智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于特征融合及动态背景去除的室内机器人语义VI-SLAM

王立鹏, 王小晨, 齐尧, 张佳鹏

引用本文:
王立鹏, 王小晨, 齐尧, 等. 基于特征融合及动态背景去除的室内机器人语义VI-SLAM[J]. 智能系统学报, 2024, 19(6): 1438-1448.
WANG Lipeng, WANG Xiaochen, QI Yao, et al. Indoor robot semantic VI-SLAM based on feature fusion and dynamic background removal[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1438-1448.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202309025

DOI: 10.11992/tis.202309025
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240910.1119.004

基于特征融合及动态背景去除的室内机器人语义 VI-SLAM

王立鹏, 王小晨, 齐尧, 张佳鹏
(哈尔滨工程大学 智能科学与工程学院, 黑龙江 哈尔滨 150001)

摘　要: 为提升室内机器人在动态场景中的定位精度, 同时构建细节丰富的三维语义地图, 提出一种基于特征融合及动态背景去除的室内机器人语义 VI-SLAM (visual-inertial simultaneous localization and mapping) 算法。首先, 改进 ORB-SLAM3 算法框架, 设计一种可以实时构建三维稠密点云地图的 VI-SLAM 算法; 其次, 将目标识别算法 YOLOv5 与 VI-SLAM 算法融合, 获取二维语义信息, 结合二维语义信息与极线约束原理去除动态特征; 再次, 将二维语义信息映射为三维语义标签, 将语义特征与点云特征相融合, 构建三维语义地图; 最后, 基于公开数据集及移动机器人平台, 在动态场景下开展三维语义地图构建实验。实验结果验证了提出的该语义 VI-SLAM 算法在动态环境下定位与建图的可行性和有效性。

关键词: 室内机器人; VI-SLAM; 特征动态去除; 语义地图; 特征融合; 稠密点云; 点云分割; 动态场景
中图分类号: TP242.6　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)06−1438−11

中文引用格式: 王立鹏, 王小晨, 齐尧, 等. 基于特征融合及动态背景去除的室内机器人语义 VI-SLAM[J]. 智能系统学报, 2024, 19(6): 1438–1448.
英文引用格式: WANG Lipeng, WANG Xiaochen, QI Yao, et al. Indoor robot semantic VI-SLAM based on feature fusion and dynamic background removal[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1438–1448.

Indoor robot semantic VI-SLAM based on feature fusion and dynamic background removal

WANG Lipeng, WANG Xiaochen, QI Yao, ZHANG Jiapeng
(College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: An indoor robot semantic VI-SLAM algorithm based on feature fusion and dynamic background removal is proposed to improve the positioning accuracy of indoor robots in dynamic scenes and build a three-dimensional (3D) semantic map with rich details. The framework of the ORB-SLAM3 algorithm is improved, and a VI-SLAM algorithm for real-time construction of 3D dense point cloud maps is designed. The algorithm fuses target recognition algorithms YOLOv5 and VI-SLAM to obtain two-dimensional (2D) semantic information. Dynamic features are then removed by combining the 2D semantic information with the epipolar constraint principle. Subsequently, the 2D semantic information is mapped into a 3D semantic tag, constructing a 3D semantic map by fusing the semantic features with the point-cloud features. Finally, experiments in 3D semantic map construction were conducted in indoor scenes using public data sets and a mobile robot platform. Results verify the feasibility and effectiveness of the semantic VI-SLAM algorithm in dynamic environments.

Keywords: indoor robot; VI-SLAM; feature dynamic removing; semantic map; feature fusion.; dense point cloud; point cloud segmentation; dynamic scene

收稿日期: 2023−09−13.　**网络出版日期:** 2024−09−10.
基金项目: 黑龙江省教育科学规划 2023 年度重点课题 (GJB1423059); 国家自然科学基金项目 (62173103); 黑龙江省自然科学基金项目 (LH2024F037); 中央高校基本科研业务费专项 (3072024XX0403).
通信作者: 王立鹏. E-mail:wanglipeng@hrbeu.edu.cn.

机器人已深入到生活中各个场景, 同步定位与建图 (simultaneous localization and mapping, SLAM) 是机器人的最基本功能, 使机器人可在复杂的环境中具有高效率。目前传统 SLAM 方法

在依赖外部环境的纯视觉方案下, 定位和建图精度容易受到光照和纹理等条件的影响。为了应对这些挑战, 通过视觉传感器与 IMU (intertial measurement unit) 传感器耦合的 VI-SLAM 方案适用范围更广, 但机器人无法理解周围环境, 始终是难以攻克的难题。为此, 语义 VI-SLAM 应运而生, 通过融合语义信息, 使机器人不仅能构建地图和定位, 还能识别环境中的物体及其关系, 提升了机器人的环境理解能力和智能化水平。

语义 SLAM 是移动机器人研究的核心技术之一, 许多研究人员重点研究语义地图构建, 同时跟踪地图中语义对象[1], 例如 Whelan 等[2] 将实例分割算法与 RGB-D SLAM 算法相结合, 基于 ElasticFusion 修改匹配目标函数, 构建室内环境精确语义地图。Mccormac 等[3] 在 ElasticFusion 的基础上引入了卷积神经网络(convolutional neural networks, CNN), 能够在帧与帧之间建立更加精确的对应关系, 从而提高了系统在复杂环境中的数据关联和匹配精度, 此外, 将多视角的语义预测方法巧妙地融合到地图构建过程中。 Wang 等[4] 提出了 QISO-SLAM, 该方法整合了数据关联、单帧椭球初始化和其他过程, 以建立高级的 3D 地图。Salas-Moreno 等[5] 提出了 SLAM++算法, 生成物体的三维模型, 将其应用于实例级面向对象的 3D-SLAM 姿态图优化算法框架中。Dame 等[6] 在 SLAM 算法运行过程中, 使用特定对象的知识来构建精确的语义地图。

对于动态环境中的 SLAM 问题, 许多学者结合语义信息来删除环境中的动态目标, 以此提高定位精度。例如, Yu 等[7] 提出 DS-SLAM, 结合语义分割网络 SegNet 与运动特征点检测来滤除每一帧中动态物体。Kaneko 等[8] 提出的 Mask-SLAM, 通过语义分割的结果来识别动态特征点的属性标签, 直接排除某些区域的特征点。Wang 等[9] 利用语义分割将一些类别区域定义为背景, 并将其他类别区域定义为可移动物体。Bescos 等[10-11] 提出的 DynaSLAM, 对 RGB-D 输入图像的动态点做了细致的处理, 提升多对象跟踪能力。Li 等[12] 提出的 DP-SLAM, 结合了几何约束和语义分割来跟踪贝叶斯概率估计框架中的动态特征点。

有些学者致力于研究如何提高语义 SLAM 实时性, 例如 Zhao 等[13] 提出了 KSF-SLAM, 通过一种关键分割帧的选择策略, 提升了 SLAM 系统的实时性。Hu 等[14] 提出的 DeepLabv3 SLAM, 利用改进的 DeepLabv3(+) 语义分割网络, 以此减少动态目标的检测时间。Huang 等[15] 利用在线三维重建过程中高效融合多视图二维特征和投影到超体素上的三维特征, 构建了基于超体素的卷积神经网络, 称为 Supervoxel-CNN。Mccormac 等[16] 提出了 Fusion++, 具有任意重构对象的 3D 图, 之后通过基于 Mask-RCNN 的深度融合逐步细化了物体的分割[17]。还有些学者研究提高绝对效率和更大环境的可伸缩性, 如 Nakajima 和 Saito 通过使用快速和可扩展的对象检测, 实现了高精度的面向对象场景实时重建, 该算法减少了计算成本和内存占用[18]。Pham 等[19] 使用高效的超体素聚类与基于结构和物体线索的高阶约束的条件随机场(conditional random fields, CRF)对三维室内场景进行实时密集重建和语义分割。

在语义 SLAM 与深度学习结合方面, Tateno 等[20] 提出的 CNN-SLAM, 使用卷积神经网络 CNN 进行深度预测, 解决了单目 SLAM 在位姿估计和环境重建中缺少绝对尺度的问题。Clark 等[21] 提出了视觉惯性网 VINet(visual-inertial net), 利用 CNN 和 RNN(recurrent neural network)构建了一个视觉惯性网络 VIO(visual-inertial odometry), 直接输出估计的位姿结果。Detone 等[22] 提出 Deep SLAM, 利用 CNN端对端的学习位姿然后完成 SLAM 中特征点的提取与匹配, 后来提出 Superpoint, 通过直接学习特征点与描述子来实现特征提取与匹配。Chen 等[23]将 DCNNs 和全连接的 CRFs 结合起来, 形成了一个端到端的系统, 叫作 DeepLab, CRFs 可以对 DCNNs 的预测结果进行优化, 使得分割结果更加平滑和精确。语义 SLAM 中闭环检测也是一个较大问题, Song 等[24] 利用语义信息进行 SLAM 数据关联与闭环检测的表检索方法。Qian 等[25] 提出的 SmSLAM+LCD 方法通过将高级 3D 语义信息与低级特征信息相结合, 显著提升了系统的性能。这种融合策略不仅实现了更为准确的闭环检测, 还有效地抑制了长期运行中的漂移问题。

即使视觉 SLAM 和语义 SLAM 研究成果较多, 但目前仍存在一些问题: 1)单帧图像特征点较少, 增加机器人 SLAM 的匹配难度; 2)视觉 SLAM 在运行过程中容易将运动物体错误地识别为地图上的静态点, 致使位姿估计误差过大; 3)语义信息与图像信息结合过程的配准和实时性问题。为解决以上问题, 本研究提出一种基于深度学习和稠密点云处理的语义 VI-SLAM 方法, 通过改进 ORB-SLAM3 算法, 添加稠密建图线程并通过点云拼接实现稠密点云地图的构建, 利用深度学习 YOLO v5s目标检测网络实时获取关键帧中

物体的类别及其位置，基于语义信息及极线约束去除环境中的动态特征点，提升系统在动态场景下的定位精度；采用点云分割的方法对稠密点云地图进行分割，然后将语义特征与分割后的点云特征融合，完成三维语义地图构建。

1　三维稠密点云地图构建

本研究通过在 ORB-SLAM3 系统原有线程的基础上增加稠密建图线程，构建稠密点云地图，利用相机成像原理，将关键帧各像素点结合其对应的深度信息映射到三维空间形成单帧稠密点云。假设图像中某一个点的像素坐标是(u,v)，其深度值为d，在三维空间中的坐标(x,y,z)的计算式为

$$\begin{cases} x=\dfrac{1}{f_x}(u-c_x)z \\ y=\dfrac{1}{f_y}(v-c_y)z \\ z=d \end{cases} \tag{1}$$

式中：f_x、f_y、c_x、c_y均为相机的内参。得到单帧稠密点云数据后，计算每个点与其邻近点之间平均距离，将离群点滤除，建立三维体素栅格，用体素中所有点的重心近似地表示体素上的全部点，进一步滤除外点及异常点，形成单帧稠密点云。

结合 ORB-SLAM3 算法中该关键帧对应的相机位姿估计结果，将所有点转换到世界坐标系中，便可建立出原始的三维稠密点云地图。假设第i个和第j个关键帧生成滤波后的稠密点云分别为C_{loud_i}和C_{loud_j}，对应的相机位姿估计分别为T_i和T_j，通过下式将 2 个关键帧点云转换到世界坐标系下：

$$C'_{\text{loud}_i}=T_i^{-1}C_{\text{loud}_i} \tag{2}$$

$$C'_{\text{loud}_j}=T_j^{-1}C_{\text{loud}_j} \tag{3}$$

按照下式拼接得到全局的点云数据：

$$C^{*}_{\text{loud}}=C'_{\text{loud}_i}+C'_{\text{loud}_j} \tag{4}$$

在拼接过程中要将各个单帧点云与该关键帧进行数据关联，局部 BA 与闭环校正后会更新关键帧的位姿估计结果，此时需要将该帧对应的点云数据从全局稠密点云地图中进行删除，并结合优化后的位姿再次转换并拼接，避免全局稠密点云图中出现重影，获得了全局稠密点云地图后再次通过体素滤波进行降采样，得到基于 ORB-SLAM3 的稠密点云地图，单帧和拼接后点云结果如图 1 所示。

(a) 单帧点云

(b) 拼接后稠密点云

图 1　三维点云图

Fig. 1　3D point cloud map

2　基于目标检测与极线约束的动态特征筛选

本研究基于 YOLOv5s 网络模型提取关键帧中的语义信息，结合语义信息与极线约束对关键帧中的动态特征点进行去除，以提升本研究算法在动态场景下的性能，考虑到 YOLOv5s 网络目标识别技术较为成熟，获得目标语义信息的过程本研究不做赘述。去除动态特征点流程如图 2 所示。

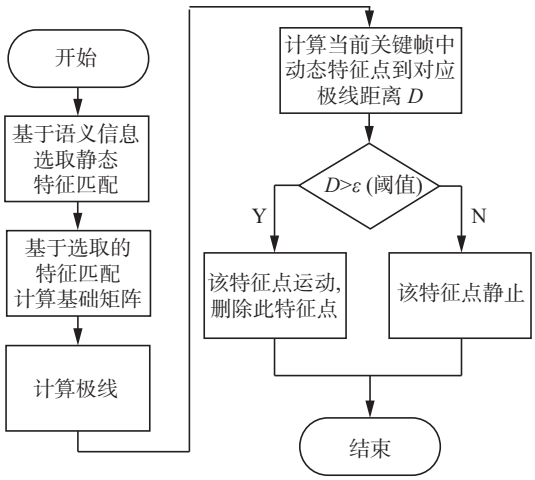


图 2　动态特征点去除流程

Fig. 2　Dynamic feature point removal flow chart

通过目标检测所获得的语义信息仅能代表该目标运动的可能性，为此本研究设计基于语义信息的动态特征初筛选策略为

$$R = P \wedge \neg Q \tag{5}$$

式中: P、Q、R均为二值逻辑的命题变量, P为特征点处于动态物体框内, Q为特征点处于静态物体框内, R为判定为动态特征点; $\wedge$为逻辑与, $\neg$为逻辑非。

通过式 (5) 的初步筛选方法, 仅当该特征点位于动态物体框且不在静态物体框内时, 将其判定为动态特征点, 能够减少部分特征点静动属性的误判断, 但是仍有部分特征点被误判断。本研究通过光流法获取匹配的特征点对, 并根据匹配的特征点对计算基础矩阵, 通过基础矩阵和特征点计算特征点对应的极线, 当特征点到极线的距离大于一定值, 即判断为动态特征点。假设一对匹配特征点$\boldsymbol{P}_1$和$\boldsymbol{P}_2$的齐次坐标为$\boldsymbol{P}_1 = [u_1, v_1, 1]$, $\boldsymbol{P}_2 = [u_2, v_2, 1]$。其中$u$和$v$为像素横纵坐标, 根据基础矩阵 $\boldsymbol{F}$, 前一个关键帧中特征点$\boldsymbol{P}_1$与当前关键帧中极线l的关系可表示为下式

$$l = \boldsymbol{F}\boldsymbol{P}_1 = [A \quad B \quad C]^{\mathrm{T}} \tag{6}$$

式中: A、B、C为三维空间中极线l的方程系数, 当前关键帧上匹配的特征点$\boldsymbol{P}_2$, 其到极线的距离D为

$$D = \frac{\left|\boldsymbol{P}_2^{\mathrm{T}}\boldsymbol{F}\boldsymbol{P}_1\right|}{\sqrt{\|A\|^2 + \|B\|^2}} \tag{7}$$

若式 (7) 计算的距离D大于阈值ε, 认定该点为正在发生运动的动态特征点, 加入到需要去除的动态特征点集合, 予以去除。

本研究选择动态场景中相邻的 2 个关键帧, 经过基于语义信息初步筛选后, 对每一个动态特征点开展本研究提出的极线约束动态特征检测, 检测及动态特征去除结果如图 3 所示。图 3(a)中, 红色特征点和绿色特征点分别为识别到的动态特征点和静态特征点, 紫色线框中特征点为极线约束回收的动态特征点, 最终去除结果如图 3(b)所示, 图 3(a)到图 3(b)行人发生了一定的位置移动, 部分回收特征点被遮挡。

(a) 极线约束回收部分静态特征点

(b) 动态特征点去除结果

图 3　动态特征点去除结果

Fig. 3　Dynamic feature point removal results

3　基于语义信息与点云分割的三维语义地图构建

本研究将结合 VI-SLAM 算法获取的二维语义信息与稠密点云地图构建线程, 将其映射到三维空间形成三维语义标签, 利用空间点云的颜色属性保存其对应的语义数据, 并与三维稠密点云地图的分割结果相融合, 进而构建出三维语义地图。

3.1　三维语义标签构建方法

通过训练后的 YOLOv5s 网络模型对关键帧进行检测, 获得关键帧中物体类别与其在图像中的位置, 完成对环境的二维语义信息提取, 将关键帧对应的 RGB 图中该检测框内所有像素点颜色对应的 RGB 值, 更改为三维语义标签对应的颜色, 结合关键帧对应的位姿与深度图进行稠密点云地图构建, 便可将二维语义信息转换为三维语义标签。

3.2　基于聚类及最小割的点云分割算法

基于超体素聚类的分割方法, 对原始的稠密点云地图开展分割, 根据点的相似性, 将无规则的点云转换为面结构, 经过超体素分割之后形成的曲面都有一个质心与一个法向量, 点云分割可被定义为最小割问题, 采用最小割的方法进行进一步分割。

对本研究构建的稠密点云地图进行超体素聚类分割, 设置 Rvoxel=0.008(体素大小), Rseed=0.05 m(种子分辨率), 根据实验室环境调整颜色、空间距离和法向量所占的权重, 获得的超体素聚类分割结果的邻接图。由于设置的种子是同时进行生长的, 最后的分割结果中会造成一个物体被分割成多个区域, 为了解决这种问题, 将对超体素聚类分割的结果基于表面块的几何信息进行最

小割，以获得更为精确的物体目标点云。

使用 RANSAC 算法处理曲面块以生成候选的实际平面 $\boldsymbol{P}_{\mathrm{C}}=\{p_{\mathrm{c}_1},p_{\mathrm{c}_2},\cdots,p_{\mathrm{c}_m}\}$，计算 $d(c_i,p_{\mathrm{c}_m})$，即各个曲面块质心 c_i 到候选平面 p_{c_m} 的距离，增加一个阈值 δ，用于筛选所有到平面 p_{c_m} 距离在 δ 内的曲面块，将满足条件的曲面块视为集合 $\Pi=\{v_i\in V|d(c_i,p_{\mathrm{c}_m})<\delta\}$。定义

$$D(p_{\mathrm{c}_m})=\begin{cases}1-\dfrac{\Pi}{\eta}, & \Pi<\eta\\ \exp\left(\dfrac{\Pi}{\eta}\right), & \Pi\geqslant\eta\end{cases}\tag{8}$$

式中：η 表示区分前景物体与背景的约束条件，可以理解为一个合格的实际平面应该至少有 η 个曲面块满足到该平面的距离小于 δ；$D(p_{\mathrm{c}_m})$ 为平面 p_{c_m} 是否为实际平面的权重，值越大，代表到该平面距离近的曲面块数量越多，则 p_{c_m} 更可能为实际平面，应该将满足条件的曲面块分配给 p_{c_m}。在实验中，设置 $\eta=30$ 和 $\delta=0.02$ m。图分割问题的最小化能量 P^* 可表示为

$$P^*=\arg\min E(P),P\subset P_{\mathrm{C}}\tag{9}$$

$E(P)$ 为拟合的能量：

$$E(P)=\sum_{P_{\mathrm{c}m}\in E}D(p_{\mathrm{c}_m})\tag{10}$$

基于上述算法，可获得满足条件的预选实际平面集合 P 与对应的曲面块集合 M，利用图割法最小化能量函数，将曲面块分配到实际平面。上述算法过程中 V 与 E 为点云最小割中的顶点与边的集合，对应了聚类分割邻接图中的 $G=\{v,\varepsilon\}$，基于点云最小割理论，找出分割代价最小的分割线进行分割，即能量 $E(P)$ 最小化时的边，便可根据上述的约束条件实现将各个曲面块基于空间几何信息合并为实际平面，得到更为精确的点云分割结果，实现将超体素聚类分割结果中属于同一类物体的曲面块合并。

3.3　融合语义特征与点云特征的地图构建方法

在点云分割结果中，每个聚类均有一个属性，以最终随机赋予的颜色作为区分，构建的三维语义标签与点云分割结果中点云的数量与空间坐标是完全一致的，其中语义标签中的颜色信息代表了三维语义信息，将二者融合便可获得更为精准的三维语义地图。

基于语义信息对点云分割的结果进行优化，优化策略主要有以下 2 点。

1) 当点云分割结果中成功将该目标物体分割出来，此时该聚类中带有三维语义标签的点云超过 75%，便会将整个聚类中的点云都赋予该三维语义标签。

2) 当点云分割结果中未能成功分割出目标物体，仍然会结合三维语义标签为点云添加语义信息，但此时不会更改该聚类中全部点云的属性。

设置一个阈值 α，当该聚类中带有不同颜色属性点云数超过 α，就将该聚类基于颜色信息拆分为 2 个聚类；若点云数未超过 α，会删除这些点云带有的语义信息，即将颜色改为聚类的颜色。通过上述方法，最终得出的三维语义地图能够去除三维语义标签中的错误信息，并能优化点云分割结果，将点云分割中未分割出来的部分目标点云，基于三维语义标签分割出来，最终语义地图如图 4 所示。

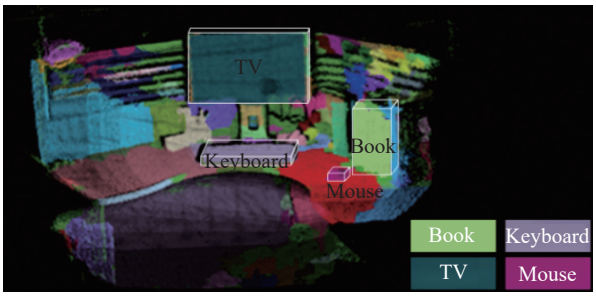


图 4　三维语义地图

Fig. 4　Three-dimensional semantic map

由图 4 可以看出，通过语义标签与点云分割结果的融合，物体 TV 能够分割得更为完整并且成功添加了语义信息，背景墙面上的三维语义标签也被成功去除。基于三维语义地图的优化策略，小物体鼠标也被成功添加了语义信息。

4　基于机器人平台语义 VI-SLAM 实验

首先，基于公开数据集 EuRoC 来验证本研究的 VI-SLAM 算法精度，并与 ORB-SLAM3 的位姿估计结果进行对比，分析本研究算法在 RGB-D 信息与 IMU 信息输入时的定位精度；然后，基于公开数据集 TUM RGBD 来验证本研究算法在动态场景中的定位精度，并与 ORB-SLAM3、DS-SLAM 算法进行对比；最后，搭建移动机器人作为实验平台，基于真实场景中的室内环境进行语义地图构建实验，并对构建的语义地图进行分析。

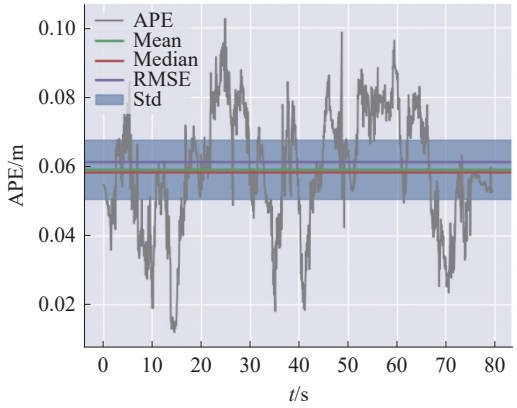
4.1　静态场景中基于 RGB-D 及 IMU 数据的机器人定位实验

由于公开的 RGB-D 与 IMU 融合的数据集并未提供机器人的真实轨迹位姿，为解决此问题，本研究实验将采用 EuRoC 公开的数据集进行测试，EuRoC 数据集是由一个双目惯性相机测量得到的，相机频率是 20 Hz，IMU 频率为 200 Hz。基于三角测量法根据左右眼的灰度图生成深度图并

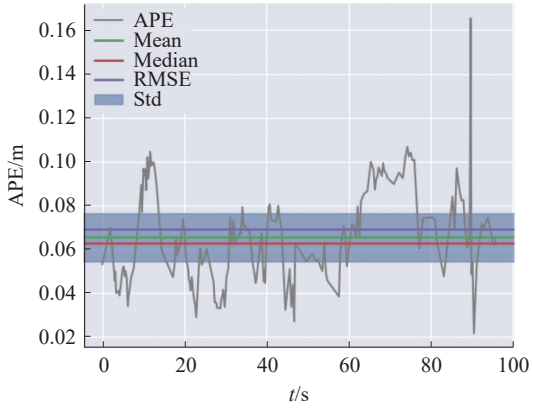
与右眼图像进行对齐, 将预处理后的 EuRoC 数据集室内场景用于系统的性能分析, 该数据集于静态场景中录制。

4.1.1　静态场景性能分析实验

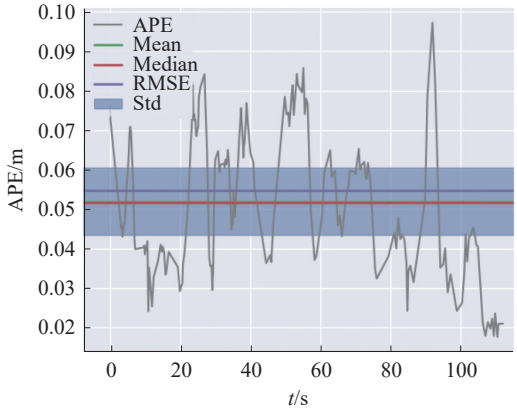
EuRoC 提供的视觉惯性数据集分为简单、中等、困难 3 个等级, 在所有室内场景中进行测试, 部分实验结果如图 5 所示, 曲线数值是算法估计的机器人运动轨迹与数据集提供的真实轨迹的绝对轨迹误差(absolute pose error, APE), 所有数据集下的误差对比见表 1。



(a) V1_02_medium 数据集



(b) V1_03_difficult 数据集



(c) V2_01_easy 数据

图 5　不同静态数据集轨迹与 APE 误差

Fig. 5　Trajectory and APE error of different static data sets

表 1　各个数据集 APE 误差指标表

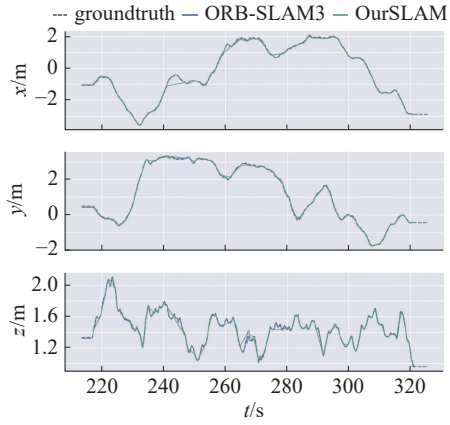
Table 1　APE error index table for each data set　m

数据集	Mean	Median	Std	RMSE
V1_02_medium	0.059	0.058	0.018	0.061
V1_03_difficult	0.065	0.063	0.022	0.069
V2_01_easy	0.054	0.056	0.026	0.060
V2_02_medium	0.052	0.051	0.017	0.054
V2_03_difficult	0.170	0.108	0.231	0.287
平均值	**0.080**	**0.067**	**0.063**	**0.106**

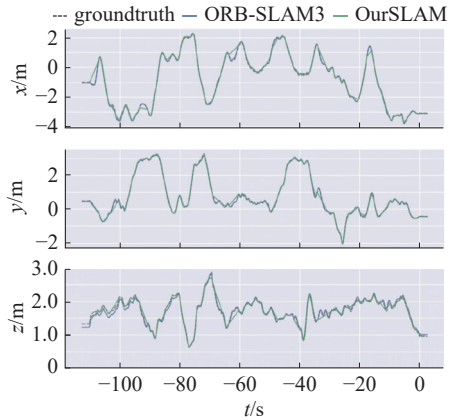
如表 1 所示, RMSE 为 APE 均方根误差, 其平均值为 0.106 m, 其值越小说明位姿估计的结果越好。由此可知, 本研究设计的 VI-SLAM 系统在通过视觉惯性传感器数据融合后, 于静态场景中位姿估计的平均误差在 0.080 m, 并且无论是在简单、中等还是困难模式下, 系统都能正常运行。

4.1.2　静态场景性能对比实验

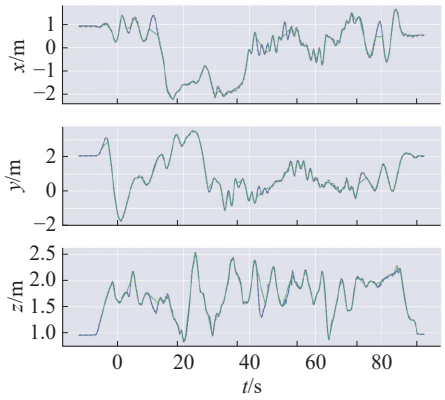
为了验证本研究方法在静态场景中基于视觉惯性的定位性能, 设计如下对比实验: 将本研究算法基于 RGB-D 信息与 IMU 信息输入时的位姿估计结果, 与 ORB-SLAM3 系统双目惯性模式下直接使用该数据集的位姿估计结果进行对比分析。图 6 分别为简单、中等、困难 3 种室内场景中的位姿估计的结果对比, 其他算法与本研究算法在各个室内场景中的均方根误差对比, 见表 2。



(a) V2_01_easy 数据集



(b) V2_02_medium 数据集



(c) V1_03_difficult 数据集

图 6　不同静态数据集中与 ORB-SLAM3 位姿估计对比

Fig. 6　Comparison with ORB-SLAM3

表 2　各系统均方根误差表

Table 2　Root-mean-square error table for each system

m

数据集	OKVIS	VINS-Fusion	ORB-SLAM3	本系统
V1_02_medium	0.200	0.129	0.048	0.061
V1_03_difficult	0.240	0.188	0.071	0.069
V2_01_easy	0.178	0.127	0.061	0.060
V2_02_medium	0.193	0.143	0.083	0.054
V2_03_difficult	0.319	0.237	0.208	0.287
平均值	**0.271**	**0.165**	**0.094**	**0.106**

由表 2 可知，本研究算法基于视觉惯性定位的均方根误差平均值为 0.106 m，与 ORB-SLAM3 算法的定位精度接近，与 OKVIS 与 VINS-Fusion 相比，定位精度更高。

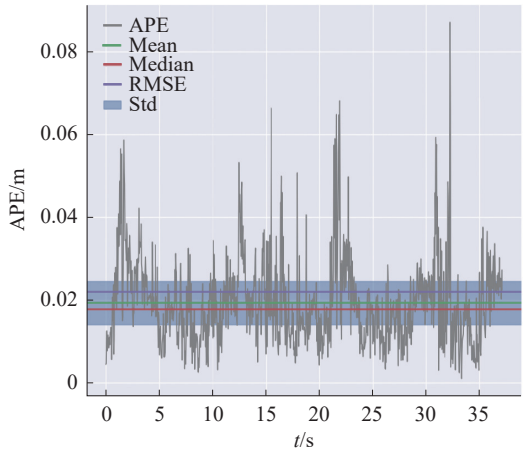
4.2　动态场景中基于 RGB-D 信息的机器人定位实验

本研究中动态场景分为 2 种，分别是低动态场景和高动态场景。低动态场景为运动物体少、物体运动速度慢、背景和环境结构相对稳定的场景；高动态场景为运动物体多、物体运动速度快、背景和环境结构会存在显著形变或运动的场景。考虑到目前没有同时录制 RGB-D 与 IMU 的数据集，本研究开展只有 RGB-D 信息输入时也能进行语义信息提取与动态特征去除工作，其中 TUM RGB-D 数据集是包含 RGB 图与深度图且在动态场景办公室环境中录制的数据集，sitting 数据集中人的动作幅度较小，对应了低动态场景，而 walking 数据集中，人始终处于行走状态，对应高动态场景。

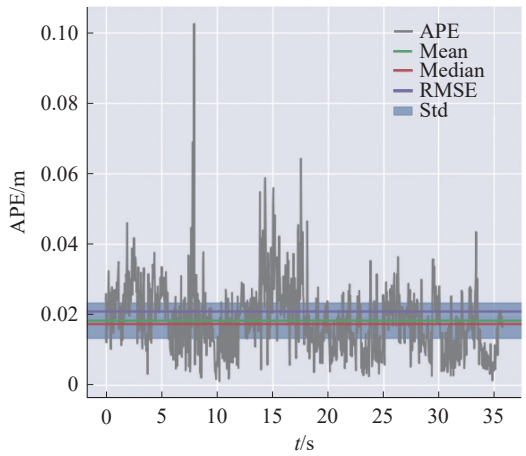
4.2.1　动态场景性能分析实验

本研究算法在不同动态场景中的误差结果如图 7 所示。

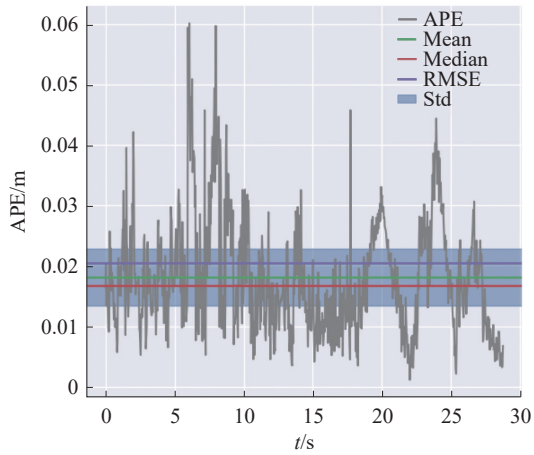
由图 7(a)可知，在低动态场景 s_halfsphere 中，本研究算法的 APE 平均值为 0.019 m，APE 的 RMSE 为 0.022 m。而由图 7(b)和图 7(c)可知，高动态场景中的 APE 平均值均为 0.018 m，相应的 RMSE 分别为 0.021 m 与 0.020 m。其他动态场景误差见表 3。



(a) s_halfsphere 数据集



(b) w_halfsphere 数据集



(c) w_xyz 数据集

图 7　不同动态数据集下 APE 误差

Fig. 7　APE error graph under different dynamic data sets

表 3　各个动态场景误差指标表

Table 3　Table of error indicators in each dynamic scenario

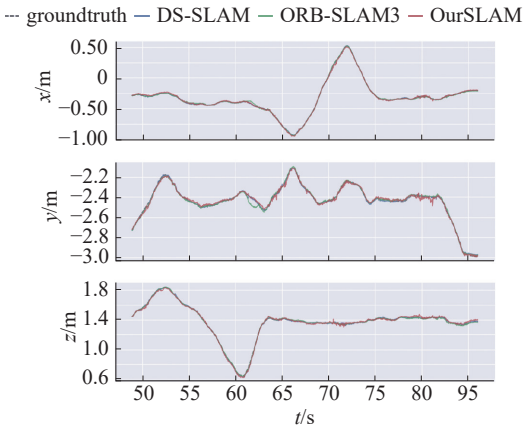
m

数据集	RMSE	Median	Std	Mean
s_halfsphere	0.022	0.018	0.010	0.019
s_rpy	0.026	0.022	0.011	0.024
s_static	0.006	0.004	0.003	0.005
w_half	0.021	0.017	0.010	0.018
w_rpy	0.033	0.022	0.019	0.027
w_static	0.006	0.005	0.002	0.005
w_xyz	0.020	0.016	0.009	0.018
平均值	**0.019**	**0.015**	**0.009**	**0.017**

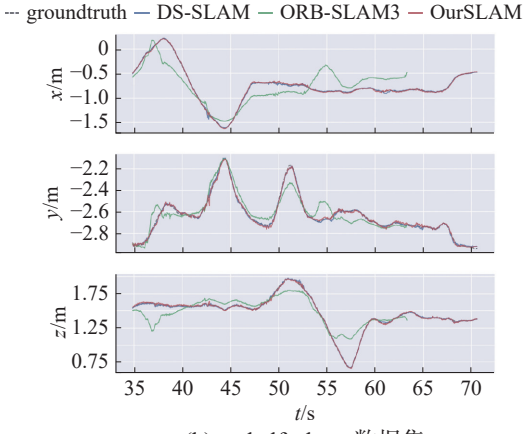
由表 3 可知, 本研究所设计的动态去除方法在高、低动态场景均有较好表现, 在动态场景中绝对轨迹误差的平均值为 0.017 m, 均方根误差为 0.019 m。

4.2.2　动态场景性能对比实验

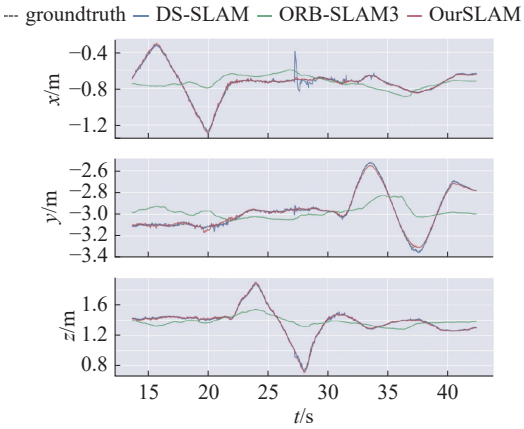
基于高动态场景与低动态场景的数据集, 对比了 ORB-SLAM3、DS-SLAM 与本研究算法在动态场景中的全局轨迹误差, 如图 8 所示。



(a) s_halfsphere 数据集



(b) w_halfsphere 数据集



(c) w_xyz 数据集

图 8　不同动态数据集中对比结果

Fig. 8　Compare the results of different dynamic data sets

由图 8 可看出, 加入本研究算法之后, 系统在动态场景中的位姿估计精度得到了明显提升, 与基于 ORB-SLAM2 和 SegNet 网络分割算法的 DS-SLAM 算法相比, 在定位精度上相差不大, 在低动态场景即 sitting 数据集中, 本研究算法相较于 ORB-SLAM3 算法在精度上基本一致, 在高动态场景即 walking 数据集中, 本研究算法相较于 ORB-SLAM3 算法误差均值和均方根误差都大幅降低, 绝对轨迹误差的均方根误差改进值平均为 85%。本研究算法在动态场景中的精度略高于 DS-SLAM, 绝对轨迹均方根误差为 0.020 m, DS-SLAM 为 0.034 m, 因此可验证本研究结合语义信息进行动态特征去除方法的有效性。

综上所述, 本研究改进后的 VI-SLAM 算法适用于绝大多数室内场景, 且整体的定位精度无论是在动态场景还是静态场景中都很高, 完全能够满足室内移动机器人的定位需求。

4.3　基于 RGB-D 及 IMU 数据的机器人建图实验

4.3.1　移动机器人平台

搭建室内移动机器人平台开展实验, 将 D435i 相机固定在机器人顶部, 使其在运动过程中能够获取桌面以上空间中的环境信息。利用 ROS 主从通信机制, 通过 ROS 话题发布 D435i 视觉惯性传感器采集的 RGB-D 信息与 IMU 信息, 机器人平台和实验环境如图 9 所示。

4.3.2　基于移动机器人的室内语义地图构建实验

基于室内机器人进行了三维语义地图构建实验, 根据 SLAM 系统关键帧对应的 RGB 图、深度图和位姿信息, 构建出实验室的全局稠密点云地图。在实验过程中, 总共生成了 312 个关键帧, 并通过将映射到空间中的三维语义标签与经过精细处理和点云分割后的稠密点云地图进行融合, 从而最终生成具有丰富语义信息的稠密点云地图,

即语义地图，如图 10 所示。

通过实验结果可以得知，本研设计的三维语义地图构建方法能够精准地将语义信息添加到对应的物体点云上。然而，由于深度相机提供的深度值存在误差，导致三维稠密点云中部分墙面不在同一个平面上，最终点云分割结果中会出现部分杂乱的点云聚类块。另外，为了使机器人能够直接使用该语义地图进行导航，将大场景下的三维语义地图以八叉树地图的形式存储，只保留并染色具有语义信息的点云的颜色。

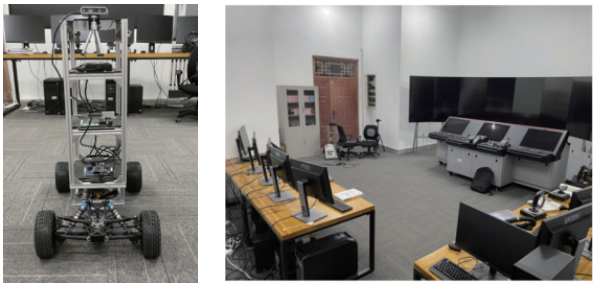


图 9　机器人平台和实验环境

Fig. 9　Robot and experimental environment

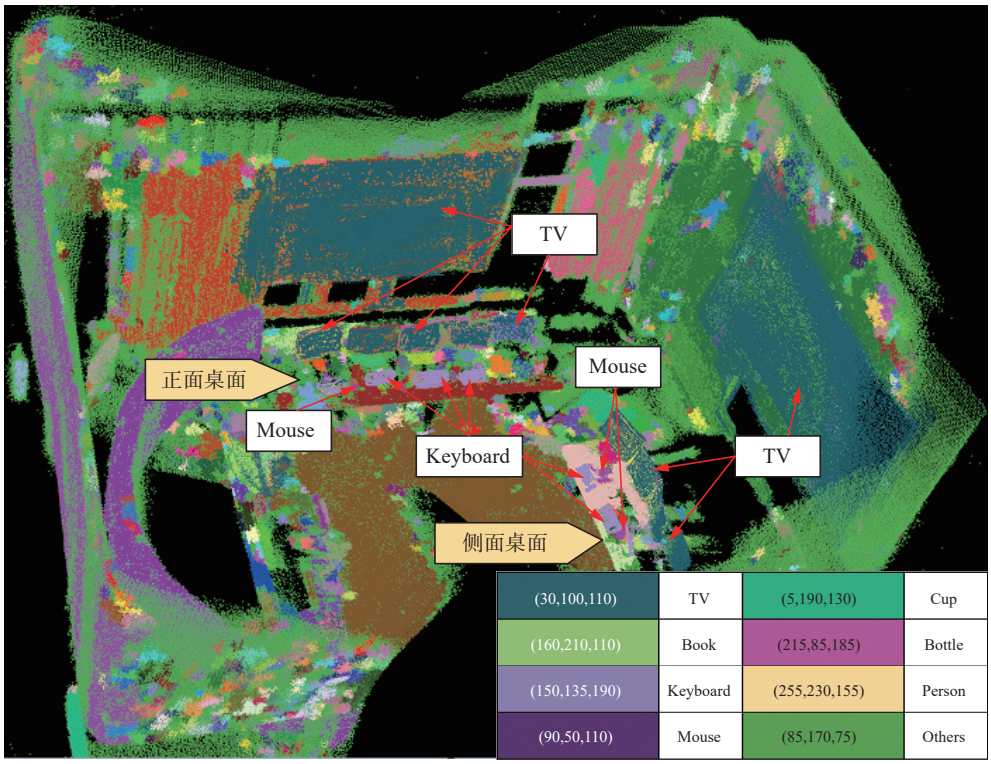


图 10　综合三维语义地图

Fig. 10　Integrate 3D semantic map

5　结束语

本研究提出了一种基于特征融合及动态背景去除的室内机器人语义 VI-SLAM 算法，旨在提升机器人在复杂环境下的定位精度和环境理解能力。该算法充分利用了 RGB-D 信息和 IMU 传感器的位姿估计数据，能够实时构建具有语义信息的室内三维点云地图。主要结论如下：1）通过去除关键帧中的动态特征点，该方法在公开数据集上的性能评估中表现出色，尤其是在高动态场景中，定位精度显著提升，达到近 85% 的提升幅度，绝对轨迹误差控制在 0.020 m 以内，这表明该算法在处理复杂场景时具有较高的鲁棒性和精确性。2）算法将关键帧中的二维语义信息映射到三维空间，并与三维稠密点云分割结果进行融合，确保了语义地图的实时构建。在本研究设定的实验中，YOLO v5s 目标检测模型达到了 45 Hz 的高帧率，与传统的 ORB-SLAM3 系统相比，表现出明显的性能优势。此外，虽然本研究所使用的深度相机在像素深度测量范围上存在一定限制，但未来工作将考虑采用双目相机，以进一步提高系统的深度感知能力和适用范围。

参考文献：

[1]　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. New York: IEEE, 2017: 5079−5085.

[2] WHELAN T, SALAS-MORENO R F, GLOCKER B, et al. ElasticFusion: real-time dense SLAM and light source estimation[J]. The international journal of robotics research, 2016, 35(14): 1697−1716.

[3] MCCORMAC J, HANDA A, DAVISON A, et al. SemanticFusion: dense 3D semantic mapping with convolutional neural networks[C]//2017 IEEE International Conference on Robotics and Automation. New York: IEEE, 2017: 4628−4635.

[4] WANG Yutong, XU Bin, FAN Wei, et al. QISO-SLAM: object-oriented SLAM using dual quadrics as landmarks based on instance segmentation[J]. IEEE robotics and automation letters, 2023, 8(4): 2253−2260.

[5] SALAS-MORENO R F, NEWCOMBE R A, STRASDAT H, et al. SLAM: simultaneous localisation and mapping at the level of objects[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2013: 1352−1359.

[6] DAME A, PRISACARIU V A, REN C Y, et al. Dense reconstruction using 3D object shape priors[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2013: 1288−1295.

[7] YU Chao, LIU Zuxin, LIU Xinjun, et al. DS-SLAM: a semantic visual SLAM towards dynamic environments[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. New York: IEEE, 2018: 1168−1174.

[8] KANEKO M, IWAMI K, OGAWA T, et al. Mask-SLAM: robust feature-based monocular SLAM by masking using semantic segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE, 2018: 371−378.

[9] WANG Kai, LIN Yimin, WANG Luowei, et al. A unified framework for mutual improvement of SLAM and semantic segmentation[C]//2019 International Conference on Robotics and Automation. Montreal: IEEE, 2019: 5224−5230.

[10] 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.

[11] BESCOS B, CAMPOS C, TARDÓS J D, et al. DynaSLAM II: tightly-coupled multi-object tracking and SLAM[J]. IEEE robotics and automation letters, 2021, 6(3): 5191−5198.

[12] LI Ao, WANG Jikai, XU Meng, et al. DP-SLAM: a visual SLAM with moving probability towards dynamic environments[J]. Information sciences, 2021, 556: 128−142.

[13] ZHAO Yao, XIONG Zhi, ZHOU Shuailin, et al. KSF-SLAM: a key segmentation frame based semantic SLAM in dynamic environments[J]. Journal of intelligent & robotic systems, 2022, 105(1): 3.

[14] HU Zhangfang, ZHAO Jiang, LUO Yuan, et al. Semantic SLAM based on improved DeepLabv3^{+} in dynamic scenarios[J]. IEEE access, 2022, 10: 21160−21168.

[15] HUANG Shisheng, MA Zeyu, MU Taijiang, et al. Supervoxel convolution for online 3D semantic segmentation [J]. ACM transactions on graphics, 2021, 40(3): 1−15.

[16] MCCORMAC J, CLARK R, BLOESCH M, et al. Fusion: volumetric object-level SLAM[C]//2018 International Conference on 3D Vision. New York: IEEE, 2018: 32−41.

[17] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. New York: IEEE, 2017: 2980−2988.

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

[19] PHAM Q H, HUA B S, NGUYEN T, et al. Real-time progressive 3D semantic segmentation for indoor scenes[C]// 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2019: 1089−1098.

[20] TATENO K, TOMBARI F, LAINA I, et al. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6565−6574.

[21] CLARK R, WANG Sen, WEN Hongkai, et al. VINet: visual-inertial odometry as a sequence-to-sequence learning problem [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2017: 3995−4001.

[22] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: self-supervised interest point detection and description[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City : IEEE, 2018: 337−33712.

[23] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. New York: IEEE, 2018: 834−848.

[24] SONG Chengqun, ZENG Bo, SU Tong, et al. Data association and loop closure in semantic dynamic SLAM using the table retrieval method[J]. Applied intelligence, 2022, 52(10): 11472−11488.

[25] QIAN Zhentian, FU Jie, XIAO Jing. Towards accurate loop closure detection in semantic SLAM with 3D semantic covisibility graphs[J]. IEEE robotics and automation letters, 2022, 7(2): 2455−2462.

作者简介：

王立鹏，副教授，博士生导师，主要研究方向为语义 SLAM、非线性控制、复杂系统建模。主持国家自然科学基金面上项目、青年项目、黑龙江省自然科学基金、民品横向项目 8 项。获授权发明专利 9 项。获省部级科技进步特等奖、一等奖。发表学术论文 30 余篇。E-mail: wanglipeng@hrbeu.edu.cn。

王小晨，硕士研究生，主要研究方向为多机器人协同、视觉 SLAM。E-mail: 13593593764@163.com。

齐尧，硕士研究生，主要研究方向为深度学习、视觉惯性 SLAM。E-mail: qiyao0208@163.com。

第二届中国具身智能大会（CEAI 2025）

The 2nd China Embodied AI Conference

由中国人工智能学会主办，CAAI 具身智能专委会（筹）、中国科学院计算技术研究所、同济大学和上海交通大学承办的中国具身智能大会（CEAI 2025）定于 2025 年 3 月 28—30 日在北京市举行。本次大会聚焦具身智能领域的最新科研进展和产业应用前沿，以构建广泛覆盖学术界、产业界、政策制定部门以及社会公众的高水平交流与合作平台为目标，推动技术创新、成果转化与产业协同发展。大会立足具身智能技术发展的全局需求，围绕科学研究、技术突破与产业实践的关键议题，致力于促进国内外专家学者深入交流，强化学术界与产业界的互动合作，形成跨领域、多维度的协同创新体系。

CEAI 2025 诚邀国内外顶级专家学者参与，阵容强大，涵盖院士、行业领军人物以及一线科研人员，共同探讨具身智能领域的未来发展。CAAI 名誉理事长、中国工程院李德毅院士，CAAI 理事长、中国工程院戴琼海院士担任大会荣誉主席；中国工程院高文院士，CAAI 监事长、中国工程院蒋昌俊院士，中国工程院于海斌院士共同担任大会主席；中国科学院计算技术研究所蒋树强研究员，上海交通大学卢策吾教授，清华大学刘华平教授，浙江大学杨易教授共同担任程序主席。

大会亮点

1. CEAI 2025 汇聚全球具身智能领域的权威专家、学术领军人物和青年学者，围绕具身智能的基础科学问题、关键技术和发展方向等热点议题发表主旨学术演讲和学术报告；

2. 设置多个专题学术论坛、圆桌讨论、博士生论坛和讲习班等，探讨具身智能技术在学术、经济、社会和伦理领域的深远影响；

3. 面向全球征集具身智能领域的研究论文并评选出青年优秀论文、优秀海报奖，助力参会学者展示科研成果，激励具身智能领域的后起之秀；

4. 邀请国内外知名企业、新型研发机构参展，展示具身智能领域的最新产品与实际应用案例，为学术界与产业界搭建合作桥梁，探索从基础研究到技术落地的全链条模式。

联系我们

大会秘书处：CAAI 具身智能专委会（筹）

联系方式：caaiembodiedai@163.com, 18503@tongji.edu.cn