文章编号　1004-924X(2024)06-0857-11

# 基于实例分割与光流的动态场景RGB－D SLAM

王成根[1]，史金龙[1*]，诸皓伟[1]，白素琴[1]，孙蕴翰[2]，卢加文[1]，黄树成[1]
(1. 江苏科技大学 计算机学院，江苏 镇江 212000；
2. 南京大学 计算机软件新技术国家重点实验室，江苏 南京 210046)

**摘要**：为了提高动态场景RGB-D SLAM中相机位姿精度，基于实例分割与光流算法，提出一种高精度RGB-D SLAM方法。首先，通过实例分割算法检测出场景中的物体，删除非刚性物体并构造语义地图。接着，通过光流信息计算运动残差，检测场景中动态刚性物体，并在语义地图中追踪这些动态刚性物体。然后，删除每一帧中非刚性物体和动态刚性物体上的动态特征点，利用其他稳定的特征点优化相机位姿。最后，通过TSDF模型重建静态背景，并以点云的形式显示动态刚性物体。在TUM和Bonn数据集中测试表明，本文方法与当前最先进的SLAM工作ACEFusion相比相机精度提升约43%。消融实验结果表明，保留动态刚性物体处于静止状态下的特征点对相机位姿估计结果提升约37%。稠密建图实验结果表明，本文方法在动态场景中重建结果优于当前先进的工作，平均重建误差为0.042 m。代码开源在https://github.com/wawcg/dy_wcg。

**关　键　词**：动态场景；同步定位与地图构建；实例分割；光流
**中图分类号**：TP394.1；TH691.9　　**文献标识码**：A　　**doi**：10.37188/OPE.20243206.0857

# RGB-D SLAM method of dynamic scene based on instance segmentation and optical flow

WANG Chenggen[1], SHI Jinlong[1*], ZHU Haowei[1], BAI Suqin[1], SUN Yunhan[2],
LU Jiawen[1], HUANG Shucheng[1]

(1. *School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang* 212000, *China*;
2. *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210046, *China*)
* *Corresponding author, E-mail: shi_jinlong*@163.*com*

**Abstract**: A new method for improving the accuracy of camera pose estimation in RGB-D SLAM of dynamic scenes was proposed. This method was based on instance segmentation and optical flow. The first step was to detect objects in the scene using instance segmentation, eliminate non-rigid objects, and construct a semantic map. The second step involved calculating motion residuals through optical flow information, detecting dynamic rigid objects, and tracking them in the semantic map. Next, dynamic feature points on non-rigid objects and dynamic rigid objects in each frame were removed, and the camera pose was optimized using stable feature points. Finally, the static background was reconstructed using the TS-

**收稿日期**：2023-09-07；**修订日期**：2023-11-14.
**基金项目**：国家自然基金资助项目(No.62276118, No.61772244)；中国民航大学民航智慧机场理论与重点实验室开放基金资助(No.SATS202207)

DF model, and the dynamic rigid objects were displayed as point clouds. Tests conducted on the TUM and Bonn datasets demonstrate that Compared with the most advanced work ACEFusion, the method proposed in this article improves camera accuracy by approximately 43%. The results show that retaining feature points of dynamic rigid objects in a static state can significantly improve camera pose estimation results. The dense mapping experiments show that our method outperforms better in dynamic 3D reconstruction, the average reconstruction error is 0.042 m. Our code is available athttps://github.com/wawcg/dy_wcg.

**Key words**: dynamic scenes; SLAM; instance segmentation; optical flow

## 1 引　言

基于深度相机的RGB-D SLAM技术已经成为多个领域内的研究热点[1],其在虚拟维修、智能制造、AR及VR等场景中被广泛应用。由于消费级深度相机工作距离限制,通常被应用于室内场景,如办公室、餐厅、客厅等,往往包含运动的物体,包括一些发生较小形变的运动刚性物体如被搬动的椅子、箱子等和可能发生较大形变的运动非刚性物体包括行走的人,随风摆动的窗帘等[2-3]。这些运动物体与环境的交互会严重影响相机位姿的计算,这给动态场景RGB-D SLAM带来了重大挑战。

为了解决这一难题,动态场景RGB-D SLAM往往结合二维实例分割网络分析场景中物体的语义信息,或结合光流检测运动物体,最后利用三维重建技术生成场景的三维地图[4]。首先,现有的SLAM方法中具有代表性的是ORB-SLAM系列工作[5-7],ORB-SLAM2[5]在ORB-SLAM[6]的基础上添加全局优化模块,使相机位姿优化更具鲁棒性。ORB-SLAM3[7]在ORB-SLAM2的基础上引入Atlas[8]模块,保存琐碎的子地图,进一步提高大场景中相机位姿优化精度。其次,当前经常与SLAM技术结合的实例分割方法中,主流的是基于候选框的R-CNN(Region-CNN)系列工作[9-11]和无候选框的SOLO系列工作[12-13],Mask R-CNN[10]在Faster R-CNN[11]网络中添加预测物体掩码信息的分支,实现了图像的目标检测和实例分割。SOLOv2[12]对SOLO[13]进行改进,动态地分割场景中每个实例,与Mask R-CNN相比实现了更高的分割精度。此外,近年来主流的光流方法研究中,PWC-net[14]在保证光流估计精度和速度的同时,极大地减少了光流网络训练的空间开销。RAFT[15]采用了一种基于LSTM[16](Long Short-Term Memory)网络的可逆映射方法,并且在不同的分辨率下处理图像,从而提高光流估计的准确性和鲁棒性,达到了更优的效果。动态场景中RGB-D SLAM方法一般通过栅格化的方式表示三维场景。KinectFusion[17-18]率先使用TSDF(Truncated Signed Distance Function)实时重建三维场景,其利用高速的GPU线程计算每个栅格中的三维信息,Voxel Hashing[19]采用哈希索引策略管理GPU中每一个栅格,实现大场景三维地图。

现有的动态场景RGB-D SLAM工作与二维实例分割结合。MaskFusion[20]将Mask R-CNN与ElasticFusion[21]结合,在动态场景中分割物体,追踪动态刚性物体,并重建动态三维场景。DynaSLAM[22]将Mask R-CNN与ORB-SLAM2结合,剔除场景中动态区域,再通过静态三维场景补全剔除的动态区域,获得更高的相机位姿优化精度。Mid-Fusion[23]使用Mask R-CNN提取场景中的实例信息,通过Point-Plane ICP[24]策略优化相机位姿,追踪运动的刚性物体,并重建动态三维场景。YOLO-SLAM[25]在YOLO[26]的基础上提出一个轻量级Darknet19-YOLOv3模块提取图像中的实例信息,结合ORB-SLAM2计算相机位姿。虽然上述方法通过实例先验信息能判断和删除动态物体,但却忽略了运动物体也会有静止的状态。这样的策略导致相机位姿优化中,物体静止时表面的特征没有被充分利用。

为了解决上述问题,一些工作与光流结合,检测物体的运动状态,为相机位姿优化提供更多稳定的特征。Alcantarilla[27]等人将SLAM技术与密集场景流结合,提高动态场景三维建模的精度和鲁棒性。FlowFusion[28]引入PWC-net,通过光流残差找到场景流中运动不一致的点云,并在追踪相机和重建场景时去除这些点云。Occlu-

sionFusion[29]利用 RAFT 提取场景中物体的运动信息，通过图神经网络[30]推断物体被遮挡区域的运动，获得更精确的动态场景 RGB-D 三维重建结果。虽然上述工作可以检测出场景中的动态物体，但往往存在动态物体边界不清晰，部分动态物体上的特征点被用于相机位姿计算，导致相机位姿优化不够精准。此外，ACEFusion[31]结合基于 DNN（Dynamic Neural Network）的实例分割模块和光流，将场景中检测的动态物体边缘刻画得更精确，但仍然忽略了运动物体静止时表面的特征也可以用于相机位姿优化。

针对上述问题，本文做了以下工作：

（1）基于先进的相机位姿优化模块、实例分割网络和光流网络，提出一种新颖的动态场景 RGB-D SLAM 方法；

（2）考虑到场景中可能会包含运动的非刚性物体，利用实例分割结果剔除场景中非刚性物体，并根据刚性物体语义信息构造语义地图；

（3）针对运动的刚性物体，通过光流网络估计相邻帧之间的光流，利用光流计算语义地图中物体的运动残差，找到动态刚性物体并追踪，再剔除动态刚性物体上的特征。通过前两部分，可以得到用于计算相机位姿的稳定特征点，实现更加精确的相机位姿估计和三维场景；

（4）本文在 TUM[32]和 Bonn[33]公开数据集共 30 个 RGB-D 序列中进行测试，结果显示：本文方法较现有最优的方法具有更高的相机位姿优化精度；本文消融实验结果显示：本文方法较仅利用先验信息剔除物体特征点的方法具有更高的相机位姿精度；本文稠密建图实验结果显示：本文重建的场景模型较当前先进的工作具有更高的三维重建精度。此外，通过点云的方式实时显示动态刚性物体，使得重建的场景更加完整。

## 2　动态 RGB-D SLAM 框架

图 1 为系统框架，本文在 ORB-SLAM3[7]相机追踪、局部建图和回环检测三个模块的基础上，增加动态检测模块和稠密建图模块。动态检测模块包括非刚性物体检测和刚性物体检测与追踪两个部分。非刚性物体检测部分识别场景中非刚性物体并剔除物体对应区域的 ORB 特征点。刚性物体检测与追踪部分检测场景中运动刚性物体，追踪并剔除物体对应区域的 ORB 特征点。稠密建图模块构造语义地图，并重建静态背景和动态刚性物体。

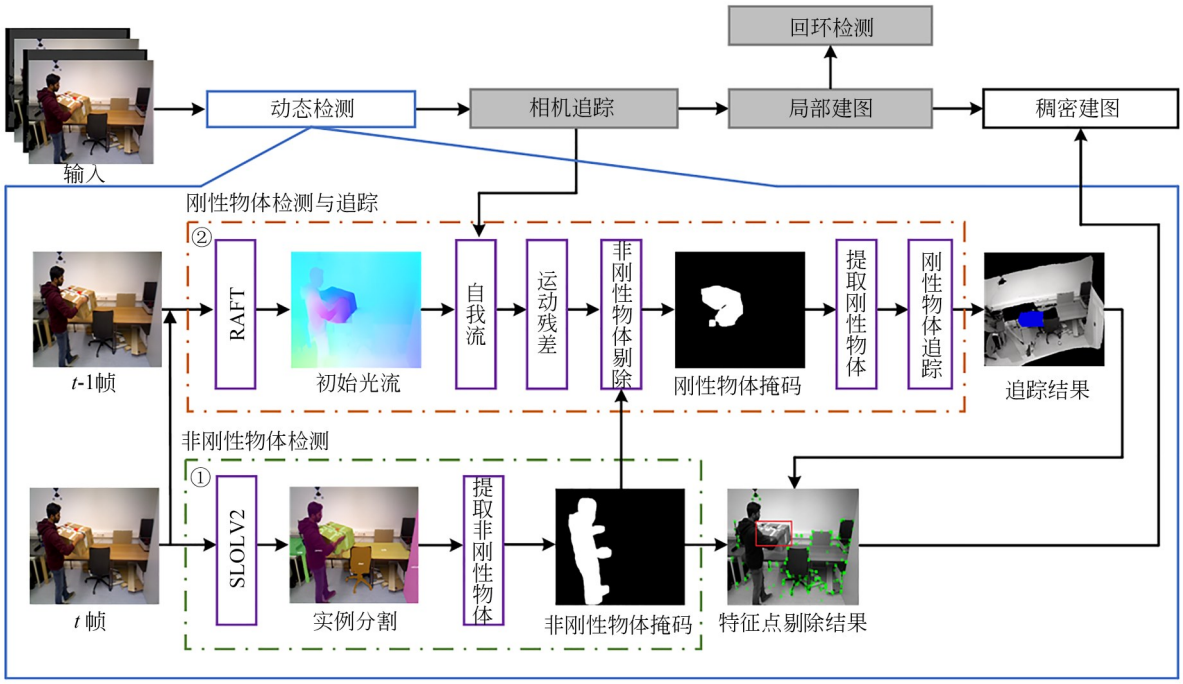



图 1　系统框架图

Fig. 1　System framework diagram

具体来说，方法将RGB-D图像序列的每一帧输入到动态检测模块。该模块首先使用SOLOv2[12]进行非刚性物体检测，提取场景中所有实例信息，并生成非刚性物体掩码图。接着，将非刚性物体外的实例信息融合到稠密建图模块构造的地图中，生成具有实例信息的语义地图。然后，使用RAFT[15]光流网络计算前后两帧图像之间的光流信息，并通过光流和自我流提取场景的运动残差。再根据运动残差提取实例点云并追踪其6D位姿，实现动态刚性物体检测与追踪。

当动态检测完成后，剔除非刚性物体掩码区域和动态刚性物体在当前帧投影区域的ORB特征点，将剩余的ORB特征点输入ORB-SLAM3的相机追踪、局部建图、回环检测模块优化相机位姿。

最后，将优化后的相机位姿输入到稠密建图模块中构造语义地图，并管理地图中的实例物体。使用带符号截断距离函数重建静态背景，并以点云的形式重建动态刚性物体。

## 3　动态检测

### 3.1　非刚性物体检测

当第 $t$ 帧 $I_t$ 到达时，将RGB图像输入到SOLOv2[12]网络，预测二维实例掩码图 $Mask_t$，如图2(a)所示。$Mask_t$ 中记录了每个像素所属的实例类别，即 $Mask_t(x,y)=m$，其中 $m\in[0,w]$ 为像素坐标 $(x,y)$ 处的掩码值，$w$ 为 $Mask_t$ 中的实例总数。设场景中非刚性物体的掩码值为 $\theta$，提取 $Mask_t$ 中值为 $\theta$ 的部分，得到 $I_t$ 初始非刚性物体掩码图 $K_t$，如图2(b)所示。由于网络预测的实例结果会存在过分割和漏分割的情况，随后该模块将 $K_t$ 腐蚀并膨胀，得到处理后掩码图 $K_t'$，结果如图2(c)所示。最后将非刚性物体外的实例信息融合到语义地图 $M$ 中，如图2(d)所示，同时在 $M$ 中分割实例物体 $O_j(\forall j\in\{0,1,\cdots,N\})$，$N$ 为物体的总数)，具体分割算法见第3节。

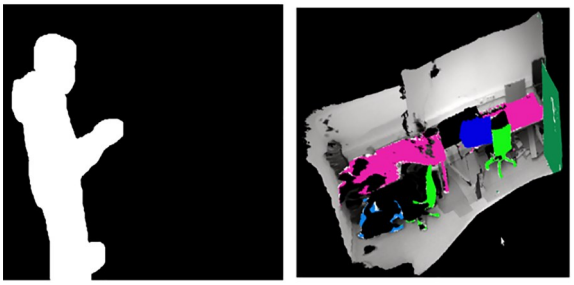
### 3.2　刚性物体检测与追踪

#### 3.2.1　运动刚性物体检测

非刚性物体检测完成后，先通过光流和自我流计算 $M$ 中每一个实例的2D运动残差。光流、自我流关系如图3所示。

(a) 实例分割Mask$_t$
(a) Instance segmentation Mask$_t$

(b) 原始掩码$K_t$
(b) Raw mask $K_t$



(c) 处理后掩码$K'_t$
(c) Post-processing mask $K'_t$

(d) 语义地图$M$
(d) Semantic map $M$

图2　非刚性物体检测效果

Fig. 2　Effect diagram of non rigid object detection

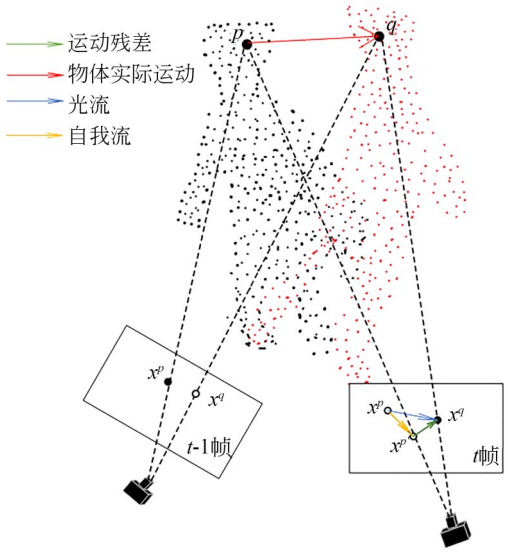



图3　光流与自我流关系

Fig. 3　Diagram of relationship between optical flow and self flow

其中 $x^p$ 为 $t-1$ 帧中运动物体三维点 $p$ 对应的像素坐标，$x^q$ 为 $t$ 帧中三维点 $q$ 对应的像素坐标，$q$ 为 $p$ 在 $t$ 时刻的位置。$x^{p'}$ 为 $x^p$ 投影到 $t$ 帧中对应的像素坐标，2D运动残差即为 $\overrightarrow{x^p x^q}-\overrightarrow{x^p x^{p'}}$。

具体来说，首先通过RAFT[15]光流网络计算

出 $I_t$ 和 $I_{t-1}$ 之间的光流值 $f_{t,t-1}^{x}$（$f_{t,t-1}^{x}\in\boldsymbol{R}^2, x\in\boldsymbol{R}^2$ 为任意像素点），描述了 $x$ 在 $I_t$ 与 $I_{t-1}$ 之间的运动，其中包含了相机运动与物体自身运动。

接着，设 $x$ 对应的 2D 运动残差为 $R_{t,t-1}^{x}\in\boldsymbol{R}$，并将非刚性物体像素区域的运动残差 $R_{t,t-1}^{x_0}$ 置为 0，$x_0\in K_t$。

随后，根据相机位姿 $T_t$ 和 $T_{t-1}$，将 $I_t$ 中非刚性物体外的像素 $x_t\notin K_t$ 投影到 $I_{t-1}$，计算出相机自我流 $e_{t,t-1}^{x_t}\in\boldsymbol{R}^2$，用于描述 $x_t$ 在 $I_t$ 与 $I_{t-1}$ 之间的相机运动，如公式(1)所示：

$$e_{t,t-1}^{x_t}=x_t-\pi(T_{t-1}^{-1}\cdot T_t\cdot\pi^{-1}(x_t)). \tag{1}$$

公式(1)中，$\pi^{-1}$ 将 $x_t$ 反投影到三维空间，$\pi$ 将三维空间中的点投影到二维平面，$T_{t-1}^{-1}$ 为 $I_{t-1}$ 相机位姿的逆。

然后，根据 $f_{t,t-1}^{x}$ 和 $e_{t,t-1}^{x_t}$ 相减的二范式计算 $R_{t,t-1}^{x_t}$，如公式(2)所示：

$$R_{t,t-1}^{x_t}=\left\| f_{t,t-1}^{x_t}-e_{t,t-1}^{x_t}\right\|_2. \tag{2}$$

最后提取 $R_{t,t-1}^{x_t}>\rho$（$\rho$ 为运动残差阈值）的部分作为潜在运动区域 $c_t$，将语义地图 $M$ 中每个实例 $O_j$ 投影到 $I_t$，当投影区域和 $c_t$ 重合时，利用 MarchingCubes[34] 算法从 $M$ 中提取 $O_j$ 的点云 $L_j$，即为运动刚性物体。

#### 3.2.2　刚性物体追踪

检测出运动的刚性物体 $L_j$ 后，通过非线性优化估计 $L_j$ 从 $t-1$ 时刻到 $t$ 时刻变化的 6D 位姿 $\xi_{t,t-1}^{j}\in se3$。

具体来说，先根据 $T_{t-1}$ 将 $L_j$ 任意三维点 $p_j^k$（$k$ 为三维点索引）投影到 $I_{t-1}$ 中得到对应的像素坐标 $x_j^k$ 的光流值 $f_{t,t-1}^{x_j^k}$。计算 $x_j^k$ 在 $I_{t-1}$ 相机空间下坐标 $C_{t-1}^{x_j^k}=\pi^{-1}(x_j^k)$，$C_{t-1}^{x_j^k}\in\boldsymbol{R}^4$。计算 $x_j^k$ 在 $I_t$ 中对应的像素坐标 $x_j^{k\prime}=x_j^k+f_{t,t-1}^{x_j^k}$，再将 $x_j^{k\prime}$ 反投影到 $I_t$ 相机空间下，即 $C_t^{x_j^{k\prime}}=\pi^{-1}(x_j^{k\prime})$，$C_t^{x_j^{k\prime}}\in\boldsymbol{R}^4$。

然后构造能量函数，优化 $L_j$ 从 $t-1$ 时刻到 $t$ 时刻的 6D 位姿 $\xi_{t,t-1}^{j}$，如公式(3)所示：

$$E_{\text{total}}=E_{\text{flow}}+E_{\text{depth}}. \tag{3}$$

$E_{\text{total}}$ 为非线性最小二乘问题的能量函数，其中 $E_{\text{flow}}$ 为光流项，如公式(4)所示：

$$E_{\text{flow}}=\sum_{p_j^k\in L_j}\left\|\pi(\exp(\xi_{t,t-1}^{j})\cdot C_{t-1}^{x_j^k})-\pi(C_t^{x_j^{k\prime}})\right\|_2^2. \tag{4}$$

公式(4)构造了 $p_j^k$ 在优化中的预测值与其对应光流值的残差，$\exp(\xi_{t,t-1}^{j})$ 将 $\xi_{t,t-1}^{j}$ 转化为 $\boldsymbol{R}^{4\times4}$ 变换矩阵。$E_{\text{depth}}$ 为深度项，如公式(5)所示：

$$E_{\text{depth}}=\sum_{p_j^k\in L_j}\left\|n_t^{x_j^k}\cdot(\exp(\xi_{t,t-1}^{j})\cdot C_{t-1}^{x_j^k}-C_t^{x_j^{k\prime}})\right\|_2^2, \tag{5}$$

其中：公式(5)为 Point-Plane ICP[24] 项，$n_t^{x_j^k}\in\boldsymbol{R}^{4\times1}$ 为 $p_j^k$ 对应法线向量，其使得 $p_j^k$ 在优化中的预测值与 $C_t^{x_j^{k\prime}}$ 在法线方向上对齐。

### 3.3　相机位姿优化

在 $I_t$ 中完成动态检测后，得到非刚性物体掩码图 $K_t$ 以及追踪的动态刚性物体点云 $L_j$。将 $p_j^k$ 投影到 $I_t$，得到动态刚性物体掩码图 $U_t^j$。然后剔除 $K_t$ 和 $U_t^j$ 区域的 ORB 特征点，接着将剩余的特征点输入到 ORB-SLAM3[5] 相机追踪模块、局部建图模块和回环检测模块进行相机位姿优化。

## 4　稠密建图

RGB-D 图像 $I_t$、相机位姿 $T_t$ 和二维实例掩码 $\text{Mask}_t$ 作为稠密建图模块的输入，构造基于 Voxel Hashing[19] 的语义地图 $M$。$M$ 内任意体素 $v$ 包含 SDF 值 $d_v$、权重 $w_v$、颜色值 $c_v$ 和物体编码 $m_v$。

随着相机运动，视野范围内可见体素中 $d_v$，$w_v$，$c_v$，$m_v$ 的值被不断更新。假设 $\{v\}_t$ 为 $M$ 在 $I_t$ 视野范围内可见的体素集合，$v_i\in\{v\}_t$。$v_i$ 中 $d_{v_i}$，$w_{v_i}$，$c_{v_i}$ 采用 Voxel Hashing 的方法进行更新。为了分割背景和每一个实例物体，本文先给 $\text{Mask}_t$ 中每一个实例赋予物体编码值，再将 $\text{Mask}_t$ 融合到 $M$ 中更新 $m_{v_i}$，具体算法步骤如下：

输入：二维实例掩码图 $\text{Mask}_t$、$I_t$ 的相机位姿 $T_t$、语义地图 $M$。

输出：体素的物体编码值 $m_v$。

Step1：将 $\text{Mask}_t$ 中二维实例 $e_t^l$（$l\in[0,w]$，$w$ 为 $\text{Mask}_t$ 实例总数）的每个像素反投影到 $M$ 中，找到其对应的体素集合 $\{v\}_l$。

Step2：若 $M$ 中存在实例物体 $O_j$，$O_j$ 对应体素与 $\{v\}_l$ 重合数量大于 $\sigma$（$\sigma$ 为重合阈值），则将 $e_t^l$ 物体编码置为 $j$；若不存在或重合数量小于 $\sigma$，赋予 $e_t^l$ 新的物体编码。

Step3：根据 $I_t$ 的相机位姿将 $v_i$ 投影到 $\text{Mask}_t$，找到与投影点欧式距离最小的像素坐标 $x_{v_i}$，将 $\text{Mask}_t$ 对应的物体编码 $Mask_t^{x_{v_i}}$ 赋值给 $m_{v_i}$。

此外，对于运动的刚性物体，通过 3.2.2 节中的方法追踪 $O_j$ 并更新 $L_j$，更新方法如公式(6)

所示：

$$p_j^{k'}=\exp(\xi_{t,t-1}^{j})\cdot p_j^{k}, \tag{6}$$

其中：$p_j^k\in L_j$，$p_j^{k'}\in \boldsymbol{R}^4$为更新后的三维点坐标。

## 5 实验对比与分析

本文实验平台为Intel Xeon Silver 4214 CPU、32 G内存、RTX2080TiGPU的PC，操作系统采用Ubuntu18.04。实例分割与光流网络的训练与测试均使用Python编写，SLAM部分、非刚性物体检测、动态刚性物体检测与追踪部分使用C++编写。

### 5.1 实验数据与网络训练

Bonn[33]数据集与TUM[32]数据集均为动态RGB-D SLAM领域广泛使用的公开数据集。Bonn中共有24个RGB-D序列，其中包含了人体走动、人与箱子交互、人与气球交互等多个动态场景，并提供了每一帧相机位姿真值。TUM数据集中有6个动态序列，其可根据人体的运动程度划分为3个低动态场景和3个高动态场景。低动态场景中，人坐在椅子上进行办公和交流；高动态场景中，人在办公室内走动。TUM同样提供了每一帧对应的相机位姿真值。

为了识别到场景中存在的语义，文章在Bonn数据集中选取900张图片并标注，其中包含人体、椅子、桌子、气球、小车、柜子和箱子共7个类别。使用SOLOv2[12]训练，迭代600次，主干网络使用Resnet[35]。RAFT[15]光流网络采用其预训练模型raft-sintel。

### 5.2 相机位姿精度对比与分析

相机位姿精度是衡量RGB-D SLAM工作的重要指标，文章采用绝对轨迹误差(ATE-RMSE)评估相机位姿精度，绝对轨迹误差通常用于评估相机轨迹的全局一致性，表明了每一帧相机位姿估计值与真值之间的差值，数值越小，相机位姿精度越高。

在Bonn数据集24个动态序列上，将本文方法与StaticFusion[36]，DynaSLAM[22]，ReFusion[33]和ACEFusion[31]对比，分别用SF，DS，RF和AF表示。

实验结果如表1所示，文章提出的方法在14个动态序列中达到了最先进的相机位姿优化精度。其中，Kidnapping2，Moving No Box，Moving O Box2，Placing No Box等序列均有显著提升。经研究发现，这些序列均为人与箱子交互的场景，而本文方法将箱子处于静态时稳定的特征点用于相机位姿优化。

**表1 Bonn数据集绝对轨迹误差对比**

Tab. 1 Comparison of absolute trajectory error of Bonn dataset (m)

| 序列 | SF[36] | DS[22] | RF[33] | AF[31] | OURS |
|---|---|---|---|---|---|
| Balloon | 0.233 | 0.030 | 0.175 | 0.028 | **0.028** |
| Balloon2 | 0.293 | **0.029** | 0.254 | 0.030 | 0.029 |
| Balloon Tracking | 0.221 | 0.049 | 0.302 | 0.045 | **0.041** |
| Balloon Tracking2 | 0.366 | 0.035 | 0.322 | **0.033** | 0.059 |
| Crowd | 3.586 | 0.016 | 0.204 | **0.016** | 0.027 |
| Crowd2 | 0.215 | 0.031 | 0.155 | **0.027** | 0.028 |
| Crowd3 | 0.168 | 0.038 | 0.137 | **0.023** | 0.038 |
| Kidnapping | 0.336 | 0.029 | 0.148 | 0.030 | **0.029** |
| Kidnapping2 | 0.263 | 0.035 | 0.161 | 0.030 | **0.025** |
| Moving No Box | 0.141 | 0.232 | 0.071 | 0.070 | **0.025** |
| Moving No Box2 | 0.364 | 0.039 | 0.179 | **0.029** | 0.037 |
| Moving O Box | 0.331 | **0.044** | 0.343 | 0.343 | 0.262 |
| Moving O Box2 | 0.309 | 0.263 | 0.528 | 0.443 | **0.134** |
| Person Tracking | 0.484 | 0.061 | 0.289 | 0.070 | **0.041** |
| Person Tracking2 | 0.626 | 0.078 | 0.463 | 0.071 | **0.056** |
| Placing No Box | 0.125 | 0.575 | 0.106 | 0.088 | **0.021** |
| Placing No Box2 | 0.177 | 0.021 | 0.141 | **0.020** | 0.022 |
| Placing No Box3 | 0.256 | 0.058 | 0.174 | 0.051 | **0.043** |
| Placing O Box | 0.330 | 0.225 | 0.571 | 0.324 | **0.177** |
| Removing No Box | 0.136 | **0.016** | 0.041 | 0.020 | 0.017 |
| Removing No Box2 | 0.129 | **0.021** | 0.111 | 0.025 | 0.025 |
| Removing O Box | 0.334 | 0.291 | 0.222 | 0.314 | **0.197** |
| Synchronous | 0.446 | 0.015 | 0.410 | 0.014 | **0.012** |
| Synchronous2 | 0.037 | 0.009 | 0.022 | 0.010 | **0.009** |

注：**表中加粗数据表示该序列的相机位姿精度最优值**

在TUM数据集上，将本文方法与DynaSLAM，ReFusion，RigidFusion[37]和ACEFusion对比，分别用DS，RF，Rigid和AF表示。

实验结果如表2所示，其中Sitting Static，Sitting XYZ，Sitting Halfsphere，Walking Static，Walking XYZ等序列中，本文方法比现有的方法精度更高，这得益于文章将SOLOv2实例分割网络和ORB-SLAM3相结合，将非刚性物体分割得

更精确，参与相机位姿优化的特征点更可靠。

**表 2　TUM 数据集绝对轨迹误差对比**

Tab. 2　Comparison of absolute trajectory error of TUM dataset　(m)

| 序列 | DS[22] | RF[33] | Rigid[37] | AF[31] | OURS |
|---|---|---|---|---|---|
| Sitting Static | 0.007 | 0.011 | 0.019 | 0.028 | **0.006** |
| Sitting XYZ | 0.015 | 0.026 | 0.054 | 0.021 | **0.014** |
| Sitting Halfsphere | 0.028 | 0.038 | 0.129 | 0.035 | **0.019** |
| Walking Static | 0.007 | 0.014 | 0.018 | 0.011 | **0.007** |
| Walking XYZ | 0.017 | 0.074 | 0.090 | 0.025 | **0.015** |
| Walking Halfsphere | **0.026** | 0.048 | 0.076 | 0.035 | 0.027 |

注：**表中加粗数据表示该序列的相机位姿精度最优值**

此外，文章选取部分实验中的相机轨迹结果与真值对比，如图 4 所示。

图 4 中直观地展示了文章提出的方法在 Bonn 和 TUM 数据集中相机绝对轨迹误差对比。图 4 共有 9 个子图，每个子图中黑色线代表数据集中真实轨迹，蓝色线代表算法估计的相机位姿，红色线代表两者误差（彩图见期刊电子版）。从图中可以看出，在大多数序列中，算法估计的相机位姿与真实轨迹的偏差较小，再次证明本文所提出方法在各个场景中都有优越表现。

### 5.3　相机位姿精度对比与分析

如前文所述，当动态刚性物体处于静止状态时，其表面特征较为稳定，能够很好的辅助相机位姿优化。为了证明这一点，文章选取 Bonn 数据集中含有动态刚性物体的序列，对比计算全程剔除和动态剔除两种策略下的绝对估计误差。

其中 11 个序列本文提出的动态剔除方法性能更好，其他 4 个序列 Balloon，Balloon2，Moving O Box 与 Moving O Box2 均未有提升，这是由于场景中的动态刚性物体如气球、箱子始终处于运动状态，无静止状态。

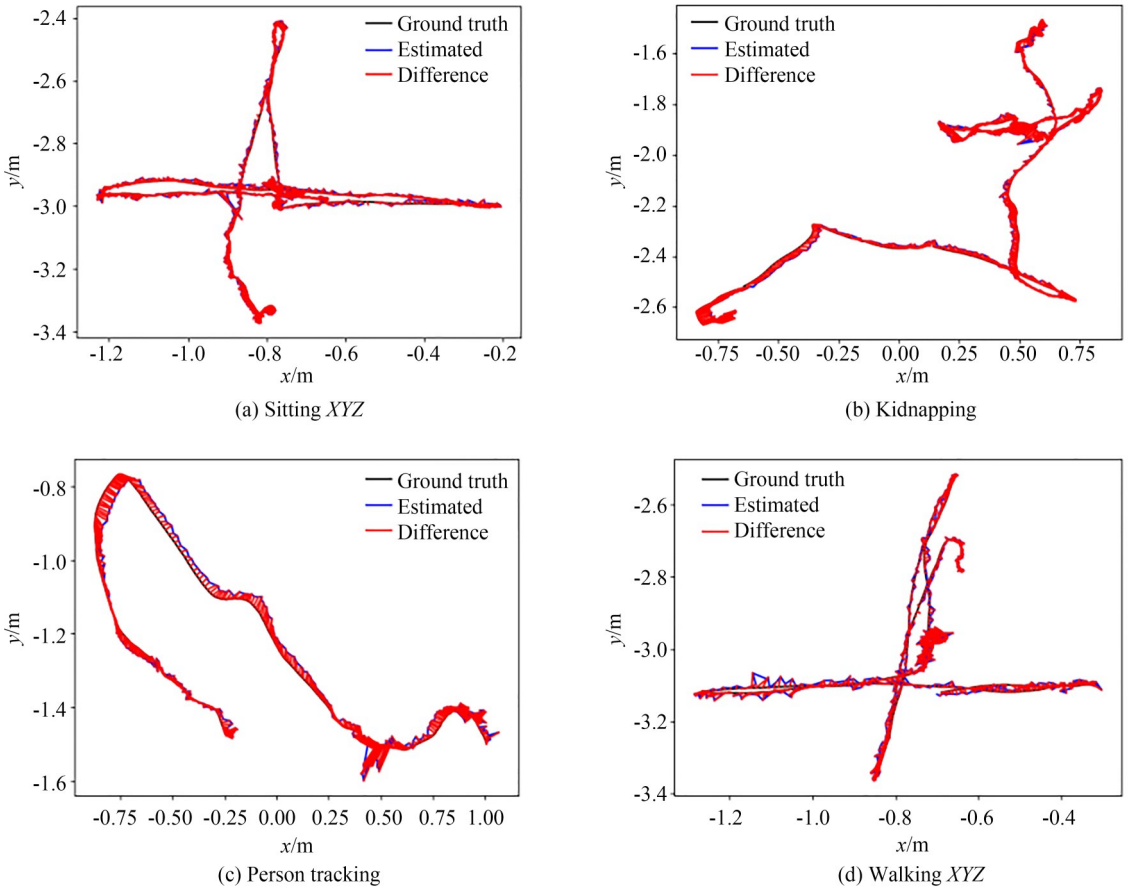



图 4　绝对轨迹误差对比图

Fig. 4　Diagram of Absolute trajectory error comparison

**表 3　Bonn 数据集不同策略的绝对轨迹误差对比**

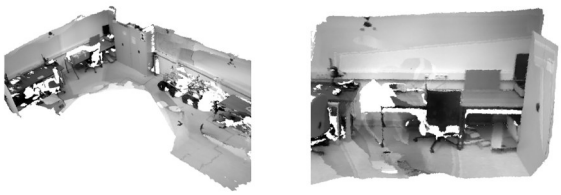
Tab. 3　Comparison of Absolute Trajectory Errors for Different Strategies in the Bonn Dataset　(m)

| 序列 | 全程剔除 | 动态剔除 |
|---|---|---|
| Balloon | 0.028 | 0.028 |
| Balloon2 | 0.029 | 0.029 |
| Balloon Tracking | 0.049 | **0.041** |
| Balloon Tracking2 | 0.089 | **0.059** |
| Moving No Box | 0.035 | **0.025** |
| Moving No Box2 | 0.042 | **0.037** |
| Moving O Box | 0.262 | 0.262 |
| Moving O Box2 | 0.134 | 0.134 |
| Placing No Box | 0.022 | **0.021** |
| Placing No Box2 | 0.025 | **0.022** |
| Placing No Box3 | 0.045 | **0.043** |
| Placing O Box | 0.174 | **0.177** |
| Removing No Box | 0.019 | **0.017** |
| Removing No Box2 | 0.028 | **0.025** |
| Removing O Box | 0.197 | **0.191** |

注：**表中加粗数据表示该序列的相机位姿精度最优值**

### 5.4　稠密建图实验与分析

在稠密建图系统将 RGB-D 序列帧和优化后的相机位姿输入到稠密建图模块中，使用 TSDF 模型重建静态背景，重建结果如图 5 所示。



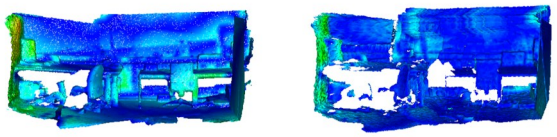
(a) Person tracking　(b) Moving no box

图 5　背景重建

Fig. 5　Background reconstruction map

图 5 中 5(a)为 Bonn 数据集中 Person Tracking 序列的背景重建结果，此序列中仅包含非刚性物体，即行走的人。5(b)为 Bonn 数据集中 Moving No Box 序列的背景重建结果，此序列包含非刚性物体和运动刚性物体，即被人搬动的箱子。图 6 中展示了本文方法与 ReFusion 在 Bonn 数据集 Moving No Box 序列的重建结果对比。

本文方法与 ReFusion 均可重建出背景模型，但 ReFusion 平均重建误差为 0.066 m，在墙面和



(a) ReFusion　(b) 本文方法 (b) Proposed method

图 6　重建结果对比

Fig. 6　Comparison of reconstruction results

桌面上误差较大。本文方法平均重建误差为 0.042 m，与 ReFusion 相比重建精度更高。

此外，场景中动态刚性物体以点云的形式重建，如图 7 所示。

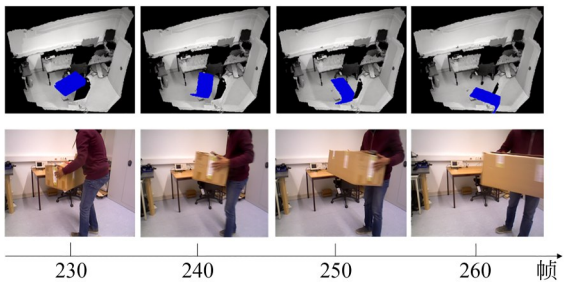



图 7　动态刚性物体融合

Fig. 7　Dynamic rigid object fusion

图 7 为 Bonn 数据集中 Moving No Box 序列动态刚性物体融合结果，图中以时间戳的形式展示方法在 230，240，250 和 260 帧时动态刚性物体融合效果。场景中采用蓝色点云凸显追踪的动态刚性物体，从图中看出方法可将融合动态物体的三维场景重建完整(彩图见期刊电子版)。

## 6　结　论

为了提高 RGB-D SLAM 方法在室内动态场景中的精度，文章提出一种基于实例分割与光流的 RGB-D SLAM 方法，在 TUM 和 Bonn 数据集上测试表明与当前最先进的工作 ACEFusion 相比相机精度提升约 43%。本文通过语义信息和光流信息，检测出场景中的非刚性物体和动态刚性物体，剔除非刚性物体和动态刚性物体特征点，重建静态背景以及动态刚性物体，本文方法最终平均重建误差为 0.042 m。

但方法仍存在的局限性，首先，仅仅通过光流检测出场景中动态刚性物体，没有将光流信息

用于静态部分的相机位姿优化中。其次，直接剔除场景中非刚性物体，没有追踪和重建非刚性物体。在下一阶段的工作中，进一步使用光流信息用于相机位姿优化，以提升SLAM系统的相机位姿精度，进一步的考虑非刚性物体追踪和重建方法，提高动态场景模型重建精度。

**参考文献：**


[1] 张裕，张越，张宁，等. 基于逆深度滤波的双目折反射全景相机动态SLAM系统[J]. 光学 精密工程，2022，30(11)：1282-1289.
ZHANG Y，ZHANG Y，ZHANG N，*et al*. Dynamic SLAM of binocular catadioptric panoramic camera based on inverse depth filter[J]. *Opt. Precision Eng.*，2022，30(11)：1282-1289.(in Chinese)

[2] 郭道亮. 可变形物体的全局非刚性配准与重建[D]. 天津：天津大学，2018.
GUO D L. *Global Non-Rigid Registration and Reconstruction of Deformable Objects*[D]. Tianjin：Tianjin University，2018.(in Chinese)

[3] NEWCOMBE R A，FOX D，SEITZ S M. DynamicFusion：reconstruction and tracking of non-rigid scenes in real-time[C]. 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston，MA，USA*. IEEE，2015：343-352.

[4] 刘东生，陈建林，费点，等. 基于深度相机的大场景三维重建[J]. 光学 精密工程，2020，28(1)：234-243.
LIU D S，CHEN J L，FEI D，*et al*. Three-dimensional reconstruction of large-scale scene based on depth camera[J]. *Opt. Precision Eng.*，2020，28(1)：234-243.(in Chinese)

[5] 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.

[6] MUR-ARTAL R，MONTIEL J M M，TARDÓS J D. ORB-SLAM：a versatile and accurate monocular SLAM system[J]. *IEEE Transactions on Robotics*，2015，31(5)：1147-1163.

[7] CAMPOS C，ELVIRA R，RODRÍGUEZ J J G，*et al*. ORB-SLAM3：an accurate open-source library for visual，visual-inertial，and multimap SLAM[J]. *IEEE Transactions on Robotics*，2021，37(6)：1874-1890.

[8] AAD G，ANDUAGA XS，ANTONELLI S，*et al*. The ATLAS experiment at the CERN large hadron collider[J]. 2008.

[9] GIRSHICK R，DONAHUE J，DARRELL T，*et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014 *IEEE Conference on Computer Vision and Pattern Recognition. Columbus，OH，USA*. IEEE，2014：580-587.

[10] HE K M，GKIOXARI G，DOLLÁR P，*et al*. Mask R-CNN[C]. 2017 *IEEE International Conference on Computer Vision (ICCV). Venice，Italy*. IEEE，2017：2980-2988.

[11] REN S Q，HE K M，GIRSHICK R，*et al*. Faster R-CNN：towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*，2017，39(6)：1137-1149.

[12] WANG X L，ZHANG R F，KONG T，*et al*. SOLOv2：Dynamic and Fast Instance Segmentation[EB/OL]. 2020：*arXiv*：2003.10152. http://arxiv.org/abs/2003.10152

[13] WANG X L，KONG T，SHEN C H，*et al*. *SOLO*：*Segmenting Objects by Locations*[M]. Computer Vision – ECCV 2020. Cham：Springer International Publishing，2020：649-665.

[14] SUN D Q，YANG X D，LIU M Y，*et al*. PWC-net：CNNs for optical flow using pyramid，warping，and cost volume[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，UT，USA*. IEEE，2018：8934-8943.

[15] TEED Z，DENG J. *RAFT*：*Recurrent All-Pairs Field Transforms for Optical Flow*[M]. Computer Vision – ECCV 2020. Cham：Springer International Publishing，2020：402-419.

[16] CHO K，VAN MERRIENBOER B，BAHDANAU D，*et al*. On the Properties of Neural Machine Translation：Encoder-Decoder Approaches[EB/OL]. 2014：*arXiv*：1409.1259. http://arxiv.org/abs/1409.1259

[17] IZADI S，KIM D，HILLIGES O，*et al*. KinectFusion：Real-Time 3D reconstruction and interaction using a moving depth camera[C]. *Proceedings*

*of the 24th annual ACM symposium on User interface software and technology. Santa Barbara California USA*. ACM, 2011: 559-568.

[18] NEWCOMBE R A, IZADI S, HILLIGES O, *et al*. KinectFusion: Real-Time dense surface mapping and tracking[C]. 2011 10*th IEEE International Symposium on Mixed and Augmented Reality. Basel, Switzerland*. IEEE, 2011: 127-136.

[19] NIEßNER M, ZOLLHÖFER M, IZADI S, *et al*. Real-time 3D reconstruction at scale using voxel hashing[J]. *ACM Transactions on Graphics*, 2013, 32(6): 1-11.

[20] RUNZ M, BUFFIER M, AGAPITO L. MaskFusion: Real-Time recognition, tracking and reconstruction of multiple moving objects[C]. 2018 *IEEE International Symposium on Mixed and Augmented Reality* (*ISMAR*). *Munich, Germany*. IEEE, 2018: 10-20.

[21] WHELAN T, LEUTENEGGER S, SALAS MORENO R, *et al*. ElasticFusion: dense SLAM without a pose graph[C]. *Robotics: Science and Systems XI. Robotics: Science and Systems Foundation*, 2015, 11(3).

[22] 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.

[23] 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, QC, Canada*. IEEE, 2019: 5231-5237.

[24] RUSINKIEWICZ S, LEVOY M. Efficient variants of the ICP algorithm[C]. *Proceedings Third International Conference on 3-D Digital Imaging and Modeling. Quebec City, QC, Canada*. IEEE, 2002: 145-152.

[25] WU W X, GUO L, GAO H L, *et al*. YOLO-SLAM: a semantic SLAM system towards dynamic environment with geometric constraint[J]. *Neural Computing and Applications*, 2022, 34(8): 6011-6026.

[26] KIM D S, FIGUEROA K W, LI K W, *et al*. Profiling of dynamically changed gene expression in dorsal root Ganglia post peripheral nerve injury and a critical role of injury-induced glial fibrillary acidic protein in maintenance of pain behaviors[J]. *Pain*, 2009, 143(1/2): 114-122.

[27] ALCANTARILLA P F, YEBES J J, ALMAZÁN J, *et al*. On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments[C]. 2012 *IEEE International Conference on Robotics and Automation. Saint Paul, MN, USA*. IEEE, 2012: 1290-1297.

[28] ZHANG T W, ZHANG H Y, LI Y, *et al*. FlowFusion: dynamic dense RGB-D SLAM based on optical flow[C]. 2020 *IEEE International Conference on Robotics and Automation* (*ICRA*). *Paris, France*. IEEE, 2020: 7322-7328.

[29] LIN W B, ZHENG C W, YONG J H, *et al*. OcclusionFusion: occlusion-aware motion estimation for real-time dynamic 3D reconstruction[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *New Orleans, LA, USA*. IEEE, 2022: 1726-1735.

[30] SCARSELLI F, GORI M, TSOI A C, *et al*. The graph neural network model[J]. *IEEE Transactions on Neural Networks*, 2009, 20(1): 61-80.

[31] BUJANCA M, LENNOX B, LUJÁN M. ACEFusion-accelerated and energy-efficient semantic 3D reconstruction of dynamic scenes[C]. 2022 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). *Kyoto, Japan*. IEEE, 2022: 11063-11070.

[32] 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-Algarve, Portugal*. IEEE, 2012: 573-580.

[33] PALAZZOLO E, BEHLEY J, LOTTES P, *et al*. ReFusion: 3D Reconstruction in dynamic environments for RGB-D cameras exploiting residuals[C]. 2019 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), *China*. IEEE, 2019: 7855-7862.

[34] LORENSEN W E, CLINE H E. Marching cubes: a high resolution 3D surface construction algorithm[J]. *ACM SIGGRAPH Computer Graphics*, 1987, 21(4): 163-169.

[35] TARG S, ALMEIDA D, LYMAN K. Resnet in ResNet: Generalizing Residual Architectures[EB/

OL]. 2016: *arXiv*: 1603.08029. http://arxiv.org/abs/1603.08029

[36] SCONA R, JAIMEZ M, PETILLOT Y R, *et al*. StaticFusion: background reconstruction for dense RGB-D SLAM in dynamic environments [C]. 2018 *IEEE International Conference on Robotics and Automation* (*ICRA*). *Brisbane*, *QLD*, *Australia*. IEEE, 2018: 3849-3856.

[37] WONG Y S, LI C J, NIEßNER M, *et al*. RigidFusion: RGB-D scene reconstruction with rigidly-moving objects [J]. *Computer Graphics Forum*, 2021, 40(2): 511-522.

**作者简介：**

**王成根**(1998－)，男，江苏省盐城人，硕士研究生，主要从事计算视觉、SLAM 方面的研究。E-mail: 15751018227@163.com

**通讯作者：**

**史金龙**(1976－)，男，黑龙江宾县人，博士，教授，硕士生导师。主要从事计算机视觉、人工智能、软件工程方向的研究与教学工作，主持或参与科技部国家重点研发计划子课题、江苏省科技厅产学研前瞻项目、军工课题、企业课题等 20 余项。E-mail: shi_jinlong@163.com