arXiv:2602.01266v1 [cs.RO] 1 Feb 2026

用于自主导航主动感知的强化学习

Grzegorz Malczyk*、Mihir Kulkarni 和 Kostas Alexis

摘要——本文探讨了在复杂未知环境中进行自主导航时主动感知所面临的挑战。本文重新审视主动感知的基本原理，提出一种端到端强化学习框架：机器人不仅要在避开障碍物的同时到达目标，还要主动控制机载相机，以增强态势感知。策略接收的观测包括机器人状态、当前深度帧，以及根据短时历史深度读数构建的精细局部几何表示。为将无碰撞运动规划与信息驱动的主动相机控制相结合，我们在导航奖励中加入了一种基于体素的信息度量。借助这一方法，空中机器人能够学习一种稳健策略，在目标导向运动与探索性感知之间取得平衡。大量评估结果表明，与固定式非主动相机基线相比，我们的方法能够实现更安全的飞行，同时还能促发内在探索行为。

I. 引言

空中机器人日益广泛地部署于复杂且杂乱的环境中，在这些环境下，安全导航和有效感知对于基础设施巡检、搜救和环境监测等任务至关重要。传统上，机器人任务通常由一系列航路点组成，机器人需要在避开障碍物的同时到达这些目标。在此类点对点导航过程中，感知往往被视为一种被动过程，即“仅仅”消耗机器人运动过程中采集的数据，而传感器则固定安装在机器人上。可控相机在导航研究中仍不常见。然而，这种分离忽视了一个关键事实：感知本身是一个主动过程。它涉及决定感知什么、何时感知以及在哪里感知，以提升态势感知和任务性能 [1]。

主动感知的概念早在 [2]–[3]中就已提出，并于近期在 [4],中重新探讨，强调感知应以目标为导向。早期机器人系统 [1]展示了中央凹相机和可移动传感器在改善任务相关感知方面的潜力，但这些方法受到当时硬件和计算能力的限制。在此基础上，近期有关下一最佳视点（NBV）规划和主动建图[5], [6] 的研究主要聚焦于选择能够最大化信息增益或覆盖范围的视点。然而，这些方法的主要目标是建图和探索目标，对如何优化视点选择以导航至特定目标关注有限。即使在那些将.....表述为

本研究得到挪威研究理事会奖项 NO-338694 和地平线欧洲计划资助协议编号 101119774 的支持。作者来自挪威科技大学（NTNU）工程控制论系。

*通讯作者。电子邮箱：grzegorz.malczyk@ntnu.no

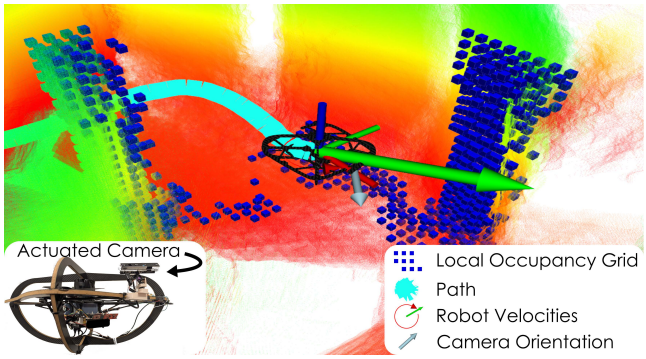


图 1. 配备可驱动相机系统的四旋翼平台。局部占据栅格为机器人提供附近障碍物的信息，而相机则朝向杂乱环境中的新区域进行探索。

内在注意力目标，例如 [7], 中的研究；大多数文献仅涉及固定在机器人机架上的非驱动相机。

与此同时，强化学习（RL）已成为在模拟环境 [8] 和真实环境 [9]中进行导航的强大范式。这些方法学习将高维感知输入（包括 RGB 图像 [10], [11] 和深度数据 [12]–[14], ）直接映射为运动指令，通常采用演员-评论家或策略梯度方法。尽管已有研究探索了多目标强化学习，以联合优化目标到达和探索目标 [15],，但大多数基于强化学习的导航框架都假设传感器刚性安装且不可驱动，其相对于机器人具有固定朝向，这限制了智能体将感知装置主动指向与任务相关区域的能力 [16]。语义感知和覆盖驱动的强化学习近期进展已引入用于探索[17] 或表面覆盖 [18]的内在奖励。尽管 [18]将问题扩展至三维视觉检测，但并未考虑主动式传感和主动感知赋能的导航。

主动感知理论与现代基于强化学习的导航系统之间的这一差距带来了重大机遇。我们认为，有效的自主性需要耦合两个相互交织的目标：（i）安全的目标导向运动，确保机器人在无碰撞的情况下到达其目标；以及（ii）内在动机驱动的信息性视点选择，使机器人能够主动改进场景理解。

A. 贡献

在本研究中，我们考虑配备主动（可驱动）相机的飞行机器人，并提出一种新颖的强化学习框架，联合优化安全的目标导向运动与信息增益，使智能体能够在执行

导航任务时主动决定如何移动以及将相机指向何处，如图1所示。为增强韧性，该方法不假设定位能够长期保持一致，而仅依赖即时传感器读数、局部平滑的里程计，以及机器人周围紧凑而富有表现力的局部几何表示。后者对三维自由空间和障碍物进行编码，从而更好地帮助智能体推理附近的几何结构以实现避碰。我们不仅展示了主动感知在导航中的优势，还设计了一种多目标奖励函数，将 (i) 反映目标进展、成功和避碰情况的传统导航奖励，与 (ii) 信息增益项相结合；该项鼓励智能体最大限度地进行环境探索并提升感知理解，同时不损害安全性或导航任务的完成。我们首先通过大量仿真验证该方法，结果表明，与依赖机体固定式非驱动相机的基线方法相比，该方法具有较高的目标到达成功率和更高的地图完整度。此外，我们还在飞行机器人上对所提出的训练策略进行了实验部署，证明该方法能够从仿真迁移至实体硬件，并在三维环境中成功实现无碰撞目标到达和主动感知。为支持可复现性，该方法已开源于
h
t
t
p
s://github.co
m/ntnu-arl/active-perception-RL-navigation

本文其余部分安排如下：第二节和第三节分别介绍问题表述和所提出的方法。评估结果见第四节，结论见第五节。

II. 问题表述

本文研究的是未知环境中空中机器人的主动感知赋能三维导航问题，即寻找一种控制策略，使机器人能够安全、高效地到达指定目标位置，同时利用主动式传感增强态势感知。我们将其建模为一个逐步推导无碰撞路径 $\mathcal{P}_i$ 至目标位置 $\mathcal{G}_i$ 的问题，并假设仅可获取以下信息：a) 机器人在时间 t的里程计局部一致估计$\mathbf{s}_t$ ； b) 当前深度图像$\mathbf{D}_t$；c) 相关的相机朝向$\mathbf{c}_t$，表示相机坐标系 $\mathcal{C}$ 相对于机体固定坐标系 $\mathcal{B}$的相机俯仰角 β_t 和偏航角 γ_t ；以及 d) 与车辆坐标系 $\mathcal{V}$对齐的自我中心局部占据栅格$\mathbf{m}_t^o$，如图 2 所示。车辆坐标系在偏航方向上与机体固定坐标系对齐，其x-y 平面平行于惯性坐标系 $\mathcal{I}$。估计的机器人状态

定义如下：

$$\mathbf{s}_t = [\mathbf{p}_t, \mathbf{q}_t, \mathbf{v}_t, \boldsymbol{\omega}_t], \tag{1}$$

其由三维位置$\mathbf{p}_t$、以关联四元数的四维向量形式表示的姿态$\mathbf{q}_t$ （在 $\mathcal{I}$中表示）组成，而三维线速度$\mathbf{v}_t$ 和三维角速度 $\boldsymbol{\omega}_t$则在 $\mathcal{B}$中表示。深度图像$\mathbf{D}_t$ 可由机载RGB-D相机设备获取（如本研究中所述），局部自我中心占据栅格$\mathbf{m}_t^o$ 则基于测距传感器读数和局部平滑里程计在线构建。给定一个三维目标位置

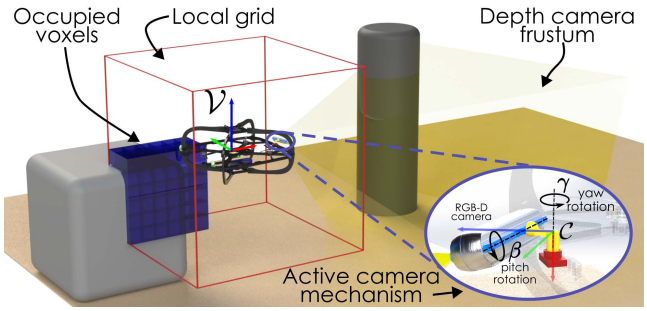


图 2. 配备可驱动RGB-D相机系统的四旋翼平台。

在未知环境中，目标是迭代计算一个融合导航、避障和主动感知的最优动作向量。动作向量定义为：

$$\mathbf{a}_t = [\underbrace{\mathbf{v}_t^r,\ \omega_{t,z}^r}_{\mathbf{a}_t^{\text{nav}}},\ \underbrace{\mathbf{c}_t^r}_{\mathbf{a}_t^{\text{cam}}}], \tag{2}$$

其中$\mathbf{v}_t^r \in \mathbb{R}^3$ 和 $\omega_{t,z}^r$ 表示以 $\mathcal{V}$为单位的指令线速度和偏航角速度，而$\mathbf{c}_t^r = \{\beta_t^r, \gamma_t^r\}$ 表示以 $\mathcal{B}$为单位的相机指令俯仰角和偏航角。这些量可分为低级机器人控制器$\mathbf{a}_t^{\text{nav}}$的指令参考值，以及可控相机$\mathbf{a}_t^{\text{cam}}$的方向设定值。最优动作向量$\mathbf{a}_t$ 必须同时满足一组相互耦合的目标，即：(i) 在未知环境中导航并到达目标；(ii) 在存在障碍物的情况下保持无碰撞飞行；(iii) 通过联合优化机器人运动和相机朝向，主动获取对环境的认知。最后一项目标能够提升性能，同时不损害导航和安全目标。

III. 方法

我们将支持主动感知的无碰撞导航建模为一项强化学习任务。我们将状态空间 $\mathcal{S}$ 定义为智能体和环境在离散时刻 t的所有可能状态的集合，其中 $\mathbf{s}_t \in \mathcal{S}$ 。动作空间表示为 $\mathcal{A}$ ，其中$\mathbf{a}_t \in \mathcal{A}$。观测空间表示为 $\mathcal{O}$ ，其中智能体接收到的每个观测记为 $\mathbf{o}_t \in \mathcal{O}$。最后，$\mathcal{R}$ 表示奖励函数。随后，我们将定义如何构建并推导强化学习问题中的各个量，以实现支持主动感知的安全导航。

A. 自我中心局部占据栅格

为实现安全导航，我们的方法采用紧凑的局部表示来描述机器人周围的即时环境。与基于全局占据地图$\mathbf{M}_t^o$ 进行规划的传统规划方法 [19] 不同，这类地图容易随着时间推移受到定位漂移的影响；我们的强化学习框架则基于局部的自我中心占据栅格$\mathbf{m}_t^o$运行。这一设计选择增强了在多样化环境中的泛化能力，并显著提高了计算效率。

以 $\mathcal{I}$ 表示的稠密地图会持续整合深度测量，并使用（可能存在漂移的）里程计估计值。据此，在时间 t，从机器人附近提取局部占据栅格$\mathbf{m}_t^o$ 。为了构建这一

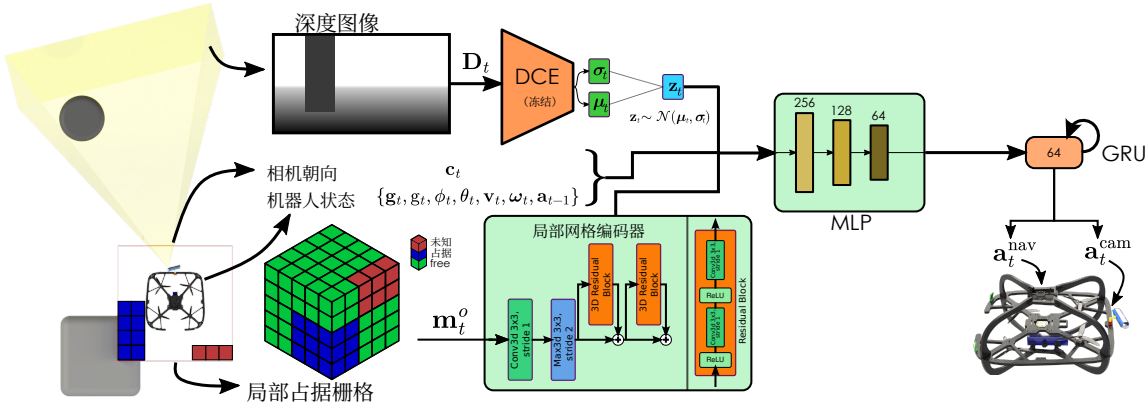


图 3. 所提出的用于主动感知安全导航的网络架构。该网络通过专用编码器模块处理深度图像和局部占据栅格，利用多层感知机和门控循环单元模块整合机器人状态与相机朝向，并为机器人指令动作。

表示，首先将深度传感器测量从相机坐标系变换到车辆坐标系。对于深度图像中的每个像素，沿射线投射至最大距离 d_m （此处为 3 m），将包含对应障碍物点的体素标记为占据，同时将射线经过的体素标记为空闲。从未被观测到的体素则保持标记为未知。该建图过程在机器人周围的局部区域内增量进行，并随着机器人移动在时间上累积信息。时间整合有助于减轻传感器噪声，相比仅依靠瞬时深度测量，能够更完整地表示局部环境。同时，由于该地图仅针对机器人周围的短距离范围构建并迭代更新，因此不需要长期一致的里程计。总体而言，与仅假设能够访问瞬时相机数据 [13], [20], [21]的方法相比，这种局部表示为强化学习智能体提供了改进避碰所需的关键信息。它使策略能够更加稳健，主动避开位于传感器视锥之外的障碍物。图 3 展示了局部自我中心栅格。

B. 策略学习

a) 状态与观测空间：底层状态$\mathfrak{s}_t \in \mathcal{S}$ 编码了机器人的六自由度位姿、目标位置以及周围环境的完整几何信息。由于无法直接观测状态，智能体接收观测值 $\mathfrak{o}_t \in \mathcal{O}$，其定义如下：

$$\mathfrak{o}_t = \{\mathbf{g}_t, \mathrm{g}_t, \phi_t, \theta_t, \mathbf{v}_t, \boldsymbol{\omega}_t, \mathbf{c}_t, \mathbf{a}_{t-1}, \mathbf{z}_t, \mathbf{m}_t^o\}, \tag{3}$$

其中，$\mathbf{g}_t \in \mathbb{R}^3$ 是指向目标位置的单位向量，以 $\mathcal{V}$ 表示，对应距离为 $\mathrm{g}_t \in \mathbb{R}$。机器人的姿态通过俯仰角 ϕ_t 和横滚角 θ_t表示，以 $\mathcal{V}$为单位。此外，还包含以 $\mathcal{B}$ 表示的线速度$\mathbf{v} \in \mathbb{R}^3$和角速度 $\boldsymbol{\omega} \in \mathbb{R}^3$，以便让策略了解系统动力学。通过俯仰角和偏航角$\mathbf{c}_t = \{\beta_t, \gamma_t\}$观测受控相机的朝向，同时将先前的控制动作$\mathbf{a}_{t-1} \in \mathbb{R}^6$也附加到状态中。高维深度图像$\mathbf{D}_t$ 经深度碰撞编码器（DCE） [22], 处理后压缩为潜在嵌入$\mathbf{z}_t$ ，并附加到观测向量中。DCE 专注于保留碰撞信息，同时对输入深度图像进行大幅压缩。除了向量化特征外，智能体还会接收以 $\mathcal{V}$表示的局部占据栅格$\mathbf{m}_t^o$ ，从而增强空间上下文信息，以便避障。

b) 动作空间：动作空间共同涵盖导航与主动感知。在每个控制步，策略都会输出控制空中机器人的运动命令（$\mathbf{a}_t^{\mathrm{nav}}$）以及可控相机的期望朝向（$\mathbf{a}_t^{\mathrm{cam}}$）。第一个动作 $\mathbf{a}_t^{\mathrm{nav}} \in \mathbb{R}^4$ 经参数化后，表示所命令的线速度 $\mathbf{v}_t^r$ 和偏航角速度ω_t^r, z。因此，我们允许智能体充分利用各种可能的运动方式，高效探索环境，而不限制命令必须严格处于深度相机的视场角（FOV）范围内 [13]。第二个动作$\mathbf{a}_t^{\mathrm{cam}} := \mathbf{c}_t^r$规定了可控相机的期望俯仰角和偏航角。这些动作限制在硬件范围内 $\beta_t^r \in [-\beta_{\max}, \beta_{\max}]$, $\gamma_t^r \in [-\gamma_{\max}, \gamma_{\max}]$，以确保其符合舵机驱动的可行性。因此，整体动作向量定义为：

$$\mathbf{a}_t = \left[\mathbf{a}_t^{\mathrm{nav}}, \mathbf{a}_t^{\mathrm{cam}}\right] \in \mathbb{R}^6, \tag{4}$$

使智能体能够同时控制机器人的运动并主动调整相机朝向。这种统一的动作空间使策略能够确保无碰撞导航，同时高效驱动相机，以最大限度地获取有助于无碰撞导航和探索环境的信息。

c) 奖励设计：智能体在每次状态转移时获得奖励$\mathcal{R}(\mathfrak{s}_t, \mathfrak{a}_t)$，并据此学习一个策略 π ，该策略将观测和信念状态映射为动作：$\mathbf{a}_t = \pi(\mathfrak{o}_t, \mathfrak{b}_t)$，其中 $\mathfrak{b}_t$ 表示智能体在给定观测历史下对可能环境状态的信念分布。在训练过程中，环境的全局占据栅格$\mathbf{M}_t^o$ 作为特权知识，用于：(i) 计算奖励信号，(ii) 依据障碍物位置的真实值评估安全违规，以及 (iii) 确定全局环境中哪些体素已被智能体的传感器观测到。这使得学习过程中能够进行准确的奖励计算和适当的监督。然而，这些全局信息被有意对智能体的策略隐瞒，该策略必须仅基于局部

观测 $\mathfrak{o}_t$ 仅提供附近障碍物和自由空间的局部、含噪测量。该训练范式确保习得策略在无法获得全局环境知识的真实场景中仍然可以部署。奖励函数定义为：

$$\mathcal{R}(\mathfrak{s}_t, \mathbf{a}_t) = r_t + l_t + n_t + p_t, \tag{5}$$

其中，每一项都发挥着不同的作用。 r_t项根据当前距离定义，用于奖励智能体接近目标位置：

$$r_t = \lambda_d(d_{t-1} - d_t) + \lambda_e(e^{-\alpha d_t^2}), \tag{6}$$

其中， d_t 是机器人在时间 t的当前位置与目标之间的欧氏距离，而 λ_d， λ_e， $\alpha \in \mathbb{R}_+$是缩放因子。指数项鼓励智能体在接近目标时更快地取得进展。 l_t 项会因机器人和相机的急促或大幅运动而惩罚智能体，从而促进平滑运动。其定义如下：

$$l_t = -\lambda_a||\mathbf{a}_t - \mathbf{a}_{t-1}||, \tag{7}$$

其中， $\lambda_a \in \mathbb{R}_+^6$ 是缩放因子。 n_t 项鼓励智能体主动探索并发现新信息。此内在奖励根据全局占据栅格中的特权信息计算（不同于策略所使用的局部占据栅格$\mathbf{m}_t^o$ ），与在时间步$t-1$ 和 t之间状态从未知 转变为空闲 或占据的体素数量成正比。这会促使智能体将相机重新转向先前未观测到的区域，从而提升其态势感知能力。 n_t 定义如下：

$$n_t = \lambda_{\mathcal{G}} \sum_{i \in \mathbf{M}_t^o} \mathbb{I}[\text{state}(i)_t \neq \text{unknown} \wedge \text{state}(i)_{t-1} = \text{unknown}], \tag{8}$$

其中， $\lambda_{\mathcal{G}} \in \mathbb{R}_+$ 是缩放因子， $\mathbb{I}[\cdot]$ 是指示函数。求和遍历全局地图中的所有体素；如果一个体素从未知状态转变为已知状态（空闲 或占据），则指示函数的值为 1，否则为 0。请注意，除非明确指定，否则默认不使用此项。最后，距离惩罚 p_t旨在防止碰撞。它与机器人临界碰撞距离 $d_{coll} > 0$ 内占据和未知体素的数量成正比。 d_{coll} 反映了机器人的物理尺寸。惩罚项 p_t 的计算方式如下：

$$p_t = -\lambda_p \sum_{i \in \text{sphere}(d_{coll})} \mathbb{I}[\text{voxel}_i \in \{\text{occupied, unknown}\}], \tag{9}$$

其中， $\lambda_p \in \mathbb{R}_+$ 是缩放因子。p_t 的值较高表示机器人距离障碍物较近，因此会促使智能体转向以避开障碍物。

C. 实现

a) 可控相机：我们采用与实际舵机相匹配的一阶动力学来建模相机执行器的响应。给定指令角 β^r, γ^r，关节动力学为：

$$\dot{\beta}_t = \frac{1}{\tau_\beta} \operatorname{sat}_{[-\beta_{\max}, \beta_{\max}]} (\beta_t^r - \beta_t), \tag{10}$$

$$\dot{\gamma}_t = \frac{1}{\tau_\gamma} \operatorname{sat}_{[-\gamma_{\max}, \gamma_{\max}]} (\gamma_t^r - \gamma_t), \tag{11}$$

其中 τ_β，τ_γ 是根据实体舵机辨识得到的时间常数，而 $\operatorname{sat}_{[a,b]}(\cdot)$ 表示将数值限制在区间 $[a, b]$ 内的饱和函数。对于一阶系统，10–90% 上升时间为 $t_r \approx 2.2\,\tau$，我们利用它根据阶跃响应测量结果校准时间常数 τ。仿真中实现的离散时间表述采用前向欧拉积分法：

$$\beta_{t+1} = \beta_t + \frac{\Delta t}{\tau_\beta}\left(\tilde{\beta}_t^r - \beta_t\right), \tag{12}$$

$$\gamma_{t+1} = \gamma_t + \frac{\Delta t}{\tau_\gamma}(\tilde{\gamma}_t^r - \gamma_t), \tag{13}$$

$\tilde{}$，其中命令经过饱和处理， $\beta_t^r = \operatorname{sat}_{[-\beta_{\max}, \beta_{\max}]}(\beta_t^r)$和$\tilde{\gamma}_t^r = \operatorname{sat}_{[-\gamma_{\max}, \gamma_{\max}]}(\gamma_t^r)$。

b) 网络架构： 我们采用 [23]中的异步近端策略优化（APPO）算法，训练一个深度神经网络策略，使机器人能够利用主动感知实现无碰撞导航。与 [18], 类似，在二维模型架构的基础上，我们通过采用 [24] 中的超参数调整基于ResNet的编码器，将该方法扩展到三维，从而有效处理来自局部三维占用栅格输入的空间信息 $\mathbf{m}^o$。大小为 $n \times n \times n$的自我中心占用图（其中 $n = 21$）使用 $r_V = 0.1\text{m}$的栅格分辨率进行离散化，该分辨率经过选择，可在碰撞距离 d_{coll} 设为0.4 m的情况下，结合机器人的实际尺寸提供足够的局部空间细节。相应地，这也意味着局部占据栅格的尺寸为$2.1 \times 2.1 \times 2.1$ m^3。奇数个栅格单元确保机器人始终对称地位于局部占据栅格中间的体素处。ResNet编码器从这些局部占据栅格中学习空间特征的压缩潜表示，使策略能够高效地针对物体学习避碰行为，即使这些物体已不再位于深度传感器视场角内。

预训练并冻结的 DCE 所生成的压缩表征，与栅格编码器以及机器人和相机状态的表征相结合。随后，这些信息被输入多层感知机（MLP）。该 MLP 由三个全连接层组成，每层分别有 256、128 和 64 个神经元，并采用 ELU 激活层，后接一个隐藏层大小为 64 的门控循环单元（GRU）模块。给定观测向量 $\mathfrak{o}_t$，策略输出一个 6 维动作指令 $\mathbf{a}_t$。线速度指令的范围为 $\pm$ 1m/s，偏航角速度的范围为$\pm$ 1rad/s。类似地，相机朝向指令分别按照俯仰角和偏航角的最大范围进行缩放 $\pm\,\beta_{\max}$ 和 $\pm\,\gamma_{\max}$。随后，这些指令会发送至机器人机载的底层速度控制器和可控相机，如图 3 所示。同时，俯仰角和偏航角的相机角度指令以$\beta_{\max} = \gamma_{\max} = \frac{\pi}{2}$ rad 为参考输入伺服控制器。 –

D. 训练环境

在训练过程中，我们使用 Aerial Gym 模拟器 [25]，它提供环境和接口，用于训练深度强化学习策略，使其能够在各种环境中导航。该模拟器支持在配备外感传感器的空中机器人上进行大规模并行仿真。仿真中的飞行平台使用来自 [26]的速度控制器，并配备一台深度相机，其水平和垂直视场角为 $\{86, 57\}^\circ$，感知范围为 0.2 至 10m。

我们在仿真器中生成环境，其中包含走廊状场景；场景中随机放置了由不同尺寸的基本几何体构成的静态障碍物，如图 4 所示。采用基本几何体与导航的泛化能力目标有关，因为这些基本形状能够提供多样的几何挑战，同时不会引入特定领域的复杂性。走廊形环境的尺寸为$L \times W \times H$，范围为 $[10, 12] \times [5, 8] \times [4, 6]$ 米。对于每个回合，起始位置和目标位置都会在环境两端相对的位置随机采样。智能体的初始偏航角从 $\mathcal{U}\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ 中均匀采样，以避免对直线轨迹产生偏置。考虑到环境尺寸和机器人的最大速度，每个回合持续 10 秒。

为了增强网络性能对真实世界不确定性的鲁棒性，我们引入了多种随机化和噪声来源。令 $\mathcal{N}$，$\mathcal{U}$ 分别表示正态分布和均匀分布。将扰动力旋量 $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I})$ 施加到仿真平台上，其中 $\sigma_w = 5$N；同时，方程（3）中的观测会受到噪声$\boldsymbol{\epsilon}_s \sim \mathcal{U}(-\delta_s, \delta_s)$ 的扰动，位置噪声为 $\delta_s = 0.1$m，速度分量噪声为 δ_s =0.05 m s^{-1}。相机传感器的位置和方向分别受到 $\boldsymbol{\epsilon}_p \sim \mathcal{U}(-5\text{cm}, 5\text{ cm})$ 和 $\boldsymbol{\epsilon}_\theta \sim \mathcal{U}(-5^\circ, 5^\circ)$ 的扰动。此外，速度控制器参数随机变化 $\tau_{nominal} \pm 12\%$，以改变阶跃响应特性；深度图像加入高斯噪声 $\boldsymbol{\epsilon}_d \sim \mathcal{N}(0, \sigma_d^2(z))$，并以 $p = 0.01$ 的概率随机丢弃像素，以模拟真实的深度传感器特性 [27]。最后，我们将图像采集和控制频率设为 10 Hz，而物理仿真以 100 Hz 的频率运行。策略在 NVIDIA RTX A6000 上训练约需两小时。

IV. 评估研究

我们通过涵盖高保真仿真和真实世界机器人部署的全面评估，展示了我们启用主动感知的导航策略的有效性。实验验证了该方法在多种环境配置下的性能，展现出其强大的泛化能力，并成功弥合了仿真到现实差距。

A. 消融研究

为评估本方法各项关键组件的单独贡献，我们在 Aerial Gym 模拟器中开展全面的消融研究，以分离主动相机控制与局部网格表示的作用。这些研究

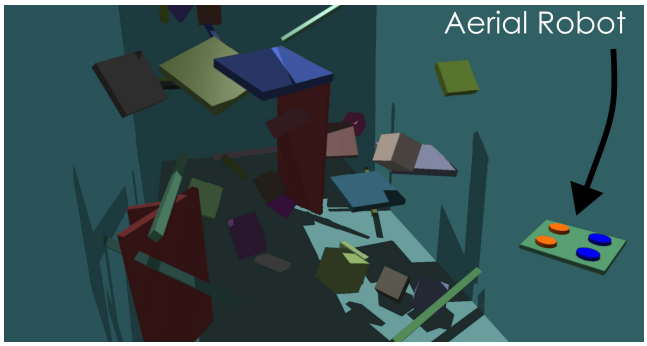


图 4. Aerial Gym 训练环境，通过随机化的基本障碍物确保场景多样性和仿真到现实的迁移能力。

考察了不同障碍物密度环境下的九种配置，从而量化每种架构选择所带来的性能增益。所评估的方法包括：（i）静态 基线，使用固定的相机朝向，策略可在任意方向规划导航动作，包括当前相机视场之外（从而有意为感知—动作协调制造挑战），（ii）静态+视场角，由 [13],提出，其中导航动作被限制为仅沿当前相机视场内可见的方向移动（即智能体不能在传感器视锥之外下达速度指令），（iii）静态+网格 在保持固定相机朝向的同时纳入局部占据栅格表示，（iv）静态+网格+视场角 将局部网格输入与受视场角约束的动作相结合，并使用静态相机，（v）主动 启用动态相机控制，（vi）主动+视场角 在瞬时相机视场内引入动作约束，并具备主动重新调整相机朝向的能力，（vii）主动+网格 将主动相机控制与局部占据栅格表示相结合，（viii）主动+网格+视场角将前一种方法限制为仅沿可控相机视场内当前可见的方向移动，最后（ix）主动+网格+n_t 在主动+网格的基础上增加探索性感知奖励。

表 I 所示的结果表明，在复杂度各异的环境中，主动感知和局部网格表示都会对导航性能产生显著影响。请注意，成功定义为在 1 m 范围内到达目标；碰撞表示与障碍物发生碰撞；若 10s 内既未成功也未发生碰撞，则判定为超时。在无障碍环境（没有漂浮障碍物的走廊）中，所有方法的成功率都接近完美（≥99.3%），表明各项配置均已具备扎实的基础导航能力。然而，随着环境复杂度增加，各方法的性能差异逐渐显著。基线静态相机方法的性能随障碍物密度增加而大幅下降，在最复杂的场景（30 个障碍物）中成功率仅为 65.3%，碰撞率则高达 32.4%，令人担忧。将动作限制在当前相机视场内的方法带来了小幅改进，在密集环境中的成功率提高至 71.5%。引入局部网格表示带来了最显著的提升，大幅降低了碰撞率，使其降至 4.8%，同时提高了成功率。引入

表 I 静态与主动相机配置下的导航和环境探索性能比较。本文研究了一组主动相机策略，最后一种策略进一步聚焦于场景探索。

障碍物数量	0			10			20			30			
	成功	超时	崩溃	成功	超时	崩溃	成功	超时	崩溃	成功	超时	崩溃	探索
静态	99.3%	0.1%	0.6%	82.4%	7.2%	10.4%	70.7%	4.8%	24.5%	65.3%	2.3%	32.4%	26.5%
静态+视场角	99.4%	0.6%	0.0%	90.4%	2.2%	6.6%	79.6%	3.2%	17.2%	71.5%	8.9%	19.6%	24.0%
静态+网格	99.6%	0.4%	0.0%	91.3%	6.7%	2.0%	82.4%	14.7%	2.9%	85.5%	9.7%	4.8%	29.6%
静态+网格+视场角	99.8%	0.2%	0.0%	91.4%	4.4%	4.2%	87.2%	5.7%	7.1%	86.0%	8.9%	5.1%	28.3%
主动	99.7%	0.2%	0.1%	92.5%	1.7%	5.8%	86.4%	0.4%	14.0%	83.2%	0.5%	16.3%	41.5%
主动+视场角	99.9%	0.1%	0.0%	97.8%	1.7%	0.5%	95.9%	2.1%	2.0%	94.9%	2.4%	2.7%	41.2%
主动+网格	99.9%	0.0%	0.1%	98.4%	0.9%	0.7%	96.2%	1.6%	2.2%	95.4%	2.0%	2.6%	43.4%
主动+网格+视场角	99.8%	0.2%	0.0%	98.2%	1.5%	0.3%	95.9%	2.0%	2.1%	95%	2.5%	2.5%	42.6%
▷ 主动+网格+n_t	99.9%	0.1%	0.0%	97.4%	1.5%	1.1%	96.0%	1.7%	2.3%	94.3%	2.8%	2.9%	63.4%

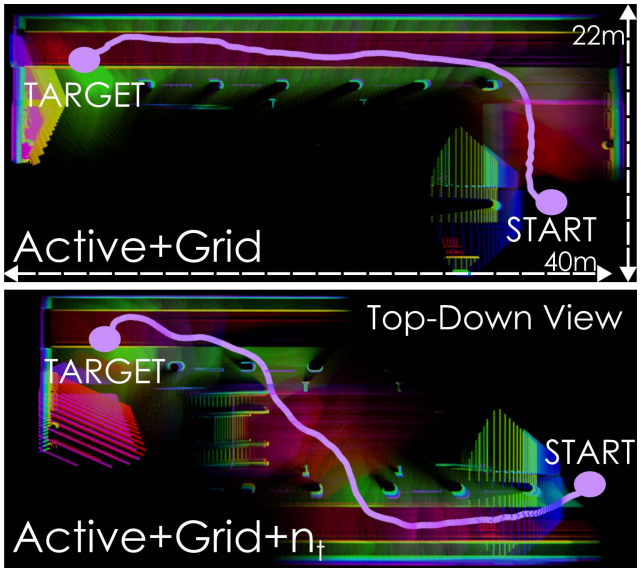


图 5. Gazebo 火车站导航实验。俯视图对比了两种方法的轨迹：一种不包含 n_t，另一种包含 n_t。后者展现出更强的空间感知能力，在朝目标导航的同时扫描了更多环境区域。

局部占据栅格对成功率的提升效果，远超将动作限制在传感器视场角内所带来的影响。

显然，引入配备局部网格的主动相机，或通过视场角限制运动，能使机器人在所有复杂度等级下显著降低碰撞率（≤2.7%），同时保持较高的成功率。有趣的是，只有不采用局部网格表示的主动相机方法，其碰撞率高于采用局部占据栅格的静态相机方法。引入局部占据栅格后，所有消融实验中的碰撞率均有所下降，凸显了它的必要性。最后，主动+网格+n_t 方法在分别存在 10、20 和 30 个障碍物时，取得了 97.4%、96.0% 和 94.3% 的导航成功率，表现与其他方法相当。这表明，主动感知与有效的空间表示相结合，能够实现鲁棒的导航性能，并能良好地适应环境复杂度的提升。除了导航成功率，消融研究还揭示了空间探索效率存在显著差异，如表 I 最右列所示。我们通过测量导航任务期间已探索环境体积的百分比来量化探索性能，具体计算方式是：从未知状态转变为自由或占据状态的体素数量与基于特权地面真值占据栅格的总体环境体积之比。评估在含有 30 个障碍物的环境中进行。静态相机配置的环境感知能力相对有限，在不同输入变体下，探索范围为 24.0% 至 29.6%。然而，主动相机方法展现出显著更强的环境探索能力。基础主动相机配置最多可探索 43.4%，凸显了主动感知对于理解环境的重要性。引入局部网格表示对这一指标没有显著影响，这表明它的主要作用是提升机器人安全性，对探索行为的影响有限。方法 (ix) 主动+网格+n_t 的探索率最高，达到 63.4%，比静态基线方法提升了 139%。这表明，主动感知不仅能提升导航安全性和成功率，还能显著增强机器人在任务执行过程中收集环境信息的能力，从而改善环境感知，并有望在复杂场景中支持更明智的决策。

B. 仿真研究

为进一步研究内在探索奖励 n_t 的影响，我们在 Gazebo 中使用火车站环境开展仿真实验（图 5）。四旋翼飞行器的任务是在避开碰撞的同时到达目标位置。我们比较了策略的两个变体，即主动+网格 和主动+网格 +n_t。后者加入了术语 n_t，鼓励智能体主动探索并发现新的空间信息。为进行受控比较，我们制定了一套标准化协议：机器人始终从同一位置（俯视图中下方轨道旁）出发，前往火车站站台另一端、另一条轨道旁约 42.5 米处的同一目标位置。此设置确保性能差异归因于奖励函数设计，而非环境变化；同时表明该方法并不局限于其训练所用的直线走廊状环境。在这两种情况下，均未向智能体提供任何关于环境的先验信息。

表 I 汇总的结果证实，加入奖励 n_t 并未损害导航性能，因为两种策略变体的目标到达

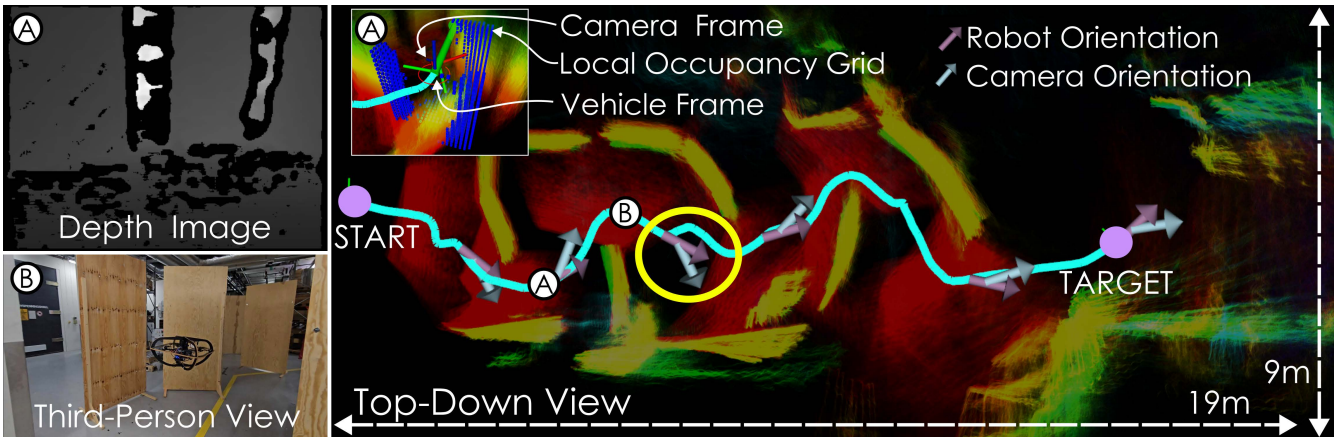


图 6. 机器人穿过杂乱走廊时的俯视图。青色轨迹和点云显示了机器人的路径与感知结果。灰色和紫色箭头表示相机朝向偏离机器人航向，以增强空间感知。图中展示了 A 时刻的网络输入（深度图像、局部占据栅格）以及 B 时刻的第三人称任务视图。

成功率相近。同时，加入n_t的策略显著提升了对环境的探索。在 Gazebo 火车站场景（图 5）中，采用主动相机控制并使用 n_t 的智能体探索了约 61% 的环境，而未使用 n_t 的策略变体在 5 次运行中（起点和目标位置相同）探索了约 47.5% 的环境。这些发现表明，内在奖励能促进更丰富的场景理解和更广泛的空间探索，同时不降低目标到达能力，从而验证了 n_t 在训练过程中作为一种有价值信号的作用。

C. 真实世界评估

本文考虑一种定制的四旋翼平台，如图 2 所示，该平台配备了惯性测量单元 (IMU)、用于里程计估计的雷达传感器，以及可驱动RGB-D相机系统。主动感知系统由一台 Intel RealSense D455 相机组成，该相机安装在机器人机架边缘的双轴驱动机构上。两个伺服电机采用云台配置，可对相机的俯仰角和偏航角进行独立的关节位置控制。每个伺服电机都配有集成电位计，可实时反馈其关节位置。这使相机能够实现精确的闭环位置控制，确保准确达到所需朝向。该机构可沿偏航轴旋转 $\pm 45^\circ$，沿俯仰轴旋转 $\pm 60^\circ$。RGB-D相机以最高 10 Hz 的频率输出同步的彩色与深度数据流，从而支持具备空间感知能力的场景理解和避障等任务。平台由机载 NVIDIA Jetson Orin NX 16GB 模块供电，该模块提供运行高层导航模块所需的计算能力。低层姿态稳定和电机控制由 PX4飞行控制器负责。高层系统与飞行控制器之间通过 ROS中间件通信。本实验选用主动+网格+n_t 策略，因为它兼具安全导航与探索的优势。该策略以 10 Hz 的频率在机器人上执行。

a) 杂乱走廊中的机动：为验证我们方法的实际适用性并展示仿真到现实迁移能力，我们在杂乱的室内走廊环境中部署主动+网格+n_t方法，如图 6 所示。实验设置包括一条狭窄的走廊（长 15 m，宽 3.0m），其中布置了障碍物和结构元素，构成了具有挑战性的导航场景，需要精确机动和避碰。该策略表现出色，能够成功穿行于狭窄空间并绕过障碍物，同时在穿越环境时转动相机进行观察。如图 6 中的黄色圆圈所示，智能体会主动将相机转向未探索区域，以全面理解周围空间，而不受当前导航朝向的影响。

*b) T*形走廊中的空间感知：为展示主动感知在另一种真实场景部署中的优势，我们在T形走廊环境中进行了一项实验（图7）。在此场景中，四旋翼飞行器需要在T形交叉口做出导航决策，以抵达目标位置。主动感知策略通过重新调整可驱动的RGB-D传感器方向，扫描走廊的侧向分支，从而在决定转向动作之前获取重要的空间信息，并成功适应了这一场景。因此，机器人避免了过早或错误的导航决策，同时持续向目标前进，最终成功抵达目标。我们在同一环境中将主动+网格+n_t 策略与静态+视场角方法 [13]进行比较。如图7右侧所示，静态+视场角方法无法在该环境中导航，无法继续朝目标位置前进，尤其会困在狭窄区域。为便于与 [13],进行比较，我们不得不通过移除两个面板来修改环境。即使在这种障碍较少的配置中，静态相机方法仍在行进途中与环境发生碰撞（红圈标记处），尽管最终到达了目标位置。这项实验表明，在具有挑战性的环境中，主动感知能够帮助系统建立更丰富的态势感知，而当相机始终固定朝向前方时则无法做到这一点。

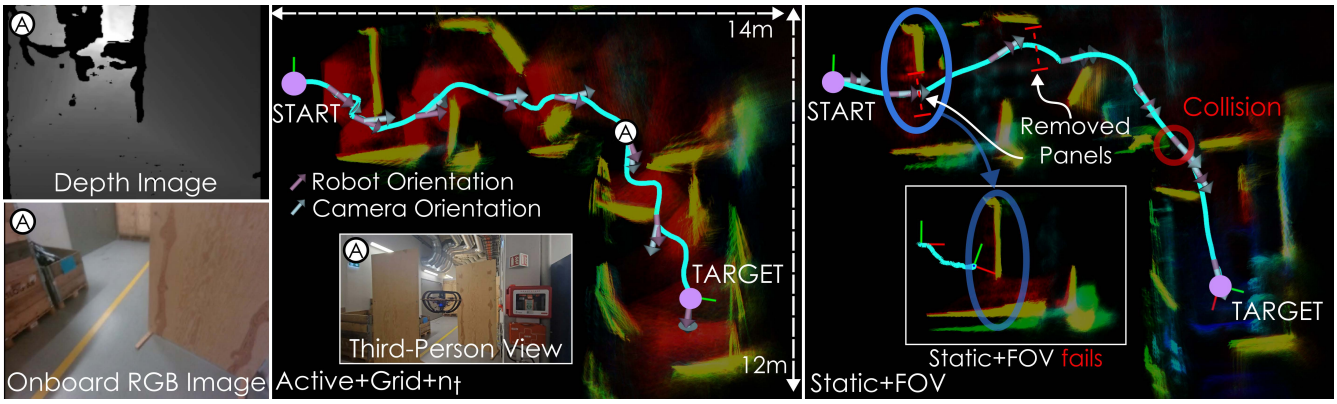


图7. T形走廊中的导航任务，对比了（主动+网格+n_t）与静态+视场角 [13]。轨迹展示了导航性能方面的差异。对于静态+视场角 方法，为了实现成功导航，对环境进行了调整以减少杂乱程度。红色圆圈表示与环境发生碰撞的位置。

环境中，而当相机始终固定朝向前方时，则无法做到这一点。

五、结论

本文提出了一种用于自主空中导航的新型强化学习框架，将主动感知整合到控制策略中。针对运动规划与信息驱动的视点选择相结合这一挑战，我们设计了一个多目标奖励函数，将导航目标与由探索驱动的信息增益项相结合。我们的方法使机器人不仅能够主动控制相机以增强态势感知、实现安全导航，还能有效探索环境。我们通过仿真和真实环境部署中的大量研究，展示了该框架的有效性。结果表明，与静态纯导航基线相比，所提出的策略能够提高成功率、减少碰撞，并改善环境观测。未来，我们计划将语义信息和推理纳入该框架，使智能体能够主动搜索特定的感兴趣对象。此外，我们还希望提升该方法在动态环境中的泛化能力，并探索实时策略适应。

参考文献

[1] S. Chen 等，《机器人系统中的主动视觉：近期进展综述》，《国际机器人研究杂志》，第30卷，第11期，第1343–1377页，2011年。[2] Y. Aloimonos 等，“主动视觉”，载于*DARPA*图像理解研讨会论文集，1987年，第552–573页。[3] R. Bajcsy，“主动感知”，《IEEE会刊》，第76卷，第8期，第966–1005页，1988年。[4] R. Bajcsy 等，“重新审视主动感知”，《自主机器人》，第42卷，第2期，第177–196页，2018年。[5] J. I. Vasquez-Gomez 等，“具有定位误差的三维物体重建体积式次优视点规划”，《国际先进机器人系统杂志》，第11卷，第10期，第159页，2014年。[6] A. Bircher 等，“三维探索的滚动时域次优视点规划器”，载于*IEEE*国际机器人与自动化会议。IEEE，2016年，第1462–1468页。[7] T. Dang 等，“应用于空中机器人的视觉显著性感知滚动时域自主探索”，载于*IEEE*国际机器人与自动化会议。IEEE，2018年，第2526–2533页。[8] M. G. Mateus 等，“基于深度强化学习的无人机主动感知”，载于*2022*年拉丁美洲机器人研讨会（*LARS*）、*2022*年巴西机器人研讨会（*SBR*）和*2022*年机器人教育研讨会（*WRE*）。IEEE，2022年，第1–6页。

[9] P. Mirowski等，“复杂环境导航学习”，arXiv预印本 arXiv:1611.03673，2016年。[10] Y. Zhu等，“使用深度强化学习在室内场景中进行目标驱动视觉导航”，载于IEEE国际机器人与自动化会议。IEEE，2017年，第3357–3364页。[11] D. S. Chaplot等，“使用目标导向语义探索进行物体目标导航”，《神经信息处理系统进展》，第33卷，第4247–4258页，2020年。[12] Y. Zhang等，“通过可微物理学习基于视觉的敏捷飞行”，《自然·机器智能》，第1–13页，2025年。[13] M. Kulkarni等，“利用深度碰撞编码实现无碰撞飞行的强化学习”，载于IEEE国际机器人与自动化会议。IEEE，2024年，第15 781–15 788页。[14] J. Lee、A. Rathod、K. Goel、J. Stecklein和W. Tabib，“利用特权信息进行强化学习的四旋翼导航”，arXiv预印本 arXiv:2509.08177，2025年。[15] X. Chen等，“用于空地无人系统导航的结合搜索式优化的多阶段深度强化学习”，《应用科学》，第13卷，第4期，第2244页，2023年。[16] T. Sun等，“通过因果强化学习实现无人机自主避障”，《显示器》，第87卷，第102966页，2025年。[17] L. Bartolomei等，“使用深度强化学习的无人机语义感知主动感知”，载于IEEE/RSJ智能机器人与系统国际会议。IEEE，2021年，第3101–3108页。[18] G. Malczyk等，“用于巡检路径规划的语义驱动深度强化学习”，IEEE机器人与自动化快报，2025年。[19] H. Oleynikova等，“用于在线无人机重规划的连续时间轨迹优化”，载于IEEE/RSJ智能机器人与系统国际会议。IEEE，2016年，第5332–5339页。[20] H. Nguyen等，“使用深度神经网络的不确定性感知视觉注意导航”，《国际机器人研究杂志》，第43卷，第6期，第840–872页，2024年。[21] A. Loquercio等，“在野外学习高速飞行”，《科学·机器人学》，第6卷，第59期，第eabg5810页，2021年。[22] M. Kulkarni等，“基于深度图像的碰撞编码任务驱动压缩”，载于国际视觉计算研讨会。Springer，2023年，第259–273页。[23] A. Petrenko等，“Sample factory：以100000帧每秒的速度通过像素进行自我中心三维控制与异步强化学习”，载于国际机器学习会议。PMLR，2020年，第7652–7662页。[24] L. Espeholt等，“IMPALA：采用重要性加权演员-学习者架构的可扩展分布式深度强化学习”，载于国际机器学习会议。PMLR，2018年，第1407–1416页。[25] M. Kulkarni等，“Aerial Gym模拟器：用于空中机器人高度并行化仿真的框架”，IEEE机器人与自动化快报，2025年。[26] T. Lee等，“四旋翼无人机在SE(3)上的几何跟踪控制”，载于第49届IEEE决策与控制会议（CDC）。IEEE，2010年，第5420–5425页。[27] K. Khoshelham，“Kinect深度数据的精度分析”，《国际摄影测量、遥感与空间信息科学档案》，第38卷，第133–138页，2012年。