浙江理工大学学报，2022，47(5)：764-773
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2022.05.017

基于语义信息与动态特征点剔除的SLAM算法

潘海鹏，刘培敏，马　淼
（浙江理工大学 机械与自动控制学院，杭州 310018）

摘　要：传统的同时定位与地图构建(Simultaneous localization and mapping，SLAM)算法在现实场景中易受动态物体及背景的影响，针对该问题提出了一种将语义分割与动态特征点剔除相结合的动态SLAM算法，以实现动态场景地图的构建。首先，根据多层通道注意力和空间注意力机制，构造特征融合网络MulAttenNet(Multilayer attention network)，并进行语义分割，剔除场景中运动概率大的物体，粗略估计相机位姿；其次，根据相机位姿和深度信息剔除动态区域；最后，利用剔除后的特征点进行地图的构建。对MulAttenNet网络和动态SLAM算法进行实验，以验证算法的有效性，实验结果表明：该算法构造的MulAttenNet网络能有效提高语义分割的准确性，平均像素准确度提高4.05%，均交并比提高2.60%；将该算法构建的动态SLAM算法与现有SLAM算法相比，建图的绝对位姿误差和相对位姿误差都有所缩小。该算法能在动态场景下构建高精度的语义地图。

关键词：同时定位与地图构建；动态环境；动态特征点剔除；注意力机制；损失函数
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673-3851 (2022) 09-0764-10

SLAM algorithm based on semantic information and the elimination of dynamic feature points

PAN Haipeng, *LIU Peimin*, *MA Miao*
(School of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Traditional simultaneous localization and mapping (SLAM) algorithms are easily influenced by dynamic objects and their backgrounds in real scenes. In this paper, a dynamic SLAM algorithm that combines semantic segmentation techniques with dynamic feature point rejection has been proposed to establish dynamic scene maps. Firstly, the feature fusion network MulAttenNet based on multi-channel attention and spatial attention was built to perform semantic segmentation by using RGB information, which eliminated objects with high motion probability in the scene and roughly estimated camera poses. Secondly, dynamic areas were eliminated according to camera poses and depth information. Finally, the map was established according to the static feature points. In order to verify the effectiveness of the proposed algorithm in this paper, validation experiments were conducted for the established MulAttenNet network and dynamic SLAM algorithm respectively. The experimental results have shown that the network built in this paper could effectively improve the accuracy of semantic segmentation, the MPA value was increased by 4.05% and the MIoU value was increased by 2.60%. In addition, compared with the existing SLAM algorithms, the dynamic SLAM algorithm established by this algorithm reduced the

收稿日期：2022－03－24　　网络出版日期：2022－06－02
基金项目：浙江省自然科学基金项目(LQ19F030014)
作者简介：潘海鹏(1965－　)，男，河南濮阳人，教授，硕士，主要从事智能信息处理方面的研究。
通信作者：马　淼，E-mail：mamiao@zstu.edu.cn

absolute pose errors and the relative pose errors. The algorithm in this paper can construct highly accurate semantic maps in dynamic scenes.

Key words: simultaneous localization and mapping (SLAM); dynamic environment; dynamic feature point elimination; attention mechanism; loss function

0 引　言

同时定位与地图构建(Simultaneous localization and mapping, SLAM)[1]是指机器人从未知环境的未知位置出发,在运动过程中通过观测到的地图特征识别并定位自身位置,再根据自身的姿态和轨迹构建出完整的场景地图。目前使用在SLAM算法上的传感器主要分为激光雷达和摄像头两大类,它们的核心都是获取颜色信息和深度信息。近年来,视觉 SLAM 领域出现了许多优秀算法,例如:LSD-SLAM[2]利用深度滤波器获取像素点的深度,能够直接通过图像获得准确的位姿,构建大尺度的半稠密地图;ORB-SLAM2[3]提取 ORB 特征点,并在不同图像中寻找特征匹配,利用匹配点的信息计算相机的位姿,从而实现特征点跟踪、局部建图和回环检测。

上述 SLAM 算法都假设机器人所处的环境为静止状态,因此无法对复杂且变化的真实场景进行识别,尤其是当场景中出现移动的物体时,定位、建图的精度和准确度会大大降低。Newcombe 等[4]提出的 DynamicFusion 算法,是较早能够实时处理的动态 SLAM 算法,它通过几何变化不断更新建立的模型,减少动态物体对建图的影响,使得获得的场景逐渐趋于真实场景。Bescos 等[5]提出了 Dyna-SLAM 算法,利用深度信息的一致性剔除由动态物体运动而产生的动态特征点,并在 ORB-SLAM2 的基础上结合实例分割和多视图几何关系提取动态区域,利用前后几帧的投影对动态物体引起的遮挡背景进行填充,以解决场景中的动态干扰。Chen 等[6]提出了 Suma++算法,将语义分割的结果投影至激光数据形成的球形面,为每一帧提取的地图点添加标签,通过检查当前帧与关联帧中语义标签的一致性对动态物体进行过滤。Zhang 等[7]提出的 VDO-SLAM 算法,是使用光流法进行动态特征点剔除的典型方法,利用光流联合估计来估计相机位姿和运动,最大化跟踪运动物体上的点,提取出运动物体模型,再结合光流和语义标签分配目标物体唯一标识符,以保证持续跟踪、计算位姿。Wimbauer 等[8]提出了 MonoRec 算法,使用深度 MVS 和单目深度估计算法计算静态和动态物体的准确深度,减少动态物体对系统的影响。

现有的动态 SLAM 算法通常利用光流(如 VDO-SLAM[7])或者深度不确定方法(如 Dyna-SLAM[5]),进行动态特征点的剔除。光流法通过计算特征点在前后两帧图像二维平面中运动的情况来区分动态特征点和静态特征点;深度不确定法将关键帧特征点投影到相邻关键帧中,计算该特征点在相邻关键帧中的深度和投影深度差,判断其是否运动。当场景中物体只在坐标的单一维度(如只在 z 轴方向移动)或两个维度范围内(如只在 xy 轴平面中移动)运动时,上述方法无法有效识别动态点,因此结合运动物体的位姿及深度变化,设计切实可行的剔除动态物体的算法,可以减少物体运动对系统的影响,有效解决现有 SLAM 算法难以剔除动态物体、难以识别场景中物体的缺点,从而提高 SLAM 算法构建地图的准确性和有效性。

本文首先引入多层全局和局部注意力机制,构建了一种多尺度的特征融合网络 MulAttenNet (Multilayer attention network),更有针对性地提取并融合全局特征图和局部特征图;然后提出了针对不同维度的动态物体进行几何剔除的方法,通过计算运动物体的位姿及深度变化,有效剔除场景中的动态特征点,并根据区域中动态点的数量判断区域是否为动态;最后在上述基础上搭建了动态 SLAM 算法,以识别场景中的动态区域并剔除,减少动态对象的影响。

1 动态语义 SLAM 算法

1.1 系统框架

本文在 ORB-SLAM2 模块的基础上增加了语义分割模块和动态剔除模块,构建了动态 SLAM 算法,其系统框架如图 1 所示。

本文算法使用 ORB-SLAM2 模块进行特征提取和匹配,语义分割网络用于识别输入图像中的物体,动态剔除模块用于判断特征点是否运动,两者结合可以剔除场景中的动态物体,减少动态物体的遮挡和干扰。通过场景静态物体中的静态特征点进行相机的位姿估计,实现场景的重建。具体步骤如下:

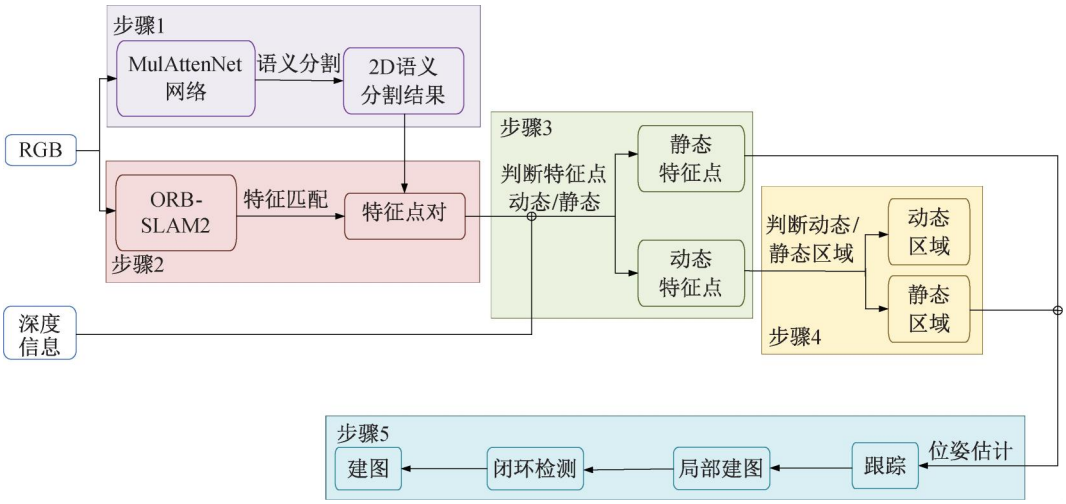


图 1 动态语义 SLAM 算法系统框架

步骤 1:构建 MulAttenNet 网络对机器人采集的 RGB 信息进行语义分割,识别环境中的物体,并根据先验知识剔除场景中运动概率大的物体,如人、车、猫等。

步骤 2:利用 RGB 信息提取 ORB 特征点并进行匹配,粗略计算相机位姿,利用相机位姿进行空间点云分割,并构建点云的空间关系八叉树,生成用于描述未知场景的三维点云图;进而,将场景的二维语义信息与三维点云图相结合,使语义分割的结果对应到点云的相应空间位置,得到带有语义信息的三维点云图。

步骤 3:结合相机位姿和深度信息,计算从参考帧到当前帧的运动变换和深度变化,得到每个特征点的动态概率,进而判断提取到的特征点是否运动,并剔除动态概率高的特征点。

步骤 4:计算语义分割的区域范围内动态特征点的比例,判断物体区域是否为动态,剔除运动物体,减少由运动物体遮挡对建图的影响。

步骤 5:最后利用静态特征点的匹配进行位姿估计,并建立局部地图和优化位姿,实现静态物体的建图。

1.2 语义分割模块

现有的语义分割网络需要有足够的深度才能有较好的学习能力,无法在不同尺寸物体识别中同时保持较好的性能。为解决上述问题,本文构造了基于多层通道注意力和空间注意力的融合网络 MulAttenNet,MulAttenNet 网络框架如图 2 所示。该网络采用一种多级特征融合网络,结合了多层局部和全局注意力机制[9],可以加权处理以强化重要的通道特征和空间特征,挖掘浅层和深层特征关联信息,显著提高网络特征提取的性能[10]。

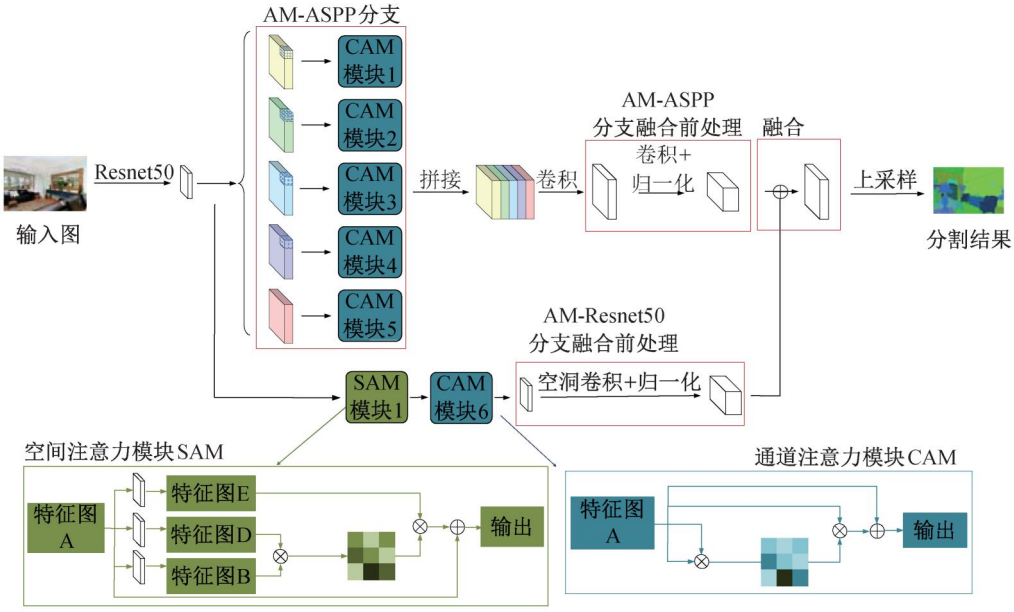


图 2 MulAttenNet 网络框架

MulAttenNet 网络使用 Resnet50 作为基础特征提取网络,搭建带有多层空间注意力机制(Spatial attention module, SAM)和通道注意力机制(Channel attention module,CAM)的 AM-Resnet50 提取全局特征图,搭建带有通道注意力机制的空间金字塔 AM-ASPP 提取不同尺度的特征图,最后将 AM-Resnet50 分支和 AM-ASPP 分支融合后得到最终的输出。

空间注意力的结构如图 3 所示,空间注意力通过主干网络得到的局部特征图 $\boldsymbol{A}(C,H,W)$,C、H 和 W 分别表示为特征图 $\boldsymbol{A}$ 的长、宽、通道数。主干网络后接一个卷积层,得到同尺寸的特征图 $\boldsymbol{B}$、$\boldsymbol{D}$、$\boldsymbol{E}$。改变 $\boldsymbol{B}$ 的形状为 (C,N),其中 $N=H\times W$,同时将 $\boldsymbol{B}'$ 转置为 $\boldsymbol{B}^{\mathrm{T}}(N,C)$,然后改变 $\boldsymbol{D}$ 的形状为 $\boldsymbol{D}'(C,N)$,将 $\boldsymbol{D}'$ 与 $\boldsymbol{B}^{\mathrm{T}}$ 做矩阵乘法,最后使用 Softmax 层计算出空间注意力矩阵 $\boldsymbol{S}(N,N)$,改变 $\boldsymbol{E}$ 的形状 $\boldsymbol{E}'(C,N)$,并与 $\boldsymbol{S}$ 相乘得到(C,N),乘以系数 α 并改变形状,与 $\boldsymbol{A}$ 相加后得到最终的输出。

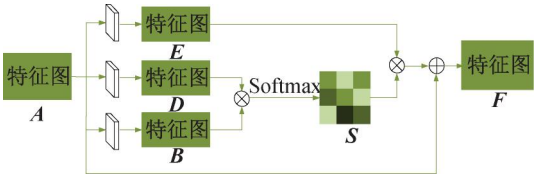


图 3　空间注意力结构

通道注意力的结构如图 4 所示,通道注意力通过主干网络得到的局部特征图 $\boldsymbol{A}$,$\boldsymbol{A}$ 的尺寸为 $\boldsymbol{A}(C,H,W)$,改变 $\boldsymbol{A}$ 的形状为 $\boldsymbol{A}'(C,N)$,同时将 $\boldsymbol{A}$ 转置得到 $\boldsymbol{A}^{\mathrm{T}}$,将 $\boldsymbol{A}'$ 与 $\boldsymbol{A}^{\mathrm{T}}$ 做矩阵乘法并通过 Softmax 层得到通道注意力矩阵 $\boldsymbol{X}(C,C)$,与 $\boldsymbol{A}$ 相乘后乘以系数 β,调整形状后与 $\boldsymbol{A}$ 相加后得到最终的输出。

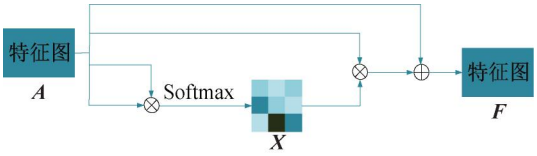


图 4　通道注意力结构

AM-Resnet50 分支带有空间注意力,使得这一分支能够自适应寻找网络中最重要的部位并处理,保留关键信息;同时带有通道注意力,合理分配 AM-Resnet50 分支与 AM-ASPP 各个卷积通道之间的资源,调整最终输出对全局特征图的依赖程度。AM-ASPP 分支采用一个卷积层、三个不同扩张率的空洞卷积[11]和一个全局平均池化层对特征图进行处理,五个并行的卷积模块都串有一个通道注意力,挖掘不同尺度的卷积特征的同时,对不同的图像通道采用不同的关注程度;将五个不同尺度的特征在通道维度串联在一起,得到 AM-ASPP 分支的特征图。AM-ResNet50 和 AM-ASPP 分支分别利用不同比例的真值标签进行训练,并将两分支特征图以对应通道相加的方式进行融合,不同尺度特征图的特征元素赋予不同 Mask 权重,保留浅层和深层特征信息,能够更有针对性地提取关键特征。

MulAttenNet 网络的不同分支处理的是多分类问题,因此设计带有权重的 Softmax 交叉熵损失函数(Softmax cross entropy loss function)[12]进行损失计算,实现对每一个分支的有效训练。损失函数 $\mathcal{L}$ 的计算公式为:

$$\mathcal{L}=-\sum_{t\in\{1,2\}}\lambda_t\frac{1}{x_t y_t}\sum_{x=1}^{x_t}\sum_{y=1}^{y_t}\log\frac{\mathrm{e}^{\mathcal{F}^t_{x,y,\hat{n}}}}{\sum_{n=1}^{\mathcal{N}}\mathrm{e}^{\mathcal{F}^t_{x,y,n}}}\tag{1}$$

其中:t 表示分支索引号,λ_t 表示 AM-ResNet50 和 AM-ASPP 两个分支损失的权重,$\mathcal{N}$ 表示类别数($\mathcal{N}=150$,本文实验中为 ADE20k 数据集[13]分类类别数)。在分支 t 中,预测的特征图 $\mathcal{F}^t$ 的大小为 $x_t\times y_t\times\mathcal{N}$,其中特征图第三维对应 $\mathcal{N}$ 个分类类别。特征图 $\mathcal{F}^t$ 中位置(x,y,n)的值为 $\mathcal{F}^t_{x,y,n}$,表示分支 t 的特征图在平面位置(x,y)处属于类别 n 的概率。$\hat{n}$ 表示标签真值,$\mathcal{F}^t_{x,y,\hat{n}}$ 为位置(x,y)的输出标签与标签真值一致的概率。

1.3　动态剔除模块

为解决 ORB-SLAM2 在动态场景下丢失目标以及现有动态 SLAM 算法无法剔除全部动态物体的问题,本文设计了动态剔除模块,针对不同维度的特征点进行几何剔除。通过计算特征点在当前帧和投影帧位置及深度的变化,有效剔除动态特征点,结合语义分割的结果,能够很好地识别出动态的区域并剔除,减少动态物体对建图的影响。

将关键帧中的特征点 $\boldsymbol{X}_{\mathrm{KF}}$ 根据相机的相对位姿投影到当前帧中,特征点投影方法如图 5 所示,获得匹配点 $\boldsymbol{X}_{\mathrm{Cur}}$ 在当前帧中的投影位置 $\boldsymbol{X}_{\mathrm{proj}}(x_{\mathrm{proj}},y_{\mathrm{proj}})$ 和深度 z_{proj},再根据当前帧 RGB 图及深度图获得该匹配点在当前帧实际测量的位置$(x_{\mathrm{Cur}},y_{\mathrm{Cur}})$和深度 z_{Cur},计算投影位置、深度与实际测量的位置、深度差,计算公式为:

$$\begin{cases}\Delta d=\sqrt{(x_{\mathrm{proj}}-x_{\mathrm{Cur}})^2+(y_{\mathrm{proj}}-y_{\mathrm{Cur}})^2}\\ \Delta z=\left|z_{\mathrm{proj}}-z_{\mathrm{Cur}}\right|\end{cases}\tag{2}$$

其中:Δd 为关键帧中的特征点在当前帧的投影及

实际测量的位置差，Δz 为深度差。Δd 或 Δz 大于阈值时，则认为该特征点在这两帧中运动，为动态特征点，否则为静态特征点。当位置差的阈值设置为 0.12，深度差的的阈值设置为 0.7 时，实验中对动态物体的剔除效果最好，获得的轨迹误差最小，因此，将 Δd 的阈值 τ_d 设置为 0.12，Δz 的阈值 τ_z 设置为 0.7。

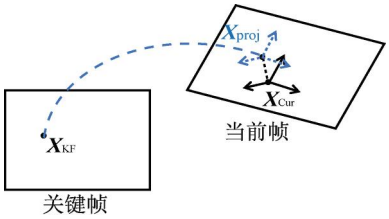


图 5 特征点投影

将计算的特征点与语义分割的结果结合起来，每个分割区域包含特征点集 $\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\cdots,\boldsymbol{X}_n\}$，若该分割区域中 30%的特征点为动态特征点，则认为该物体为运动物体，在后续计算中将该动态区域剔除。

2 实验结果与讨论

为验证本文搭建的语义分割网络的有效性，在公开数据集 ADE20k[13] 上对本文所搭建的网络进行分割实验验证，并与其他三个前沿语义分割网络 FCN 网络[14]、PSPNet 网络[15] 和 DeepLabv3 网络[16] 进行对比实验。ADE20k 数据集是 MITComputer Vision 团队发布的最大的用于语义分割和场景解析的开源数据集，其有 150 个生活中常见的室内室外场景类别，每幅图像均对应一幅由灰度图表示的语义信息，且其中稀疏注释了场景中的物体。

为验证本文提出的动态语义 SLAM 算法的有效性，使用 TUM RGB-D 数据集[17] 中两个数据包 fr2_desk_with_person(以下简称为场景 fr2)和 fr3_walking_xyz(以下简称为场景 fr3)进行系统性能的评估，并与 ORB-SLAM2、Dyna-SLAM 和 VDO-SLAM 算法进行对比实验。TUM RGB-D 由德国慕尼黑工业大学 TUM 计算机视觉组于 2012 年发布，是目前应用最为广泛的 RGB-D 数据集，主要为纹理丰富的办公室场景和仓库场景，数据集中含有彩色信息和由 Microsoft Kinect 传感器采集的深度图像。

本文的实验在 CPU 为 Intel TM Core i5-9400F @2.6 Hz，GPU NVIDIA GeForce RTX2060 的计算机，Ubuntu16.04 操作系统上进行。

2.1 语义分割算法有效性验证

本文使用平均像素准确度(Mean pixel accuracy，MPA)和均交并比(Mean intersection over union，MIoU)[18] 作为评价指标，对搭建的 MulAttenNet 网络进行有效性验证，并与其他前沿语义分割网络进行对比。各语义分割算法的性能指标对比见表 1 所示，语义分割对比如图 6 所示。

表 1 语义分割性能指标

语义分割算法	MPA 值/%	MIoU 值/%
FCN	65.40	34.00
PSPNet	78.01	41.68
DeepLabv3	79.97	45.66
MulAttenNet	84.02[a]	48.28[a]

注：标注 a 的数值为最优值。

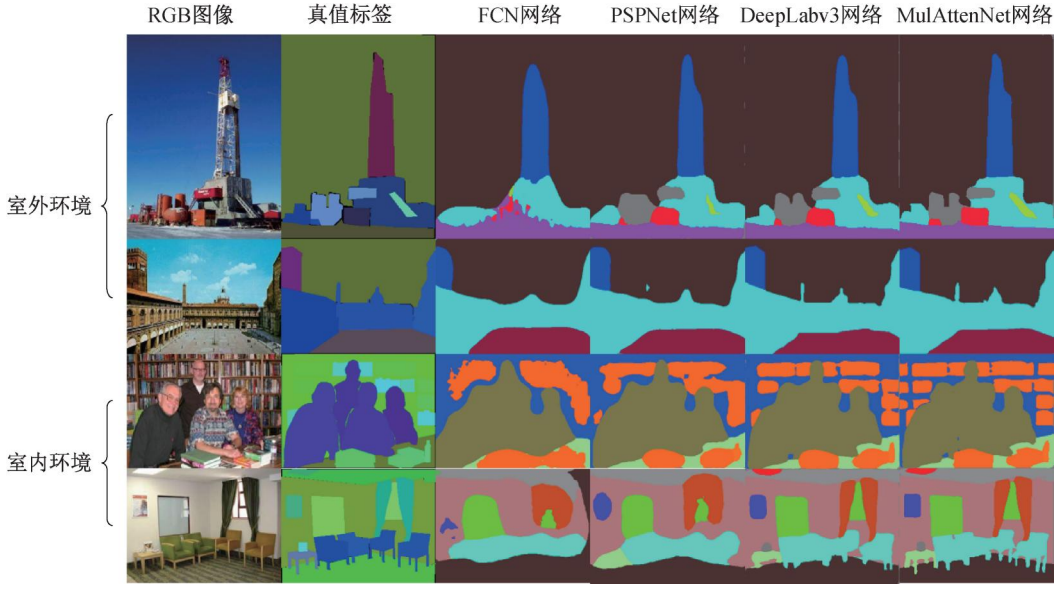


图 6 语义分割结果示例

从表 1 可以看出,本文的 MulAttenNet 网络与其他语义分割网络相比,MPA 值及 MIoU 值都有所提升。其中,FCN 网络的 MPA 值及 MIoU 值分别为 65.40% 和 34.00%,与比本文网络 MulAttenNet 相比分别低了 18.62% 和 14.28%,FCN 网络池化过程丢失部分细节信息,导致最后的分割结果不够精细;而本文网络采取了多种上采样方法,同时使用双线性插值和多种扩张率的空洞卷积,使得分割的结果不会丢失过多信息,更加注重对细节的处理。PSPNet 网络经过多次下采样,且解码过程简单恢复图像尺寸,降低了分割的质量;本文的 MulAttenNet 网络结合了空洞卷积可在不丢失分辨率的情况下扩大卷积核的感受野,使得本文网络优于 PSPNet 网络,与 PSPNet 网络(MPA 值及 MIoU 值分别为 78.01%、41.68%)相比,MulAttenNet 网络的 MPA 值及 MIoU 值分别高了 6.01%和 6.60%。DeepLabv3 搭建了空间金字塔来进行特征的提取,但是没有挖掘局部特征图和全局特征图之间的关联性,导致无法兼顾不同尺寸物体的分割;而 MulAttenNet 网络在此基础上搭建了带有注意力机制的两分支融合结构,使得训练过程能够聚焦重点、增加关联性,并且 MulAttenNet 网络对两个分支采用不同比例的标签进行训练和融合。上述操作使得本文的 MulAttenNet 网络比 DeepLabv3 网络(MPA 值为 79.97%、MIoU 值为 45.66%)的 MPA 值及 MIoU 值分别高了 4.05%和 2.62%。通过上述分析,本文搭建的 MulAttenNet 网络能有效提高语义分割的精度和准确度。

不同语义分割算法对室外及室内环境的语义分割结果如图 6 所示,其中第 1 至 6 列分别为 RGB 图像、真值标签(Ground truth)、FCN 网络、PSPNet 网络、DeepLabv3 网络和本文网络的分割结果。由图 6 可以看出:FCN 网络的分割结果比较粗糙,存在混乱的错误色块;PSPNet 网络没有出现错误色块,但分割的结果仍然不够清晰,小尺寸的物体被忽略;DeepLabv3 网络相较于 FCN 网络和 PSPNet 网络有了较大的提升,大尺寸的物体边缘分割清晰,但小尺寸物体交错在一起,无明显边缘;而本文搭建的 MulAttenNet 网络对不同尺寸的物体都能准确分割,能更好地处理细节信息,边缘清晰。由此可见,本文搭建的 MulAttenNet 网络在各个方面的表现均优于其他网络。

2.2　动态语义 SLAM 算法有效性验证及分析

本文使用轨迹对比、绝对位姿误差(Absolute pose error,APE)和相对位姿误差(Relative pose error,RPE)[19]作为评价指标,对构建的动态语义 SLAM 算法进行验证实验。轨迹对比用于表示估计位姿和真实位姿的直接差值,可以直观地反应算法精度和轨迹全局一致性。绝对位姿误差是估计位姿和真实位姿差异的精度,用于评价 SLAM 轨迹的整体一致性。相对位姿误差主要用于描述相隔固定时间差的两帧位姿相比真实位姿差的差值,相当于直接测量里程计的误差,适合于估计系统的漂移。图 7—图 10 为使用本文算法对两个不同场景的处理过程示例,以及使用本文算法及不同算法对两个场景进行建图的轨迹和精度对比情况,表 2 列出了使用本文算法及对比算法对两个场景进行建图的评价指标数据对比情况。

图 7 展示了本文算法对场景 fr2 的处理过程中具有代表性的 10 帧图像的实时分割结果和动态剔除情况,第一行至第三行分别为视频序列图像、实时分割结果和动态剔除情况,第三行中黑色阴影部分为剔除的动态物体。由图 7 可知,本文搭建的语义分割网络 MulAttenNet 对场景中大部分物体都能准确分割,包括场景中细长的杆子、混乱的书桌、边缘交错的植物等,但对场景中距离摄像机较远,且光线过亮的墙上的物体无法正确识别。场景中大部分动态物体能准确剔除,如第 1084 帧、第 1835 帧、第 2125 帧中移动的人,第 2663 帧中移动的人和书,第 3863 帧中移动的人和键盘。但对于场景部分帧中静止的动态物体无法剔除,如第 136 帧中的腿和第 4067 帧中的人,由于该物体在前后帧中都未发生移动,系统将其识别为静止物体。

图 8 为 ORB-SLAM2[4]、Dyna-SLAM[5]、VDO-SLAM[6]以及本文算法分别对场景 fr2 的建图轨迹和精度对比情况,其中图 8(a)—8(c)分别为轨迹对比、绝对位姿误差和相对位姿误差,第一列至第四列分别为 ORB-SLAM2、Dyna-SLAM、VDO-SLAM 和本文算法的建图结果。从图 8 可以看出,由于 ORB-SLAM2 无法识别动态物体并剔除,建图结果存在较大的错误,并且丢失部分动态物体所占比例过大的帧,建图结果不完整;Dyna-SLAM 和 VDO-SLAM 相较于 ORB-SLAM2 有了较大的提升,但是对无法剔除所有的动态物体,丢失了部分帧的建图结果;本文算法对场景中绝大部分帧都能准确识别,对剔除动态物体的场景建图结果准确且完整,轨迹的准确度和误差较其他 SLAM 算法有较大提升。

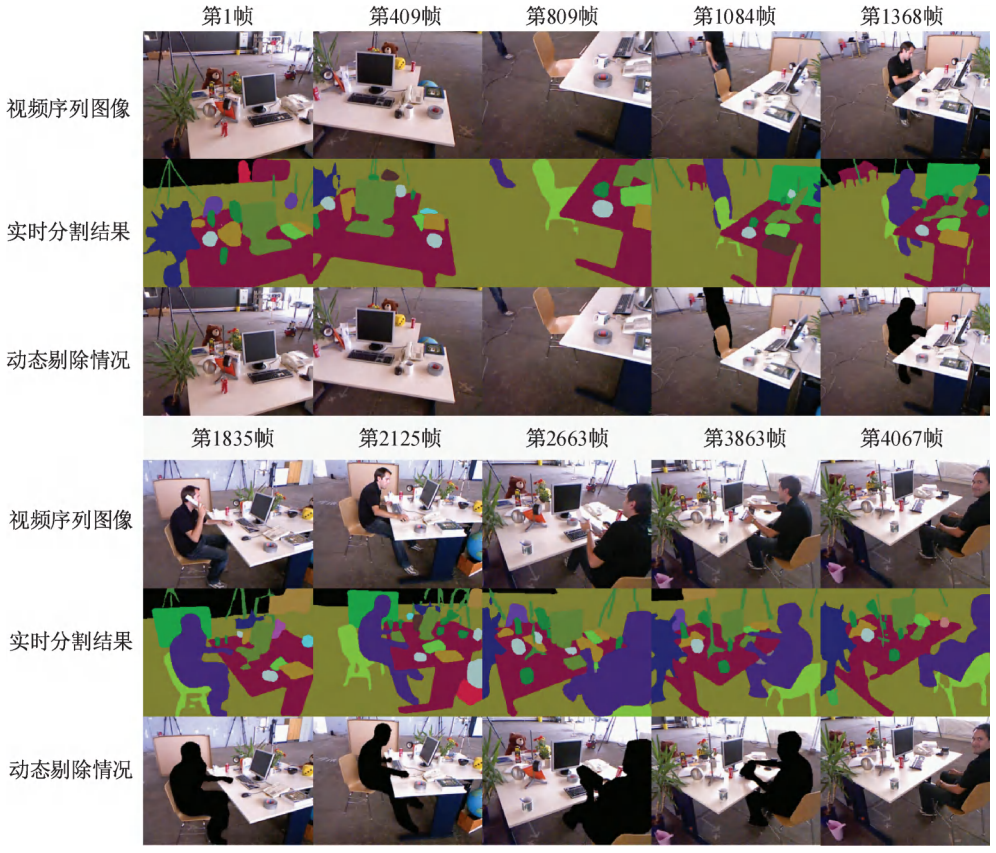
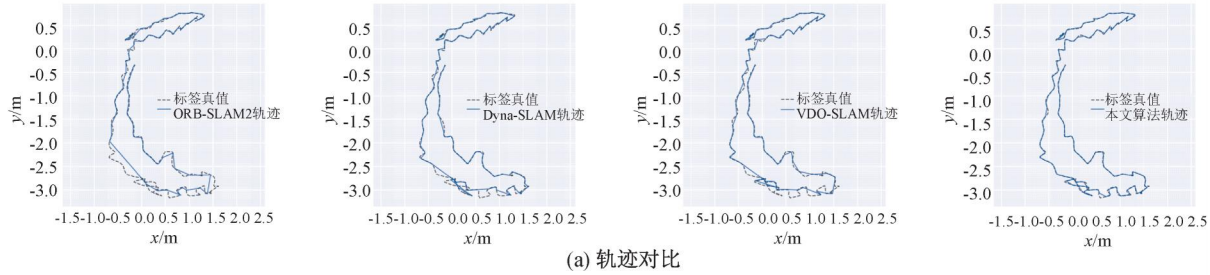
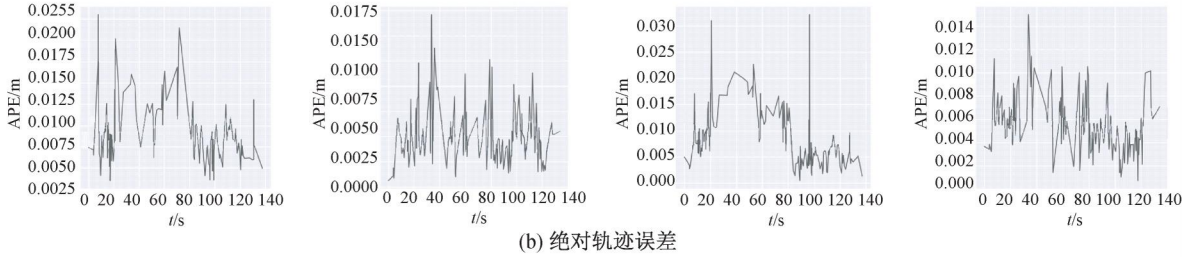


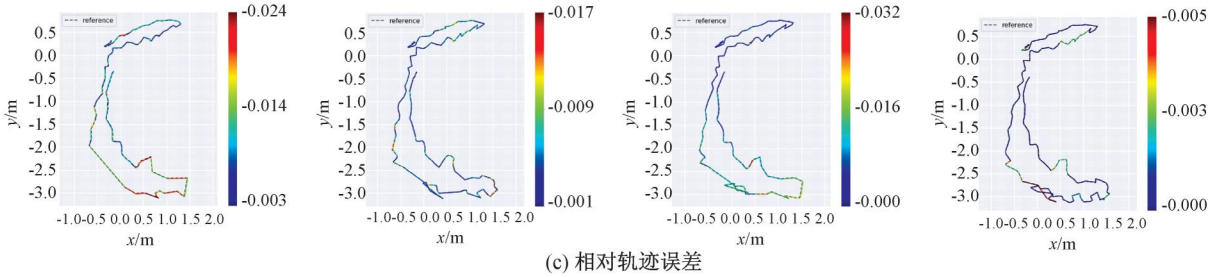
图 7 本文算法对 fr2 的处理过程示例



(a) 轨迹对比



(b) 绝对轨迹误差



(c) 相对轨迹误差

图 8 不同算法对 fr2 的建图轨迹和精度

图 9 展示了本文算法对场景 fr3 的处理过程示例。由图 9 可知,本文算法能准确分割场景 fr3 中的物体,边缘清晰准确,可准确分割堆满办公物品的桌面、混乱的墙壁和天花板、细长交错的杆子等。对于场景中频繁移动的物体,包括第 362 帧中体积所占比例很小的人头,本文算法都能较好地剔除,且未出现与人体靠近的物体,如椅子、办公用品等的错误剔除。

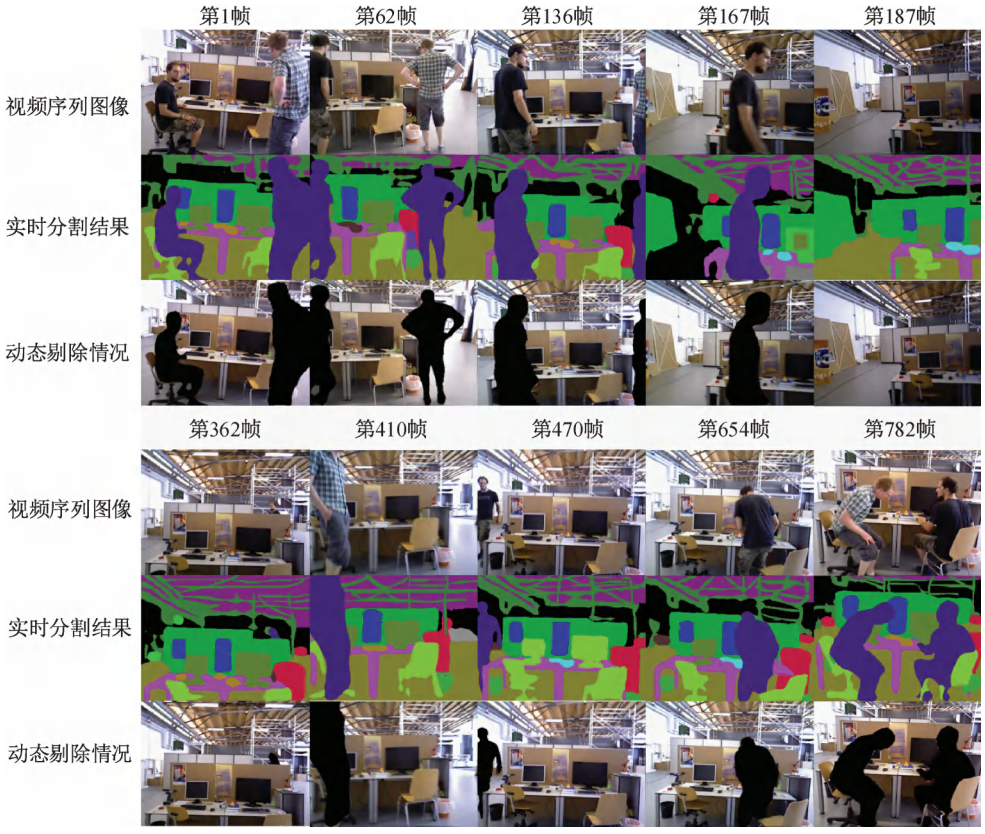


图 9　不同算法对 fr3 的建图轨迹和精度

图 10 为 ORB-SLAM2、Dyna-SLAM、VDO-SLAM 以及本文算法,分别对场景 fr3 的建图轨迹和精度对比。由于场景中绝大多数帧中都出现了动态物体,且动态物体在每帧图像中所占比例较大,ORB-SLAM2 出现了严重的建图错误,丢失了部分帧中的轨迹信息,建图结果不完整,Dyna-SLAM 和 VDO-SLAM 也出现了部分帧轨迹的偏移,而本文算法对场景 fr3 的建图结果仍然完整准确,未出现轨迹的偏移,建图的误差明显缩小。

表 2 为不同算法对场景 fr2、fr3 建图结果的评价指标,由表 2 可知:由于无法剔除动态物体,ORB-SLAM2 的建图精度最低,对有较多动态物体的场景识别效果较差,且丢失部分场景信息,地图不完整;Dyna-SLAM 和 VDO-SLAM 相较于 ORB-SLAM2 有了一些提升,未出现较大的错误,地图包含的信息比较完整,但是由于无法剔除场景中所有的动态物体,对建图精度的提升有一定限制;本文算法的建图精度最高,误差最小,绝对位姿误差 APE 值相较于前面三种算法分别缩小了 0.014815、0.00502、0.005804,相对位姿误差 PRE 值分别缩小了 0.013199、0.007793、0.008415,从而证明建图精度提升显著。

表 2　建图结果评价指标

场景	ORB-SLAM2		Dyna-SLAM		VDO-SLAM		本文算法	
	APE	PRE	APE	PRE	APE	PRE	APE	PRE
fr2	0.009594	0.009986	0.004489	0.009337	0.007243	0.009622	0.000198[a]	0.000828[a]
fr3	0.026171	0.024020	0.011677	0.013858	0.010500	0.014817	0.005938[a]	0.006781[a]

注:标注 a 的数值为最优值。

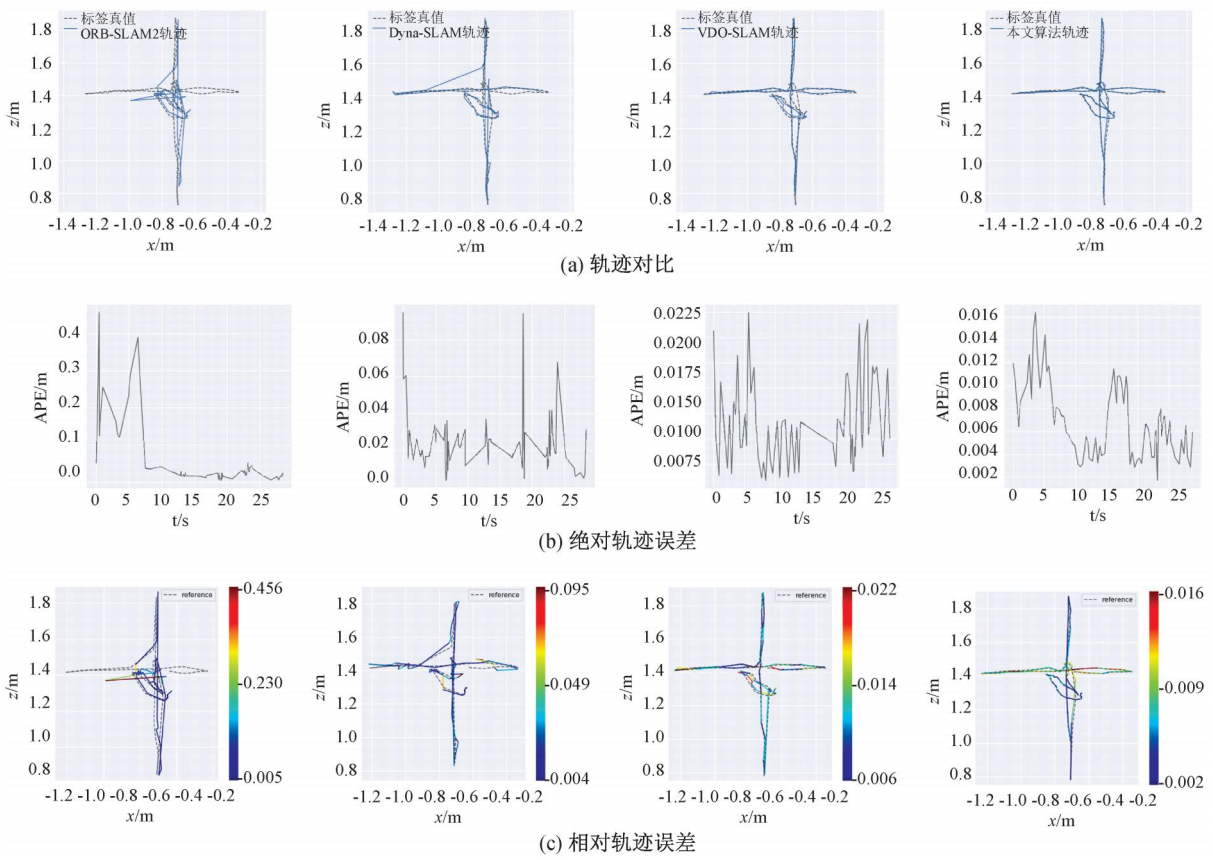


图 10 不同算法对 fr3 的建图轨迹和精度

3 结 论

针对 SLAM 算法中无法识别不同尺度物体,以及无法消除场景中动态物体影响等问题,本文提出了基于语义信息与动态特征点剔除的 SLAM 算法。本算法构造了 MulAttenNet 网络识别场景中物体,通过有效结合 ORB-SLAM2 模块并提出动态特征点剔除法,将影响建图准确性的动态特征点剔除,有效排除场景中动态物体的干扰,提高了 SLAM 算法在动态环境下的建图准确性。MulAttenNet 网络在 ADE20k 数据集上的实验结果表明,本文提出的语义分割算法的语义分割准确性有所提升;动态 SLAM 算法在 TUM RGB-D 数据集上的实验结果表明,SLAM 算法的建图效果较其他算法有较大提升。

参考文献:

[1] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1052-1067.

[2] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C]//European Conference on Computer Vision (ECCV 2014). Cham: Springer, 2014: 834-849.

[3] Mur-Artal R, Tardós 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.

[4] Newcombe R A, Fox D, Seitz S M. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 343-352.

[5] Bescos B, Fácil J M, Civera J, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes [J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4076-4083.

[6] 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, China: IEEE, 2019: 4530-4537.

[7] Zhang J, Henein M, Mahony R, et al. VDO-SLAM: A visual dynamic object-aware SLAM system[EB/OL]. (2020-5-22) [2022-03-24]. https://arxiv.org/abs/2005.11052v2.

[8] Wimbauer F, Yang N, Von Stumberg L, et al. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021: 6108-6118.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017: 5998-6008.

[10] 邓远远, 沈炜. 基于注意力反馈机制的深度图像标注模型[J]. 浙江理工大学学报(自然科学版), 2019, 41(2): 208-216.

[11] Ma J J, Dai Y P, Tan Y P. Atrous convolutions spatial pyramid network for crowd counting and density estimation[J]. Neurocomputing, 2019, 350: 91-101.

[12] Sudre C H, Li W, Vercauteren T, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations [C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS 2017, Québec City, QC, Canada. Cham: Springer, 2017: 240-248.

[13] Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017: 633-641.

[14] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 3431-3440.

[15] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 6230-6239.

[16] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV 2018), Glasgow, United Kingdom. Cham: Springer, 2018: 833-851.

[17] Schubert D, Goll T, Demmel N, et al. The TUM VI benchmark for evaluating visual-inertial odometry[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018: 1680-1687.

[18] Liu L J, Zhao G P, Bo Y M. Point cloud based relative pose estimation of a satellite in close range [J]. Sensors, 2016, 16(6): 824.

[19] 王涛, 张恩政, 刘翠苹, 等. 基于改进神经网络的机器人逆解与轨迹精度提高方法[J]. 浙江理工大学学报(自然科学版), 2021, 45(5): 624-632.

(责任编辑:康　锋)