DOI: 10.16452/j.cnki.sdkjzk.2024.03.010 文章编号:1672-3767(2024)03-0094-11

# 基于深度学习实例分割的室内环境物体语义建图

许 鹏[1],龚士博[1],田德旺[1],苑 晶[2],孙凤池[1]
(1. 南开大学 软件学院,天津 300350;2. 南开大学 人工智能学院,天津 300350)

**摘 要**:为提升家庭服务机器人面向物体语义建图的精确性,提出一种改进的语义建图方法。该方法基于深度学习对单帧图像进行实例分割,利用其结果指导基于几何关系的聚类,使聚类结果既能保留深度学习方法识别的物体语义信息,又能贴近物体的几何边界。针对现有方法所建地图中物体标签缺乏可靠性的问题,提出一种语义地图优化方法,对地图中距离相近、重合度较高、类别相似的重复物体进行合并,并删除错误物体。实验验证表明,提出的方法能够提升物体边界分割的准确性,得到的物体数目更加准确。

**关键词**:物体语义建图;实例分割;深度学习;室内环境
**中图分类号**:TP391.41; TP242 **文献标志码**:A

## Object-oriented semantic mapping for indoor environments based on instance segmentation with deep learning

XU Peng[1], GONG Shibo[1], TIAN Dewang[1], YUAN Jing[2], SUN Fengchi[1]
(1. College of Software, Nankai University, Tianjin 300350, China;
2. College of Artificial Intelligence, Nankai University, Tianjin 300350, China)

**Abstract**: To enhance the fidelity of the object-oriented semantic maps of domestic service robots, this paper proposed an improved semantic map building method. The method performed instance segmentation on the images with a deep learning method and instructed a geometric relationship-based clustering method to segment objects from the point clouds. The clustering results not only preserved the semantic information identified by the deep learning method, but also described geometric boundaries of objects more precisely. To solve the problem of the lack of reliability for object labels in existing approaches, this paper also proposed a semantic map optimization process. It merged instances which had close distances, high overlapping rates, and similar label scores. Meanwhile, it removed false objects. The experimental results show that the proposed method can improve the quality of the map built by the system in terms of the accuracy of the boundaries and the number of instances.

**Key words**: object-oriented semantic mapping; instance segmentation; deep learning; indoor environments

未来的家庭服务机器人需要处理各种任务,这些任务要求与环境中的不同物体进行多种交互,如抓取、移动、操作等。因此,机器人需要提前对环境中的物体建模,进而构建关于物体的语义地图(简称物体地图)。传统的物体语义建图方法获取精确且可靠的物体边界较为困难,使得地图中能建模的几何信息较少,通常限于位置和大致范围,难以对物体的具体形状、边界等信息进行建模。随着深度学习方法在计算机视觉领域的发展,基于深度学习的目标检测、实例分割等方法使得物体的语义信息与边界可以被更可靠地获取,对物体

**收稿日期**:2022-09-19
**基金项目**:国家自然科学基金项目(61873327,62073178);国家自然科学基金区域创新发展联合基金重点项目(U21A20486)
**作者简介**:许 鹏(1997—),男,河南信阳人,硕士研究生,主要从事移动机器人位置识别、语义建图的研究.
孙凤池(1973—),男,山东滨州人,教授,博士,主要从事机器学习、智能机器人系统的研究,本文通信作者. E-mail: fengchisun@nankai.edu.cn

精确边界建模的语义建图得到发展。

当前,依据是否以物体为中心可将语义建图方法分为两类。第一类为标注物体的语义建图,该类方法主要基于表面元(surface element,surfel)地图,在图中每个 surfel 上标注其属于各个标签的概率。SuMa++基于 SuMa 建立 surfel 地图[1-2],利用 RangeNet++[3]进行语义分割,输出逐点的语义标签概率分布,并以此标注 surfel 的标签。相比之下,Lu 等[4]使用迭代最近点(iterative closest point,ICP)估计相机位姿,通过实例分割与超像素合并提取物体,建立标有物体标签的 surfel 地图。上述两种方法分别使用了语义标签和物体标签,语义标签对应着不同的物体类别,而物体标签则在此基础上进一步区分了同一类别下的不同物体实例。第二类方法是面向物体的语义建图,将物体作为建图主体。Sünderhauf 等[5]使用 ORB-SLAM2[6]提供相机位姿,结合目标检测算法得到物体框,并利用基于图的几何方法与平面凹凸关系判定标准进行实例分割,建立包含物体观测结果及类别标签置信度的地图。文献[7-8]延续了这种目标检测配合几何方法的思路。在文献[5]的流程基础上,Voxblox++使用 Furrer 等[9-10]提出的物体数据库存储物体并进行关联。相较于通过 surfel 的标签分布隐式表达物体边界的第一类方法,这类方法能对物体进行更加直观地建模,并且能进行更细致的处理。如 MID-fusion[11]将分割的物体片段与物体模型关联,加入物体模型的运动残差模型,修正分割结果。但该类方法缺少点云分布的约束,容易造成物体重复。

上述两类建图方法严重依赖物体分割边界的正确性,在物体分割不准确时,建图精确性较低且难以优化,而且对地图中存在错误或需要后续修改的物体实例缺少相应的处理手段。为此,本研究在文献[5]的基础上,提出一种新的面向物体的语义建图方法,主要创新点包括:

1) 提出一种由图像实例分割技术指导的点云实例分割方法。由于地图质量依赖物体分割的边界,通过将深度学习实例分割与基于加权图的超体素聚类相结合,改善分割结果的精确性。

2) 提出一种基于关联计数的地图优化方法。针对现有建图方法对物体标签缺乏后续处理的缺陷,通过统计地图中物体被关联的次数,判断重复或错误物体,并分别进行合并或删除。

3) 提出一个面向物体的语义建图方法,结合以上两种方法,从 RGB-D 图像序列建立物体地图。

## 1　方法概述

本研究提出的语义建图方法主要包括同步定位建图(simultaneous localization and mapping,SLAM)、点云分割、物体关联及地图优化四个处理过程,工作流程如图 1 所示。其中,绿色图形代表处理过程,蓝色图形代表这些过程的输入输出数据。本研究将所建立的物体地图称为统一图。

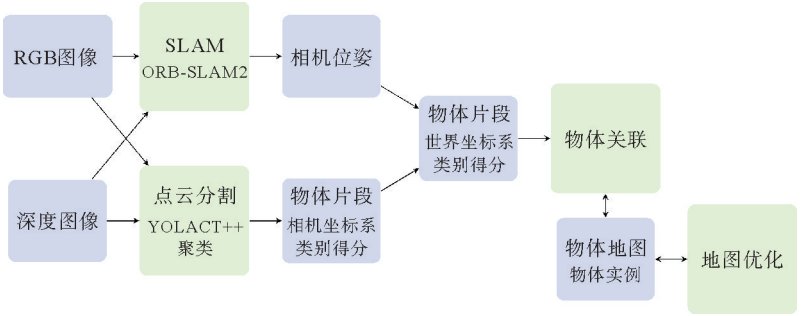



**图 1　地图构建方法工作流程**

Fig. 1　The working process of the map building system

本方法将 RGB-D 图像(RGB 图像与深度图像)序列作为输入,并从中建立物体地图。当输入一帧 RGB-D 图像时,首先由 SLAM 过程与点云分割过程对其处理。SLAM 过程采用 ORB-SLAM2 计算当前帧的相机位姿估计。点云分割过程使用 YOLACT++(you only look at coefficients)[12]指导的超体素聚类方法,分割出当前帧中的物体片段。然后,通过相机位姿估计将这些物体片段变换到世界坐标系中,物体关联过程负责将物体片段与地图中的物体实例建立关联关系,以此更新物体地图。在此期间,地图优化过程通过检查

关联过程中的计数,对地图中重复、错误的物体实例进行合并或删除,提升物体地图的可靠性。

## 2 图像实例分割技术指导的点云实例分割

本节主要对单帧 RGB-D 图像进行点云分割处理,分割算法流程如图 2 所示。对于每帧 RGB-D 图像,一方面,使用基于深度学习的图像实例分割技术在 RGB 图像上进行分割,获取物体标签;另一方面,通过深度图像将 RGB 图像转化为点云,并在点云中划分超体素;然后,利用物体标签信息以及超体素间的几何关系信息,对超体素聚类,获得若干准物体片段;最后,统计准物体片段的物体标签信息,合并标签相同的准物体片段,获得最终的物体片段。该分割方法既保留了深度学习方法识别到的物体语义信息,又能产生贴近几何边界的分割结果,为地图提供单帧图像中边界更加精确的物体。

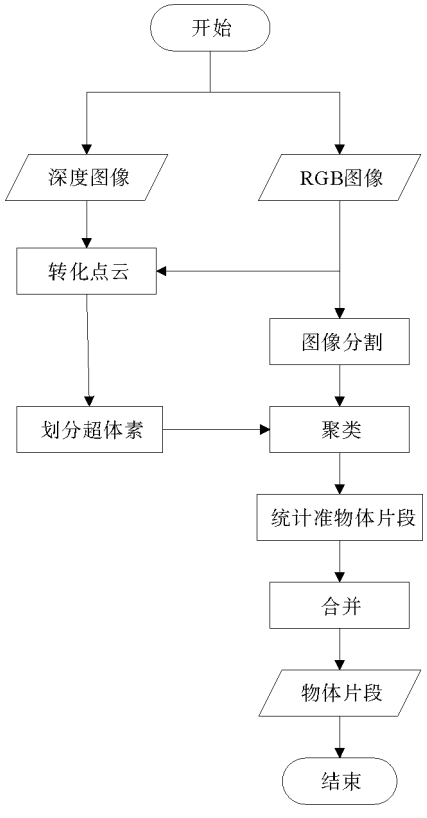



**图 2 点云实例分割流程**

Fig. 2 The workflow of the point cloud instance segmentation

### 2.1 图像实例分割技术

本研究借助基于深度学习的图像实例分割来获取物体语义信息。在综合考虑精度和速度两方面的性能之后,选用 YOLACT++网络对图像进行实例分割。YOLACT++网络接受一帧 RGB 图像,预测图像中所含有的物体,并输出物体标签、标签置信度及其掩模。

原始的 YOLACT++使用 Fast-NMS(Fast non-maximum suppression)[12]进行非极大值抑制操作,从而去除重复的检测结果。但应用在室内环境中时,由于物体更容易出现堆叠、聚集等情况,该操作会错误地丢弃物体,产生过度抑制的问题。为提升 YOLACT++在室内环境下的实例分割效果,本研究将 Fast-NMS 改为 Soft-NMS[13]。相较前者达到阈值就去除的抑制机制,后者能够更"光滑"地对物体框重合度较高的情况进行抑制,对较密集的物体有更好的检测能力。

Soft-NMS 在每次选出得分最高的物体候选框 $b_{\max}$ 后,将减小其他有重叠的待选物体框的得分。本研究选择高斯函数来达到"光滑"目的,即对于每个其他待选候选框 $b_i$,计算两候选框交并比 $\mathrm{IoU}(b_{\max},b_i)$,并按照式(1)更新其标签得分 $s_i$,

$$s_i'=s_i\exp\left(-\frac{\mathrm{IoU}(b_{\max},b_i)}{\sigma^2}\right)。\tag{1}$$

式中:$\sigma$ 为所选择高斯函数标准差,$i$ 为候选框编号。

### 2.2 结合物体标签信息和几何信息的聚类权重

由于 YOLACT++只在 RGB 图像上进行分割,未考虑深度图像中蕴含的几何信息。而且,为统一网络结构,YOLACT++将图像缩放至 550×550 后输入,生成 138×138 大小的掩模,使得掩模中的每个像素都对应原图像中的一个像素块,导致物体的边缘不够精确。因此,本研究利用深度图像中的几何信息挖掘物体边界信息,弥补 RGB 图像的不足。

首先,将 RGB-D 图像转换为点云数据;然后,利用体素云连通分割算法(voxel cloud connectivity segmentation,VCCS)[8],在点云中等间距放置种子,并依据各点的颜色、空间位置、快速点特征直方图(fast point feature histograms,FPFH)划分出超体素。将超体素记为 $v_i$,则原 RGB-D 图像可以视为一张图 $G=\langle V,E\rangle$,其中 $V=\{v_1,v_2,\cdots,v_{n_t}\}$ 为超体素的集合,$n_t$ 为超体素的个数,$E$ 为超体素之间相邻关系构成的无向边集合。

为 $E$ 中每条边赋予权重,使 $G$ 成为一张加权图。超体素 $v_i$、$v_j$ 间边的权重定义为:

$$w(i,j)=w_{\text{sem}}(i,j)+w_{\text{geo}}(i,j)。\tag{2}$$

式中,$w_{\text{sem}}$ 和 $w_{\text{geo}}$ 分别为边上的语义权重和几何权重,分别取决于两个顶点上的物体标签信息和几何信息。边上的权重越小,意味着两个超体素越相似,越应被归入同一类。

2.2.1　语义权重

语义权重度量两个超体素在实例分割中物体标签的一致性。对于超体素 $v_i$,找到与该超体素重叠比例最大的掩膜,将掩膜的标签作为该超体素的物体标签,记为 $m_i\in\{0,1,2,\cdots,N_t\}$,其中 0 代表背景,其他标签分别代表该帧图像分割出的 $N_t$ 个物体,则语义权重表达为:

$$w_{\text{sem}}(i,j)=\begin{cases}0, & m_i=m_j\neq 0;\\ w_{\text{diff}}, & m_i\neq m_j,m_im_j\neq 0;\\ w_{\text{bg}}, & m_im_j=0。\end{cases}\tag{3}$$

式中,$w_{\text{diff}}>w_{\text{bg}}>0$。当两个超体素属于同一物体时,赋予最小权重 0;当两个超体素为不同物体时,赋予最大权重 $w_{\text{diff}}$;当一个是物体、一个是背景时,赋予较小权重 $w_{\text{bg}}$。

2.2.2　几何权重

几何权重度量两相邻超体素的交界对应邻接边界的概率,该权重依赖超体素的支撑平面及两超体素间的凹凸关系。对于每个超体素 $v_i$,当其曲率较小时,则认为该超体素可能构成平面,并使用几何一致性平面抽取算法提取支撑平面,标记每个超体素 $v_i$ 所处支撑平面的平面标签,记为 $l_i\in\{0,1,2,\cdots,N_p\}$,其中 0 代表这一超体素不处于任何支撑平面,而非 0 则代表这一超体素处于对应编号的支撑平面上。

超体素间的凹凸关系使用局部凸连接(locally convex connected patches,LCCP)方法[14]进行判断,主要衡量两相邻超体素平面交界向外凸起还是向内凹陷。首先,定义超体素 $v_i$、$v_j$ 之间的广义凸关系为:

$$(\boldsymbol{n}_i-\boldsymbol{n}_j)\cdot(\boldsymbol{c}_i-\boldsymbol{c}_j)>0。\tag{4}$$

式中:$\boldsymbol{c}_i$、$\boldsymbol{c}_j$ 和 $\boldsymbol{n}_i$、$\boldsymbol{n}_j$ 分别为 $v_i$、$v_j$ 的中心和法向量。进一步地,超体素 $v_i$、$v_j$ 需要满足以下三点才能真正构成凸关系:①$v_i$、$v_j$ 构成广义凸关系;②存在公共相邻超体素 $v_k$,使得 $v_i$、$v_k$ 与 $v_j$、$v_k$ 均构成广义凸关系;③$\boldsymbol{c}_i-\boldsymbol{c}_j$ 与 $\boldsymbol{n}_i\times\boldsymbol{n}_j$ 之间的夹角足够接近 90°。第②点用于降低噪声的影响,第③点则保证了两个超体素之间存在交界,同理可定义凹关系。关于凹凸关系的技术细节可参考文献[14]。

与文献[5]中所述权重类似,本研究将几何权重定义为:

$$w_{\text{geo}}(i,j)=\begin{cases}0, & l_i=l_j;\\ 1, & l_il_j=0,l_i\neq l_j;\\ \dfrac{1}{4}(1-\boldsymbol{n}_i\cdot\boldsymbol{n}_j)^2, & l_i\neq l_j,\text{且构成凸关系};\\ 1-\boldsymbol{n}_i\cdot\boldsymbol{n}_j, & l_i\neq l_j,\text{且构成凹关系}。\end{cases}\tag{5}$$

对于大部分的物体,其表面近似于多个平面,且这些平面的交界向外凸起。因此,对于两个相邻超体素,如果处于同一平面,或者处于不同平面但构成了凸关系,那么该相邻超体素更有可能属于同一物体,因此其交界对应邻接边界的概率更低,将被赋予较小权重。

### 2.3　准物体片段合并

经过以上两步,一帧 RGB-D 图像已被处理为加权图,此时进行基于加权图的聚类。然而,该聚类方法的停止条件鲁棒性较差,易出现过分割或欠分割问题。实验证明,调整控制聚类大小偏好的参数 $K$,当参数值较小时,能让聚类方法较早结束,并产生轻微过分割的结果,而当 $K$ 取值为 2 时,能够得到最好的实验结果,不会发生欠分割问题。由于存在过分割,将此时得到的聚类结果称为准物体片段,并提出以下准物体片段的合并算法。

首先,根据超体素中心在 RGB 图像中的位置,从 YOLACT++的结果中获得每个超体素的物体标签;然后,统计每个准物体片段中超体素的物体标签,将出现次数最多的物体标签作为该准物体片段的标签;最后,合并标签相同的准物体片段,得到最终的物体片段。

## 3 面向物体的语义地图构建

本节对多帧的图像序列进行处理,利用第 2 节的点云实例分割算法,从每帧图像中得到若干物体片段,再通过 SLAM 算法估计的相机位姿,将这些物体片段变换到世界坐标系中,然后基于变换后的物体片段,利用物体关联和地图优化两个处理过程,将物体片段与地图中的物体实例进行关联整合,并对地图中重复或错误的物体进行合并或删除,建立一个更加准确的物体地图。

### 3.1 物体地图的结构

本研究构建的物体地图结构如图 3 所示,用 $M$ 表示物体实例的有序集合,

$$M=\{o_1,o_2,\cdots,o_{N_o}\}。\tag{6}$$

式中:$N_o$ 表示物体实例个数,$o_i$ 代表第 $i$ 个物体实例,

$$o_i=\langle n_i,\boldsymbol{c}_i,P_i\rangle。\tag{7}$$

式中:合并计数 $n_i$ 表示已合并入该物体实例的物体片段个数,向量 $\boldsymbol{c}_i=(c_{i,1},c_{i,2},\cdots,c_{i,C})^{\mathrm{T}}$ 表示该物体实例的标签置信度累计得分,$P_i$ 是该物体实例的历史观测点在世界坐标系下的坐标集合。

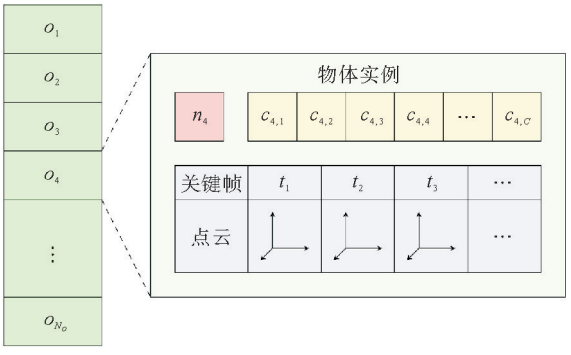



**图 3 物体地图结构**

Fig. 3 The structure of the object-oriented map

### 3.2 单帧数据与地图间的物体关联

点云分割获得的物体片段,需要与地图中已建立的物体实例进行关联。关联需要进行合并操作,若发生错误的关联则难以恢复。因此,本研究提出以下物体关联方法。

计算每个物体片段的点云 $Q_{t,i}$ 在世界坐标系下的重心$\bar{\boldsymbol{q}}_{t,i}$,与每个物体实例的点云 $P_{t,j}$ 的重心$\bar{\boldsymbol{p}}_{t,j}$ 相比较,筛选出足够近的实例作为候选实例,即重心距离满足 $\|\bar{\boldsymbol{q}}_{t,i}-\bar{\boldsymbol{p}}_{t,j}\|<d_{\mathrm{assoc}}$($d_{\mathrm{assoc}}$ 为距离阈值)。对于候选实例,由近到远依次计算点云重合度,并将第一个重合度足够大的实例标记为合并目标,即满足:

$$\frac{|\{\boldsymbol{q}\in Q_{t,i}\mid\exists\,\boldsymbol{p}\in P_{t,i},\|\boldsymbol{p}-\boldsymbol{q}\|<d_{\mathrm{point}}\}|}{|Q_{t,i}|}>r。\tag{8}$$

式中:$d_{\mathrm{point}}$ 为距离阈值,当物体片段的点到实例点云的最小距离小于阈值时,则该点能够落在实例中;$r$ 为重合比例阈值,当点的重合比例大于阈值时,表示物体片段和物体实例代表同一物体,可以进行合并。实验中,$r$ 过大容易导致物体不能被正确合并,过小则会合并错误的目标,因此选择 $r=50\%$作为重合比例阈值。

对于有合并目标的物体片段,将该片段的点云合并到目标点云中,合并计数加 1,并将标签置信度得分向量相加;对某个片段,若所有候选实例均不满足重合度条件,则认为该物体片段没有合并目标,是新观测到的物体,因此在地图中为其建立一个新的物体实例。

### 3.3 基于关联计数的地图优化方法

由于 RGB-D 传感器视野受限,单帧图像难以获取当前环境下物体的整体信息,因此物体关联过程未必能得到正确且一致的结果。同样,若当前已观测序列仍不足以推测某一物体的完整情况,则需要进行后续处理。为此,本研究提出基于关联计数的地图优化方法,以找出地图中需要合并或删除的物体。地图优化每隔一段时间进行一次,关联计数是指两次地图优化间隔中,物体实例在物体关联过程中被选为候选实例的次数。

删除实例的依据是物体实例难以被关联,因此度量标准为关联计数与最后观测时间。由于物体可能暂时离开视野,而关联计数会在每一次地图优化后清零,因此可使用合并计数代表历史关联情况。对于实例 $o_i$,计算该实例关联计数 $s_i$ 与合并计数 $n_i$ 的和,若小于地图优化间隔帧数的 1/10,则该实例需要被删除。同时,如果一个物体刚刚进入传感器视野,是最近几帧才被观测到的最新物体,那么地图中为其新建物体实例,其关联计数与合并计数一定比较小。为了避免错误删除这些物体实例,需要保留最后一次观测时间与当前帧时间戳之差小于阈值的实例。分析不同物体实例的关联次数、最后观测时间以及是否在上次优化后建立,可将实例分为三种状态,如图 4 所示。

触发地图优化过程时,对于处于“未稳定”与“新创建”状态的实例,首先检查其关联计数,若关联计数达到要求,则将实例的状态更改为“稳定”状态,否则检查其最后观测时间;若最后观测时间与当前帧时间戳之差大于阈值,则将这一实例删除,否则将其状态更改为“未稳定”状态。

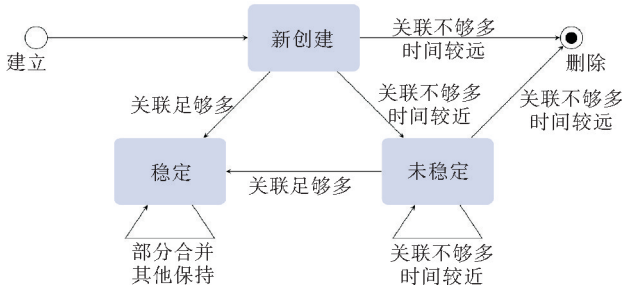



**图 4　物体实例的状态转移**

Fig. 4　The state transition of the object instances

对于处于“稳定”状态的实例,检查是否需要合并。本研究将合并实例的标准细化为四项。第一项,状态标准:要求两个实例均已处于“稳定”状态,避免将分割错误产生的实例合并,干扰实例的边界信息。第二项,共同关联计数:若两个实例的共同关联计数较多,超过任一实例的关联计数某一比例(如 70%),表示两实例容易同时成为候选实例,因此重心位置接近。第三项,几何相似度:要求两实例的点云重合度大,用于排除位置接近的不同物体。第四项,语义相似度:要求两实例标签置信度得分的分布接近,以排除两物体因表面较贴近造成点云重合度大的情况。由于检查所有实例的全部观测点的时间消耗过大,本研究首先通过前两项,即对“稳定”物体进行共同关联计数,筛选出可能合并的实例对,然后使用第三、四项进行确认。

几何相似度即点云的重合度使用式(8)衡量,而语义相似度则通过两实例标签置信度平均得分的 JS 散度(Jensen-Shannon divergence)衡量,即

$$\mathrm{JS}(\bar{\boldsymbol{c}}_i,\bar{\boldsymbol{c}}_j)=\frac{1}{2}(\mathrm{KL}(\bar{\boldsymbol{c}}_i\parallel\boldsymbol{m})+\mathrm{KL}(\bar{\boldsymbol{c}}_j\parallel\boldsymbol{m}))。\tag{9}$$

式中:$\bar{\boldsymbol{c}}_i=(\bar{c}_{i,0},\bar{c}_{i,1},\cdots,\bar{c}_{i,C})^{\mathrm{T}}=\left(\bar{c}_{i,0},\frac{1}{n_i}c_{i,1},\cdots,\frac{1}{n_i}c_{i,C}\right)^{\mathrm{T}}$ 代表实例 $o_i$ 的得分增广向量,由平均得分向量加入“其他”类别分量 $\bar{c}_{i,0}=1-\frac{1}{n_i}\sum_{k=1}^{C}c_{i,k}$ 组成,而 KL(Kullback-Leibler)定义为:

$$\mathrm{KL}(\bar{\boldsymbol{c}}_i\parallel\boldsymbol{m})=\sum_{k=0}^{C}\bar{c}_{i,k}\ln\frac{\bar{c}_{i,k}}{m_k}。\tag{10}$$

式中,$\boldsymbol{m}=\frac{1}{2}(\bar{\boldsymbol{c}}_i+\bar{\boldsymbol{c}}_j)$为两向量的平均值。

通过前两项筛选出的实例对,进一步验证第三、四项标准后,如果均达标,则进行合并操作。

### 3.4　建图方法的可行性

本研究的建图方法涉及 SLAM、点云分割、物体关联及地图优化四个过程。其中,点云分割过程使用的实例分割方法和几何聚类方法都在原文献中表现出不错的时间性能,对每帧图像的分割可以与实时运行的 SLAM 系统相匹配,因此本建图方法的时间消耗主要在物体关联和地图优化过程。

假设地图中已有的物体实例数为 $a$,当前帧中分割出的物体片段数为 $b$,物体关联过程需要把当前帧中观测到的物体片段与历史帧中出现过的所有物体实例依次进行比较,因此时间复杂度为 $O(ab)$。而地图优化过程需要将所有历史帧中的物体实例两两比较,寻找需要合并或删除的物体,因此时间复杂度为 $O(a^2)$。由于是按物体实例数目的平方级别增长,当物体实例较多时,对所占内存及运行时间都有较大需求,因此执行地图优化过程的频率不宜过高,通过 4.3 节实验部分可知,选择地图优化间隔的帧数为 50 时效果最佳。

在实际使用中,本研究方法还可以进一步优化。首先,在两次地图优化过程之间,如果物体实例没有关联更多的物体片段,则不需要进行重复比较,因此地图优化过程只需比较合并计数发生变化的物体实例即可,由此可大大减少比较的次数;其次,在计算点云重合度时,通过引入 K-D 树索引来查找每个观测点 $\boldsymbol{q}$ 的最近点,也可进一步降低算法时间。

## 4 实验

### 4.1 分割方法的实验验证

点云分割过程的实验基于公开的 Stanford 2D-3D-Semantics 数据集进行。该数据集在类别上与 YOLACT++的原训练集有一定区别,因此,为适应本方法的应用需求,需要对网络进行微调。由于该数据集中除了序列 3 之外,其他序列均含有大量与原训练集不一致的物体实例,且考虑到计算资源与训练成本,实验选择序列 3 中 20%的图像(736 张)作为评估测试集,并将该序列其他图像与序列 1、2、5a 一同作为训练数据,共计 35 270 张。训练时,仅使用 board、bookcase、sofa、chair、table、clutter 这 6 个类别的标注作为物体,而将其他各个类别,包括房屋的结构元素等一同作为背景。同时,本实验改变了原 YOLACT++检测头最后一层中的类别数,并在此使用原学习率,而将其他部分调整为 1/10 学习率,以对模型进行微调。根据收敛情况,实验选择第 10 个 epoch 后的模型作为对比用模型。

表 1 列出了定量对比的结果,对比中包括对文献[5]方法的复现。需要指出的是,常见的用于衡量实例分割性能的指标,如平均精度均值(mean average precision,mAP)等,往往基于物体标签置信度的精确率-召回率曲线设计,用于衡量所有类别分割的正确性。本研究对分割方法的改进主要是结合深度学习实例分割与几何分割,增加掩膜对物体边界的贴近程度,用于提高建图的准确性,所以并不改变标签置信度,因此该评价指标不能有效评估本研究的分割方法。本实验选择的定量评价指标为在 bookcase、sofa、chair、table 四种类别上计算出重叠权重(weighted overlap,WOv)[14],计算方法为:

$$W_{\mathrm{Ov}}=\frac{1}{\sum_i |G_i|}\sum_i |G_i|\,O_{\mathrm{v}_i},\ O_{\mathrm{v}_i}=\max_j \frac{G_i\cap S_j}{G_i\cup S_j}。\tag{11}$$

式中:$W_{\mathrm{Ov}}$ 为重叠权重,$G_i$ 表示物体的真值(ground truth)点集,$S_j$ 表示预测物体的点集。

计算结果如表 1 所示。由于文献[5]方法仅通过几何关系并结合投影的方式进行分割,对比 YOLACT++方法可知,基于深度学习方法的实例分割能够获得更好的结果。对比有无微调的 YOLACT++可知,无微调的 YOLACT++在辨认测试集中部分物体时会出现困难,而经过微调后,这一问题得到改善。

由表 1 还可看出,使用 YOLACT++时,无论有无微调,结合 Soft-NMS 与聚类的方法,与仅使用 YOLACT++方法相比,总得分均得到提高,边界分割的精确性得到提升。图 5 展示了部分分割结果。

**表 1 不同实例分割方法的 WOv 得分**

Table 1 WOv scores of instance segmentation methods

| 方法 | WOv/% |
|---|---|
| 文献[5]方法 | 20.20 |
| YOLACT++ | 47.57 |
| YOLACT++/Soft-NMS | 48.09 |
| YOLACT++/聚类 | 49.48 |
| YOLACT++/Soft-NMS/聚类 | **49.90** |
| YOLACT++(有微调) | 52.84 |
| YOLACT++(有微调)/Soft-NMS | 52.93 |
| YOLACT++(有微调)/聚类 | 54.47 |
| YOLACT++(有微调)/Soft-NMS/聚类 | **54.64** |

(a) 原图1 (b) YOLACT++结果1 (c) 本研究方法结果1 (d) 原图2 (e) YOLACT++结果2 (f) 本研究方法结果2

**图 5 不同方法的分割结果**

Fig. 5 Segmentation results of different methods

由图 5 可以看出,对于形状棱角分明的物体,本研究的分割方法比直接使用 YOLACT++更有优势,能够得到更加接近物体真实边界的分割结果,如图 5(c)中的椅子、图 5(f)中的桌子等。这表明与深度学习的结果相比,超体素划分的边界更加符合真实的颜色、深度变化。

然而,对于较为扁平、边界几何特征不强的物体,超体素划分则难以准确地捕捉边界。图 6(a)～6(d)展示了应用于该类数据时的部分结果。由于超体素分割使用固定间隔选取体素的种子并聚类,聚类结果的边界来源于超体素本身的边界,若几何特征不明显,将会获得相邻体素间的不规则组合,这样的结果在图像上会投影为不规则片段,其边界多为锯齿状。如图 6(b)中台灯(蓝色框)与白板(黄色框)所在区域的超体素划分情况,白板区域最终会产生锯齿状边界,如图 6(d)所示。

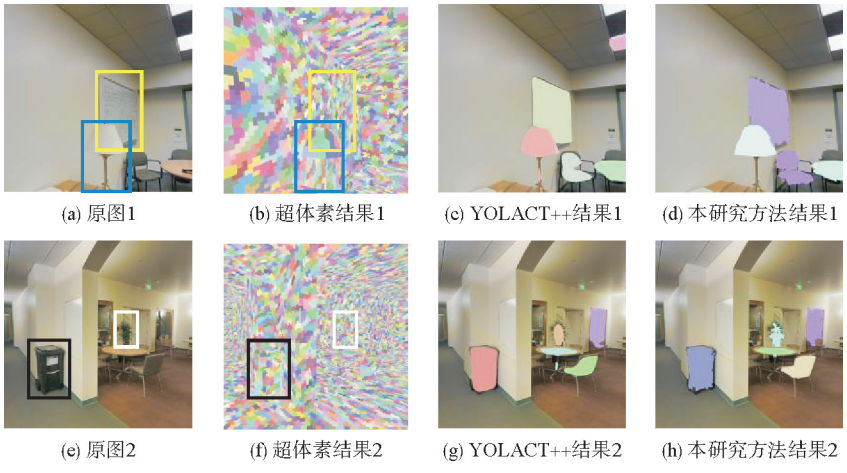



**图 6　物体边界几何特征对超体素分割步骤的影响**

Fig. 6　Effects of the geometric characteristic of the object boundaries on the supervoxel segmentation step

同时,对于绿植而言,由于几何方法和深度学习均难以确定其边界,结合后同样不够精确。如图 6(f)中白色框区域所示,超体素无法正确划分绿植的边界,因此在图 6(h)中分割出的边界不够精确。

### 4.2　地图优化的实验验证

为验证地图优化过程的有效性,本实验从 TUM-RGBD[15] 的 freiburg2/desk 序列中连续抽取 104 帧 RGB-D 图像作为测试数据,并在第 50 帧、第 100 帧各触发一次地图优化过程。此外,由于 SLAM 的结果具有一定的随机性,需要重复多次实验并统计数据。实验中,通过对优化过程中合并、删除操作的不同启用,对比所建立的地图中物体实例数目的变化,结果如图 7 所示。图 7 表明,地图优化过程的合并和删除操作均会减少地图中物体实例的数目,且当二者同时启用时,实例数目的减少更加明显。

为了证明实例数的减少对物体地图质量提升的有效性,表 2 列出了每种操作启用时 3 次实验中部分类别的实例数目。其中,将系统能够成功分割出的物体作为预期数据,与建图结果中的物体进行比较。可以看出,未经优化的地图中含有的实例数目超过预期数目较多,而使用优化后实例数目更接近预期。此外,地图优化中合并、删除操作均会使实例数目减少,但不会出现少于预期的情况,表明地图中实例数减少是因为实例数据更贴近真实的物体数据。

### 4.3　地图优化间隔选择的实验验证

本实验验证不同的地图优化间隔对地图质量的影响,使用的数据仍然是 4.2 节中的 104 帧图像,地图优化间隔的帧数分别选择 10、20、33、50、100。多次重复实验后统计结果如图 8 所示。

图 8 表明,地图优化的间隔对建图结果的影响并不明显,尤其当间隔大于 33 帧时,建立的地图差别较小。当间隔小于 33 帧时,建图结果有偏离真实数据的趋势。这是因为间隔过短,导致删除操作的阈值较小,使得部分的错误识别物体因关联计数大于该值而未能被删除。

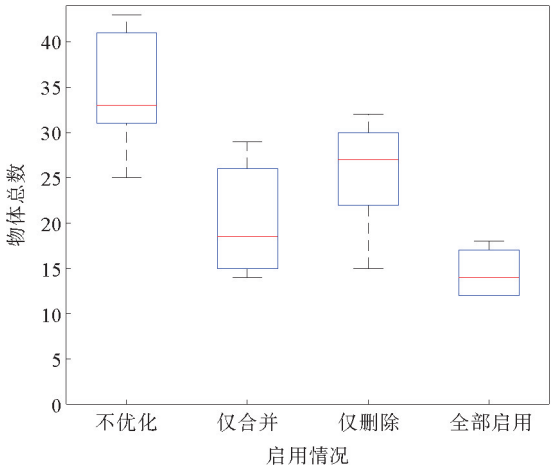



**图 7 地图优化过程对建图实例数目的影响**

Fig. 7 Effects of map optimizing process on the number of the mapped instances

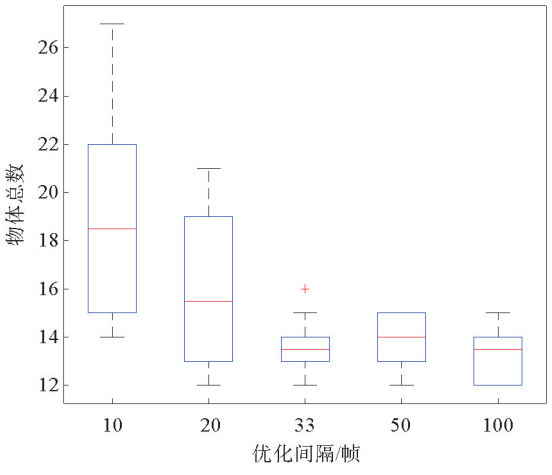



**图 8 地图优化间隔对建图实例数目的影响**

Fig. 8 Effects of the interval of map optimizing process on the number of the mapped instances

为进一步证明实例数目的差异主要来源于地图优化过程,图 9 列出了每种间隔下,3 次建图过程中实例数目的变化过程。从图 9 可以看出,优化前的地图中确实存在重复、错误的新物体实例,而根据本研究提出的优化方法,这些实例已被合并或删除。

当帧数累积到 85 帧前后时,图 9 中各序列均出现了物体数目突然增加的情况。事实上,这一帧相机位姿误差较大,因此导致了部分小物体重复建立。此外,33 帧以上的间隔都在第 100 帧(33 帧为第 99 帧)处发生物体数量减少的情况,而 10 和 20 帧的间隔没有这一情况,这说明更大的间隔更容易检查出可能存在错误的数据。

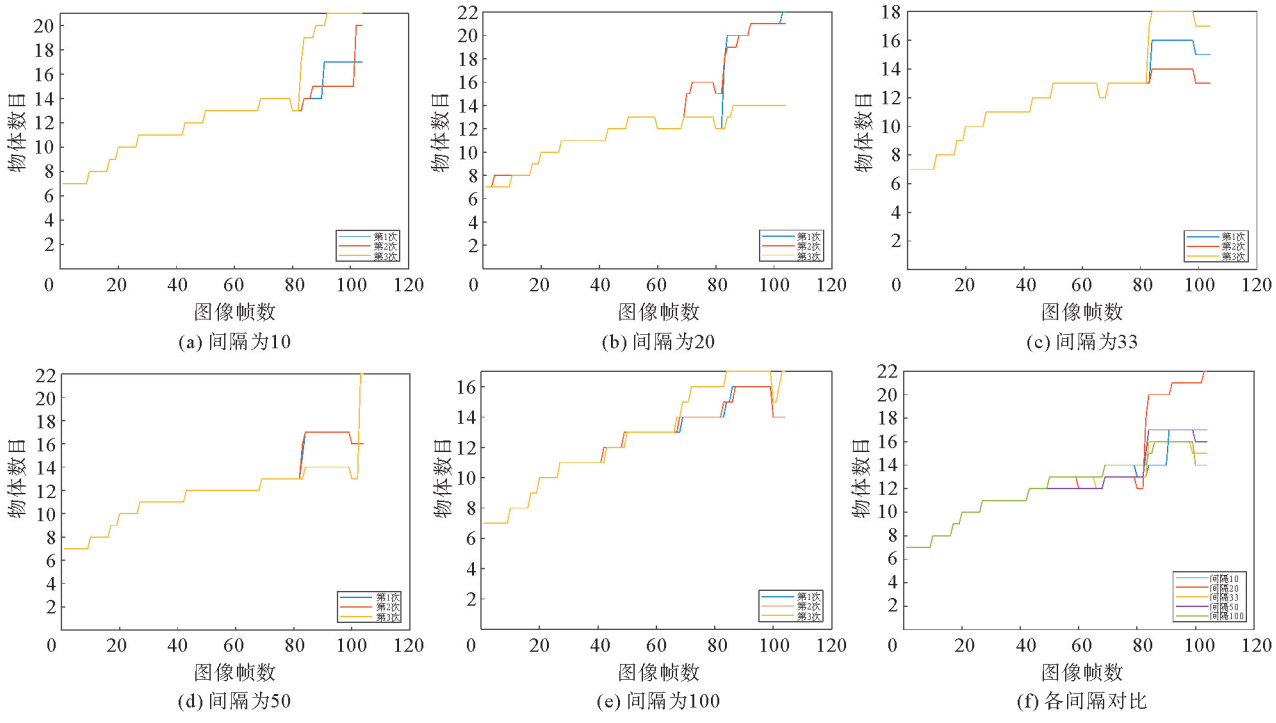
**表 2 地图优化过程对建图中典型分类实例数目的影响**

Table 2 Effects of map optimizing process on the number of the mapped instances of typical categories

| 启用情况 | 典型类别实例数目 | | | | | 物体总数 |
|---|---|---|---|---|---|---|
| | 碗 | 杯子 | 餐桌 | 盆栽 | 电视 | |
| 不优化 | 4 | 4 | 4 | 5 | 3 | 31 |
| | 5 | 5 | 5 | 6 | 3 | 34 |
| | 6 | 5 | 6 | 5 | 5 | 43 |
| 仅合并 | 4 | 3 | 2 | 3 | 4 | 26 |
| | 2 | 2 | 1 | 2 | 3 | 16 |
| | 5 | 3 | 2 | 3 | 3 | 29 |
| 仅删除 | 5 | 4 | 3 | 4 | 3 | 31 |
| | 5 | 4 | 3 | 4 | 3 | 32 |
| | 2 | 2 | 2 | 3 | 1 | 15 |
| 全部启用 | 3 | 2 | 1 | 2 | 2 | 18 |
| | 3 | 3 | 1 | 2 | 2 | 17 |
| | 2 | 2 | 1 | 1 | 1 | 12 |
| 预期 | 2 | 2 | 1 | 1 | 1 | — |

### 4.4 建图过程的实验验证

图 10 展示了本研究方法与文献[5]方法在建图结果上的对比。由图 10 可以看出,本研究方法建立的地图在实例的观测点、物体标签两方面比文献[5]方法均有提升。首先,本研究的分割方法能够获得更加精确的边界,减少了地面观测点被误标为物体的情况。因此,对重心等数据的计算更加准确,进一步提升了物体关联的准确性。其次,地图的优化过程合并了被重复建立的实例,如图 10(a)中建立了位置接近的 3 个键盘实例,且出现了含有便携电脑标签的错误实例,而图 10(b)中键盘位置被正确地建立为 1 个实例。且不含有标注为便携电脑的错误实例。比较建图结果表明,本研究提出的建图方法是可行的。




(a) 间隔为10　(b) 间隔为20　(c) 间隔为33

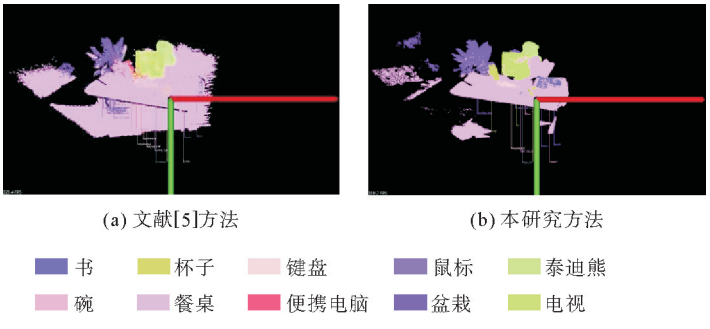
(d) 间隔为50　(e) 间隔为100　(f) 各间隔对比

**图 9　不同地图优化间隔下实例数目的变化过程**

Fig. 9 The variation of the number of the mapped instances with optimizing process at different intervals

## 5　总结

本研究提出一种面向物体的语义建图方法。首先设计了由 YOLACT++ 指导超体素聚类的点云实例分割方法，得到了具有语义信息并贴合几何边界的分割结果；然后增加了对地图中物体实例进行后续优化的过程，弥补了物体地图中建立物体标签后缺少纠正的问题。实验表明，本研究提出的方法能够从物体实例数、物体的边界范围等方面提升所建地图的可靠程度。




(a) 文献[5]方法　(b) 本研究方法

**图 10　建图结果示例**

Fig. 10　An example of map building results

本研究依赖超体素边界确定物体边界，但现有方法对几何特征不明显的区域鲁棒性较差，难以提供准确的物体边界。因此，未来工作将根据物体类别标签，调整几何信息的重要性，使几何边界更准确时重要性更高，反之则需要抑制几何信息对结果的影响，由此提升分割边界与物体真实边界的贴合性，提升所建地图的可靠性。此外，将地图信息反馈给 SLAM 过程，提升相机位姿估计的准确性，也能进一步提升地图的质量。

**参考文献：**

[1] CHEN X, MILIOTO A, PALAZZOLO E, et al. SuMa++: Efficient LiDAR-based semantic SLAM[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Macau: IEEE, 2019: 4530-4537.

[2] BEHLEY J, STACHNISS C. Efficient surfel-based SLAM using 3D laser range data in urban environments [C]//Robotics: Science and Systems XIV. Pittsburgh: Robotics: Science and Systems Foundation, 2018: 147-156.

[3] MILIOTO A, VIZZO I, BEHLEY J, et al. RangeNet++: Fast and accurate LiDAR semantic segmentation[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Macau: IEEE, 2019: 4213-4220.

[4] LU F X,PENG H T,WU H Y,et al. InstanceFusion:Real-time instance-level 3D reconstruction using a single RGBD camera[J]. Computer Graphics Forum,2020,39(7):433-445.

[5] SÜNDERHAUF N,PHAM T T,LATIF Y,et al. Meaningful maps with object-oriented semantic mapping[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Vancouver:IEEE,2017:5079-5085.

[6] 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.

[7] NAKAJIMA Y,SAITO H. Efficient object-oriented semantic mapping with object detector[J]. IEEE Access,2019,7:3206-3213.

[8] JU Q,LIU F F,LI G C,et al. Semantic map generation algorithm combined with YOLOv5[C]//2021 International Conference on Computer Engineering and Application(ICCEA). Kunming:IEEE,2021:7-10.

[9] GRINVALD M,FURRER F,NOVKOVIC T,et al. Volumetric instance-aware semantic mapping and 3D object discovery [J]. IEEE Robotics and Automation Letters,2019,4(3):3037-3044.

[10] FURRER F,NOVKOVIC T,FEHR M,et al. Incremental object database:Building 3D models from multiple partial observations[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Madrid:IEEE,2018:6835-6842.

[11] XU B B,LI W B,TZOUMANIKAS D,et al. MID-fusion:Octree-based object-level multi-instance dynamic SLAM[C]//2019 International Conference on Robotics and Automation(ICRA). Montreal:IEEE,2019:5231-5237.

[12] BOLYA D,ZHOU C,XIAO F,et al. YOLACT++:Better real-time instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,44(2):1108-1121.

[13] BODLA N,SINGH B,CHELLAPPA R,et al. Soft-NMS:Improving object detection with one line of code[C]//2017 IEEE International Conference on Computer Vision(ICCV). Venice:IEEE,2017:5562-5570.

[14] STEIN S C,SCHOELER M,PAPON J,et al. Object partitioning using local convexity[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus:IEEE,2014:304-311.

[15] 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. Vilamoura:IEEE,2012:573-580.

(责任编辑:傅 游)