标系定义

定义了上述 UVW 轴后，两个法线之间的差可以由一组角度值表达：

$$\alpha = V \cdot n_t \tag{4.4}$$

$$\phi = U \cdot (p_t - p_s)/d \tag{4.5}$$

$$\theta = \arctan(W \cdot n_t, U \cdot n_t) \tag{4.6}$$

其中 d 是两个点之间的欧氏距离$d = \|p_t - p_s\|_2$，对某点与其 k 近邻中的点形成的点对计算$\langle \alpha, \phi, \theta, d \rangle$四元组，就将其两个点 12 个特征值（XYZ 和法线信息）减少到了 4 个。要用四元组来描述点云局部特征，还需要将其集合根据值的范围划分为多个区间，并计算直方图。

该方法的计算复杂度是$O(k^2)$，k 为某点附近的近邻点数量。对于定位这类实时的应用，PFH 方法耗时太多，需要进一步改进。因此有了快速点直方图方法（FPFH），该方法的计算复杂度降低到了$O(nk)$，并且还保留了 PFH 的大部分判别能力。

FPFH 与 PFH 同样计算某点自身与近邻的一个元组，在该方法中，这是个三元组$\langle \alpha, \phi, \theta \rangle$，又称简化点特征直方图（SPFH）。确定了该三元组之后，在对该点的 k 个近邻计算它们的 SPFH 值，最后进行如下加权：

$$FPFH(p_q) = SPFH(p_q) + 1/k \cdot \sum_{i=1}^{k} 1/d \cdot SPFH(p_k) \tag{4.7}$$

图 4.6 展示了 FPFH 的计算过程。对于一点p_q，算法首先通过这点与其 k 个近邻点计算 SPFH 值，这个过程为加粗红线部分。然后对这 k 个近邻点重复此操作，并计算它们的 SPFH 值并加权得到p_q的 FPFH 值。由于加权方案而产生的额外的 FPFH 值在图中由加粗黑线标出。

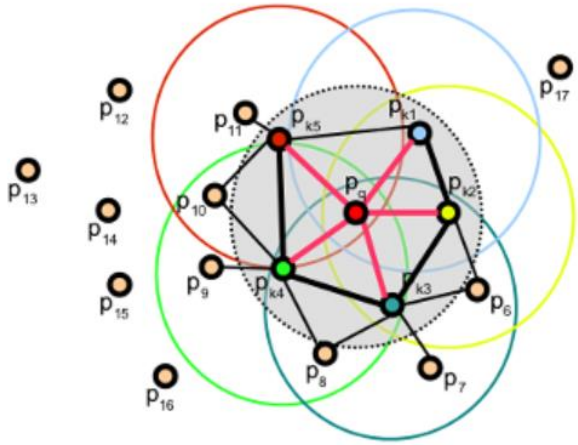

图 4.6　FPFH 的计算示意图

相比 PFH，FPFH 没有完全互连所有近邻，因此可能缺少一些有助于描述周围几何特征的值，但 FPFH 因此也获得了更快的计算速度。

FPFH 对于点云的这种快速描述特点使得其能够在物体存储之前提供相似性判断，只需要 FPFH 直方图相似，并且点云尺寸大小相近，就可以认为两个物体具有相似性。

4.3.2 基于频次的混合点云存储方法

面向中小规模的室内场景语义建图，本文提出一种基于语义物体出现频率的混合点云存储方法，兼顾效率与存储规模。

在建图过程中，物体模型存储前会进行一次相似性判断，若判定为相似物体，则该物体不被存储，但它会增加该相同物体出现的频次。在遍历完整个语义地图后，所有物体模型按照频次大小进行排序，这个次序作为一个先验的信息来决定模型以哪种方式存储。出现频次高的前几个模型在定位等在线应用中会被频繁调用的可能性较大，则优先保存，在线使用地图时将会直接读入内存。而出现频次排名靠

后的物体，由于不会经常被调用，故保存成点云文件而不读入内存，可以节省内存的开销。

在线使用地图时，系统会重新维护一个出现频次排序表，并一直以此作为指导来进行存储方式的选择。

4.3.3 实验结果

针对上述算法，本文分别进行了实验，实验结果呈现如下。

A．相似性判断实验

提取数据集中的某一帧 RGB-D 数据，经由语义分割网络得到语义标注，将其转化为三维点云，找到帽子的语义点云，经过之前的提取分割优化，得到图 4.7 中的红色点云。绿色点云表示模型库中的点云。

通过计算两者 FPFH 进行匹配，可以将下图左边位姿区别很大的情况，配准到右边位姿接近的情况，再计算尺寸大小，判断两者相似，则不予保存。

FPFH 还可以用在机器人的语义定位中，设想将上述过程扩展到当前帧所有语义物体与全局语义地图中相似物体进行配准，计算得到的变换矩阵就是当前帧相对于全局地图原点的位姿。

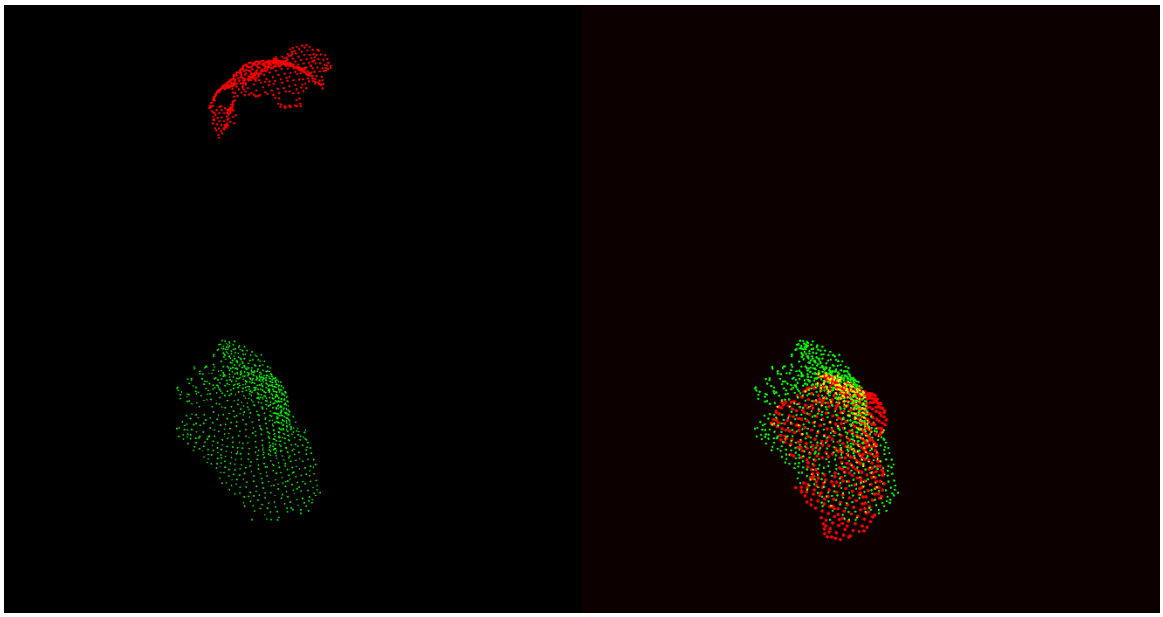

图 4.7　红色点为当前帧中提取的帽子点云，绿色点为数据库中的帽子点云

B．中小规模数据存储

由于本文采用的数据集中，出现的物体较少，本身对内存的影响较小，同一类物体也都较为相似，导致只有极少数物体在线存储，其他的被存储为点云文件。

针对图 4.8 构建的语义地图，频次如表 4.1：

表 4.1　物体模型频次统计

	频次	存储方式
转椅	1	点云文件
椅子	1	点云文件
桌子	1	点云文件
咖啡杯	1	点云文件
易拉罐	1	点云文件
帽子	2	内存

C．物体地图构建效果

从语义点云地图中提取物体点云并优化，形成物体模型数据库，构成了物体层面的语义地图，效果如图 4.8 所示，上图为构建的语义地图，下图为对应的地图中提取的物体数据库。

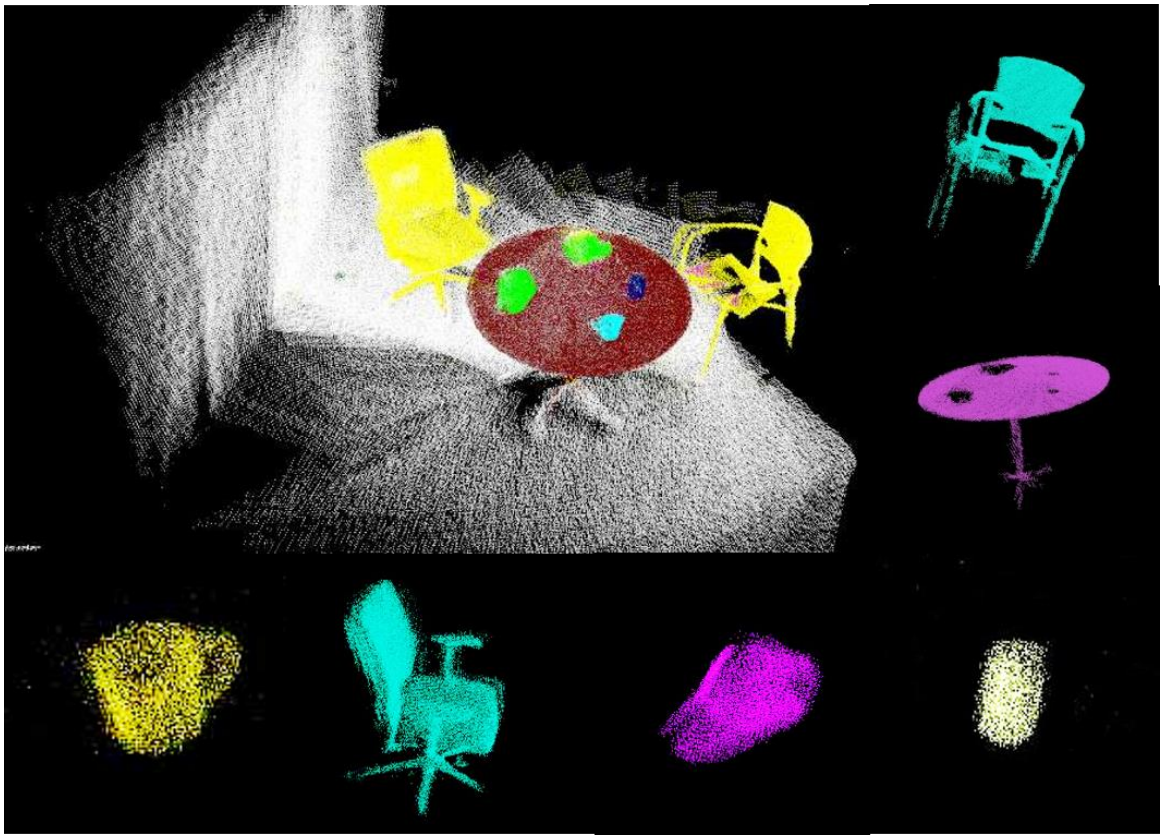

图 4.8　左上图为构建的语义地图，其他为提取的物体模型

4.4 本章小结

本章阐述了如何将点云层面的语义地图转换到物体层面的语义地图，以提供更好的可理解性。该过程中涉及物体模型提取以及模型存储的多种算法，并最终挑选并实现适合本文情况的方案。

5 结论与展望

5.1 结论

本次毕业设计主要完成的工作内容可分为三部分。

第一部分为语义分割网络的搭建。修改了循环神经网络中数据关联循环单元的数据关联方式，以结合多视角的物体信息，帮助提高语义标注的准确度。实验证明，通过 ORB-SLAM2 框架进行数据关联，引入多视角的信息能够提高网络语义标注的能力，平均像素 IoU 接近 90%，相比全卷积网络提高了约 4%。

第二部分为语义标注融合及优化。将语义信息加入到传统的度量地图构建框架中实现语义融合。针对融合后地图出现的一类错误标注，根据近邻点语义概率更新优化，成功剔除了大部分错误标注，得到一致性较好的语义点云地图。

第三部分为物体语义地图的构建。在语义标注融合得到的点云地图基础上，提取物体模型，并提出了采用 FPFH 特征匹配的相似性判断方法以及中小规模点云数据存储方案，可以节省内存开销，间接提高存储效率，为机器人定位与导航应用提供便利。

5.2 展望

本次毕业设计为后续开展语义地图构建奠定了基础，进一步可以开展的工作包括：

（1）语义分割网络目前无法达到实时标注的程度，未来可以通过优化网络结构，用 C 语言重构等方式达到实时性。

（2）未来可以利用构建的物体语义地图进行定位，实现语义定位进而实现语义导航的应用。

（3）目前整个系统还处于离线分步测试的阶段，未来可将所有环节集成为一个类似 ORB-SLAM2 的易用框架，应用于机器人实际移动过程中。

参考文献

[1] Thrun S. Robotic mapping: A survey[J]. Exploring artificial intelligence in the new millennium, 2002(February): 31.

[2] Nuchter A, Hertzberg J. Towards semantic maps for mobile robots[J]. Robotics and Autonomous Systems, Elsevier B.V., 2008, 56(11): 915–926.

[3] Salas-Moreno R F, Newcombe R A, Strasdat H, et al. SLAM++: Simultaneous localisation and mapping at the level of objects[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013: 1352–1359.

[4] Fioraio N, DI Stefano L. Joint detection, tracking and mapping by semantic bundle adjustment[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013: 1538–1545.

[5] Stuckler J, Biresev N, Behnke S. Semantic mapping using object-class segmentation of RGB-D images[J]. IEEE International Conference on Intelligent Robots and Systems, 2012: 3005–3010.

[6] Hermans A, Floros G, Leibe B. Dense 3D semantic mapping of indoor scenes from RGB-D images[J]. Proceedings - IEEE International Conference on Robotics and Automation, 2014: 2631–2638.

[7] Sunderhauf N, Pham T T, Latif Y, et al. Meaningful maps with object-oriented semantic mapping[J]. IEEE International Conference on Intelligent Robots and Systems, 2017, 2017-Septe: 5079–5085.

[8] Xiang Y, Fox D. DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks[J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6565–6574.

[9] Newcombe R A, Davison A J, Izadi S, et al. KinectFusion: Real-time dense surface mapping and tracking[C]//2011 10th IEEE International Symposium on Mixed and Augmented Reality. IEEE, 2011(November 2014): 127–136.

[10] Schonberger J L, PollefeysS M, Gerger A, et al. Semantic Visual Localization[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 6896–6906.

[11] Wang P, Yang R, Cao B, et al. DeLS-3D: Deep Localization and Segmentation with a 3D Semantic Map[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 5860–5869.

[12] Boykov Y Y, Jolly M-P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images[C]//2002.

[13] Rother C, Kolmogorow V, Blake A. 《GrabCut》 - Interactive foreground extraction using iterated graph cuts[J]. Proceedings of ACM SIGGRAPH, 2004.

[14] Lafferty J, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[M]. ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001, 8(June).

[15] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[16] 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, 2017, 33(5): 1255–1262.

[17] Abadi M, Agarwal A, Paul Barham E B, et al. TensorFlow: Large-scale machine learning on heterogeneous systems[J]. Methods in Enzymology, 1983.

[18] Jia Deng, Wei Dong, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009.

[19] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. 2014: 1–14.

[20] Lai K, Bo L, Fox D. Unsupervised feature learning for 3D scene labeling[C]//Proceedings - IEEE International Conference on Robotics and Automation. 2014.

[21] Wasenmuller O, Stricker D. Comparison of kinect v1 and v2 depth images in terms of accuracy and precision[J]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, 10117 LNCS: 34–45.

[22] Lancaster P, Sakkauskas K. Surfaces Generated by Moving Least Squares Methods[J]. Mathematics of Computation, 2006, 37(155): 141.

[23] Rusu R B, Blodow N, Beetz M. Fast Point Feature Histograms (FPFH) for 3D registration[C]//2009.

[24] Rusu R B, Marton Z C, Blodow N, et al. Persistent point feature histograms for 3D point clouds[C]//Intelligent Autonomous Systems 10, IAS 2008. 2008.

附录

1 投稿论文