第 41 卷 第 22 期/2021 年 11 月/光学学报 研究论文

光学学报

基于全景环带成像的语义视觉里程计

陈浩[1]，杨恺伦[2]，胡伟健[1]，白剑[1]，汪凯巍[1*]

[1]浙江大学国家光学仪器工程技术研究中心，浙江 杭州 310058；

[2]德国卡尔斯鲁厄理工学院人类与机器人研究所，德国 卡尔斯鲁厄 76131

摘要 视觉里程计在智能机器人、自动驾驶等领域有着广泛的应用。但是基于有限视场(FOV)针孔相机的经典视觉里程计算法容易受到环境中运动物体和相机快速旋转的影响，在实际应用中鲁棒性和精度不足。针对这一问题，提出全景环带语义视觉里程计。通过将具有超大视场的全景环带成像系统应用到视觉里程计，并将基于深度学习的全景环带语义分割所提供的语义信息耦合到算法的各个模块，减小运动物体和快速旋转的影响，提高在应对这两种挑战性场景时算法性能。实验结果表明，相较于经典的视觉里程计，所提算法在实际环境下可以实现更加精确和鲁棒的位姿估计。

关键词 机器视觉；视觉里程计；全景环带镜头；语义分割；位姿估计

中图分类号 TP242.6 **文献标志码** A **doi**：10.3788/AOS202141.2215002

Semantic Visual Odometry Based on Panoramic Annular Imaging

Chen Hao[1], Yang Kailun[2], Hu Weijian[1], Bai Jian[1], Wang Kaiwei[1*]

[1]*National Engineering Research Center of Optical Instrumentation, Zhejiang University, Hangzhou, Zhejiang* 310058, *China*;

[2]*Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe* 76131, *Germany*

Abstract Visual odometry is commonly used in various applications including intelligent robots and self-driving cars. However, traditional visual odometry algorithms based on the pinhole camera with a limited field of view (FOV) are usually fragile to moving objects in the environment and fast rotation of the camera, resulting in insufficient robustness and accuracy in practical use. This paper proposes panoramic annular semantic visual odometry as a solution to this problem. Using the panoramic annular imaging system with ultra-wide FOV into visual odometry and coupling semantic information provided by the panoramic annular semantic segmentation based on deep learning into each module of the algorithm, the effect of moving objects and fast rotation is reduced; then, the performance of visual odometry in dealing with these challenging scenarios can be improved. Compared with traditional visual odometry systems, experimental results show that the proposed algorithm achieves more accurate and robust pose estimation in realistic scenarios.

Key words machine vision; visual odometry; panoramic annular lens; semantic segmentation; pose estimation

OCIS codes 150.5758; 150.0155; 150.1135

1 引 言

视觉里程计(VO)是一种利用相机所拍摄的连续图像来估计相机 6 自由度位姿的技术[1]，被广泛应用于智能机器人、自动驾驶、虚拟/增强现实等应用中。其中，由于单目相机具有价格低廉、结构简单、标定方便等优点，单目视觉里程计更加受到研究者的青睐。

收稿日期：2021-03-18；**修回日期**：2021-04-13；**录用日期**：2021-06-03

基金项目：浙江大学舜宇智慧光学研究中心项目(2020-03)

通信作者：*wangkaiwei@zju.edu.cn

近些年来，国内外研究者们已经提出了很多优秀的算法，例如 DSO[2]、SVO[3]、ORBSLAM[4] 等，在一些开源数据集上取得了不错的效果。但是，当应用于真实场景的时候，这些方法依然存在一些问题。首先，现有的视觉里程计算法大多基于静态环境假设，针对有限视场（FOV）的针孔相机而设计，这使得当环境中存在运动物体的时候，可能出现错误的位姿估计[5]；其次，在真实的应用场景中（比如智能机器人），很容易出现比较快速的旋转运动。有限的 FOV 会使得相邻帧之间的共视区域由于快速旋转而迅速减小，位姿估计不准确。鉴于此，提出全景环带语义视觉里程计（PASVO）。通过使用全景环带镜头（PAL）成像，并结合基于深度学习的全景环带语义分割，提高视觉里程计在真实应用环境下的精度和鲁棒性。

PAL 是一种可以在单次成像中获得 360°全景感知的透镜[6]。得益于 PAL 的超大 FOV，PASVO 在面对真实应用场景时，可以保证充足的帧间共视关系，保障了位姿估计的连续和稳定。而语义分割指对图像中的每个像素根据其所属的类别来分配语义标签[7]。随着深度学习的发展，一些语义分割网络已经取得了出色的效果。利用语义分割所获得的稠密语义信息，可以帮助 PASVO 在位姿估计过程中有效降低运动物体的干扰。

目前，已有一些研究人员注意到了更大的 FOV 对视觉里程计的重要意义，并基于一些超大 FOV 物镜进行了相关探索。Lemaire 等[8]、Rituerto 等[9]、Pretto 等[10]学者相继提出了基于折反射式全景系统（catadioptric system）的定位算法；基于鱼眼镜头（fisheye lens）的全景成像方案也同样受到研究者的青睐，如南开大学提出的 CubemapSLAM[11]、基于比较出名的开源框架扩展而来的 SVO2.0[12]、Omnidirectional DSO[13]、ORBSLAM3[14] 等；也有研究者基于拼接式的全景系统进行了一些工作，如 Seok 等提出的 ROVO[15] 和 OmniSLAM[16]、上海交通大学提出的基于 Richo Theta V 全景相机的 PVO[17]、武汉大学提出的 PanoSLAM[18] 等。此类方法依靠全景视角实现了比经典视觉里程计更好的精度，但由于没有理解环境的语义，依旧会受到运动物体的影响。此外，这些方案所采用的全景系统也有着一些不足之处：折反射系统需要支撑结构，且对反射镜的面型精度要求较高而无法实用；鱼眼镜头有着严重的负畸变而压缩了边缘分辨率；拼接式全景系统无法凝视成像且依赖拼接算法。相较于这些，PAL 在实现对 360°周遭环境凝视成像的基础上，具有紧凑的光学结构、更好的畸变控制、更加均匀的像面相对照度，这些优点使得 PAL 更加适合应用于视觉里程计。

在语义信息和视觉里程计结合方面，近些年亦有不少工作，如语义信息用于选取特征点[19-21]、用于去除环境中的动态目标[22-24]、用于减小漂移提高定位精度[25-26]、用于构建语义地图[27] 等。但这些工作大都是基于透视图像提出的，在实际应用中它们依然受到相机 FOV 的限制，存在对相机快速旋转不够鲁棒的问题。本文提出基于全景环带成像的语义视觉里程计，通过综合应用 PAL 和全景环带语义分割，减小运动物体和快速旋转的影响，提高在应对挑战性场景时算法性能。本文的贡献可以归纳为三点：1）基于全景环带成像系统和全景环带语义分割，提出了 PASVO 算法；2）提出了基于语义信息校验的稀疏直接法位姿估计，及语义引导的关键点选择策略和极曲线搜索策略；3）在搭载 PAL 相机的遥控车所采集的真实场景数据集上测试了所提算法。

2 PAL 相机模型

在详细阐述 PASVO 算法流程之前，首先介绍一下所使用的 PAL 相机模型。PAL 通过对入射光线进行两次反射，产生周遭 360°场景的环形像，再由后面的中继透镜（relay lens）成像在图像传感器上。通过这种方式，PAL 将围绕光轴的三维圆柱区域投影到图像传感器，形成二维平面上的环带，亦被称作平面圆柱投影，如图 1 所示。

由于 PAL 不满足透视投影，所以针孔相机模型显然是不适用的。为了准确描述全景环带成像中三维空间中的物点（3D 坐标）与其在图像平面上的投影位置（2D 坐标）之间的数学关系，使用

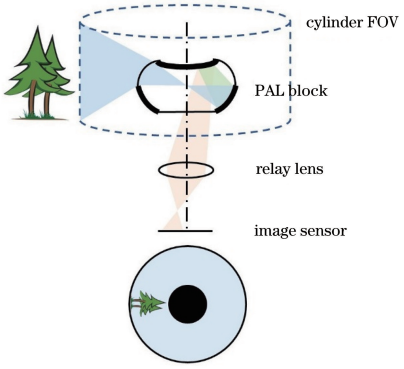

图 1 全景环带成像

Fig. 1 Panoramic annular imaging

Scaramuzza 等[28-29]于 2006 年提出的一种“泰勒模型”。之所以叫泰勒模型，是因为作者使用泰勒多项式来拟合相机的参数模型。该模型可以非常精确地模拟折反射式全景系统、鱼眼镜头等全景成像系统，为研究者所广泛采用。

如图 2 所示，$\boldsymbol{P}$ 是环境中某个三维点，记 $\boldsymbol{P}$ 的投影点在图像平面的坐标为 $\boldsymbol{u}=[u,v]^{\mathrm{T}}$[如图 2(b)所示]，在传感器平面的坐标为 $\boldsymbol{u}'=[u',v']^{\mathrm{T}}$[如图 2(c)所示]，则 $\boldsymbol{u}$ 与 $\boldsymbol{u}'$ 之间的关系可用仿射变换 $\boldsymbol{u}'=\boldsymbol{A}\boldsymbol{u}+\boldsymbol{t}$ 描述，其中 $\boldsymbol{A}\in\mathbf{R}^{2\times2}$，$\boldsymbol{t}\in\mathbf{R}^{2\times1}$。那么，泰勒模型可以完整表述为

$$\boldsymbol{T}\cdot\boldsymbol{P}=\lambda\boldsymbol{p}=\lambda g(\boldsymbol{u}')=\lambda g(\boldsymbol{A}\boldsymbol{u}+\boldsymbol{t}),\lambda>0,\tag{1}$$

式中：$\boldsymbol{p}$ 是模长为 1 的方向向量；$\boldsymbol{T}$ 为世界坐标系到相机坐标系的位姿变换矩阵。函数 $g(\boldsymbol{u}')$ 的形式为

$$g(\boldsymbol{u}')=\begin{bmatrix}u'\\v'\\f_{\mathrm{b}}(\rho')\end{bmatrix},\tag{2}$$

$$f_{\mathrm{b}}(\rho')=\alpha_0+\alpha_2\cdot\rho'^2+\cdots+\alpha_N\cdot\rho'^N,\tag{3}$$

$$\rho'=\sqrt{u'^2+v'^2}。\tag{4}$$

由二维坐标计算得到其对应物点的三维坐标(2D→3D)，这个过程称为相机的反向投影，使用 $\pi^{-1}(\cdot)$ 表示。值得注意的是，对于单一视图，三维点 $\boldsymbol{P}$ 的深度(从 $\boldsymbol{P}$ 到光学中心 O 的距离)是不可知的。故而通过反向投影仅可以获得方向向量 $\boldsymbol{p}$，对应于 $\boldsymbol{P}$ 在全景球面(panoramic sphere)上的投影点

$$\boldsymbol{p}=\pi^{-1}(\boldsymbol{u}')=\frac{g(\boldsymbol{u}')}{\|g(\boldsymbol{u}')\|}=\frac{1}{\sqrt{u'^2+v'^2+f_{\mathrm{b}}^2(\rho')}}\begin{bmatrix}u'\\v'\\f_{\mathrm{b}}(\rho')\end{bmatrix}。\tag{5}$$

相应的，已知物点的三维坐标，计算其在图像传感器平面上投影位置(3D→2D)的过程称为相机的正向投影，使用 $\pi(\cdot)$ 来表示。记 $\boldsymbol{T}\cdot\boldsymbol{P}=\boldsymbol{P}'=[x',y',z']^{\mathrm{T}}$，那么正向投影可以表示为

$$\boldsymbol{u}'=\pi(\boldsymbol{P}')=f_{\mathrm{p}}(\theta)\cdot h(\boldsymbol{P}')=f_{\mathrm{p}}(\theta)\cdot\begin{bmatrix}\dfrac{x'}{\sqrt{x'^2+y'^2}}\\[2ex]\dfrac{y'}{\sqrt{x'^2+y'^2}}\end{bmatrix},\tag{6}$$

$$f_{\mathrm{p}}(\theta)=\beta_0+\beta_2\cdot\theta^2+\cdots+\beta_N\cdot\theta^N,\tag{7}$$

$$\theta=\arctan\frac{z'}{\sqrt{x'^2+y'^2}},\tag{8}$$

式中：多项式 $f_{\mathrm{b}}(\rho)$ 和 $f_{\mathrm{p}}(\theta)$ 的系数和次数需要通过标定来获得。

值得注意的是，对于泰勒模型，可以很容易地使用链式法则来计算投影函数 $\pi(\boldsymbol{P}')$ 的雅可比矩阵，这在视觉里程计的优化过程中十分重要。如(8)式所示，它是一个 2×3 的矩阵：

$$\frac{\mathrm{d}\pi(\boldsymbol{P}')}{\mathrm{d}\boldsymbol{P}'}=\left[\frac{\partial\pi}{\partial h}\right]_{1\times1}\cdot\left[\frac{\partial h}{\partial\boldsymbol{P}'}\right]_{2\times3}+\left[\frac{\partial\pi}{\partial f_{\mathrm{p}}}\right]_{2\times1}\cdot\left[\frac{\partial f_{\mathrm{p}}}{\partial\theta}\right]_{1\times1}\cdot\left[\frac{\partial\theta}{\partial\boldsymbol{P}'}\right]_{1\times3},\tag{9}$$

式中：$[\cdot]_{m\times n}$ 表示 $m\times n$ 矩阵；$[\cdot]_{1\times1}$ 表示标量。

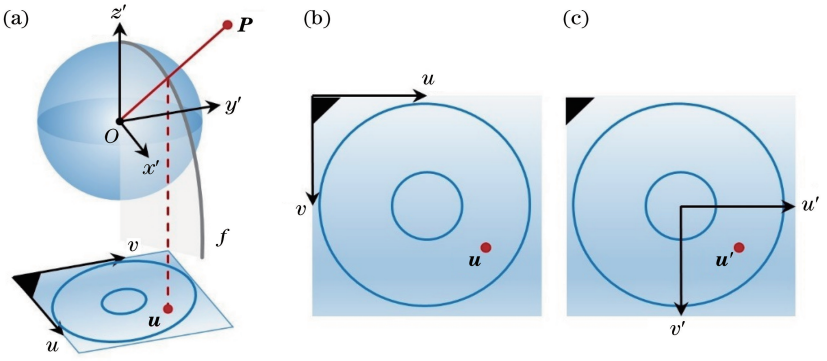

图 2　相机模型。(a) PAL 相机模型；(b) 图像平面；(c) 传感器平面

Fig. 2　Camera model. (a) PAL camera model; (b) image plane; (c) sensor plane

3　PASVO 算法流程

PASVO 算法流程如图 3 所示。新的图像首先被送入到全景环带语义分割网络，获得语义图像。随后，语义图像和原始图像一同进入位姿估计流程，通过由粗到精的两步位姿计算，获得相机的 6 自由度位姿。如果该帧被确定为关键帧，则会暂时存储在由最近数个关键帧组成的地图中，并在该帧图像上初始化一些关键点。无论该帧是否为关键帧，都会在该帧图像上执行沿极曲线的匹配搜索，以恢复地图中关键点的深度。

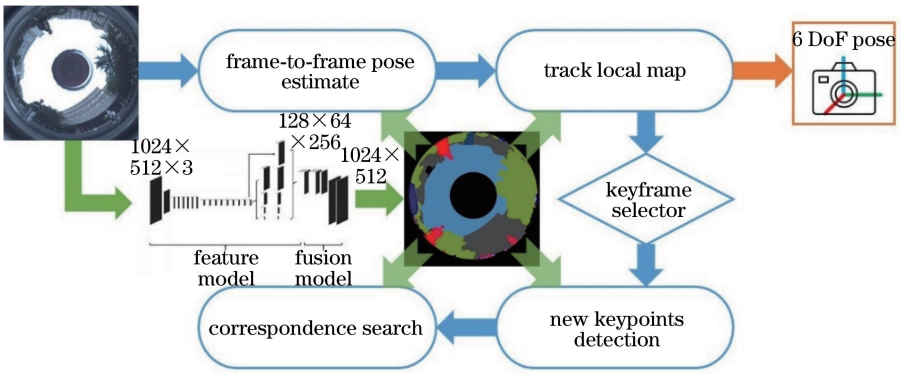

图 3 PASVO 算法流程图

Fig. 3 Flowchart of PASVO algorithm

3.1 全景环带语义分割

语义分割任务由基于全景环带图像在线展开的语义分割框架(PASS)来完成[30-31]。目前主流的基于深度学习的语义分割大多使用透视图像(如 KITTI、Cityscapes 等数据集)来训练网络,当直接应用于全景环带图像时,性能会有着明显的下降。PASS 框架解决了这一问题。具体来说,在训练阶段,依然利用传统的透视图像数据集对实时语义分割网络 ERF-PSPNet[32]进行训练;而当用网络来预测时,将全景环带图像在线展开并进行分段,利用训练好的网络来分别预测不同分段中具有语义信息的特征图,然后将它们融合以完成全景场景的解析,输出像素级的稠密语义信息,如图 4 所示。

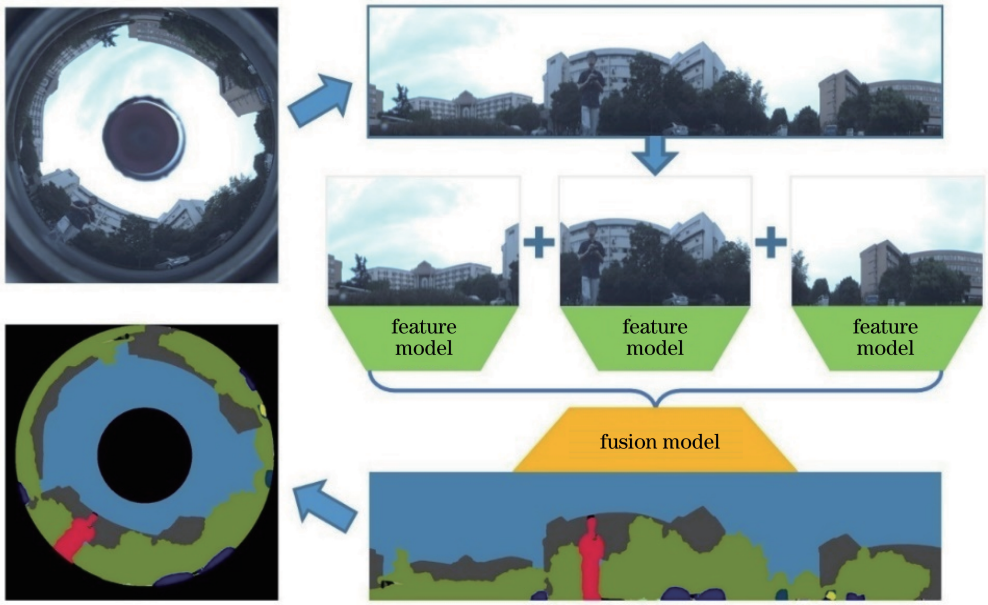

图 4 全景环带语义分割(PASS)

Fig. 4 Panoramic annular semantic segmentation (PASS)

关于语义类别,在所提 PASVO 中,采用 Cityscapes 数据集中所有标注的 19 类,其中包含了道路、建筑、植被等静态物体,行人、自行车、汽车等动态物体。

3.2 由粗到精的位姿估计

位姿估计的过程包含由粗到精(coarse-to-fine)两个步骤。首先,基于同一路标点在不同的图像中亮度不变的假设,通过稀疏点直接对齐法(称为“稀疏直接法”),即优化前后两帧图像在稀疏的关键点处的光度误差,来获取帧到帧的粗略位姿估计。具体来说,对于前一帧第 i 个关键点(设像素位置为 $\boldsymbol{u}_i$,对应的三维点为 $\boldsymbol{P}_i$),根据初始的帧间位姿变换,投影到当前帧,得到 $\boldsymbol{u}_i'$:

$$\boldsymbol{u}_i'=\pi(\hat{\boldsymbol{T}}_{t,t-1}\cdot\boldsymbol{P}_i)=\pi[\hat{\boldsymbol{T}}_{t,t-1}\cdot d_i\cdot\pi^{-1}(\boldsymbol{u}_i)]。\tag{10}$$

基于此,可以计算前后两帧的像素亮度残差:

$$\delta \boldsymbol{I}(\boldsymbol{u}_i,\hat{\boldsymbol{T}}_{t,t-1})=\boldsymbol{I}_t(\boldsymbol{u}'_i)-\boldsymbol{I}_{t-1}(\boldsymbol{u}_i),\tag{11}$$

式中：$\boldsymbol{I}_t$ 和 $\boldsymbol{I}_{t-1}$ 分别为 t 和 $t-1$ 时刻的图像；$\boldsymbol{I}_t(\boldsymbol{u}')$为 t 时刻图像在像素 $\boldsymbol{u}'$ 处的亮度；d_i 为第 i 个关键点的深度；$\boldsymbol{T}_{t,t-1}$ 代表 $t-1$ 时刻到 t 时刻之间的相机相对位姿变换；$\hat{}$ 指待估计的变量。由于单一像素亮度具备较大的偶然性，故而对 $\boldsymbol{u}_i$ 及其邻域 $\mathbb{N}(\boldsymbol{u}_i)$中的所有像素均计算亮度残差，作为关键点 $\boldsymbol{u}_i$ 处的光度误差(计算时假设邻域像素和中心像素具有相同深度)：

$$e^{\text{photo}}(\boldsymbol{u}_i,\hat{\boldsymbol{T}}_{t,t-1})=\frac{1}{2}\sum_{\Delta\boldsymbol{u}\in\mathbb{N}(\boldsymbol{u}_i)}\|\delta\boldsymbol{I}(\boldsymbol{u}_i+\Delta\boldsymbol{u},\hat{\boldsymbol{T}}_{k,k-1})\|^2。\tag{12}$$

接下来调整$\hat{\boldsymbol{T}}_{t,t-1}$，使得前后两帧所有关键点处的光度误差之和最小。由于关键点的投影位置可能处于运动物体等原因，某些误差项对于相机位姿是错误约束。对此利用语义信息作为先验，通过校验重投影前后语义一致性来去除错误的约束项：

$$\boldsymbol{T}^*_{t,t-1}=\underset{\hat{\boldsymbol{T}}_{t,t-1}}{\arg\min}\sum_i W(\boldsymbol{u}_i,\boldsymbol{u}'_i)\cdot e^{\text{photo}}(\boldsymbol{u}_i,\hat{\boldsymbol{T}}_{t,t-1}),\tag{13}$$

式中：$W(\boldsymbol{u}_i,\boldsymbol{u}'_i)$为语义校验函数。其具体的计算方式为

$$W(\boldsymbol{u}_i,\boldsymbol{u}'_i)=\begin{cases}1, & S_t(\boldsymbol{u}'_i)=S_{t-1}(\boldsymbol{u}_i)\\ 0, & S_t(\boldsymbol{u}'_i)\neq S_{t-1}(\boldsymbol{u}_i)\end{cases},\tag{14}$$

式中：S_t 代表对 t 时刻图像进行语义分割所得到的语义图。据此可以查询像素 $\boldsymbol{u}'_i$所属的语义类别 $S_t(\boldsymbol{u}'_i)$。

(12)式的优化问题对$\hat{\boldsymbol{T}}_{t,t-1}$ 来说是非线性的，所以利用高斯-牛顿法进行迭代优化。设中间变量 $\boldsymbol{P}'_i=\hat{\boldsymbol{T}}_{t,t-1}\cdot\boldsymbol{P}_i$，对残差函数(10)式，求当前状态下的雅可比，可得

$$\boldsymbol{J}_i=\frac{\partial\boldsymbol{I}_t}{\partial\boldsymbol{u}'_i}\cdot\frac{\partial\boldsymbol{u}'_i}{\partial\boldsymbol{P}'_i}\cdot\frac{\partial\boldsymbol{P}'_i}{\partial\hat{\boldsymbol{T}}_{t,t-1}},\tag{15}$$

其中，第一项为图像梯度：

$$\frac{\partial\boldsymbol{I}_t}{\partial\boldsymbol{u}'_i}=\frac{1}{2}\begin{bmatrix}\boldsymbol{I}_t(\boldsymbol{u}'_i+[1\quad 0]^{\mathrm{T}})-\boldsymbol{I}_t(\boldsymbol{u}'_i+[-1\quad 0]^{\mathrm{T}})\\ \boldsymbol{I}_t(\boldsymbol{u}'_i+[0\quad 1]^{\mathrm{T}})-\boldsymbol{I}_t(\boldsymbol{u}'_i+[0\quad -1]^{\mathrm{T}})\end{bmatrix},\tag{16}$$

第二项$\frac{\partial\boldsymbol{u}'_i}{\partial\boldsymbol{P}'_i}$可由(8)式计算。最后一项根据扰动模型，易得

$$\frac{\partial\boldsymbol{P}'_i}{\partial\hat{\boldsymbol{T}}_{t,t-1}}=[\boldsymbol{I}\quad -[\boldsymbol{P}'_i]_\times],\tag{17}$$

式中：$[\boldsymbol{P}'_i]_\times$ 表示三维向量 $\boldsymbol{P}'_i$对应的反对称矩阵。叠加所有关键点处的雅可比，可得正规方程：

$$\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}\xi=-\boldsymbol{J}^{\mathrm{T}}\delta\boldsymbol{I},\tag{18}$$

式中：$\xi\in\mathfrak{se}(3)$为更新步长，解之并对当前状态进行更新：

$$\hat{\boldsymbol{T}}_{t,t-1}\leftarrow\exp([\xi]_\times)\cdot\hat{\boldsymbol{T}}_{t,t-1}。\tag{19}$$

经过数次迭代，可得最优的 $\boldsymbol{T}^*_{t,t-1}$，基于此可以计算得到 t 时刻相机位姿的粗略估计：

$$\hat{\boldsymbol{T}}_t=\boldsymbol{T}^*_{t,t-1}\cdot\boldsymbol{T}_{t-1}。\tag{20}$$

接下来进行相机位姿的精细化。利用$\hat{\boldsymbol{T}}_t$，可以将地图中(不只是前一帧)的三维路标点投影到当前帧。由于粗位姿还不十分精确，路标点直接投影的位置也不会完全准确。采用类似于 Lucas-Kanade 光流[33]的方法，对路标点在当前帧的投影位置进行优化调整。随后通过最小化重投影误差，来实现当前帧相机位姿估计的精细化。对于三维路标点 $\boldsymbol{P}_i$，误差项为

$$e^{\text{proj}}(\boldsymbol{P}_i,\hat{\boldsymbol{T}}_t)=\frac{1}{2}\|\pi(\hat{\boldsymbol{T}}_t\cdot\boldsymbol{P}_i)-\boldsymbol{u}'_i\|^2,\tag{21}$$

式中：$\boldsymbol{u}'_i$为第 i 个路标点调整后的投影位置(像素坐标)；$\boldsymbol{P}_i$ 为路标点的三维坐标。同样的，在优化过程中引入语义校验函数：

$$\boldsymbol{T}^*_t=\underset{\hat{\boldsymbol{T}}_t}{\arg\min}\sum_i W(\boldsymbol{u}_i,\boldsymbol{u}'_i)\cdot e^{\text{proj}}(\boldsymbol{P}_i,\hat{\boldsymbol{T}}_t),\tag{22}$$

式中：$\boldsymbol{u}_i$ 为 $\boldsymbol{P}_i$ 在其他关键帧中(若有多个，以离当前帧最近的关键帧为准)的投影位置。同样利用高斯-牛顿法求解使重投影误差最小的 $\boldsymbol{T}^*_t$，并将其作为当前帧最优的位姿进行输出。

3.3 语义引导的极曲线搜索

位姿估计后，会进入地图构建流程，即计算关键点的三维坐标，使其成为路标点。由于 PAL 的特殊成像方式，极线在全景环带图像上表现为一条曲线，称为“极曲线”。欲计算关键点的三维坐标，需要利用沿极曲线搜索的方法在后续图像中寻找关键点的匹配点，进而使用三角化来恢复关键点的深度。常规的搜索方法是沿着极曲线，以固定的搜索步长(通常设置为 1 个像素)，寻找匹配分数最小的位置。这种固定步长的搜索方法在宽基线的情况下，搜索次数会比较大。事实上，在极曲线上某些区域，其语义

类别与关键点不一致，则在此区域搜索到匹配点的概率非常小；若此时仍旧用固定的步长，会导致搜索效率的下降。故此使用语义信息来引导极曲线搜索：初始的搜索步长设为 1 个像素；若连续 n 个搜索的点与关键点语义类别不一致，则下一步的搜索步长增加为初始值的 1.2^n 倍，以提高搜索效率；反之，若当前搜索的点与关键点语义类别相同，则将搜索步长重新设为初始值，提高搜索精度，如图 5 所示。

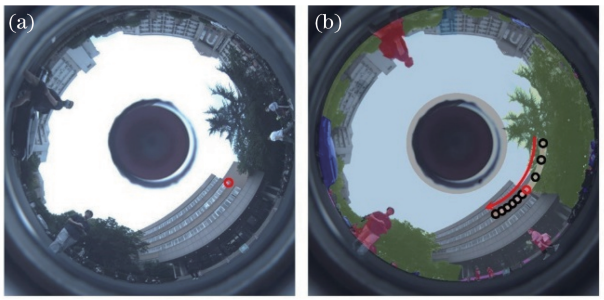

图 5 语义引导的极曲线搜索。(a)一个关键点；(b)语义引导下的沿极曲线搜索匹配点

Fig. 5 Correspondence search along epipolar curve under semantic guidance. (a) One of the keypoints; (b) searching matching points along polar curve under semantic guidance

3.4 语义引导的关键点选择

当一帧被选择为关键帧时，会在该帧图像上初始化一些关键点。而根据上文所述，运动物体上的关键点极大可能会给位姿优化带来错误的约束，所以是需要尽量避免的。事实上，从深度恢复的角度考虑，除了运动物体之外，以下三类点一般也认为是不可靠点。第一，天空中可能检测到的关键点，此类无限远点由于深度无法准确建模，需要尽量避免其参与到跟踪流程。第二，位于物体轮廓上的关键点，这类点有着前景和背景的二义性，深度不确定性较大。第三，在运动物体的轮廓周围的关键点，此类点在后续帧极大可能被运动物体所遮挡，亦不利于深度恢复。综合这些考虑，通过语义分割获得的语义标签定义图像中的“不可靠区域(UR)”，来指导关键点的检测。具体来说，利用语义先验信息，可以分辨出不利于位姿估计的区域，比如人(person，rider)、车辆(car，truck，bus，motorcycle，bicycle)等运动物体、天空(sky)等无限远区域(括号中的内容为 Cityscapes 数据集中定义的语义类别)。可以根据到“不可靠区域”的距离，定义关键点的置信度，以此引导关键点的选择：

$$\mathcal{P}(\boldsymbol{u})=1-\mathrm{e}^{-\frac{D(\boldsymbol{u})}{\sigma}}, \tag{23}$$

式中：σ 为预设参数；$\mathcal{P}(\boldsymbol{u})$ 为 $\boldsymbol{u}$ 作为关键点的置信度；$D(\boldsymbol{u})$ 为 $\boldsymbol{u}$ 到最近的不可靠区域的距离。

$$D(\boldsymbol{u})=\min\{\|\boldsymbol{u}-\boldsymbol{u}^-\|_2,\ \forall \boldsymbol{u}^-\in\Omega_{\mathrm{UR}}\}。 \tag{24}$$

如图 6 所示，距离“不可靠区域”越近的位置，置信度越低，所以减少了关键点的检测；反之，在离“不可靠区域”越远的位置，则会保持或增加关键点的检测。

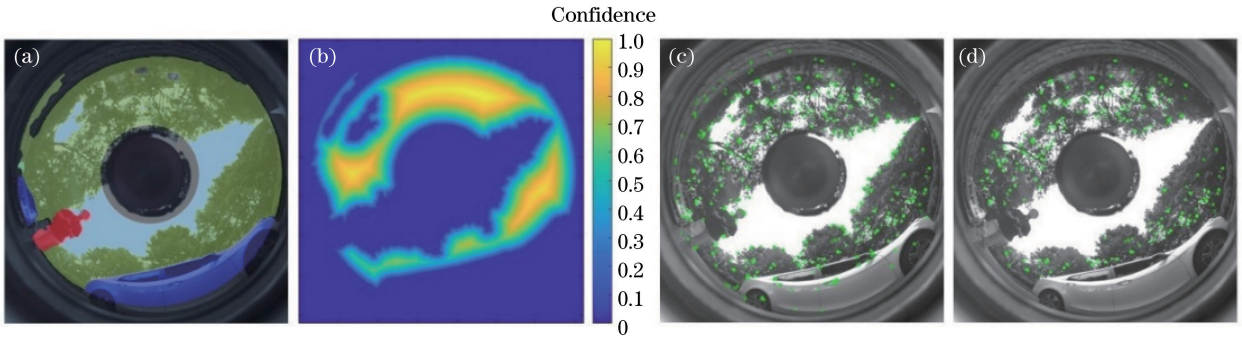

图 6 语义引导的关键点选择。(a)语义分割结果；(b)置信度分布；(c)原始的关键点检测；(d)语义引导的关键点检测

Fig. 6 Keypoints extraction under semantic guidance. (a) Semantic segmentation result; (b) confidence distribution; (c) original keypoints extraction; (d) keypoints extraction under semantic guidance

4 实 验

4.1 实验设置

为验证所提算法的性能，利用图 7(a)所示的遥控车在校园中采集了一系列真实场景数据集进行实验。遥控车上搭载了一款自主设计的具有 360°×(30°～90°)FOV 的 PAL 相机和采集数据用的微型电脑。同时，RealSense T265 传感器[34]也被搭载在遥控车前部，其输出的轨迹作为计算绝对位移误差(ATE)的参考值。为了与基于常规针孔相机的算法相比较，对 PAL 图像进行部分展开，形成具有 90°横向 FOV 的透视图像，如图 7(b)所示，作为基于针孔相机 VO 算法的输入。具体方法为：假设存在一个虚拟的针孔相机，其光心与 PAL 相机重合，且与 PAL 相机坐标系

之间的旋转为 $\boldsymbol{R}$；那么对于 PAL 图像上的某个像素 $\boldsymbol{u}_i$，其在透视图像上对应的像素坐标为

$$\boldsymbol{u}'_i=\boldsymbol{K}\cdot\boldsymbol{R}\cdot\pi^{-1}(\boldsymbol{u}_i),\tag{25}$$

式中：$\boldsymbol{K}$ 为虚拟针孔相机的内参矩阵。

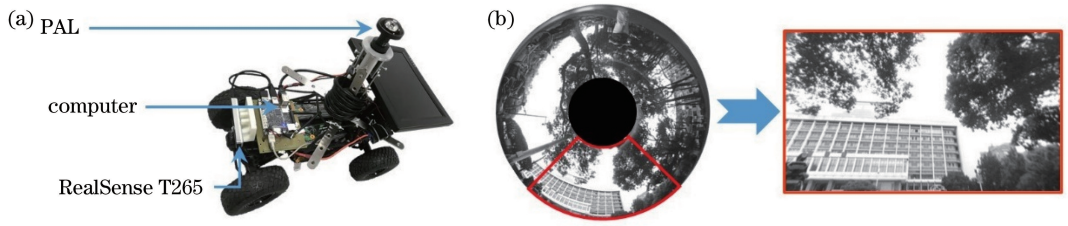

图 7　实验设置。(a)遥控车；(b) PAL 图像展开为透视图像

Fig. 7　Experiment setup. (a) Remote control vehicle; (b) PAL image expands to a perspective image

4.2　语义信息对筛除外点的作用

首先，实验验证了语义信息对筛除外点(outlier)的作用。在一段遥控车采集的含有大量行人、车辆等运动物体的 PAL 图像序列中，相隔固定数目(这里设置为 10)抽取一个关键帧。以相邻关键帧组成图像对，在前一帧上检测关键点，并利用 Lucas-Kanade 光流法追踪到下一帧；对于所有追踪成功的点对，利用八点法和随机采样一致性(RANSAC)计算本质矩阵，并统计其中内点(inlier)的比例。图 8(a)展示了采用语义引导的方法与不采用时内点率的对比。可以看出，所提语义引导的关键点选择策略使平均内点率由 63% 左右上升到了 80%。

在不同内点率下，RANSAC 收敛的概率 $\mathscr{P}_{\text{converge}}$ 与采样轮数 N' 的关系为

$$\mathscr{P}_{\text{converge}}=1-[1-(\mathscr{R}_{\text{inlier}})^8]^{N'},\tag{26}$$

式中：$\mathscr{R}_{\text{inlier}}$ 代表内点率。图 8(b)展示了内点率提升对 RANSAC 收敛概率的影响。可以看出，采用语义引导的关键点选择，在相同的 RANSAC 采样轮数时，大大增加了收敛的概率。

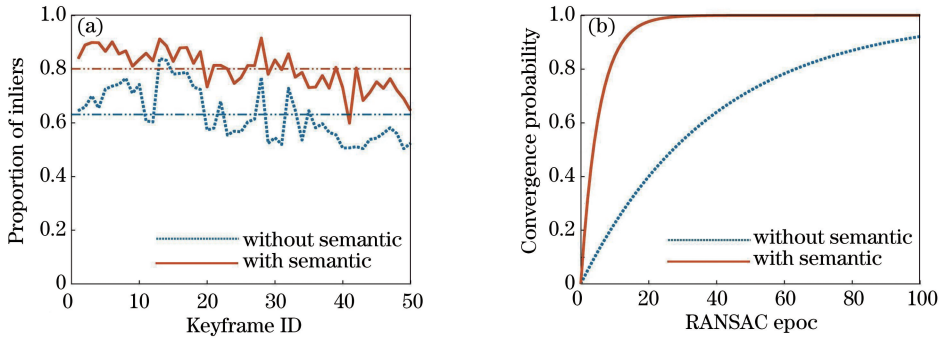

图 8　语义引导的关键点选择策略有效性测试。(a)有/无语义引导情况下的内点率；(b)有/无语义引导情况下 RANSAC 收敛的概率

Fig. 8　Effectiveness test of keypoints extraction under semantic guidance. (a) Proportion of inliers with/without semantic guidance; (b) RANSAC convergence probability with/without semantic guidance

4.3　精度测试

为了验证所提 PASVO 的精度，在真实场景数据上进行了一系列精度测试，并与同样基于全景图像的 PALVO[35] 和 CubemapSLAM[11]，基于透视图像的 SVO[3] 和 ORBSLAM2[4] 进行了对比研究。

用遥控车在室外采集了数个视野内含有大量行人、车辆等运动物体的视频，路径长度为 9～63 m。在原始图像上运行 PASVO、PALVO 和 CubemapSLAM，在展开的针孔相机透视图像上运行 SVO 和 ORBSLAM2。5 种算法所估计的轨迹与参考轨迹的误差情况如表 1 所示。可以看出：在多数数据上，基于全景图像的 3 种算法普遍比基于透视图像的 SVO 和 ORBSLAM2 的误差要小；在 S1，S2，S6，S8 和 S9 上，PASVO 都取得了最优的结果；在 S4 上，PASVO 的结果仅次于 PALVO，S7 上仅次于 CubemapSLAM；只在 S3 和 S5 上，基于透视图像的 ORBSLAM2 取得最优结果；此外，SVO 在 S3～S6 上没有运行成功。图 9 分别展示了 PASVO 在 S1(63 m)、S5(25 m)、S6(43 m)上所估计的路径，色度条表示与参考路径之间的误差。

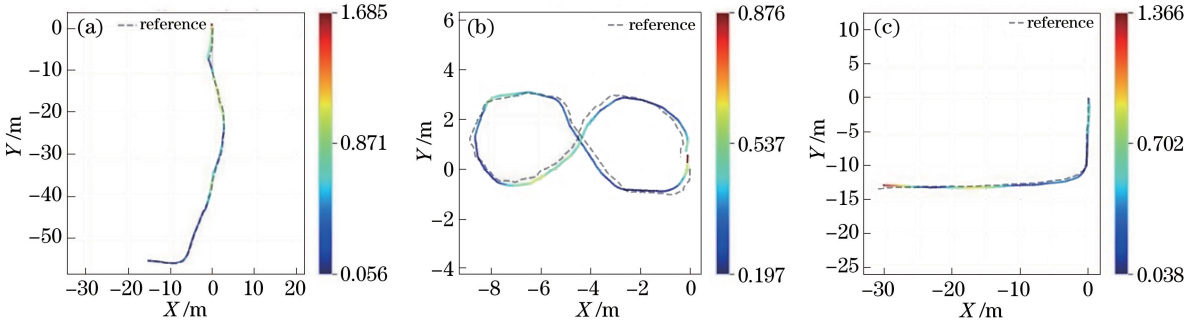

图 9 精度测试中 PASVO 所估计的路径与参考路径之间的误差。(a) S1;(b) S5;(c) S6

Fig. 9 Error between trajectory estimated by PASVO and reference trajectory in the accuracy test. (a) S1; (b) S5; (c) S6

表 1 绝对位移误差

Table 1 Absolute translation error

Dataset	Length /m	Absolute translation error /m				
		PASVO	PALVO[35]	CubemapSLAM[11]	SVO[3]	ORBSLAM2[4]
S1	63	0.6668	0.6785	0.6865	0.7350	1.4700
S2	60	0.4051	0.4373	0.4998	0.6276	0.4859
S3	9	0.1597	0.1937	0.4282		0.1513
S4	16	0.2174	0.2168	0.2405		0.3293
S5	25	0.3906	0.3943	0.4038		0.3727
S6	43	0.5451	0.5986	0.6269		0.6840
S7	19	0.0880	0.1196	0.0730	0.1178	0.1233
S8	18	0.2524	0.2616	0.2732	0.4372	0.3445
S9	41	1.4294	1.5859	1.8271	2.0401	1.7519

4.4 大尺度实际场景测试

为了检验在更大的运行尺度上所提算法的累积误差情况，采集了数个路径长度为 190～250 m，且路径起点和终点在同一位置(即路径上形成闭环)的图像序列。在这些数据上分别运行 PASVO、PALVO、CubemapSLAM、SVO 和 ORBSLAM2，并统计所估计路径的“闭环误差”：

$$e_{\text{loop}}=\frac{P_{\text{start}}-P_{\text{end}}}{L_{\text{traj}}}\times 100\%, \tag{27}$$

式中：P_{start} 和 P_{end} 分别代表算法所估计路径的起点和终点；L_{traj} 代表算法估计路径的总长度。闭环误差结果如表 2 所示。可以看出：基于全景图像的 3 种算法在所有数据上均可以成功运行，其中所提算法在“Fountain”和“Library”两个数据上闭环误差最小，在“Caolou”数据上稍逊于 CubemapSLAM；相比之下，SVO 和 ORBSLAM2 分别在“Fountain”和“Library”数据上运行失败，SVO 在“Fountain”路径的转弯处未能成功跟踪到关键点，而 ORBSLAM2 在“Library”数据上发生了严重的尺度漂移。需要说明的是，虽然路径的起点和终点在相同的位置，但是遥控车在起终点的朝向却并不相同：在“Fountain”数据中，遥控车在开始和结束时的朝向完全相反；在“Caolou”数据中，则是互相垂直的。在这种情况下，针孔相机仅具有遥控车运行前方的有限 FOV，所以透视图像的外观完全不同，故而 ORBSLAM2 中的闭环检测未能被成功触发。不同算法所估计的路径如图 10 所示。

表 2 大尺度数据集上的闭环误差

Table 2 Loop closure error in the large-scale dataset

Dataset	Length /m	Loop closure error /%				
		PASVO	PALVO[35]	CubemapSLAM[11]	SVO[3]	ORBSLAM2[4]
Caolou	190	1.1260	2.5719	0.9989	3.6702	1.2254
Fountain	200	1.0469	4.2288	4.3081		6.3477
Library	250	1.8031	3.6310	4.0794	6.5322	

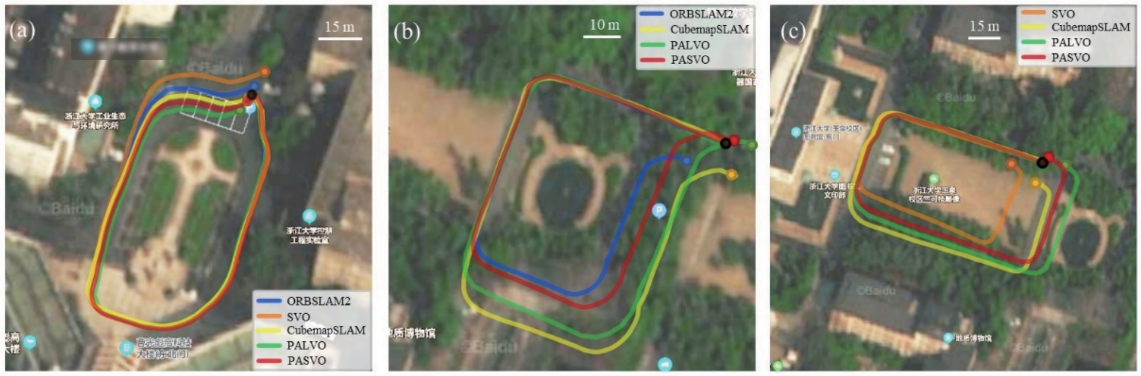

图 10 大尺度数据上不同算法所估计的路径。黑色圆点代表起点，不同颜色的圆点代表不同算法所估计路径的终点

Fig. 10 Trajectories estimated by the different algorithms on the large-scale dataset. The black dot represents the starting point, and dots with different colors denote the end points of the trajectory estimated by different algorithms

5 结 论

提出全景环带语义视觉里程计算法 PASVO，将全景环带成像系统应用到视觉里程计，并通过语义引导关键点选择和极曲线搜索，以及在位姿优化中加入语义校验的方式，将全景环带语义分割所提供的语义信息耦合入视觉里程计的各个模块。实验结果表明，所提 PASVO 在实际应用环境下实现了更加鲁棒的位姿估计，同时精度也大幅提升。基于本工作，未来可以开展关于全景语义稠密建图的研究。

参 考 文 献

[1] Scaramuzza D, Fraundorfer F. Visual odometry tutorial[J]. IEEE Robotics & Automation Magazine, 2011, 18(4): 80-92.

[2] Engel J, Koltun V, Cremers D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 611-625.

[3] Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry [C] // 2014 IEEE International Conference on Robotics and Automation (ICRA), May 31-June 7, 2014, Hong Kong, China. New York: IEEE Press, 2014: 15-22.

[4] 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.

[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] Huang Z, Bai J, Hou X Y. Design of combination of panoramic and long focal length optical system with single sensor[J]. Acta Optica Sinica, 2013, 33(4): 0422006.
黄治, 白剑, 侯西云. 单传感器全景与长焦组合光学系统设计[J]. 光学学报, 2013, 33(4): 0422006.

[7] Zhang X F, Liu J, Shi Z S, et al. Review of deep learning-based semantic segmentation[J]. Laser & Optoelectronics Progress, 2019, 56(15): 150003.
张祥甫, 刘健, 石章松, 等. 基于深度学习的语义分割问题研究综述[J]. 激光与光电子学进展, 2019, 56(15): 150003.

[8] Lemaire T, Lacroix S. SLAM with panoramic vision [J]. Journal of Field Robotics, 2007, 24(1/2): 91-111.

[9] Rituerto A, Puig L, Guerrero J J. Visual SLAM with an omnidirectional camera [C] // 2010 20th International Conference on Pattern Recognition, August 23-26, 2010, Istanbul, Turkey. New York: IEEE Press, 2010: 348-351.

[10] Pretto A, Menegatti E, Pagello E. Omnidirectional dense large-scale mapping and navigation based on meaningful triangulation [C] // 2011 IEEE International Conference on Robotics and Automation, May 9-13, 2011, Shanghai, China. New York: IEEE Press, 2011: 3289-3296.

[11] Wang Y H, Cai S J, Li S J, et al. CubemapSLAM: a piecewise-pinhole monocular fisheye SLAM system [M] // Jawahar C V, Li H D, Mori G, et al. Computer vision-ACCV 2018. Lecture notes in computer science. Cham: Springer, 2019, 11366: 34-49.

[12] Forster C, Zhang Z C, Gassner M, et al. SVO: semidirect visual odometry for monocular and multicamera systems [J]. IEEE Transactions on Robotics, 2017, 33(2): 249-265.

[13] Matsuki H, von Stumberg L, Usenko V, et al. Omnidirectional DSO: direct sparse odometry with fisheye cameras[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3693-3700.

[14] 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: 1-17.

[15] Seok H, Lim J. ROVO: robust omnidirectional visual odometry for wide-baseline wide-FOV camera systems [C] // 2019 International Conference on Robotics and Automation (ICRA), May 20-24, 2019, Montreal, QC, Canada. New York: IEEE Press, 2019: 6344-6350.

[16] Won C, Seok H, Cui Z P, et al. OmniSLAM: omnidirectional localization and dense mapping for wide-baseline multi-camera systems [C] // 2020 IEEE International Conference on Robotics and Automation (ICRA), May 31-August 31, 2020, Paris, France. New York: IEEE Press, 2020: 559-566.

[17] Lin M J, Cao Q X, Zhang H R. PVO: panoramic visual odometry [C] // 2018 3rd International Conference on Advanced Robotics and Mechatronics (ICARM), July 18-20, 2018, Singapore, Singapore. New York: IEEE Press, 2018: 491-496.

[18] Ji S P, Qin Z J. Panoramic SLAM for multi-camera rig [J]. Acta Geodaetica et Cartographica Sinica, 2019, 48(10): 1254-1265.
季顺平, 秦梓杰. 多镜头组合式相机的全景 SLAM [J]. 测绘学报, 2019, 48(10): 1254-1265.

[19] Liang H J, Sanket N J, Fermüller C, et al. SalientDSO: bringing attention to direct sparse odometry [J]. IEEE Transactions on Automation Science and Engineering, 2019, 16(4): 1619-1626.

[20] Ganti P. SIVO: semantically informed visual odometry and mapping [D]. Canada: University of Waterloo, 2018.

[21] Murali V, Chiu H P, Samarasekera S, et al. Utilizing semantic visual landmarks for precise vehicle navigation [C] // 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), October 16-19, 2017, Yokohama, Japan. New York: IEEE Press, 2017.

[22] Yu C, Liu Z X, Liu X J, et al. DS-SLAM: a semantic visual SLAM towards dynamic environments [C] // 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 1-5, 2018, Madrid, Spain. New York: IEEE Press, 2018: 1168-1174.

[23] Xiao L, Wang J, Qiu X, et al. Dynamic-SLAM: semantic monocular visual localization and mapping based on deep learning in dynamic environment [J]. Robotics and Autonomous Systems, 2019, 117: 1-16.

[24] Lu J, Liu Y H, Zhang R F. Semantic-based visual odometry towards dynamic scenes [J]. Laser & Optoelectronics Progress, 2021, 58(6): 061101.
卢金, 刘宇红, 张荣芬. 面向动态场景的语义视觉里程计 [J]. 激光与光电子学进展, 2021, 58(6): 061101.

[25] Bowman S L, Atanasov N, Daniilidis K, et al. Probabilistic data association for semantic SLAM [C] // 2017 IEEE International Conference on Robotics and Automation (ICRA), May 29-June 3, 2017, Singapore. New York: IEEE Press, 2017: 1722-1729.

[26] Lianos K N, Schönberger J L, Pollefeys M, et al. VSO: visual semantic odometry [M] // Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11208: 246-263.

[27] Zou B, Lin S Y, Yin Z S. Semantic mapping based on YOLOv3 and visual SLAM [J]. Laser & Optoelectronics Progress, 2020, 57(20): 201012.
邹斌, 林思阳, 尹智帅. 基于 YOLOv3 和视觉 SLAM 的语义地图构建 [J]. 激光与光电子学进展, 2020, 57(20): 201012.

[28] Scaramuzza D, Martinelli A, Siegwart R. A flexible technique for accurate omnidirectional camera calibration and structure from motion [C] // Fourth IEEE International Conference on Computer Vision Systems (ICVS′06), January 4-7, 2006, New York, NY, USA. New York: IEEE Press, 2006: 45.

[29] Scaramuzza D, Martinelli A, Siegwart R. A toolbox for easily calibrating omnidirectional cameras [C] // 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 9-15, 2006, Beijing, China. New York: IEEE Press, 2006: 5695-5701.

[30] Yang K L, Hu X X, Bergasa L M, et al. PASS: panoramic annular semantic segmentation [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(10): 4171-4185.

[31] Yang K L, Hu X X, Chen H, et al. DS-PASS: detail-sensitive panoramic annular semantic segmentation through SwaftNet for surrounding sensing [C] // 2020 IEEE Intelligent Vehicles Symposium (IV), October 19-November 13, 2020, Las Vegas, NV, USA. New York: IEEE Press, 2020: 457-464.

[32] Yang K L, Wang K W, Bergasa L M, et al. Unifying terrain awareness for the visually impaired through real-time semantic segmentation [J]. Sensors, 2018, 18(5): 1506.

[33] Baker S, Matthews I. Lucas-kanade 20 years on: a unifying framework [J]. International Journal of

Computer Vision, 2004, 56(3): 221-255.

[34] Bayer J, Faigl J. On autonomous spatial exploration with small hexapod walking robot using tracking camera intel RealSense T265 [C] // 2019 European Conference on Mobile Robots (ECMR), September 4-6, 2019, Prague, Czech Republic. New York: IEEE Press, 2019.

[35] Chen H, Wang K, Hu W, et al. PALVO: visual odometry based on panoramic annular lens[J]. Optics Express, 2019, 27(17): 24481-24497.