arXiv:2302.02921v1 [cs.RO] 6 Feb 2023

基于深度强化学习的自主导航系统整体训练

Linh Kästner[1], Marvin Meusel[1],Teham Bhuiyan[1],和 Jens Lambrecht[1]

摘要—近年来，深度强化学习已成为地面车辆自主导航的一种很有前景的方法，并被应用于导航的多个领域，例如巡航控制、变道和避障。然而，大多数研究要么着重于提供一种端到端解决方案，利用深度强化学习训练整个系统；要么聚焦于某个特定方面，例如局部运动规划。但这会带来一系列问题，例如灾难性遗忘、导航行为低效，以及导航栈中不同实体之间的同步不理想。本文提出一种整体式的深度强化学习训练方法，其训练过程涵盖导航栈中的所有实体。这有望增强导航栈中各实体之间的同步与相互理解，从而提升导航性能。我们使用多个不同的观测空间训练了若干智能体，以研究不同输入对智能体导航行为的影响。经过与多种基于学习的方法和经典的基于模型的导航方法进行深入评估后，我们提出的智能体在效率和安全性方面均优于基线，实现了更短的路径长度、更少的绕行和更少的碰撞。

一、引言

随着人机协作变得至关重要，拥挤环境中的移动机器人导航日益成为一个需要重点考虑的方面。在自动导引车等工业场景中，地面车辆使用的传统导航栈采用 ROS 导航堆栈 [1], , 其中包括一个全局规划器，它根据全局地图计算从起点到目标位置的最优路径；还包括一个局部规划器，它利用传感器观测执行全局规划，以避开地图中未标示的动态障碍物。借助当前使用的导航栈，静态或轻度动态环境中的导航问题可以得到解决，但在高度动态环境中进行导航仍是一项艰巨任务 [2]。智能体需要在不可预测的移动障碍物附近高效生成安全动作，以避免发生碰撞。传统的基于模型的运动规划方法通常采用人工设计的安全规则来避开动态障碍物。然而，在密集环境中，障碍物未来的运动难以预测，因此难以通过人工设计导航行为 [3]。近年来，深度强化学习（DRL）作为一种端到端方法崭露头角，并在避障方面展现出优越性

[1]Linh Kästner、Marvin Meusel、Teham Bhuiyan 和 Jens Lambrecht 任职于德国柏林柏林工业大学电气工程与计算机科学学院工业级网络与云计算讲席linhdoan@win.tu-berlin.de

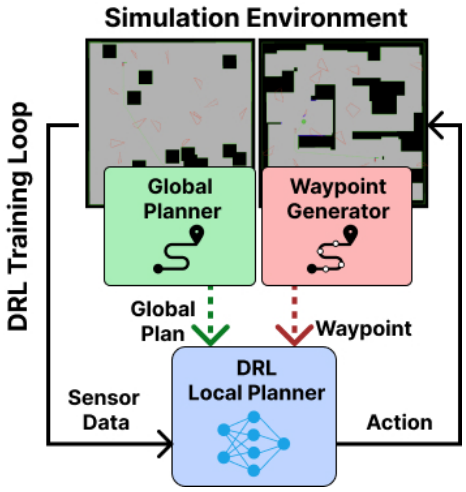

图 1：将全局规划器和航点生成器的信息作为输入提供给深度强化学习智能体，以增强高层规划器对深度强化学习智能体的理解，从而提高同步性和导航效率。

在动态环境中学习复杂的行为规则。因此，许多研究论文采用深度强化学习来解决抓取、导航或仿真等高层任务 [4], [5], [6], [7]。然而，基于深度强化学习的导航方法存在诸多问题，例如训练困难、深度强化学习智能体具有短视性，以及灾难性遗忘 [8], [9]。近期方法要么通过使用航点缩短规划时域来处理这一问题[10],[11]，要么采用混合方法 [8] [12]，在经典的基于模型的导航与深度强化学习规划器之间切换。然而，若将导航系统的各个部分分开处理，可能会导致同步问题，以及抖动运动或智能体偏离最初规划的全局路径过远等非最优行为。因此，本文提出一种整体训练方法，将全局规划器和航点生成器纳入深度强化学习训练流程。为此，我们将使用经典全局规划器（如快速扩展随机树或 A*算法），以及我们在先前工作中介绍的航点生成器，在训练过程中直接向智能体提供更多高层规划信息。这样应能提升智能体对导航栈其他组件所做决策的理解，从而使导航更加平稳、一致。我们在仿真平台 arena-rosnav 中，将不同的智能体输入与经典基线方法进行比较，并依据多种导航指标对所有智能体进行评估 [5]。本文的主要贡献如下：

- 提出一种整体训练方法，利用整个导航栈，而非采用孤立的训练流程
- 将全局规划和航点信息纳入智能体的奖励系统，以改善实体之间的同步，从而提升导航性能
- 在不同的高动态环境中进行定性和定量评估，并与基准深度强化学习及经典的基于模型的导航方法进行比较

本文结构如下。第二节介绍相关工作。随后，第三节阐述所提出的方法。第四节展示结果并进行讨论。最后，第五节总结全文并展望未来工作。

II. 相关工作

基于深度强化学习的导航方法已被证明是一种很有前景的替代方案，并已成功应用于车辆和机器人的多种导航方法，取得了显著成果。多项研究表明，基于深度强化学习的障碍物规避方法具有优势，因为它们在处理障碍物方面更加灵活，能够泛化到新的问题实例，并且无需手动设计功能即可学习更复杂的任务。因此，许多研究将深度强化学习融入导航系统，用于变道 [13],巡航控制 [14]，[15]，[16]，协同行为规划 [17]，以及避障 [6]，[7]，[5]等任务。Atouri 等人 [18]提出了一种用于自动驾驶车辆横向控制的基于深度强化学习的控制切换器。类似地，Kästner 等人提出了一种基于深度强化学习的控制切换器，用于在不同导航策略之间进行选择 [2]。Liu 等人提出了一种基于深度强化学习的方法，通过专家示范实现车辆在城市环境中的自动驾驶。

其他研究将深度强化学习用于动态障碍物规避。Everet 等人 [19] 和 Chen 等人 [7] 要求提供精确的障碍物位置，并采用基于深度强化学习的避障方法。Dugas 等人仅依靠深度强化学习进行导航 [6]。作者指出，这可能导致抖动运动，并在远距离导航时失败。由于深度强化学习智能体在大规模地图上进行远距离导航时通常只能获得稀疏奖励，因此智能体会因局部最小值而只适用于短距离导航。因此，许多研究将基于深度强化学习的局部规划与快速扩展随机树 [20] 或 A*算法 [21]等传统方法相结合。Faust 等人利用深度强化学习辅助基于概率路线图的全局规划器 [8]。类似地，Chiang 等人将深度强化学习与快速扩展随机树全局规划器结合使用。其他研究则使用航路点作为全局规划器与局部规划器之间的通信接口。这些点从全局路径中采样，并作为输入提供给深度强化学习智能体，以缩短其规划时域。Gundelring 等人 [22] 集成了一个基于深度强化学习的局部规划器与传统全局规划器，并通过采用给定静态前瞻距离的简单全局路径子采样，为深度强化学习局部规划器生成航点。类似地，Regler 等人 [23] 提出一种人工设计的子采样方法，以便将基于深度强化学习的局部规划器部署到传统导航栈中。这些方法的一个局限在于，对全局路径进行简单子采样缺乏灵活性，在复杂情况下可能造成阻碍，例如多名行人挡住通路时。

因此，其他研究采用了更智能的方式来生成航路点。Brito等人 [10] 提出了一种基于深度强化学习的航点生成方法，通过训练智能体学习代价到达模型，直接生成子目标，并由MPC规划器跟随这些子目标。更准确的代价到达值估计使MPC能够求解长期最优轨迹。类似地，Bansal等人 [24] 提出了一种名为 LB-WayPtNav 的方法，该方法使用基于监督学习的感知模块处理 RGB图像数据并输出航点。根据航点和机器人当前状态生成基于样条的平滑轨迹，并由传统的基于模型的线性反馈控制器跟踪，以导航至航点。然而，所提出的监督训练方法需要繁琐的数据采集阶段来提供带标注的训练数据。在我们先前的工作中，我们提出了多种更灵活的航点生成方法 [11], [5], [25]，并展现出在拥挤环境中进行远距离导航时更优的导航性能。

值得注意的是，前文提及的所有工作都将深度强化学习智能体作为独立实体进行训练，之后再将其集成到导航栈中，这可能导致若干问题，例如同步问题和导航行为低效。深度强化学习智能体几乎总是针对短距离避障进行训练，因此在长距离场景中会出现失效。此外，深度强化学习智能体的导航性能也在很大程度上取决于全局规划器或航点生成器的效率。基于此，本研究将导航栈中的所有实体都纳入训练流程。具体而言，由全局规划器、航点生成器和深度强化学习智能体构成的整个导航栈均部署在训练流程中。深度强化学习智能体仍负责局部避障，但会接收另外两个实体的高层输入，以更好地理解它们的决策。因此，这三个实体应能实现更好的同步与协同运作。

III. 方法

本章介绍我们所提出框架的方法。总共训练了六个使用不同输入的智能体。

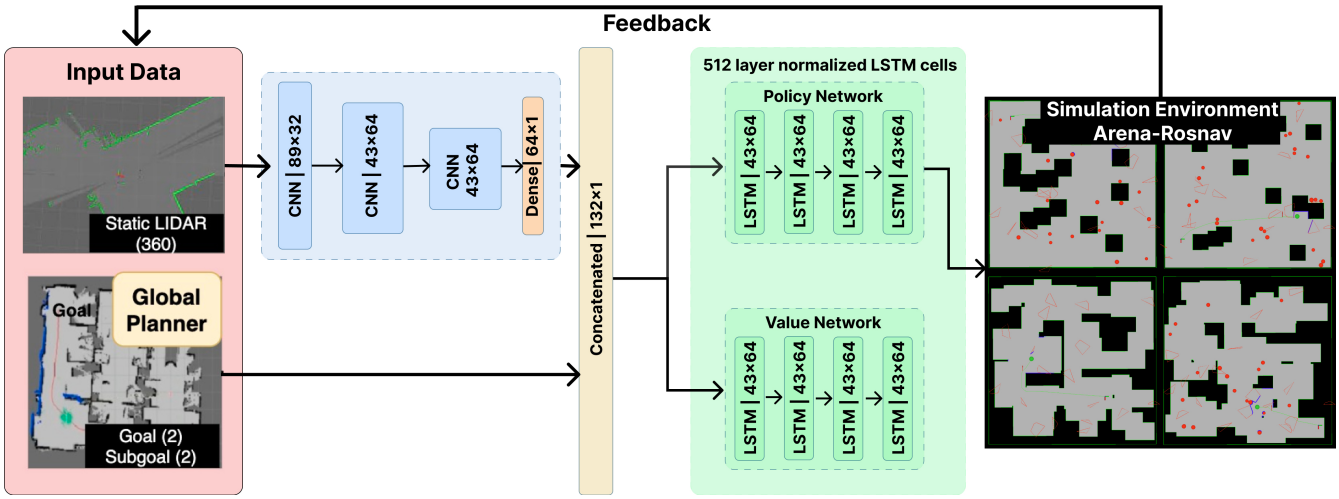

图 2：系统设计与训练流程。此处以智能体 4 为例展示输入。各智能体的具体参数和张量大小见表 I

A. 系统设计与训练过程

图 2 展示了我们方法的系统设计。深度强化学习智能体在我们的二维仿真环境 arena-rosnav [26] 中进行训练，并采用分阶段训练课程进行训练，即当智能体达到 80% 的成功率时，将加载难度递增的下一阶段。各阶段包含随机生成的动态障碍物和静态障碍物。各阶段如图 3 所示。阶段 1 是一个大小为 100×100 像素且不含任何障碍物的室外地图。阶段 2 是一个大小为 150×150 个网格、包含智能体已知静态障碍物的混合地图。阶段 3 是一个大小为 200×200 个网格、包含已知及额外未知静态障碍物的室外地图。阶段 4 是一个大小为 200×200 个网格、包含已知和未知静态障碍物的室内地图。阶段 5 是一个大小为 200×200 个网格、包含已知和未知静态障碍物以及额外未知动态障碍物的室外地图。阶段 6 是一个大小为 200×200 个网格、包含已知静态障碍物以及未知静态障碍物和动态障碍物的室内地图。阶段 7 与阶段 5 几乎相同，但包含更多未知静态障碍物和动态障碍物。

观测由深度强化学习智能体处理，并由其在环境中生成动作。与我们之前的工作 [5] 相比，训练深度强化学习智能体时并未将其与导航系统分离。相反，完整的导航栈被纳入训练循环。尽管更复杂的输入可能会增加开销并延长训练时间，但智能体应能学会更好地与全局规划器和航点生成器同步。

通常会仅使用深度强化学习训练用于局部避障的局部规划器，并将其集成到完整的导航栈中；而所提出的系统已经包含全局规划器和航点生成器，并在训练阶段使用其输入。因此，开发了 6 种不同的输入，以测试这些输入对智能体行为的影响程度。不同智能体的输入列于表 [?] 中。

B. 智能体与神经网络设计

我们总共提出了六种不同的智能体，每种智能体采用不同的观测空间，以研究不同输入的影响。智能体的输入列于表 I。所有智能体的内部架构和输出层均相同，区别仅在于输入层。

表 I：不同智能体的输入

智能体	Scan	全局目标	子目标	航路点	长度
智能体 1	0 - 359	360, 361	362, 363	364 - 463	–
智能体 2	0 - 359	360, 361	–	–	–
智能体 3	0 - 359	360, 361	362, 363	–	364
智能体 4	0 - 359	360, 361	362, 363	–	–
智能体 5	0 - 359	360, 361	–	–	362
智能体 6	0 - 359	360, 361	–	362 - 461	–

所有智能体都将激光雷达扫描数据和全局目标作为主要输入。全局目标由两个值表示，即从里程计位置到目标点的线性距离和角距离。所有点的表示方式都与全局目标相同。同样，可选的子目标表示为单个点，而全局规划则以两种不同方式表示：一种是用航路点表示。从全局规划中每隔 5 个点提取一个航点，直到总共提取 50 个点。如果规划长度不足以提取 50 个点，则用最后提取的航点填满航路点列表。第二种方式将整个规划简化为规划的总长度。

内部架构如图 2 所示。主体网络采用卷积神经网络，而行动者-评论家网络则使用长短期记忆单元进行设计，因为 [?]证明了使用记忆辅助网络进行导航的优势。智能体可能能够识别移动方向

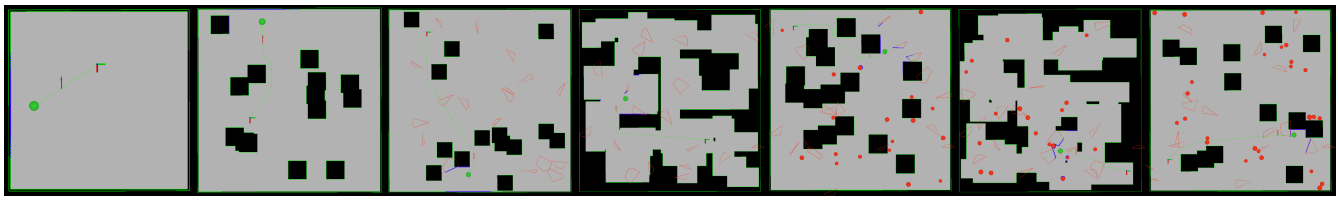

图 3：阶段一至七。

并记住较早的扫描数据，从而更好地探索该区域。

输出层由 2 个标量值组成。二者均用于为二维空间创建一条 *Twist* 消息。它由线速度和角速度组成。

C. 奖励函数

由于稀疏奖励无法使智能体快速收敛，我们将奖励函数设计为密集型，并在每次状态转移后返回奖励。只有发生碰撞，或智能体过于接近静态障碍物或动态障碍物时，才会给予负奖励。当智能体以合理的步数朝目标移动或到达目标点时，则给予正奖励：所需步数越少，奖励越高。公式 1 给出了智能体的奖励系统。奖励函数是所有子奖励之和

$$r^t = r^t_{gr} + r^t_c + r^t_{ga} + r^t_{sd} + r^t_{fgp} + r^t_{dgp} + r^t_{tc} + r^t_{adc} \tag{1}$$

其中，r^t_s 是到达目标点的成功奖励，r^t_c 是发生碰撞时的惩罚，二者都会导致回合结束。r^t_d表示接近目标点时的奖励。此外，我们引入了两项安全奖励：r^t_{ss} 用于帮助避开静态障碍物，而r^t_{sd} 则用于应对动态障碍物。

$$r^t_{gr} = \begin{cases} 45 & \text{, if goal reached} \\ 0 & \text{, otherwise} \end{cases} \tag{2}$$

$$r^t_c = \begin{cases} -50 & \text{, if collided} \\ 0 & \text{, otherwise} \end{cases} \tag{3}$$

$$r^t_{ga} = \begin{cases} 0.8 * diff^t_{robot,goal} & \text{, if } diff^t_{robot,goal} > 0 \\ 0.6 * diff^t_{robot,goal} & \text{, otherwise} \end{cases} \tag{4}$$

$$r^t_{sd} = \begin{cases} -1.25 & \text{, if } \exists o \in O : d(p^t_{robot}, p^t_o < D_s) \\ 0 & \text{, otherwise} \end{cases} \tag{5}$$

到达目标点会为智能体带来丰厚的正奖励。这是智能体的总体目的。实现这一目标的奖励设为 45。发生碰撞会导致 -50 的负奖励。靠近目标点似乎是良好的行为，尽管并非总是如此，例如遇到死胡同。因此，奖励不应设得过高。另一个目标是与障碍物保持一定的安全距离。智能体应避免行驶到距离障碍物仅 1 毫米的位置。计算取决于D_s，即智能体应保持的安全距离。该值设为 0.345 米。距离以智能体的中心为基准计算。由于智能体的半径为 0.3 米，因此智能体表面与障碍物表面之间的安全距离为 4.5 厘米。一旦智能体与障碍物的距离小于安全距离，就会给予负奖励。此外，结合全局规划器信息的奖励定义如下：

$$r^t_{fgp} = \begin{cases} 0.1 * vel^t_{linear} & \text{, if } \min_{wp \in G} d(p^t_{wp}, p^t_{robot}) < 0.5m \\ 0 & \text{, otherwise} \end{cases} \tag{6}$$

$$r^t_{dgp} = \begin{cases} 0.2 * diff^t_{robot,wp} & \text{, if } \frac{\min_{wp \in G} d(p^t_{wp}, p^t_{robot})}{diff^t_{robot,wp}} > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{7}$$

$$r^t_{adc} = -\frac{\left| vel^{t-1}_{angular} - vel^t_{angular} \right|^4}{1000} \tag{8}$$

其中 $diff_{tx,y} = d(p^{t-1}_x, p^{t-1}_y) - d(p^t_x, p^t_y)$。目标跟随奖励 r^t_{fgp} 根据线速度加权；当智能体与全局规划中的下一个点距离小于 0.5 米时，即给予该奖励。目标距离奖励是奖励智能体遵循全局规划的另一项组成部分，其作用是缩短智能体与全局规划中最近点之间的距离。此外，智能体导航旨在生成平滑路径。因此，角速度的突变会通过奖励r^t_{adc}予以惩罚。

D. 训练硬件配置

每个智能体都在两个不同系统之一上单独训练。表 II 展示了这两个系统的硬件规格。我们创建了一个 Docker 镜像，以便在这些系统上进行训练。

组件	系统 1	系统 2
CPU	Ryzen Threadripper 1950X	Ryzen R7 2700X
GPU	2x NVIDIA RTX 2080TI	1x NVIDIA RTX 2080TI
RAM	64GB	40GB

表 II：训练系统规格。此图展示了用于训练的计算机系统规格。

IV. 评估

本章介绍对智能体的评估。实验分为两类。第一部分，我们评估所有智能体的训练性能，并与不接收有关全局规划器和航点生成器的额外输入的基准智能体（智能体 2）进行比较，以评估训练所需的开销和复杂度。第二部分，我们将智能体与基准导航

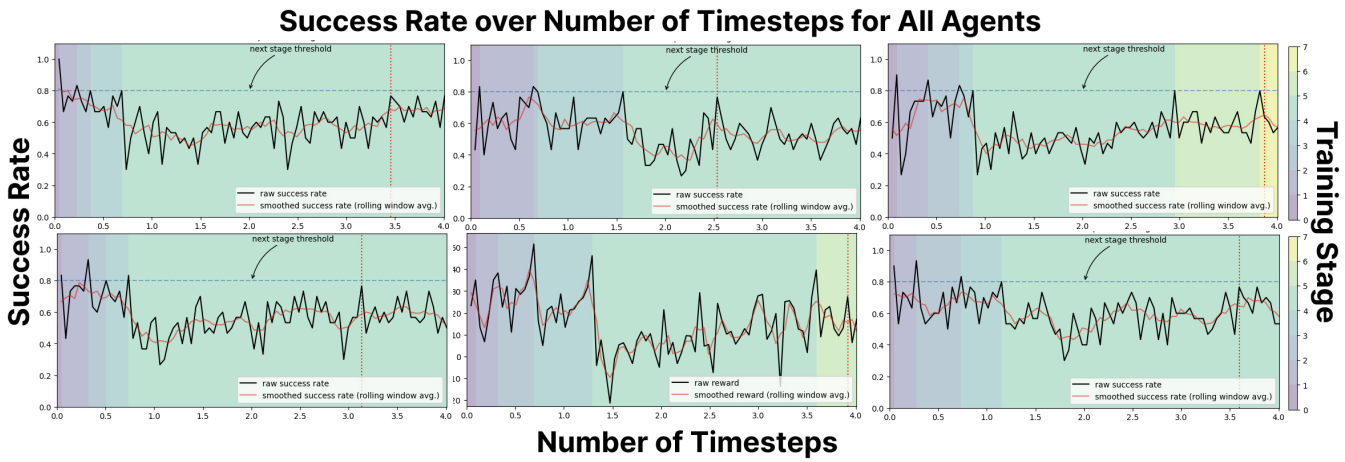

图 4：四种不同模型在三个不同场景中的想象结果。蓝色掩码覆盖的区域即为想象部分。四种模型分别基于不同的真实标注进行训练，即 60x60、60x60扩展版、100x100 和 100x100扩展版。想象结果基于观测生成，且与真实物体接近。使用扩展版真实标注训练的模型（标记为“Ext”）对物体的预测更为激进。

方法进行比较，考察路径长度、到达目标点所需时间等导航指标。因此，我们将智能体与经典局部路径规划器——定时弹性带（TEB） [27], 、动态窗口法（DWA） [28],和模型预测控制（MPC） [29] ——以及我们提出的一体化规划器进行比较。该规划器能够在经典 TEB 和深度强化学习规划之间切换 [2]。

A. 训练性能

为了评估训练过程，我们考察了无碰撞地成功完成任务的成功率。成功率如图 4 所示。图中还标出了阶段转换。总体而言，各智能体到达特定阶段的时间点差异显著。并非所有智能体都到达了最后的第 7 阶段。只有智能体 3 能够到达该阶段，而且是在训练后期。智能体 5 到达了第 6 阶段，其他所有智能体都只到达了第 5 阶段。第 5 阶段是首个包含移动障碍物的阶段。智能体 1 和智能体 4 最早到达第 5 阶段，约在训练步数 7M 时；智能体 2 则最晚，约在 15.5M 步时到达。额外输入似乎能提高较低阶段的学习速度。也就是说，使用额外输入的智能体比未使用额外输入的基准智能体 2 更早到达第 5 阶段。令人意外的是，输入相似的智能体 1 和智能体 6 之间的差异比预期更大。

正如预期，所有智能体的成功率在进入新阶段后都会下降，这表明难度有所增加。此外，所有智能体的成功率都有波动，且出现了双向的异常下降。尽管这些智能体有许多相似之处，但仍可观察到一些细微差异。例如，智能体 2 的波动似乎比其他智能体更大。此外，从第 25M 步开始，成功率似乎维持在 60% 左右。相比之下，智能体 1 在训练结束时仍呈现略微上升的趋势。此外，在最后 5M 步中，成功率约为 70%。智能体 3 是唯一达到第 7 阶段的智能体。不过，某些异常运行可能导致了这些阶段转换。

总而言之，智能体 1 的成功率最高，且最早到达第 5 阶段，因此可能是表现最佳的智能体之一，与能够到达最后一个第 7 阶段的智能体 3 并列。另一方面，基线智能体 2 最晚到达第 5 阶段，成功率也最低。此外，该智能体似乎不再有所进步。这些训练指标表明，额外输入对训练速度有积极影响，也暗示其性能有所提升。

B. 导航性能

在考察训练指标后，我们将智能体的导航性能与基线方法进行比较。这些方法包括基于模型的规划器 DWA[28], TEB [30],和 MPC [31], ，以及我们在先前工作中提出的 AIO 规划器，该规划器能够在 TEB 和深度强化学习 [2] 之间切换。评估时，在相同的随机场景和任务中，分别让每个智能体运行 150 回合。每种场景均进行了 150 回合。比较重点关注成功率、路径长度、回合时间和碰撞。为定性评估导航性能，我们在三种不同场景中测试了所有方法：a) 有 20 个障碍物；b) 有障碍物簇；c) 有跑动障碍物，且障碍物速度最高达 1m/s。场景中的起点和目标位置固定，障碍物则按照 Pedsim社会模型 [32]移动。图 6 展示了智能体 3 和 5 的示例性定性轨迹。我们每隔 100 ms 采样一次时间步，并将所有方法的采样点显示在轨迹中。障碍物轨迹上标注了起始和结束时间，单位为

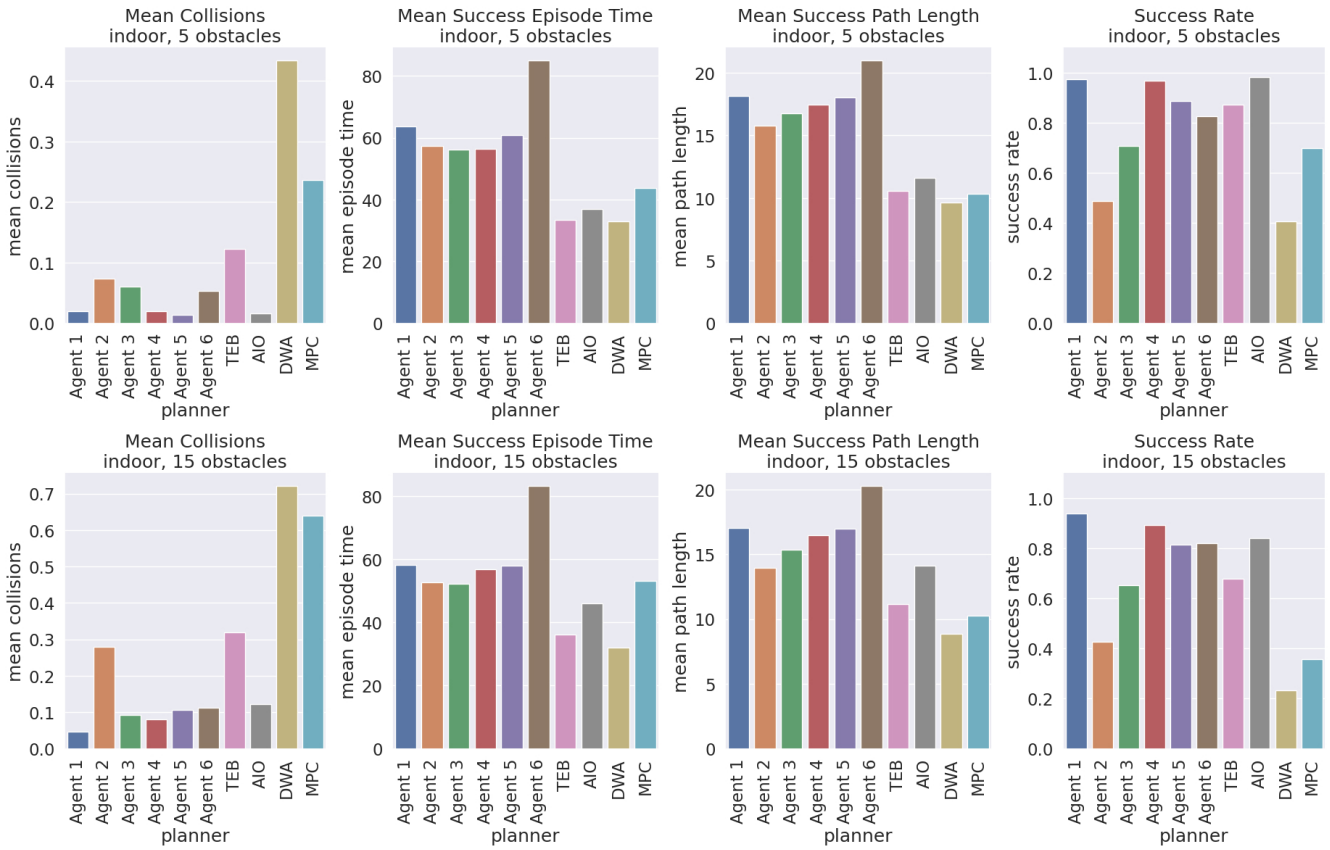

图5：四种不同模型在三个不同场景中的想象结果。蓝色掩码覆盖的区域即为想象部分。四种模型分别基于不同的真实标注进行训练，即 60x60、60x60扩展版、100x100 和 100x100扩展版。想象基于观测，并且与真实物体相近。使用扩展版真实标注（标记为“扩展版”）训练的模型对物体的预测更为积极。

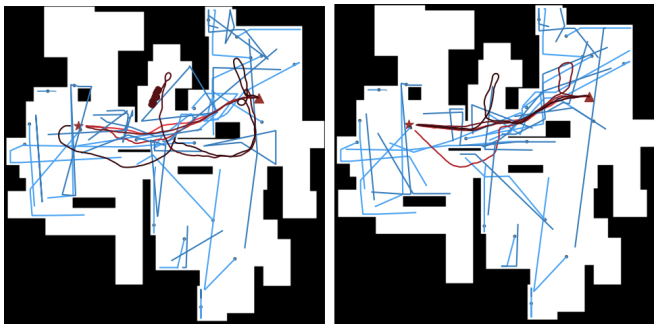

图 6：我们提出的智能体能够在拥挤环境中完成跟随人类和引导任务。它使用额外的语义信息进行训练，例如人类交谈、跑步等社交状态，从而能够推理并端到端地执行高级语义任务。

秒。发生碰撞后，该回合即结束。基准比较采用四项指标：无碰撞到达目标点的成功率、平均碰撞次数、平均路径长度和平均时间。如果智能体与障碍物发生碰撞或出现超时，则该回合被视为未成功。图 5 展示了所有规划器的结果。

结果显示，智能体 1 表现最佳，成功率为 97.3%，平均碰撞次数为 0.02；而智能体 2 的成功率最低，为 48.6%。令人意外的是，智能体 3 的表现相当差，成功率仅为 70.6%，平均碰撞次数为 0.06。其他 4 个智能体表现相近，能够与 AIO 和 TEB 规划器有力竞争。除基准智能体 2 和智能体 3 外，所有智能体的表现都优于经典规划器 DWA 和 MPC。智能体 1 的路径长度略高于智能体 2、3 和 4，但总体而言，所有智能体的路径长度都长于经典规划器。

在有 15 个障碍物的场景中，基于深度强化学习的智能体与经典基线之间的差异更加明显。6 个智能体的成功率变化不大，而基线规划器的成功率下降则更为显著。智能体 2 的平均碰撞次数从 0.94 大幅增加到 5.18。其他智能体和规划器的平均碰撞次数也有所增加，但增幅并未超过成功率下降所能解释的程度。路径长度、平均时间和速度没有发生明显变化。

总体而言，智能体 1 在 6 个智能体中表现最佳，障碍物数量增加并未显著影响其性能。在有 5 个障碍物的地图上，其表现与 AIO规划器相近；在有 15 个障碍物的地图上则略胜一筹。不出所料，智能体 2 在 6 个智能体中表现最差。令人惊讶的是，智能体 3 的表现也相当糟糕，尽管它是唯一一个达到最后训练阶段的智能体。尤其是智能体 2 和智能体 3 的路径长度都比较长。由于这两个智能体的超时次数最多，它们有时可能会在地图上徘徊，始终找不到目标点。在某些情况下，有 15 个障碍物的场景路径长度反而低于有 5 个障碍物的场景。

图 6 展示了智能体 1 与基准智能体 2 的路径对比。可以看到，获得额外全局信息的智能体会生成更直接通往目标点的轨迹，而仅使用激光雷达信息的基准智能体则会生成许多迂回路径。这表明全局规划器与局部规划器之间的协作更加有效，也使轨迹更加平滑。补充视频直观展示了我们规划器的导航行为。

V. 结论

本文提出了一种整体式深度强化学习训练流程，涵盖 ROS 导航堆栈的所有组件和实体；该导航堆栈通常用于自动导引车等工业地面车辆，以提升各实体之间的同步性。我们没有将导航栈的各个实体——全局规划器、航点生成器和局部规划器——分别处理，而是让训练涵盖所有实体，使深度强化学习智能体能够更全面地理解它们。因此，我们将全局规划和航点生成器的信息整合到所训练智能体的观测空间中。我们总共提出了六种具有不同观测组合的智能体，以探索不同输入对智能体训练和导航性能的影响。与经典及基于学习的基线方法相比，其他实体的额外信息有助于提升导航性能，实现更高的成功率、更少的碰撞和更短的路径长度。未来工作将探索行人或车辆的语义信息等额外输入参数对训练和导航性能的影响。此外，还应将这些方法部署到真实机器人上。

参考文献

[1] E. Marder-Eppstein、E. Berger、T. Foote、B. Gerkey 和 K. Konolige，“The office marathon：室内办公环境中的鲁棒导航”，载于*2010* 年 *IEEE* 国际机器人与自动化会议。IEEE，2010 年，第 300–307 页。[2] L. Kästner、J. Cox、T. Buiyan 和 J. Lambrecht，“All-in-one：一种基于深度强化学习的控制切换器，结合了最先进的导航规划器”，载于 *2022International Conferenceon RoboticsandAutomation (ICRA)*，2022 年，第 2861–2867 页。

[3] K. Qian、X. Ma、X. Dai 和 F. Fang，“服务机器人符合人类需求的导航中的社会可接受碰撞前安全策略”，*Advanced Robotics*，第 24 卷，第 13 期，第 1813–1840 页，2010 年。[4] Y. F. Chen、M. Everett、M. Liu 和 J. P. How，“基于深度强化学习的具有社会意识的运动规划”，载于*2017* 年 *IEEE/RSJ* 国际智能机器人与系统会议 *(IROS)*。IEEE，2017 年，第 1343–1350 页。[5] L. Kästner、T. Buiyan、X. Zhao、L. Jiao、Z. Shen 和 J. Lambrecht，“将基于深度强化学习的避障部署到传统自主导航系统”，*arXiv* 预印本 *arXiv:2104.03616*，2021 年。[6] D. Dugas、J. Nieto、R. Siegwart 和 J. J. Chung，“Navrep：用于动态人类环境中机器人导航强化学习的无监督表征”，*arXiv* 预印本 *arXiv:2012 04406*，2020。。[7] C. Chen、Y. Liu、S. Kreiss 和 A. Alahi，“人群-机器人交互：基于注意力机制的深度强化学习的人群感知机器人导航”，载于*2019* 年国际机器人与自动化会议 *(ICRA)*。IEEE，2019 年，第 6015–6022 页。[8] A. Faust、K. Oslund、O. Ramirez、A. Francis、L. Tapia、M. Fiser 和 J. Davidson，“PRM-RL：结合强化学习与基于采样的规划完成长距离机器人导航任务”，载于*2018* 年 *IEEE* 国际机器人与自动化会议 *(ICRA)*。IEEE，2018 年，第 5113–5120 页。[9] X. Xiao、B. Liu、G. Warnell 和 P. Stone，“使用机器学习进行移动机器人导航的运动规划与控制：综述”，*Autonomous Robots*，第 1–29 页，2022 年。[10] B. Brito、M. Everett、J. P. How 和 J. Alonso-Mora，“下一步该去哪里：学习用于动态环境导航的子目标推荐策略”，*IEEE Robotics and Automation Letters*，第 6 卷，第 3 期，第 4616–4623 页，2021 年。[11] L. Kästner、X. Zhao、T. Buiyan、J. Li、Z. Shen、J. Lambrecht 和 C. Marx，“利用航点生成器将基于深度强化学习的避障与传统全局规划器相结合”，载于*2021* 年 *IEEE/RSJ* 国际智能机器人与系统会议 *(IROS)*。IEEE，第 1213–1220 页。[12] H.-T. L. Chiang、A. Faust、M. Fiser 和 A. Francis，“使用 AutoRL 端到端学习导航行为”，*IEEE Robotics and Automation Letters*，第 4 卷，第 2 期，第 2007–2014 页，2019 年。[13] F. Ye、P. Wang、C.-Y. Chan 和 J. Zhang，“基于元强化学习的自动驾驶车辆变道策略”，载于*2021* 年 *IEEE* 智能车辆研讨会 *(IV)*。IEEE，2021 年，第 223–230 页。[14] U. Yavas、T. Kumbasar 和 N. K. Ure，“用于高级自适应巡航控制的基于模型的强化学习：混合式跟车策略”，载于*2022* 年 *IEEE* 智能车辆研讨会 *(IV)*。IEEE，2022 年，第 1466–1472 页。[15] X. Chen、J. Wei、X. Ren、K. H. Johansson 和 X. Wang，“基于近端策略优化的双向道路车辆交互自动超车”，载于*2021* 年 *IEEE* 智能车辆研讨会 *(IV)*。IEEE，2021 年，第 1057–1064 页。[16] M. Brosowsky、F. Keck、J. Ketterer、S. Isele、D. Slieter 和 M. Zöllner，“通过施加状态特定安全集实现自适应巡航控制的安全深度强化学习”，载于*2021* 年 *IEEE* 智能车辆研讨会 *(IV)*。IEEE，2021 年，第 488–495 页。[17] K. Kurzer、M. Bitzer 和 J. M. Zöllner，“结合逆强化学习与蒙特卡洛树搜索，为协同轨迹规划学习奖励模型”，载于*2022* 年 *IEEE* 智能车辆研讨会 *(IV)*。IEEE，2022 年，第 22–28 页。[18] H. Atoui、O. Sename、V. Milanés 和 J. J. Martinez，“基于强化学习的自动驾驶车辆智能控制切换”，载于*2022* 年 *IEEE* 智能车辆研讨会 *(IV)*。IEEE，2022 年，第 792–797 页。[19] M. Everett、Y. F. Chen 和 J. P. How，“利用深度强化学习在动态决策智能体之间进行运动规划”，载于*2018* 年 *IEEE/RSJ* 国际智能机器人与系统会议 *(IROS)*。IEEE，2018 年，第 3052–3059 页。[20] S. M. LaValle等，“快速探索随机树：一种新的路径规划工具”，1998 年。[21] P. E. Hart、N. J. Nilsson 和 B. Raphael，“最小代价路径启发式确定的形式化基础”，*IEEE* 系统科学与控制论汇刊，第 4 卷，第 2 期，第 100–107 页，1968 年。[22] R. Güldenring、M. Görner、N. Hendrich、N. J. Jacobsen 和 J. Zhang，

“用于室内环境中人类感知导航的局部规划器学习。”[23] P. Regier、L. Gesing 和 M. Bennewitz，“杂乱环境中的导航深度强化学习”，2020 年。[24] S. Bansal、V. Tolani、S. Gupta、J. Malik 和 C. Tomlin，“结合最优控制与学习，在新环境中实现视觉导航”，载于机器人学习会议。PMLR，2020 年，第 420–429 页。[25] L. Kästner、X. Zhao、Z. Shen 和 J. Lambrecht，“用于深度强化学习导航方法长距离引导的避障航点生成”，*arXiv* 预印本 *arXiv:2109.11639*，2021 年。[26] L. Kästner、C. Marx 和 J. Lambrecht，“动态环境中移动机器人的基于深度强化学习的语义导航”，载于*2020*年*IEEE*第*16*届自动化科学与工程国际会议 *(CASE)*。IEEE，2020 年，第 1110–1115 页。[27] C. Rösmann、F. Hoffmann 和 T. Bertram，“不同拓扑结构中的多机器人轨迹规划”，载于*2015*年欧洲移动机器人会议 *(ECMR)*。IEEE，2015 年，第 1–6 页。[28] O. Khatib，“机械臂与移动机器人的实时避障”，载于自主机器人车辆。施普林格，1986 年，第 396–404 页。[29] C. Rösmann、A. Makarow 和 T. Bertram，“基于非线性模型预测控制和非欧旋转群的在线运动规划”，*arXiv* 预印本 *arXiv:2006.03534*，2020 年。[30] C. Rösmann、F. Hoffmann 和 T. Bertram，“用于时间最优点到点非线性模型预测控制的定时弹性带”，载于*2015*年欧洲控制会议 *(ECC)*。IEEE，2015 年，第 3352–3357 页。[31] C. Rösmann，“时间最优非线性模型预测控制”，博士论文，博士论文，多特蒙德工业大学，2019 年。[32] D. Helbing 和 P. Molnar，“行人动力学的社会力模型”，《物理评论*E*》，第 51 卷，第 5 期，第 4282 页，1995 年。