arXiv:2303.00371v1 [cs.RO] 1 Mar 2023

# 基于人工智能的多目标相对状态估计与自校准能力

Thomas Jantos[1], Christian Brommer[1],Eren Allak[1],Stephan Weiss[1] 和 Jan Steinbrener[1]

**摘要——从原始传感器数据中提取特定任务所需的语义信息，是许多移动机器人应用的关键要求。例如，使用无人驾驶飞行器（无人机）对关键基础设施进行自主巡检，需要相对于待检查结构进行精确导航。近年来，基于人工智能的方法已展现出卓越能力，能够从图像中提取语义信息，例如物体的六自由度（6-DoF）位姿。**

**本文提出一种方法，将用于相机图像中物体的先进基于人工智能的位姿估计器与惯性测量单元（惯性测量单元）的数据相结合，以实现移动机器人的六自由度多目标相对状态估计。该位姿估计器可在相机图像中检测多个感兴趣物体，并估计其相对位姿。随后，利用先进的传感器融合框架，将这些测量结果与惯性测量单元数据进行融合。我们通过针对不同轨迹和数量各异、任意放置的物体开展真实世界实验，展示了所提方法的可行性。实验结果表明，得益于该方法的自校准能力，结果能够可靠地复现。**

## I. 引言

移动机器人（如无人机）依靠机载传感器提供的信息自主地在世界坐标系中导航。语义信息，即传感器数据的高层含义，可以提升机器人在周围环境中的导航能力，并使其能够执行更复杂的任务 [1]。在语义导航中，机器人会根据环境背景或任务需求移动，许多情况下还需要以场景中的感兴趣物体为参照。这类任务包括基础设施巡检 [2]或目标跟踪 [3]。后者的目标是使移动物体保持在相机的视场内，而基础设施巡检则要求机器人相对于通常静止的目标物体进行精确定位。从机器人传感器数据中提取的语义信息，即检测目标物体及其相对于机器人的位姿，是完成这项任务的重要环节。例如，监测电线杆绝缘子是否可能发生损坏，需要无人机绕目标绝缘子飞行，并从特定位置拍摄高分辨率图像，以便检测损坏或随时间发生的变化。

当前的自主任务执行通常依赖全球导航卫星系统（GNSS）进行无人机定位

[1]作者来自克拉根福大学网络化系统控制组，地址：奥地利克拉根福9020，克拉根福firstname.lastname@ieee.org本研究由联邦气候行动、环境、能源、交通、创新与技术部（BMK）通过资助协议881082（MUKISANO）提供支持。**预印本版本，2023年1月接受，DOI 将尽快公布 ©IEEE。**

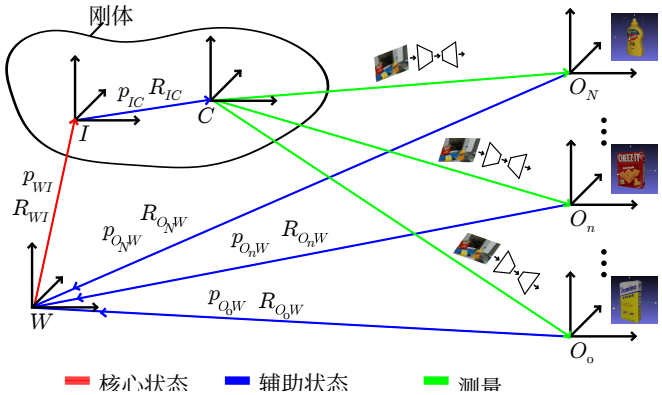


图1. 本文所用坐标系的示意图。我们估计一个固定刚体的状态，该刚体由惯性测量单元$I$和相机$C$组成，且相对于最多$N$个不同物体$O_k$，并以一个固定但任意选取的导航世界坐标系$W$为参考。除核心状态（红色）外，我们还估计惯性测量单元与相机之间的标定参数（蓝色）。我们还估计物体坐标系相对于世界坐标系的位姿（蓝色）。我们的位姿传感器由相机与物体之间基于人工智能的六自由度相对位姿测量值组成（绿色）。

。GNSS 始终提供全局位置，而非相对于感兴趣物体的相对位置。此外，其精度不足以支持精确的厘米级导航，而且在大型结构附近容易发生信号丢失。在这种情况下，通常会考虑使用其他传感器模态，例如视觉惯性里程计（VIO） [4]。利用 VIO，可以通过结合单目相机图像中几何特征（边缘、角点）的运动与惯性测量单元（IMU）的数据来估计局部位姿。传统的基于特征提取的算法不太适用于语义导航，因为它们依赖无法提供场景中任何物体信息的原始特征，并且难以应对快速或缓慢运动 [5]。

人工智能（AI）的最新进展推动了从原始传感器测量值中提取语义信息的突破，例如使用相机图像进行路径检测 [6], 使用激光扫描仪进行目标识别[7], 用于场景理解的语义分割 [8], 以及最近用于机器人抓取的物体六自由度（6-DoF）位姿估计 [9]。此外，具备人工智能能力的边缘计算设备的普及，使得此类方法能够应用于移动机器人。

本文探讨了基于人工智能的位姿估计用于移动机器人学中完整六自由度物体相对状态估计的适用性。我们考虑一种最简传感器配置，由单目相机和惯性测量单元组成，满足无人机等移动机器人平台在尺寸、重量和计算能力方面的限制。我们利用基于人工智能的位姿估计器检测、分类并

估计每幅相机图像中感兴趣物体的六自由度位姿，然后在先进的传感器融合框架中将这些信息与 IMU测量值融合，以推断机器人的六自由度物体相对位姿。图 1 展示了我们方法的示意图。我们的主要贡献可概括如下：

- 利用人工智能从图像中提取语义信息，并将这些相对位姿信息与惯性测量单元数据融合，实现精确的六自由度物体相对状态估计。
- 提出一种基于滤波器的方法，根据六自由度相对位姿测量值估计移动机器人的状态以及多个不同物体的位姿。
- 提出一种滤波器的自校准形式，无需对场景中不同物体的全局位置或相对位置作任何假设。
- 通过多项真实世界实验验证所提出的方法。实验采用一个广受欢迎的六自由度物体位姿挑战数据集中的物体，结果表明，该方法在不同轨迹和不同数量的物体下均能取得可复现的性能。

本文其余部分安排如下：第二节总结相关工作；第三节介绍如何将物体相对位姿测量融入状态估计框架；第四节讨论实验及相应结果；最后，第五节总结全文。

## II. 相关工作

对于移动机器人学中的状态估计，通常会将惯性测量单元数据与一个或多个位姿传感器（如 GNSS）融合。GNSS 可提供全局位置信息，但无法提供三维姿态信息。在没有 GNSS 信号的情况下，VIO（即单目相机与惯性测量单元数据的组合）可以利用图像中的几何特征，通过对相机位置进行三角测量来估计机器人的位姿，并使用惯性数据估计剩余的尺度因子 [4], [10]。融合多个传感器可以获得更加稳健、可靠的机器人状态估计。传感器融合主要有两种不同的方法：基于滤波的方法（递归方法）和基于优化的方法。后者能够获得更精确的状态估计，但由于需要对多个传感器测量值进行联合优化，其计算量更大 [11]。相比之下，基于滤波的方法（如扩展卡尔曼滤波器（EKF）） [12], [13],在计算上更加高效，因此非常适合移动机器人学。

GNSS 和经典视觉惯性里程计无法提供物体相对位姿测量，因此不适用于物体相对状态估计。不过，基于图像的6自由度相对物体位姿估计方法可作为状态估计框架中的位姿传感器。此类方法包括经典方法和基于深度学习的人工智能方法。经典方法分为基于模板的方法（通过为当前图像 [14], 查找匹配模板来确定物体位姿）和基于特征的方法（从图像中提取关键点，再与三维物体模型 [15]进行匹配）。另一方面，基于深度学习的方法大多是端到端学习方法，使用卷积神经网络（CNN）直接根据输入图像估计6自由度位姿。根据所使用的额外信息量，基于深度学习的方法还可进一步细分。 [9], [16] 将单张RGB图像作为网络输入，并在训练期间采用利用三维物体模型信息的对称性感知损失。也可以将三维物体模型作为额外输入提供给网络 [17], ，用于优化初始位姿估计 [18], [19], [20] ，或用于匹配网络回归得到的关键点[21]。其他形式的额外信息包括使用多张图像 [22], [23] 或深度图 [24], [25]。最近，我们提出了 PoET [26] ，这是一种6自由度多目标位姿估计框架，在基准数据集上取得了最先进的结果，并且仅将单张RGB图像作为输入，在训练和推理期间均无需任何额外信息。

物体相对状态估计的另一种方案是在物体层面进行同时定位与建图（SLAM）。[27] 使用深度图像提取六自由度物体位姿信息。与我们的方法类似， [28] 使用基于人工智能的位姿估计器，从图像中预测六自由度物体相对位姿。这两种方法都会融合来自多个视点的六自由度相对位姿信息，并结合图优化，估计相机与物体相对于地图的位姿。然而，这些方法都不使用其他传感器，例如惯性测量单元。 [29] 在SLAM方法中融合了IMU测量值、从图像中提取的几何特征以及六自由度物体位姿。对于包含所有物体位姿的图，仍需进行图优化。总的来说，SLAM对地图和优化的要求会增加移动机器人的计算负荷。我们的方法仍可在无需这些条件的情况下进行物体相对状态估计。

移动机器人学中的物体相对状态估计已在 [30], 中得到展示：通过从图像中提取几何特征并假设圆柱形基础设施的半径已知，无人机能够相对于该基础设施定位自身。类似地， [31] 使用基于颜色的椭圆检测算法，首先检测图像中的目标物体，然后利用物体尺寸、视觉外观和相机参数的信息，计算相对于该物体的位姿。与此同时， [32] 研究了用于物体相对状态估计的多种经典 6-DoF 物体位姿估计方法。他们还研究了通过在从图像中提取的经典特征上训练简单分类器，利用机器学习检测物体是否存在。

与此不同，我们在此提出一种完全基于人工智能的方法，从相机图像中提取语义信息。我们无需定义几何模型、关键点或模板，即可将图像中的物体外观映射到相对的6自由度位姿。此外，基于人工智能的模型不受特定物体几何形状的限制，也无需手工设计特征。在我们此前的工作 [26], 中，我们介绍了 PoET，这是一种采用先进人工智能方法、用于从RGB图像中估计物体6自由度位姿的位姿估计器。我们的工作主要

聚焦于针对6自由度多物体位姿估计，介绍方法的定义、训练过程、详尽的消融研究，以及在基准数据集上与其他基于深度学习的方法进行比较。在这项工作中，我们通过多项真实世界实验，结合先进的传感器融合框架，详细研究了我们的基于人工智能的物体位姿估计器作为位姿传感器，是否适用于移动机器人的6自由度物体相对状态估计。

## III. 方法

本节介绍我们方法的设计。首先，我们说明测量方程和坐标系变换所用的符号。其次，我们探讨选择何种框架进行 6自由度位姿估计和状态估计。最后，我们描述如何结合多个已知物体的估计 6 自由度信息，来估计机器人的 6自由度位姿。这包括详细说明如何扩展我们选择的传感器融合算法，使其能够纳入各个物体的 6自由度位姿测量。

### *A.* 符号

本文中，我们使用以下符号：给定三个坐标系 A、B 和 C，变换 $\mathrm{AT}_{BC}$表示在坐标系 A 中表达的、相对于坐标系 B 定义的坐标系 C。若省略左下标 A，则该变换在坐标系 B 中定义。此外，变换 $\mathrm{AT}_{AB}$可分为两部分，即 $\mathrm{Ap}_{AB}$ 和 $\mathrm{R}_{AB}$，分别表示平移和旋转。或者，也可以用四元数 $\mathrm{q}_{AB}$表示旋转。每个四元数 q 均可表示为 $\mathrm{q} = [\mathrm{q_v}\ \mathrm{q}_w]^T = [\mathrm{q}_x\ \mathrm{q}_y\ \mathrm{q}_z\ \mathrm{q}_w]^T$。四元数乘法用 $\otimes$表示。$\mathrm{I}_3$ 和 $\mathbf{0}_3$ 分别表示$\mathbb{R}^{3x3}$中的单位矩阵和零矩阵。 $[\omega]_\times$ 是反对称算子，其定义见 [33]。

### *B.* 位姿与状态估计框架

移动机器人，尤其是无人机，受到载荷限制的约束。这不仅限制了机器人可携带传感器的尺寸和数量，也限制了可用于数据处理的计算能力。因此，需要高效且计算开销低的算法。基于此，我们选择物体位姿估计框架 PoET [26]作为 6自由度位姿传感器，因为它只使用 RGB图像，不依赖深度信息和三维物体模型，既无需额外的硬件组件，也避免了处理三维模型所带来的计算负荷。首先，PoET 会检测图像中所有经过训练的目标，并预测其类别。随后，将预测的边界框和多尺度特征图输入 Transformer架构，以预测相机与每个物体之间相对且尺度不确定的 6自由度位姿。对于非对称物体，预测的旋转和平移是唯一的。对于具有一个或多个对称轴的物体，某些物体位姿的旋转或平移可能存在歧义，即有多个可能的解。这种歧义对机器人位姿估计造成的明显负面影响可以通过考虑异构配置中的多个物体，并在适当的传感器融合框架中融合各项测量值，最大限度地降低。这模拟了检查工作流程：通常，在检查结构时，多个不同的感兴趣部位会同时出现在视野中。

对于传感器融合框架，我们采用 MaRS [12]进行多传感器融合和状态估计，因为它轻量且计算效率高，并且专为移动机器人学而开发。MaRS 的设计注重模块化，将基于惯性数据的核心状态变量传播与基于各个传感器测量值的状态更新分离。它还使用与类型无关的抽象传感器类别，因此可以轻松集成新的传感器模块。对于我们的方法，我们定义了一种多位姿传感器，其单次测量由一幅 RGB图像组成。随后，我们使用 PoET 从每幅图像中提取所有检测到物体的六自由度相对位姿，并将其用于下文所述的 EKF 更新步骤。

### *C. EKF* 状态与更新

我们选择移动机器人系统的惯性测量单元（$I$）帧作为参考帧。因此，我们的目标是通过测量相机（$C$）与一组物体（$O_k$）之间的六自由度相对位姿，估计惯性测量单元（$I$）相对于世界坐标系（$W$）的位姿。不同坐标系如图 1 所示。如前所述，我们基于人工智能的位姿估计器 PoET 从图像中提取物体相对于相机的相对位姿。PoET 只会考虑训练过的物体。给定一幅 RGB图像，系统会为每个检测到的目标物体预测一个 6自由度位姿$\mathbf{T}_{CO_k}$，并根据预测的类别将预测位姿分配给相应物体。关于 PoET 架构选择的详情，请参阅 [26]。根据场景中的物体总数$N$，完整的状态向量$\mathbf{X}$ 定义如下：

$$\begin{aligned}\mathbf{X} = [&\mathbf{p}_{WI}^T, \mathbf{v}_{WI}^T, \mathbf{q}_{WI}^T, \mathbf{b}_\omega^T, \mathbf{b}_a^T, \\ &\mathbf{p}_{IC}^T, \mathbf{q}_{IC}^T, \mathbf{p}_{O_0W}^T, \mathbf{q}_{O_0W}^T, \ldots, \mathbf{p}_{O_NW}^T, \mathbf{q}_{O_NW}^T]^T\end{aligned} \tag{1}$$

状态传播所需的核心状态包括$\mathbf{p}_{WI}$ IMU的位置、其速度$\mathbf{v}_{WI}$ 和姿态$\mathbf{q}_{WI}$，以及陀螺仪偏置$\mathbf{b}_\omega$ 和加速度计偏置$\mathbf{b}_a$。位姿和速度动力学表示为 [4]

$$\dot{\mathbf{p}}_{WI} = \mathbf{v}_{WI} \tag{2}$$

$$\dot{\mathbf{v}}_{WI} = \mathbf{R}_{WI}\ (\mathbf{a}_m - \mathbf{b}_a - \mathbf{n}_a) - \mathbf{g} \tag{3}$$

$$\dot{\mathbf{q}}_{WI} = \frac{1}{2}\Omega(\omega_b - \mathbf{b}_\omega - \mathbf{n}_\omega)\ \mathbf{q}_{WI} \tag{4}$$

其中，$\mathbf{a}_m$ 表示在IMU坐标系中测得的加速度，$\mathbf{n}_a$ 表示加速度计噪声参数，$\mathbf{g}$ 表示$W$坐标系中的重力向量，$\omega_b$ 表示在IMU坐标系中测得的角速度，$\mathbf{n}_\omega$ 表示陀螺仪噪声参数，而 $\Omega(\omega)$ 是 $\omega$ 的四元数乘法矩阵。IMU偏置建模为随机游走。

此外，我们还估计惯性测量单元与相机之间的标定参数，由$\mathbf{p}_{IC}$ 和$\mathbf{q}_{IC}$表示。由于实现

方面的原因，我们假设场景中的物体数量是先验已知的，但物体的全局位姿以及物体之间的相对位姿均未知。因此，我们还估计一个物体-世界坐标变换，用于描述物体坐标系与导航世界坐标系之间的变换（$\mathbf{p}^{o_k W}$，$\mathbf{q}^{o_k W}$）。相机-IMU外参$\mathbf{p}_{IC}$，$\mathbf{q}_{IC}$ 以及物体位姿$\mathbf{p}^{o_k W}$，$\mathbf{q}^{o_k W}$均被建模为随时间保持恒定。

对于每幅图像，每个测得的相对位姿都被视为独立测量，并据此估计相机的位姿。为计算更新步骤所需的雅可比矩阵，我们考虑逆相对位姿测量 $\mathbf{T}^{o_k C}$，即相机相对于物体坐标系的位姿。根据不同坐标系之间的关系，以及独立的相对位置 $z^{\mathbf{p}o_k}$ 和姿态 $z^{\mathbf{q}o_k}$ 测量值，位置 $\tilde{z}^{\mathbf{p}o_k}$ 和姿态 $\tilde{z}^{\mathbf{R}o_k}$ 的残差可写为：

$$\begin{aligned}\tilde{z}_{\mathbf{p}o_k} &= z_{\mathbf{p}o_k} - \hat{z}_{\mathbf{p}o_k} \\ &= \mathbf{p}_{o_k C} - (\mathbf{p}_{o_k W} + \mathbf{R}_{o_k W}(\mathbf{p}_{WI} + \mathbf{R}_{WI}\ \mathbf{p}_{IC})) \end{aligned} \tag{5}$$

$$\tilde{z}_{\mathbf{R}o_k} = 2\ \frac{\tilde{z}_{\mathbf{q}\mathbf{v},o_k}}{\tilde{z}_{q_w,o_k}} \tag{6}$$

$$\tilde{z}_{\mathbf{q}o_k} = \hat{z}^{-1}_{\mathbf{q}o_k} \otimes z_{\mathbf{q}o_k} = (\mathbf{q}_{o_k W} \otimes \mathbf{q}_{WI} \otimes \mathbf{q}_{IC})^{-1} \otimes \mathbf{q}_{o_k C} \tag{7}$$

$$\tilde{z}_{O_k} = \begin{bmatrix} \tilde{z}_{\mathbf{p}o_k} \\ \tilde{z}_{\mathbf{R}o_k} \end{bmatrix} \tag{8}$$

给定这些残差，并根据来自物体 $O_k$ 的单个位姿测量，位置$H_{\mathbf{p}}$ 和姿态 $H_{\mathbf{R}}$ 相对于状态的雅可比矩阵为 [33]:

$$H_{\mathbf{p},\mathbf{p}_{WI}} = \mathbf{R}_{o_k W} \tag{9}$$

$$H_{\mathbf{p},\mathbf{R}_{WI}} = -\mathbf{R}_{o_k W}\mathbf{R}_{WI}[\mathbf{p}_{IC}]_\times \tag{10}$$

$$H_{\mathbf{p},\mathbf{p}_{IC}} = \mathbf{R}_{o_k W}\mathbf{R}_{WI} \tag{11}$$

$$H_{\mathbf{p},\mathbf{p}_{o_k W}} = \mathbf{I}_3 \tag{12}$$

$$H_{\mathbf{p},\mathbf{R}_{o_k W}} = -\mathbf{R}_{o_k W}[\mathbf{p}_{WI}]_\times - \mathbf{R}_{o_k W}[\mathbf{R}_{WI}\ \mathbf{p}_{IC}]_\times \tag{13}$$

$$H_{\mathbf{R},\mathbf{R}_{WI}} = \mathbf{R}^T_{IC} \tag{14}$$

$$H_{\mathbf{R},\mathbf{R}_{IC}} = \mathbf{I}_3 \tag{15}$$

$$H_{\mathbf{R},\mathbf{R}_{o_k W}} = \mathbf{R}^T_{IC}\mathbf{R}^T_{WI} \tag{16}$$

其中，例如，$H_{\mathbf{p},\mathbf{p}_{WI}}$ 仅考虑残差$\tilde{z}^{\mathbf{p}o_k}$ 中依赖于状态 $\mathbf{p}_{WI}$的部分。其余雅可比矩阵均等于 $\mathbf{0}_3$。由于不同物体的相对位姿测量彼此独立，因此其他（$i \neq n$）个物体-世界状态对应的雅可比矩阵，即 $H_{\mathbf{p},\mathbf{p}_{o_i W}}$、$H_{\mathbf{p},\mathbf{R}_{O_i W}}$、$H_{\mathbf{R},\mathbf{p}_{o_i W}}$、$H_{\mathbf{R},\mathbf{R}_{O_i W}}$，均等于 $\mathbf{0}_3$。对于单个物体 $O_k$，其雅可比矩阵由以下各分量堆叠而成：

$$\begin{aligned}H_{\mathbf{p},O_k} = [&H_{\mathbf{p},\mathbf{p}_{WI}}\ H_{\mathbf{p},\mathbf{v}_{WI}}\ H_{\mathbf{p},\mathbf{R}_{WI}}\ H_{\mathbf{p},\mathbf{b}_\omega}\ H_{\mathbf{p},\mathbf{b}_a} \\ &H_{\mathbf{p},\mathbf{p}_{IC}}\ H_{\mathbf{p},\mathbf{R}_{IC}}\ H_{\mathbf{p},\mathbf{p}_{O_0 W}}\ H_{\mathbf{p},\mathbf{R}_{O_0 W}} \\ &\ldots H_{\mathbf{p},\mathbf{p}_{O_N W}}\ H_{\mathbf{p},\mathbf{R}_{O_N W}}]\end{aligned} \tag{17}$$

$$\begin{aligned}H_{\mathbf{R},O_k} = [&H_{\mathbf{R},\mathbf{R}_{WI}}\ H_{\mathbf{R},\mathbf{v}_{WI}}\ H_{\mathbf{R},\mathbf{R}_{WI}}\ H_{\mathbf{R},\mathbf{b}_\omega}\ H_{\mathbf{R},\mathbf{b}_a} \\ &H_{\mathbf{R},\mathbf{p}_{IC}}\ H_{\mathbf{R},\mathbf{R}_{IC}}\ H_{\mathbf{R},\mathbf{p}_{O_0 W}}\ H_{\mathbf{R},\mathbf{R}_{O_0 W}} \\ &\ldots H_{\mathbf{R},\mathbf{p}_{O_N W}}\ H_{\mathbf{R},\mathbf{R}_{O_N W}}]\end{aligned} \tag{18}$$

$$H_{O_k} = \begin{bmatrix} H_{\mathbf{p},O_k} \\ H_{\mathbf{R},O_k} \end{bmatrix} \tag{19}$$

根据当前图像，对当前更新步骤中检测到的每个物体，分别将式 (8) 和式 (19) 中的残差与雅可比矩阵按垂直方向堆叠，从而确定用于状态更新的最终残差 $z$ 和观测矩阵 $\mathbf{H}$ 。与硬件传感器类似，我们的基于人工智能的位姿传感器也可能返回错误或不准确的测量值。类似于 [34],中所述的方法，我们基于 EKF 新息$\mathbf{S}$ 和残差进行 $\chi^2$ 检验，以检测异常值测量。该检验会针对每个物体的测量值分别执行。随后，拒绝物体 $O_k$ 的异常值测量，并通过屏蔽相应的行，重新构建最终残差和雅可比矩阵。然后，根据最终总残差及其对应的雅可比矩阵计算校正量。在更新步骤中，可以考虑每项测量的测量不确定性。在本研究中，每个物体的位置和姿态测量的不确定性分别固定为 10 厘米和 20 度。这些数值根据 [26]中报告的 PoET 在一个视频序列上的标准差确定。

各个帧的正确初始化是需要考虑的重要方面。在记录开始时，我们初始化一个任意但固定的导航世界坐标系 $W$。不失一般性，将惯性测量单元坐标系初始化于 $W$ 的原点。惯性测量单元与相机之间的初始外参标定通过视觉惯性标定[35]确定。当相机首次观测到相应物体时，便对每个物体-世界坐标系进行初始化。随后，依据当前估计的相机位姿和相对位姿测量值，将物体坐标系置于相对于世界坐标系的位置：

$$\mathbf{R}_{o_k W} = \mathbf{R}_{o_k C}\mathbf{R}^T_{IC}\mathbf{R}^T_{WI} \tag{20}$$

$$\mathbf{p}_{o_k W} = \mathbf{p}_{o_k C} - \mathbf{R}_{o_k W}(\mathbf{R}_{WI}\ \mathbf{p}_{IC} + \mathbf{p}_{WI}) \tag{21}$$

在我们的问题表述中，机器人位姿$I$ 是通过相对位姿测量相对于一组物体世界$O_k$ 进行估计的。由于机器人位姿是相对于世界坐标系定义的，而测量值是相对于相应物体坐标系定义的，因此物体世界可以相对于世界坐标系自由放置，这会导致可观测性问题。通过固定一个物体世界参考坐标系的状态，系统即可变为可观测。这个固定的物体从现在起称为主物体$O_m$，并作为物体相对六自由度状态估计的锚点。主物体坐标系相对于导航世界坐标系的位置保持不变，即相应的雅可比矩阵$H_{\mathbf{p},\ \mathbf{p}_{o_m W}}$ 和 $H_{\mathbf{R},\ \mathbf{p}_{o_m W}}$ 设为 $\mathbf{0}_3$。如果图像中看不到主物体，则直接拒绝相应的测量值。否则，在锚点不可见时估计其他物体的物体世界会导致模糊更新。

传播步骤以IMU传感器读数的频率执行，而更新步骤则以相机图像的频率进行。




图2. 序列4中使用的物体配置（左）以及 PoET 估计的物体位姿（右）。请注意，真实世界物体（左）与用于训练 PoET 的 YCB-V 物体在外观颜色上的差异。如图所示，左侧图像直接输入我们的位姿估计框架，以获得六自由度相对位姿测量。

## IV. 实验与结果

本节介绍所开展的实验，并详细讨论实验结果。我们在YCB-V数据集上训练了PoET [9]，如 [26], 所述，该数据集是用于6自由度位姿估计的基准数据集；我们选取了其中一部分物体，作为实验中的感兴趣物体。图像和惯性测量单元数据由Intel Realsense D435i记录，其RGB分辨率为1280x720，帧率为30。对图像进行去畸变后，将其裁剪为640x480分辨率，这是YCB-V数据集的标准分辨率。我们将物体放置在动作捕捉室内，并在跟踪相机机身的同时用相机围绕物体移动，以此记录我们自己的真实数据。这模拟了使用具备6自由度的移动机器人平台对一组感兴趣物体进行检查的过程。图2展示了一个物体配置示例及其图像。由于我们没有记录物体在房间中的全局位置相关信息，因此必须将通过物体相对状态估计得到的轨迹与真实轨迹对齐，才能计算误差指标。需要注意基准数据与我们自己的真实数据之间的差异。用于记录YCB-V数据集的相机在记录分辨率、视场和焦点方面均不同于实验中使用的相机。此外，一些真实世界中的物体在外观上（尺寸和颜色）与用于生成PoET训练数据集的物体略有不同。此处报告的结果均使用在YCB-V数据集上训练的PoET模型获得。我们未对模型进行重新训练或微调，以适应这些差异。

我们进行了两项不同的实验。首先，我们研究了所提方法在 8 个不同序列中的性能。这些序列在物体数量、物体的排列方式以及相机围绕物体运动的轨迹方面各不相同。对于每段序列的单次记录，我们通过比较估计轨迹与相应的真实轨迹，计算位置和欧拉角的均方根误差（RMSE）。此外，我们还计算所有轨迹的平均 RMSE 和标准差（std）。结果汇总于表I。所有序列的总体性能表明，我们的方法能够

表I 不同序列的整个轨迹上位置和姿态的均方根误差。

| 序列 | #物体 | (x, y, z) [米] | (滚转角, 俯仰角, 偏航角) [deg] |
|---|---|---|---|
| 1 | 3 | [0.066、0.163、0.034] | [3.00, 6.81, 4.52] |
| 2 | 3 | [0.162, 0.358, 0.233] | [54.38, 19.39, 20.27] |
| 3 | 2 | [0.163, 0.338, 0.136] | [25.34, 16.57, 15.13] |
| 4 | 4 | [0.045, 0.157, 0.108] | [15.38, 5.58, 10.02] |
| 5 | 1 | [0.075, 0.031, 0.054] | [61.12, 9.66, 26.19] |
| 6 | 3 | [0.061, 0.117, 0.029] | [11.00, 5.77, 7.41] |
| 7 | 4 | [0.093, 0.159, 0.099] | [19.38, 12.38, 27.24] |
| 8 | 3 | [0.127, 0.165, 0.104] | [26.19, 11.95, 10.70] |
| mean | - | [0.099, 0.186, 0.099] | [26.97, 11.0, 15.19] |
| ± 标准差 | | ± [0.043, 0.102, 0.062] | ± [19.18, 4.75, 8.01] |

表II 不同运行中 sequence4 整条轨迹的位置和姿态均方根误差。

| Run | (x, y, z) [米] | (滚转角, 俯仰角, 偏航角) [deg] |
|---|---|---|
| 1 | [0.094, 0.107, 0.119] | [4.72, 7.08, 13.16] |
| 2 | [0.077、0.121、0.103] | [1.33, 7.24, 8.05] |
| 3 | [0.049, 0.096, 0.092] | [3.14、1.72、14.49] |
| 4 | [0.065、0.119、0.101] | [7.75, 5.25, 12.90] |
| 5 | [0.069, 0.119, 0.099] | [2.79, 3.35, 11.57] |
| 6 | [0.047、0.099、0.099] | [2.53, 3.63, 12.21] |
| 7 | [0.059, 0.090, 0.088] | [1.92, 5.01, 10.89] |
| 8 | [0.057, 0.093, 0.083] | [3.68, 2.92, 11.37] |
| 9 | [0.074、0.090、0.096] | [6.85, 4.69, 13.11] |
| 10 | [0.076, 0.101, 0.118] | [7.19, 6.02, 12.10] |
| mean | [0.067、0.104、0.010] | [4.19、4.69、11.99] |
| ± 标准差 | ± [0.014, 0.012, 0.011] | ±[2.20, 1.71, 1.64] |

表III
Sequence6不同运行中，整个轨迹的位置和姿态均方根误差。

| Run | (x, y, z) [米] | (滚转角, 俯仰角, 偏航角) [度] |
|---|---|---|
| 1 | [0.061, 0.117, 0.029] | [11.00, 5.77, 7.41] |
| 2 | [0.092, 0.122, 0.052] | [17.17, 16.97, 7.16] |
| 3 | [0.072, 0.112, 0.025] | [13.08, 24.64, 5.93] |
| 4 | [0.083, 0.110, 0.037] | [12.11, 8.24, 4.58] |
| 5 | [0.074, 0.089, 0.023] | [16.02, 5.99, 6.15] |
| 6 | [0.067、0.079、0.031] | [12.29, 11.97, 7.19] |
| 7 | [0.064、0.126、0.025] | [16.64, 12.26, 7.72] |
| 8 | [0.059, 0.093, 0.036] | [15.86, 6.31, 6.41] |
| 9 | [0.067, 0.092, 0.046] | [12.38, 4.69, 4.49] |
| 10 | [0.057, 0.150, 0.031] | [12.95, 16.12, 6.89] |
| mean | [0.069、0.109、0.034] | [13.95, 11.29, 6.39] |
| ± 标准差 | ± [0.010, 0.020, 0.009] | ±[2.11, 6.09, 1.07] |

对于各种场景，给定基于人工智能的相对 6 自由度位姿测量，能够充分估计状态。对于大多数序列，位置均方根误差约为 10 厘米或更低，这对于这项相当复杂的任务而言是可接受的误差。此外，这些结果表明，我们的方法适用于物体相对状态估计。在某些情况下，例如序列 2 和 3，性能不如其他序列。这是由于 6 自由度位姿估计器的预测中存在异常值。物体部分超出图像范围、视角模糊以及图像中的运动模糊，都会导致位姿估计错误。尤其是后者，会导致单张图像中所有物体的预测都出错。虽然 $\chi^2$ test 有助于剔除此类帧，但多张连续图像中的测量值带有噪声或错误，仍会影响状态估计，因为这可能导致一段时间内没有更新、仅进行惯性测量单元传播，从而造成航位推算。在此类情况下，对不确定性进行预测

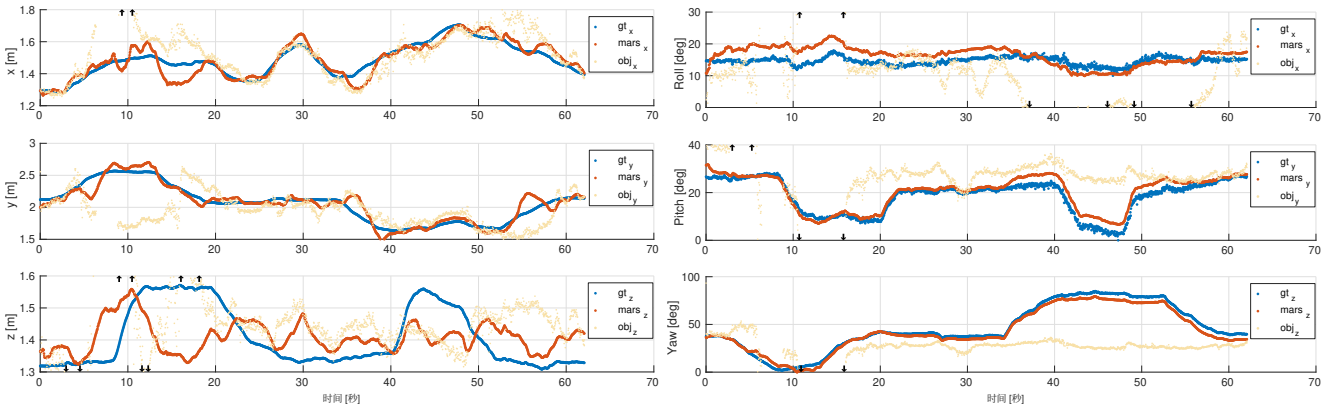


图3. 序列4的第8次运行中，估计位置和欧拉角姿态（mars）与真值（gt）的比较。对整个序列分别绘制了位置（x、y、z）和姿态（滚转角、俯仰角、偏航角）的各个分量。此外，我们还比较了由物体3（obj）的rawPoET估计值给出的重投影的IMU位姿。黑色箭头标示的区域中，重投影的IMU位姿超出了绘图范围。需要注意的是，在6.4秒至8.8秒之间，该物体未出现在相机图像中。

针对每个物体和测量值进行处理，会比采用上述固定值带来更好的结果。将 PoET 预测中的偶然不确定性和认知不确定性纳入考量，将作为未来工作。

为说明我们方法的可复现性，我们从8个序列中选择了两个（序列4和序列6），并将每个序列重复运行10次，得到相似但并非完全相同的轨迹。对于每个序列，表II和表III分别汇总了每次运行在整个轨迹上的均方根误差，以及所有运行的平均均方根误差和标准差。对于两个序列，我们的方法均能够在10次独立运行中复现其性能，并且标准差较小。这表明，基于人工智能的组件能够可靠地集成到机器人的状态估计中。

此外，我们在图3中针对一段示例录制数据，对整个轨迹中的估计位置和姿态与真实值进行了比较。该示例表明，我们的方法能够在整个录制期间可靠地估计位置和姿态。此外，图中还显示，单个物体的原始测量值有时会导致重投影的惯性测量单元位姿与真实轨迹不一致。然而，通过融合惯性测量单元信息与来自多个物体的位姿测量值，我们的方法能够在存在异常值测量的情况下，仍然可靠地估计轨迹。除此之外，我们在图4中展示了一个示例，说明我们的方法具备针对物体-世界状态的自校准能力。由于可能存在噪声测量，物体-世界在首次被观测后被错误初始化。尽管如此，该状态在5秒后收敛。

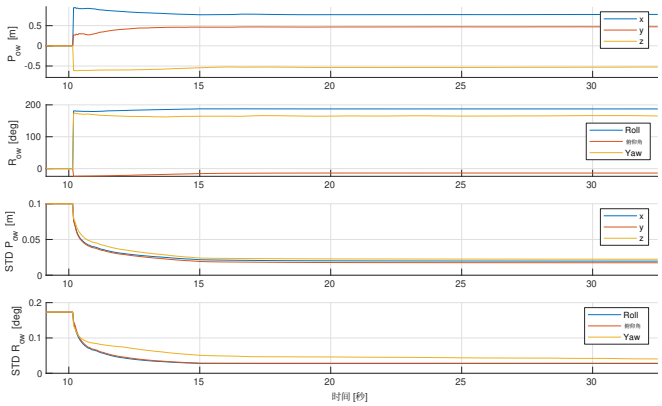


图4. 序列4第8次运行中非主物体的估计物体-世界状态（**位置**$o_kW$，**姿态**$o_kW$）及其对应状态协方差（以标准差表示）的可视化。位置分为（x、y、z），姿态则以欧拉角表示。图中绘出了从首次观测到物体时（约10秒）到状态收敛期间的状态。开始时，物体状态可能由于测量噪声而被错误初始化。然而，约5秒后状态收敛，不确定性降至最低。

## V. 结论

本文研究了移动机器人中的物体相对状态估计问题，并提出了一种基于人工智能的方法，从单个RGB图像。我们定义了一种最简传感器配置，由一个RGB相机和一个惯性测量单元组成，并设计了一个需要进行物体相对状态估计的实验场景，以模拟检查和监测任务。

我们针对给定六自由度相对位姿测量的移动机器人完整状态估计，推导并实现了一种基于滤波器的解决方案。此外，我们的方法不需要任何关于物体全局位姿和相对位姿的初始信息。通过定义物体-世界状态，我们以其中一个物体作为锚点，同时估计每个物体相对于共同导航世界坐标系的坐标系。使用我们自行采集的真实数据进行的实验表明，该方法可用于不同场景下的移动机器人状态估计，并且结果能够可靠地复现。结果显示，单个传感器提供的基于人工智能的语义信息与惯性测量单元数据结合，足以实现准确的状态估计。未来的工作将考虑在传感器融合框架中纳入基于人工智能的预测所具有的偶然不确定性和认知不确定性，以改进离群值剔除；此外，还将把我们提出的方法集成到真实无人机上，开展闭环实验。

## 参考文献


[1] J. Crespo、J. C. Castillo、O. M. Mozos 和 R. Barber，《用于机器人导航的语义信息：综述》，Applied Sciences，第 10 卷，第 2 期，第 497 页，2020年。[2] S. Jordan、J. Moore、S. Hovet、J. Box、J. Perry、K. Kirsche、D. Lewis 和 Z. T. H. Tse，《无人机检测的前沿技术》，IET Radar, Sonar & Navigation，第 12 卷，第 2 期，第 151–164 页，2018年。[3] Y. Li、C. Fu、Z. Huang、Y. Zhang 和 J. Pan，《具备关键帧感知能力的实时无人机目标跟踪》，载于《2020年IEEE国际机器人与自动化会议（ICRA）》。IEEE，2020年，第 193–199 页。[4] S. Weiss、D. Scaramuzza 和 R. Siegwart，《基于单目同时定位与建图的导航：面向GPS拒止环境中的自主微型直升机》，Journal of Field Robotics，第 28 卷，第 6 期，第 854–874 页，2011年。[5] E. Allak、A. Hardt-Stremayr 和 S. Weiss，《在无持久特征的视觉惯性里程计中，快速图像尺度变化和零运动期间的关键帧策略》，载于《2018年IEEE/RSJ智能机器人与系统国际会议（IROS）》。IEEE，2018年10月。[6] A. Giusti、J. Guzzi、D. C. Cire¸san、F.-L. He、J. P. Rodríguez、F. Fontana、M. Faessler、C. Forster、J. Schmidhuber、G. Di Caro 等，《一种面向移动机器人的森林小径视觉感知机器学习方法》，IEEE机器人与自动化快报，第 1 卷，第 2 期，第 661–667 页，2015年。[7] R. Dominguez、E. Onieva、J. Alonso、J. Villagra 和 C. Gonzalez，《面向自动驾驶车辆的基于激光雷达的感知解决方案》，载于《2011年第11届智能系统设计与应用国际会议》。IEEE，2011年，第 790–795 页。[8] M. Hofmarcher、T. Unterthiner、J. Arjona-Medina、G. Klambauer、S. Hochreiter 和 B. Nessler，《利用语义分割进行自动驾驶的视觉场景理解》，载于《可解释人工智能：深度学习的解读、解释与可视化》。Springer，2019年，第 285–296 页。[9] Y. Xiang、T. Schmidt、V. Narayanan 和 D. Fox，《PoseCNN：用于杂乱场景中六维目标位姿估计的卷积神经网络》，载于《机器人学：科学与系统（RSS）》，2018年。[10] M. Bloesch、S. Omari、M. Hutter 和 R. Siegwart，《采用基于直接扩展卡尔曼滤波的方法实现鲁棒视觉惯性里程计》，载于《2015年IEEE/RSJ智能机器人与系统国际会议（IROS）》。IEEE，2015年，第 298–304 页。[11] J. Solà、J. Vallvé、J. Casals、J. Deray、M. Fourmy、D. Atchuthan、A. Corominas-Murtra 和 J. Andrade-Cetto，《Wolf：一种基于因子图的机器人模块化状态估计框架》，IEEE机器人与自动化快报，第 7 卷，第 2 期，第 4710–4717 页，2022年。[12] C. Brommer、R. Jung、J. Steinbrener 和 S. Weiss，《MaRS：一种模块化且鲁棒的传感器融合框架》，IEEE机器人与自动化快报，第 6 卷，第 2 期，第 359–366 页，2020年。[13] T. Moore 和 D. Stouch，《面向机器人操作系统的广义扩展卡尔曼滤波器实现》，载于《智能自主系统 13》。Springer，2016年，第 335–348 页。[14] Z. Cao、Y. Sheikh 和 N. K. Banerjee，《面向无纹理物体的实时可扩展六自由度位姿估计》，载于《2016年IEEE国际机器人与自动化会议（ICRA）》。IEEE，2016年，第 2441–2448 页。[15] G. Pavlakos、X. Zhou、A. Chan、K. G. Derpanis 和 K. Daniilidis，《基于语义关键点的六自由度物体位姿估计》，载于《2017年IEEE国际机器人与自动化会议（ICRA）》。IEEE，2017年，第 2011–2018 页。[16] A. Amini、A. S. Periyasamy 和 S. Behnke，《T6d-direct：用于多物体六维位姿直接回归的Transformer》，载于《DAGM德国模式识别会议》。Springer，2021年，第 530–544 页。[17] G. Billings 和 M. Johnson-Roberson，《SilhoNet：一种用于六维目标位姿估计的RGB方法》，IEEE机器人与自动化快报，第 4 卷，第 4 期，第 3727–3734 页，2019年。[18] W. Kehl、F. Manhardt、F. Tombari、S. Ilic 和 N. Navab，《SSD-6D：让基于RGB的三维检测和六维位姿估计再创佳绩》，载于《IEEE国际计算机视觉会议论文集》，2017年，第 1521–1529 页。[19] Y. Li、G. Wang、X. Ji、Y. Xiang 和 D. Fox，《DeepIM：用于六维位姿估计的深度迭代匹配》，载于《欧洲计算机视觉会议（ECCV）论文集》，2018年，第 683–698 页。[20] Z. Li、G. Wang 和 X. Ji，《CDPN：用于实时基于RGB的六自由度目标位姿估计的基于坐标的解耦位姿网络》，载于《IEEE/CVF计算机视觉国际会议论文集》，2019年，第 7678–7687 页。

[21] G. Wang、F. Manhardt、F. Tombari 和 X. Ji，“Gdr-net：用于单目6D物体位姿估计的几何引导直接回归网络”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第16 611–16 621页。[22] C. Li、J. Bai 和 G. D. Hager，“多视图多类别物体位姿估计的统一框架”，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年，第254–269页。[23] Y. Labbé、J. Carpentier、M. Aubry 和 J. Sivic，“CosyPose：一致的多视图多物体6D位姿估计”，载于欧洲计算机视觉会议。Springer，2020年，第574–591页。[24] A. Krull、E. Brachmann、F. Michel、M. Y. Yang、S. Gumhold 和 C. Rother，“在RGB-D图像中学习用于6D位姿估计的分析合成”，载于*IEEE*国际计算机视觉会议论文集，2015年，第954–962页。[25] F. Michel、A. Kirillov、E. Brachmann、A. Krull、S. Gumhold、B. Savchynskyy 和 C. Rother，“用于6D物体位姿估计的全局假设生成”，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第462–471页。[26] T. Jantos、M. A. Hamdad、W. Granig、S. Weiss 和 J. Steinbrener，“PoET：用于单视图多物体6D位姿估计的位姿估计变换器”，载于第六届机器人学习会议论文集。PMLR，2023年。[27] R. F. Salas-Moreno、R. A. Newcombe、H. Strasdat、P. H. Kelly 和 A. J. Davison，“Slam++：物体级别的同步定位与地图构建”，载于*IEEE*计算机视觉与模式识别会议论文集，2013年，第1352–1359页。[28] N. Merrill、Y. Guo、X. Zuo、X. Huang、S. Leutenegger、X. Peng、L. Ren 和 G. Huang，“用于6自由度物体位姿估计、考虑对称性与不确定性的物体SLAM”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第14 901–14 910页。[29] S. L. Bowman、N. Atanasov、K. Daniilidis 和 G. J. Pappas，“用于语义SLAM的概率数据关联”，载于*2017*年*IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2017年，第1722–1729页。[30] J. Thomas、G. Loianno、K. Daniilidis 和 V. Kumar，“四旋翼飞行器通过悬挂于圆柱形物体实现栖停的视觉伺服”，*IEEE*机器人与自动化快报，第1卷，第1期，第57–64页，2015年。[31] G. Loianno、V. Spurny、J. Thomas、T. Baca、D. Thakur、D. Hert、R. Penicka、T. Krajnik、A. Zhou、A. Cho、等，“微型飞行器团队在极具挑战性的类沙漠环境中对磁性物体进行定位、抓取和运输”，*IEEE*机器人与自动化快报，第3卷，第3期，第1576–1583页，2018年。[32] K. Máthé、L. Bu¸soniu、L. Barabás、C.-I. Iuga、L. Miclea 和 J. Braband，“面向物体检查场景的四旋翼飞行器基于视觉的控制”，载于*2016*年国际无人机系统会议（*ICUAS*）。IEEE，2016年，第849–857页。[33] J. Sola，“误差状态卡尔曼滤波器的四元数运动学”，*arXiv*预印本 *arXiv:1711.02508*，2017年。[34] C. Brommer、C. Böhm、J. Steinbrener、R. Brockers 和 S. Weiss，“畸变磁场中的改进状态估计”，载于*2020*年国际无人机系统会议（*ICUAS*）。IEEE，2020年，第1007–1013页。[35] P. Furgale、J. Rehder 和 R. Siegwart，“多传感器系统的统一时间与空间标定”，载于*2013*年*IEEE/RSJ*智能机器人与系统国际会议。IEEE，2013年，第1280–1286页。