DOI：10.13973/j.cnki.robot.220440

基于 TS-TD3 的动态环境端到端无地图导航方法

姜 杨，曾铁文，万东东，吴成东
（东北大学机器人科学与工程学院，辽宁 沈阳 110169）

摘 要：针对基于地图的移动机器人导航框架部署在动态复杂环境时出现的问题，提出一种基于时序－双延迟深度确定性策略梯度（TS-TD3）的无地图导航方法。首先，将动态场景（具有环境部分可观测性）的导航任务定义为部分可观测马尔可夫决策过程（POMDP）。其次，引入经过长短期记忆组件处理的历史信息作为模型的输入，为策略网络的确定性策略梯度引入历史信息基准，以处理隐藏在环境观测集合中的状态信息，将关注导航动作时序关联性的评价标准引入评价网络。再次，通过专家经验网络在训练前期指导策略网络的输出，以规范导航动作。最后，建立演员－评论家框架的深度强化学习（DRL）端到端模型，根据传感器感知结果直接输出控制动作。与主流 DRL 方法进行对比实验，在仿真实验中，该方法运动轨迹自然、稳定、具有连续性，能处理多动态障碍物交汇情况，整体导航效果表现最优；在真实动态环境的测试中，模型未作调整直接部署在未知环境中，模型的导航效果和泛化性得到验证。

关键词：深度强化学习；部分可观测马尔可夫决策过程；确定性策略梯度；动态环境；无地图导航
中图分类号：TP242 **文献标识码**：A **文章编号**：1002-0446(2023)-06-0655-15

An End-to-End Mapless Navigation Method Based on TS-TD3 in Dynamic Environment

JIANG Yang，ZENG Tiewen，WAN Dongdong，WU Chengdong
(*School of Robotics Science and Engineering*, *Northeastern University*, *Shenyang* 110169, *China*)

Abstract: Aiming at the problems of map-based mobile robot navigation framework deployed in dynamic complex environment, a mapless navigation method based on TS-TD3 (time series twin delayed deep deterministic policy gradient) is proposed. Firstly, navigation tasks in dynamic scenarios (usually with partially observable environment) are defined as partially observable Markov decision process (POMDP). Secondly, the historical information processed by the long short-term memory components is introduced as the input of the model. The historical information baseline is introduced into the deterministic policy gradient of the actor network to process the state information hidden in the environmental observation set. The critic criteria concerned with the temporal correlation of navigation actions is introduced into the critic network. Thirdly, the expert experience network is used to guide the output of the actor network in the early stage of training to standardize the navigation actions. Finally, the deep reinforcement learning (DRL) based end-to-end model of the actor-critic framework is established, and the actions are controlled directly according to the sensor perception. Compared with the mainstream DRL methods, the motion trajectory obtained by the proposed method is natural, stable and continuous in the simulation experiment, the intersection of multiple dynamic obstacles can be dealed with, and the overall navigation performance is optimal. In the test in real dynamic environment, the model is directly deployed in an unknown environment without adjustment, and the navigation effect and generalization of the model are verified.

Keywords: deep reinforcement learning; partially observable Markov decision process; deterministic policy gradient; dynamic environment; mapless navigation

1 引言（Introduction）

准确、快速、安全地到达目的地是移动机器人执行更高级任务的基本前提。近年来，随着 SLAM（同步定位与地图创建）技术的快速发展，基于地图的导航技术取得了不俗的成绩[1]。然而，当机器人部署在动态、复杂场景时，基于地图的导航框架会出现一些问题：动态场景中出现大量自主移动的障碍物（如行人），机器人必须随时改变移动的路径以避开动态障碍物，动态障碍物频繁出现或是消失，上述情况均对机器人计算平台的算力和所搭载传感器的精度提出一定挑战。先验地图、全局路径

基金项目：国家自然科学基金（U20A20197）.
通信作者：姜杨，jiangyang@mail.neu.edu.cn 收稿/录用/修回：2022-12-05/2023-01-16/2023-01-18

规划器、局部避障规划器，这些独立工作的子系统在获取感知信息，传递信息，将信息整合进全局地图的过程中将不可避免地产生误差[2-3]。

此外，建立具有可靠精度的全局地图需要大量的先验知识与高性能的传感器，已有地图的维护与更新同样费时费力。

深度强化学习（DRL）技术在围棋[4]、游戏[5]等方面被证明拥有超越人类的能力。基于 DRL 的无地图导航被认为是将导航技术从流水线式的传统导航框架中解放出来的一种方式。DRL 模型通过机器人与导航环境直接交互、反复试错来探索一种到达目标地点的最佳策略，训练样本不需要人工标签[6]。采用端到端框架的深度强化学习模型可以根据传感器的环境感知结果直接输出机器人的控制动作[7]。端到端的模型架构使机器人拥有快速反应的潜力，从而可以在复杂的动态环境中导航。

现有研究中，确定性策略梯度（DPG）解决了强化学习中的连续控制问题[8]。深度确定性策略梯度（DDPG）模型[9]将 DPG 应用于深度神经网络，其核心思想在于使用神经网络 $Q_\theta(s,a)$ 来近似基于策略 π 的动作－价值函数 $Q_\pi(s,a)$，使用神经网络 $\pi_\omega(s)$ 近似确定策略 $a=\pi(s)$，而不是离散动作的概率分布 $a\sim\pi(s)$。基于演员－评论家（actor-critic）框架思想的双延迟确定性策略梯度（TD3）[10]使用裁剪双 Q 学习网络减缓了 Q 值的过高估计，使用延迟更新的目标网络（target network）减小了误差[10]。上述基于 DPG 的 DRL 模型适用于具有连续动作空间的任务，能够处理高维数据输入。

Tai 等[11]使用基于确定性策略梯度的 DRL 模型，以平面激光雷达扫描数据作为模型输入，训练出主要部署于静态、狭窄环境的移动机器人导航策略。Xie 等[12]针对现有 DRL 模型需要大量试错且训练曲线震荡难以收敛的问题，引入独立控制器在训练前期指导机器人的导航动作，其核心思想在于独立控制器能够加速模型训练而不是让机器人随机探索。Everett 等[13]提出了应用于复杂动态环境的 DRL 导航方法，但该方法对动态障碍物提出了严苛的要求，包括预先知道动态障碍物的准确位置（这在相互遮挡的动态环境中难以实现）和所有动态障碍物的预期速度。

现有 DRL 模型难以实际解决动态复杂环境的无地图导航问题，模型大多假设导航任务是马尔可夫决策过程（MDP），即决策过程具备马尔可夫性质（Markov property）[13-14]，模型可以直接获得包含机器人与环境交互的所有方面的信息（状态），状态信息将对未来的决策产生影响。这在相对静止的环境中被证明有效，但不适用于动态场景。

关于现有研究难以应用于动态复杂环境的原因，本文总结出 3 点：(1) 在 MDP 中，机器人可以直接观测包含周围环境所有信息的完整状态[13]，这显然不适用于动态环境。(2) 在 MDP 中，表征环境动态特征的状态转移函数仅与当前时刻的状态与动作有关，不考虑之前的状态与动作，导致模型不关注机器人导航动作的时序关联性[14]，使得现有模型训练获得的导航策略，存在反复探索以追求更高累积回报的问题、“机器人摆头”以追求路径最优性的问题。(3) 除此之外，DRL 模型通过机器人与环境直接交互产生训练样本，样本无需人工标签，因此需要反复试错以收集大量训练样本[15]，并存在训练曲线振荡、难以收敛的问题。

针对上述问题，本文提出一种适用于动态复杂场景的基于 TS-TD3（时序－双延迟深度确定性策略梯度）的无地图导航方法，主要贡献有：

(1) 基于 POMDP（部分可观测马尔可夫决策过程），使用环境观测代替环境状态。为 DPG 引入历史信息基准，引导确定性策略的探索，以解决动态环境的部分可观测问题。

(2) 将经过 LSTM（长短期记忆）组件（包括 LSTM 和全连接层）处理的历史信息作为模型的输入，通过 TS-TD3 模型内部的 LSTM 选择性地保留有价值的历史信息片段，使模型关注前后时刻机器人导航动作的时序关联性。

(3) 引入专家经验网络在训练前期指导时序策略网络的输出，使用符合一定规范的人工操作引导机器人的导航动作，加速模型的训练。

2 问题描述与求解框架（Problem description and solution framework）

2.1 动态环境下无地图导航任务描述及抽象化

在动态环境下，机器人搭载的传感器（如平面激光雷达）只能观测到部分环境信息，传感器采集的单帧雷达扫描数据无法捕捉动态障碍物的运动特征，如速度信息、转向信息等；动态障碍物可能会突然出现，它们之间也可能相互遮挡。因此，此时关于环境具有部分可观测性的理解是很自然的。

本文将动态复杂环境下的无地图导航任务定义为 POMDP。POMDP 是 MDP 的一般化，去除了环境状态完全可观测的假设，本文定义为 (S,A,P,R,O,H)。状态集合 S 无法被直接观测，而是从观测集合 O 接收基于隐藏状态 $p(o|s)$ 的观测结果；观测集

合 O 由单线激光雷达扫描数据和目标地点相对于机器人的极坐标拼接组成；动作集合 A 包括机器人的线速度 v_t 和角速度 ω_t；状态转移 P 表示机器人在当前状态 s_t，执行动作 a_t 后，状态 s_t 转移到下一时刻状态 s_{t+1} 的概率分布；奖励函数 r_t 表示机器人在当前状态 s_t 下执行动作 a_t 获得的奖励。

在 POMDP 条件下，状态转移 P 与当前观测 o_t 有关，o_t 与隐藏状态的转移函数 $p(o_t|s_t)$ 相关[16]，包含当前状态 s_t 的部分信息，本文在 DeepMind 模型[16] 的基础上，通过定义历史信息集合 H 来解决动态环境的部分可观测性问题，提出动态障碍物的隐藏状态（如运动学信息）与连续多个时刻的观测有关，表示为 $p(o_1|s)p(o_2|s)\cdots$，历史信息集合 H 包含了隐藏在不同时刻（障碍物被遮挡然后出现）或是连续多个时刻（障碍物连续运动）的观测值 o_t 中的状态信息，集合 H 由历史信息 h_i 构成，h_i 由一次完整导航过程中采集的每一步观测值、执行动作、环境奖励组成 $h_i=(o_1,a_1,r_1,\cdots,o_i,a_i,r_i)$。

本文将无地图导航定义为基于 POMDP 的连续控制任务，其目标是学习一个导航策略 π，π 根据当前环境观测 o_t 和历史信息 h_i 执行使期望折扣回报（expected discounted return）最大化的动作 a_t。式 (1) 中期望折扣回报 J 与连续时刻的环境奖励 r_t 有关，折扣系数 γ 制约未来时刻奖励的估计值对当前期望折扣回报的影响。

$$J=E_s\left(\sum_{t=1}^{\infty}\gamma^{t-1}r(s_t,a_t)\right) \tag{1}$$

2.2　TS-TD3 模型的总体框架

TS-TD3 算法采用演员－评论家框架，以及从经验池中抽取样本的训练方式。TS-TD3 模型由时序策略网络、双时序评价网络和专家经验网络组成，总体框架如图 1 所示。

时序策略网络（TS-actor network）的输入包括当前观测 o_t 和随机抽样的完整历史信息 h_i，由时序策略网络的 LSTM 组件对历史信息 h_i 进行总结提炼，时序策略网络的输出是一组确定动作，包括线速度 v_t 和角速度 ω_t。时序评价网络（TS-critic network）的输入包括当前观测 o_t、历史信息 h_i、执行动作 a_t。时序评价网络的输出是当前执行动作 a_t 的评价值 Q，时序评价网络的内置 LSTM 对历史信息的时序关联性进行总结，使得评价值 Q 可以反映前后时刻导航动作的关联性。专家经验网络输出专家动作 a_g，通过硬注意力机制[17] 对机器人执行动作 a_t 施加影响，用以在训练前期引导时序策略网络的输出，使机器人的导航动作符合人工操作的特定规范，加快训练速度。

TS-TD3 模型根据观测 o_t 和历史信息 h_i 执行动作 a_t，获得奖励 r_t 和下一时刻观测值 o_{t+1}，并将获得的样本 (o_t,a_t,r_t,o_{t+1}) 存储到经验缓冲区 β，同时更新历史缓冲区 β_h。经验缓冲区 β 存储序列 (o_t,a_t,r_t,o_{t+1})，历史缓冲区 β_h 存储每次完整导航过程的轨迹 $h_i=(o_1,a_1,r_1,\cdots,o_i,a_i,r_i)$。缓冲区随机抽样的特点提升了模型训练效率，使模型可以利用一组不相关的样本开展学习，由于缓冲区的规模有限，因

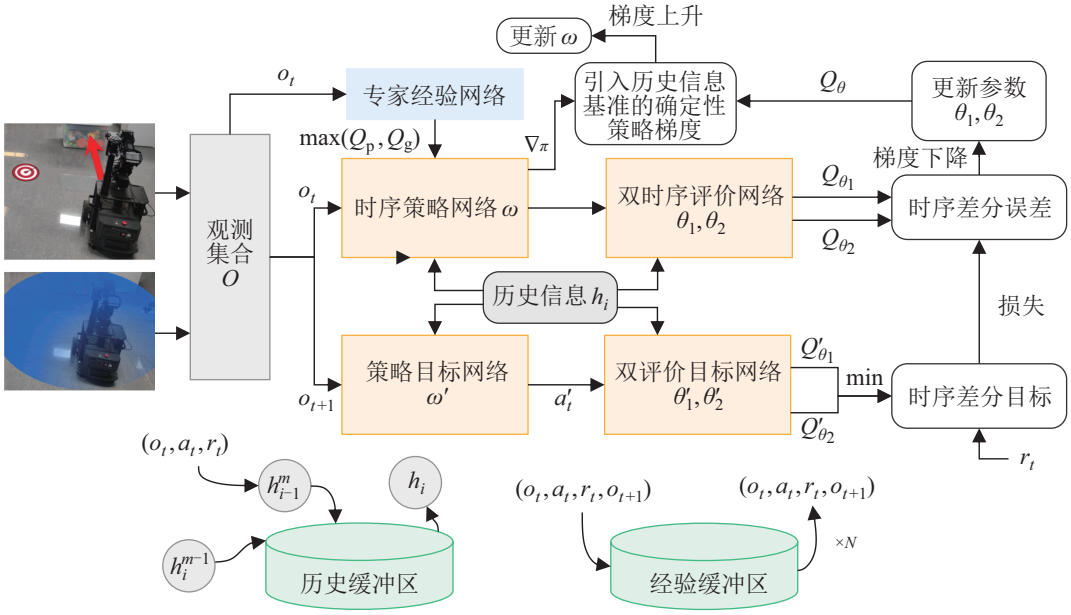

图 1　TS-TD3 算法框架

Fig.1　Framework of TS-TD3 algorithm

此样本需要不断更新[9-10]。经验缓冲区 β 和历史缓冲区 β_{h} 独立更新。

TS-TD3 模型引入历史信息折扣回报 $u(h_i)$ 作为 DPG[8] 更新的基准[18]，为 DPG 的更新提供参考，使得策略梯度探索隐藏在历史信息 h_i 中的环境状态信息，从而为解决动态环境的部分可观测问题提供一种方案。式 (2) 为引入历史信息基准的 DPG：

$$\begin{aligned}&\nabla_\omega J(\omega)\\&=\frac{1}{NT}\sum_n\sum_i\nabla_\omega\ln\pi_\omega(o_t,h_i)\nabla_{a_t}(Q_\theta(o_t,h_i,a_t)-u(h_i))\end{aligned}\tag{2}$$

公式的细节将在 3.1 节进行阐述。TS-TD3 模型并非聚焦于历史信息的完整轨迹，而是使用 LSTM 对历史信息进行有效筛选与总结，允许时序策略网络和时序评价网络学习、保留有限的历史信息。

执行动作 a_t 并非仅仅来自于时序策略网络，使用单一的策略网络处理长序列的输入会导致计算量大幅增加，模型训练缓慢且难以收敛[19]。本文引入专家经验网络对时序策略网络的输出进行指导。将专家动作 a_{g} 和时序策略网络的输出 a_{p} 分别输入进评价网络，得到评价值 Q_{g} 和 Q_{p}，通过式 (3) 评价网络的竞争机制输出执行动作：

$$a_t=\begin{cases}a_{\mathrm{p}}, & Q_{\mathrm{p}}\geqslant Q_{\mathrm{g}}\\ a_{\mathrm{g}}, & Q_{\mathrm{p}}<Q_{\mathrm{g}}\end{cases}\tag{3}$$

模型训练前期，专家动作 a_{g} 将优于时序策略网络输出的随机探索 a_{p}，随着训练量的加大，TS-TD3 模型逐渐收敛。时序策略网络输出的动作 a_{p} 逐渐精确，专家经验网络对模型训练的影响逐渐减弱。

3 方法（Method）

3.1 引入历史经验基准的 DPG

在动态场景下，机器人显然只能观测到动态障碍物的部分运动学信息，动态障碍物的动作以时间为刻度在一定范围内具有时序关联性，依靠单帧平面激光雷达扫描信息会割裂这一时序信息，时序关联性被认为隐藏在包含完整导航轨迹的历史信息 h_i[16] 中，通过转移函数 $p(s_t|h_i)$ 表征此特性。本文提出为 DPG 引入历史信息基准，历史信息基准将引导 DPG 探索动态环境的隐藏状态。

DPG 的核心思想在于通过策略梯度 $\nabla_\omega J(\omega)$ 的反向传播更新确定性策略 $a=\pi_\omega(s)$：

$$\nabla_\omega J(\omega)=E_{s'\sim p(s,\pi)}\left(\nabla_a Q_\theta(s,a)|_{a=\pi_\omega(s)}\nabla_\omega\pi_\omega(s)\right)\tag{4}$$

其中，期望值与状态转移函数 $p(s,\pi)$ 有关，在环境部分可观测的条件下，无法直接确定当前状态 s_t，而是接收当前观测 o_t，o_t 包含状态 s_t 的部分信息，表示为 $p(o_t|s_t)$。本文引入包含隐藏环境状态 s_t 的历史信息 h，在 DeepMind 方法[16] 基础上作进一步推导，因为基于当前观测 o_t 的执行动作 a_t 来源于确定性策略 π，所以历史信息 h 的子集 $\{o_1,o_2,\cdots,o_i\}$ 是状态转移函数 $p(s,\pi)$ 的无偏估计，可以将 $p(s,\pi)$ 近似表达为

$$p(s_t,\pi)\approx\prod_{i=1}p(o_i|s_t)p(s_t|h)\tag{5}$$

故最优策略 π^*、动作－价值函数 Q_π 均与历史信息 h_i 有关，引入历史信息的 DPG 有如下形式：

$$\begin{aligned}&\nabla_\omega J(\omega)\\&=E_{h_i}\left(\frac{1}{T}\sum_i\nabla_a Q_\theta(o,h_i,a)|_{a=\pi_\omega(o,h_i)}\nabla_\omega\pi_\omega(o,h_i)\right)\end{aligned}\tag{6}$$

关于 T 的取值将在下文讨论。然而，实际训练时式 (6) 并没有产生良好的效果，因为模型训练要在期望折扣回报 J 的梯度上回顾完整历史信息 h_i，延长了模型的训练时间，使模型难以收敛。

本文引入历史信息基准来代替对完整历史信息进行回顾，基准不影响期望折扣回报的期望值，能够有效减小深度强化学习模型训练过程中的方差[20-21]，引导策略梯度探索隐藏在历史信息中的环境状态信息。关于基准有如下结论[22]，该结论适用于基于 MDP 或 POMDP 的模型：

$$\begin{aligned}&E_s(b\cdot\nabla_\omega\ln\pi(s))\\&=b\cdot\sum_{i\in s}\nabla_\omega\pi(s)=b\cdot\nabla_\omega\sum_{i\in s}\pi(s)=b\cdot\nabla_\omega c=0\end{aligned}\tag{7}$$

其中，c 为常数，b 为基准，b 与当前时刻动作 a_t 无关，而是来源于历史缓冲区 β_{h} 的随机抽样 h_i，历史信息 h_i 包含完整的导航轨迹，通过 Reinforce 算法[23] 将 h_i 中机器人获得的奖励进行加权求和，得到历史信息折扣回报，如式 (8) 所示：

$$u_i=\sum_{i=1}\gamma^{I-i}r_i\tag{8}$$

其中，I 为一次完整导航过程中总的采集步数。

因为历史信息 h_i 中奖励 r_i 直接来自执行动作 a_t 和隐藏状态 s_t，故历史信息折扣回报 u_i 包含动态环境的隐藏状态信息，将其作为基准 b 引入确定性策略梯度 $\nabla_\omega J(\omega)$ 以引导策略梯度探索隐藏状态 s_t：

$$\begin{aligned}&\nabla_\omega J(\omega)\\&=E_{h_i}\left(\frac{1}{T}\sum_i\nabla_\omega\ln\pi_\omega\nabla_a(Q_\theta(o,h_i,a)|_{a=\pi_\omega}-u(h_i))\right)\end{aligned}\tag{9}$$

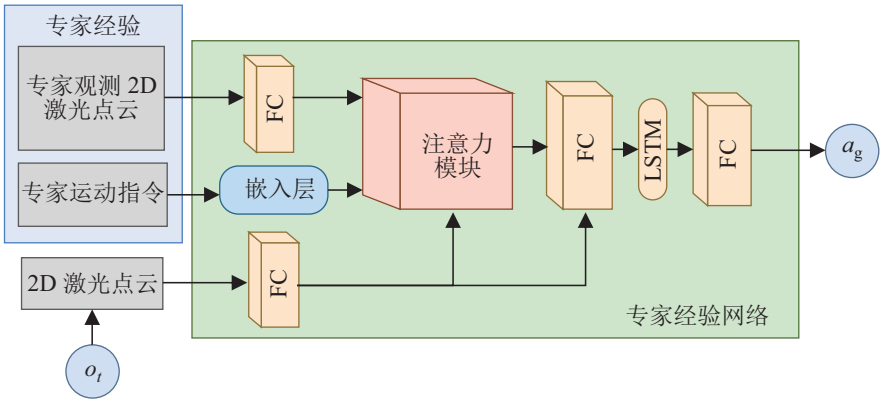

图 2　专家经验网络

Fig.2　Expert experience network

然而，训练前期的机器人运动不具备规律性，无法获得有训练价值的历史信息轨迹，机器人仍然在策略梯度上随机探索，本文将在下一节讨论在训练前期引入专家经验指导策略网络以加速模型收敛的方案。

3.2　专家经验网络引导时序策略网络学习

在实际训练中，在策略梯度上随机探索的导航动作使得采集的历史信息轨迹也是随机的，不具备时序关联性。

本文提出在训练前期引入符合一定行业规范的专家操作指导模型训练，网络框架如图 2 所示。文 [24] 证明，专业知识能够有效促进深度强化学习模型的学习。本文在深度视觉神经网络导航系统 SnapNav [25] 的基础上设计出根据专业知识学习的专家经验网络，并通过一个简单的策略切换机制在训练前期引导时序策略网络的输出。

专家经验网络使用注意力−LSTM 框架 [26]，由注意力模块、LSTM 编码器组成。使用人工采集的专家经验数据集 $g_i=\{o_1,a_1,\cdots,o_i,a_i\}$ 训练网络，单步数据集包括当前时刻的环境观测和根据观测结果执行的动作，将单步数据集按照时间顺序整合成完整数据集。训练时，将专家经验数据集分为观测集合 $O_i=\{o_1,o_2,\cdots,o_i\}$，动作集合 $A_i=\{a_1,a_2,\cdots,a_i\}$ 两个子集进行训练。测试与验证时，将当前环境观测 o_t 和观测集 O_i 分别输入到一个全连接层（FC），动作集 A_i 输入到一个线性嵌入层：

$$g_i^a=\delta(A_i|w_o^a),\quad g_i^o=\rho_1(O_i|w_o^g),\quad v_t^o=\rho_2(o_t|w_o) \tag{10}$$

如式 (10) 所示，δ 表示使用 ReLU（线性整流）激活函数的嵌入层函数，w_o^a 是嵌入层权重，ρ_1 和 ρ_2 表示使用 ReLU 激活函数的全连接层函数，w_o^g 和 w_o 是全连接层权重。

注意力模块的作用是选择与当前观测 o_t 相关度最高的专家观测 o_i 匹配的动作 a_i 进行指导：

$$(g_i^{o\prime},v_t^{o\prime})=\underset{(g_i^o,v_t^o)}{\arg\max}\ \mathrm{softmax}\left(-\|g_i^o-v_t^o\|_2\right) \tag{11}$$

遵循文 [17] 的硬注意力机制处理方式，如式 (11) 所示，注意力模块首先度量经过全连接层处理的专家观测 o_i 和当前观测 o_t 的相似性，然后通过 softmax 层进行归一化，得到的结果是一个次微分模型，并且可与模型中基于梯度的其他组件相结合。接着，将注意力模块的输出与经过嵌入层编码的当前观测 o_t 通过一个全连接层拼接，输入进一个 LSTM 网络，LSTM 提炼保留了专家经验的前后时刻关联性，当专家观测结果与当前观测结果足够相似时，注意力模块的次微分方程选择与该专家观测结果相匹配的专家动作作为输出。

简单地使用专家经验网络替代时序策略网络将导致机器人只能执行专家经验数据集中已采集的动作，只能部署在与专家经验数据集采集场景相同的场景，模型失去泛化性且容易陷入策略局部最优的情况。引入复杂的动作选择模型 [19] 将增加不必要的计算量。

本文设计了一个简单的动作选择机制，将专家经验网络输出的动作 a_g 和时序策略网络输出的动作 a_p 分别输入到评价网络中，评价网络对输入的动作进行评价得到 2 个评价结果 Q_g 和 Q_p，机器人执行价值更高的动作：

$$a_t=\begin{cases}a_p=\pi_\omega(o_t,h_i)+\varepsilon, & Q_p\geqslant Q_g\\ a_g, & Q_p<Q_g\end{cases} \tag{12}$$

式中，ε 是探索噪声，$\varepsilon\sim N(0,\tau)$ [27]；τ 是模型的超参数。由于环境奖励 r_t 直接来源于环境状态和执行动作，在训练前期，专家引导动作 a_g 的评价 Q_g

将经常高于时序策略网络的输出 a_{p} 的评价 Q_{p}，执行专家引导动作使得机器人在训练前期符合一定运动学规范与行业准则，避免机器人在策略上进行无意义的探索。随着模型训练量的增加，时序策略网络输出的动作逐渐精准，专家经验网络对策略梯度探索的影响逐渐减弱，避免机器人陷入策略局部最优的困境。

3.3 时序策略网络和时序评价网络

POMDP 的最优策略 π^*、动作－价值函数 Q_π 和状态转移函数 $p(s,\pi)$ 均与历史信息 h_i 有关。

TS-TD3 模型采用时序策略网络 $\pi_\omega(o,h_i)$ 来近似确定策略 $a=\pi(s)$，网络框架如图 3 所示，采用时序评价网络 $Q_\theta(o,h_i,a)$ 来近似动作－价值函数 $Q_\pi(s,a)$，网络框架如图 4 所示。

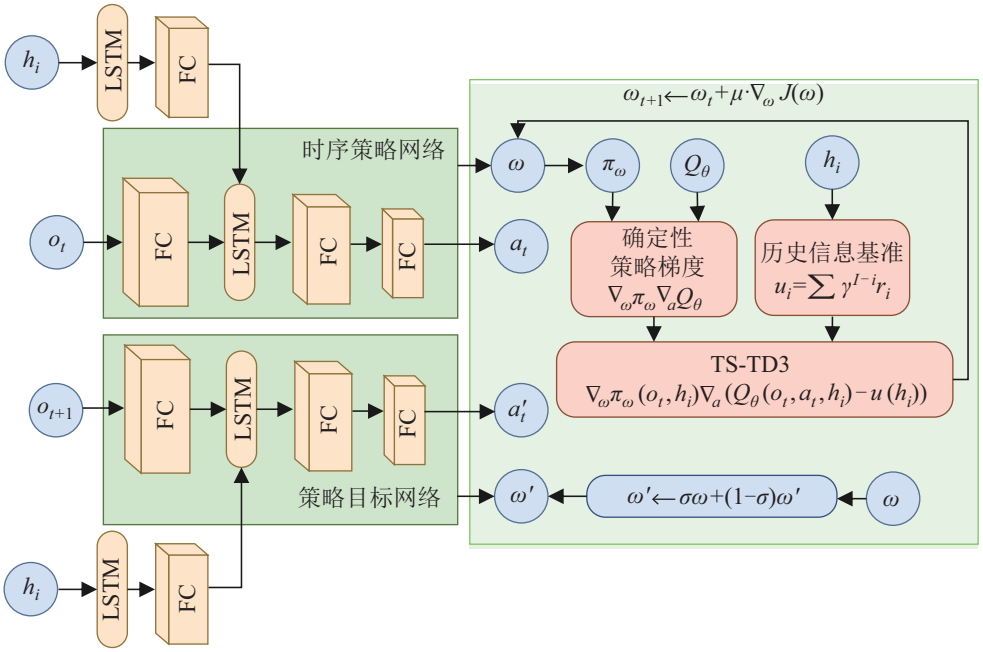

图 3 时序策略网络

Fig.3 TS-actor network

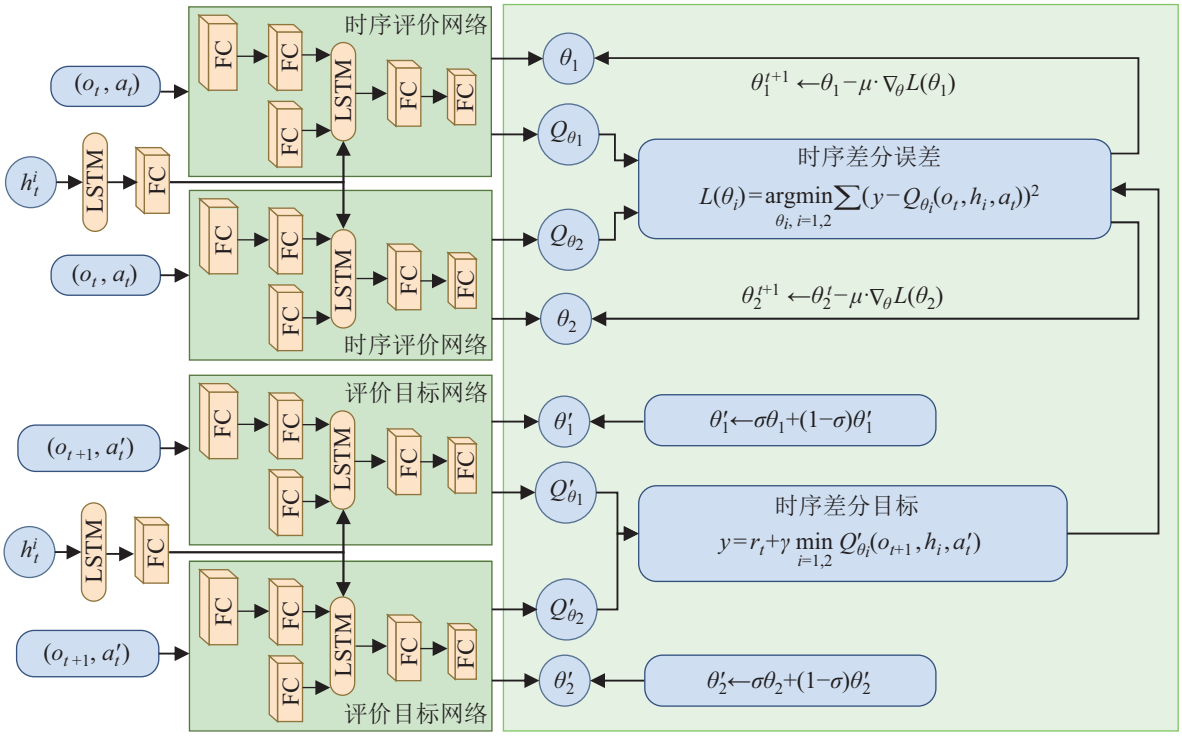

图 4 时序评价网络

Fig.4 TS-critic network

环境奖励函数设置如式 (13) 所示，包括 2 个稀疏奖励和 1 个密集奖励：

$$r_t=\begin{cases}R_{\mathrm{g}}, & d_t<\eta_{\mathrm{g}}\\ R_{\mathrm{c}}, & L_{\min}<\eta_{\mathrm{c}}\\ \cos v_t(d_{t-1}-d_t)-|\cos\omega_t-\cos\omega_{t-1}|, & \text{其他}\end{cases}\tag{13}$$

2 个稀疏奖励分别为：机器人距离目标地点的距离 d_t 小于阈值 η_{g} 时，环境给予机器人一个较大的正面奖励 R_{g}；$L_{\min}$ 代表平面激光雷达扫描数据的最小检测值，当 $L_{\min}<\eta_{\mathrm{c}}$ 时，给予机器人一个较大的负面奖励 R_{c}。

密集奖励是指 $d_{t-1}-d_t$ 鼓励机器人不断向目标地点靠近；系数 $\cos v_t$ 鼓励机器人以一个较快（不超过最大线速度）的线速度前进；常数项惩罚 $|\cos\omega_t-\cos\omega_{t-1}|$ 代表限制机器人在动态环境中频繁转向，同时常数项惩罚能够有效避免机器人反复循环，盲目追求高分环境奖励 [28] 的导航行为。

3.3.1　时序策略网络

引入历史经验折扣回报 $u(h_i)$ 作为 DPG 的基准，以解决 POMDP 条件下的部分可观测问题。然而，h_i 中的奖励没有与动作和观测产生显式关联，关联的缺失会影响模型训练，因为在同一环境下，类似的动作应该具有类似的价值 [9]。

对此，式 (14) 将 LSTM 网络添加进时序策略网络以总结、提炼历史信息 h_i 中有训练价值的片段：

$$h_i'=\rho_1\left(L_1(h_i|w_{L1})\right)\tag{14}$$

接着由式 (15) 显式地建立历史信息折扣回报中的奖励与历史信息中的动作、观测之间的联系：

$$o_t'=\rho_2(o_t),\quad o_t^h=L_2(o_t,h_i|w_{L2})\tag{15}$$

L 表示 LSTM 网络，ρ 表示使用 ReLU 激活函数的全连接层，w_L 表示 LSTM 的训练权重，训练过程中，LSTM 总结、保留 h_i 中有价值的训练片段。

时序策略网络的核心是在历史信息基准的引导下，通过梯度上升的方式更新 DPG 的参数 ω。式 (9) 给出了引入历史信息基准的 DPG 的解析解形式，在实际训练过程中，采用经验回放（experience replay）机制来提高存储序列的利用效率和稳定性 [9-10]，从经验缓冲区 β 和历史缓冲区 β_{h} 中随机抽样选取样本，通过式 (16) 使用蒙特卡洛法近似求解：

$$\nabla_\omega J(\omega)\approx\frac{1}{NT}\sum_n\sum_t\nabla_\omega\ln\pi_\omega\nabla_a\left(Q_\theta-u(h_i)\right)\tag{16}$$

其中，N 表示每轮训练过程中从经验缓冲区 β 随机抽样的样本数，样本由 o_t,a_t,r_t,o_{t+1} 组成；从历史缓冲区 β_{h} 随机抽样历史信息 h_i，T 表示 h_i 经过 LSTM 处理后的有效历史信息步数。策略训练的目的是使期望折扣回报最大化，故式 (17) 采用梯度上升的方式更新策略梯度的参数 ω：

$$\omega_{t+1}\leftarrow\omega_t+\mu\cdot\nabla_\omega J(\omega)\tag{17}$$

其中，μ 表示学习率，是超参数，μ 的取值将影响策略梯度的学习。目标网络（target network）[9] 是实现深度强化学习稳定训练的一个工具，式 (18) 使用目标策略网络（target actor network）来减小输入的长序列 o_t 和 h_i 所带来的偏差：

$$\omega'\leftarrow\sigma\omega+(1-\sigma)\omega'\tag{18}$$

其中，σ 是超参数，建议 σ 设置比较小的值，如 $\sigma=0.005$。这会延长模型训练的时间，但有助于 LSTM 总结、选择性记忆、选择性遗忘长序列的历史信息 h_i。

3.3.2　时序评价网络

时序评价网络在剪切双 Q 学习框架 [10] 的基础上，通过修改网络结构来应对动态环境的部分可观测性，引入经过 LSTM 组件处理的历史信息 h_i，式 (19) 的处理方式与时序策略网络类似：

$$h_i'=\rho_1(L_3(h_i|w_{L3}))\tag{19}$$

式 (20) 将当前执行动作 a_t（已引入专家经验网络）、当前环境观测 o_t 输入进时序评价网络：

$$o_t'=\rho_3(\rho_2(o_t)),\quad a_t=\rho_4(a_t)\tag{20}$$

其中，时序评价网络通过内部的 LSTM 网络来拼接 h_i'、o_t、a_t 三个序列，如式 (21) 所示：

$$q_t^h=L_4(o_t,a_t,h_i'|w_{L4})\tag{21}$$

时序评价网络输出对当前执行动作 a_t 的时序评价 Q_θ，Q_θ 包含对历史信息 h_i 的时序关联性的评价，并将其映射到对当前执行动作 a_t 的评价中，使时序评价网络在输出 Q_θ 时将前后时刻导航动作的时序关联性纳入评价标准，指导策略网络输出导航动作时考虑 h_i 包含的时序关联性。

采用 TD 算法来更新时序评价网络的参数，式 (22) 中双评价目标网络通过观测 o_{t+1} 和目标策略网络的输出 a_t' 来估计下一时刻的时序评价值，取其中的最小值得到 TD 目标值 y：

$$y=r_t+\gamma\min_{i=1,2}Q_{\theta_i}'(o_{t+1},h_i,a_t')\tag{22}$$

其中，a'_t 由目标策略网络输出，γ 是折扣率。a'_t 并非下一时刻的执行动作，这种估计会带来偏差，随着网络参数不断更新，这种偏差将被累积起来造成策略次优或模型发散。文 [10] 对这种累积偏差进行了证明，表达为

$$Q_\theta(o_t,h_i,a_t)=E_{s,a}\left(\sum_t \gamma^{T-t}(r_t-\delta_t)\right) \tag{23}$$

式中，Q_θ 是时序评价网络的输出，δ_t 为单次更新产生的偏差。文 [10] 证明了使用目标网络可以遏制这种偏差（式 (18)），目标网络遏制累积偏差的主要方式是通过延迟更新使得偏差无法快速累积。在目标网络中引入包含记忆与遗忘功能的 LSTM 处理 h_i 将强化目标网络的功能，式 (21) 将经过预处理的 h_i,o_t,a_t 输入进 LSTM，使得时序评价网络输出对执行动作 a_t 的评价 Q_θ 时考虑导航动作的时序关联性，总结 h_i 中值得保留和需要遗忘的部分，以形成具有时序关联性的记忆片段：

$$\begin{aligned}Q_\theta(o_t,h_i,a_t)=r+\gamma E\left(Q_\theta(o_{t+1},h_i,a'_t)\right)-\\ \delta(o_{t+1},a'_t)+q_t^h\end{aligned} \tag{24}$$

式 (21) 的输出被认为是对历史信息的有益总结，式 (24) 将式 (21) 引入时序评价网络进一步减小了单步更新的偏差，这可以理解为单步更新的偏差在一次随机抽样过程中被累积，目标网络的延迟更新减缓了偏差的累积。LSTM 通过式 (21) 总结历史信息 h_i 与观测 o_t 和动作 a_t 的映射关系，而且根据式 (24)，映射关系 q_t^h 会改善时序评价结果 Q_θ，再次减小了单步更新过程中的偏差，从而改进了时序评价网络的输出 Q_θ。

TS-TD3 完整算法的流程如算法 1 所示。

算法 1 TS-TD3 算法

输入： 机器人当前观测 o_t，历史信息 h_i

输出： 机器人执行动作 a_t

用随机参数 θ_1,θ_2,ω 初始化网络 $Q_{\theta_1},Q_{\theta_2},\pi_\omega$
初始化目标网络 $\theta'_1\leftarrow\theta_1$, $\theta'_2\leftarrow\theta_2$, $\omega'\leftarrow\omega$
初始化经验缓冲区 β，初始化历史缓冲区 $\beta_{\rm h}$
for episode(轮) = 1 to M **do**
  初始化空集历史 h_i
  存储前一集 h_i 到历史缓冲区 $\beta_{\rm h}$
  for $t=1$ to T **do**
    $\beta_{\rm h}$ 随机采样 1 个 $h_i=(o_1,a_1,r_1,\cdots,o_i,a_i,r_i)$
    时序策略网络输出 $a_{\rm p}\sim\pi_\omega(o_t,h_i)+\varepsilon$
    根据 o_t 从专家经验网络选取动作 $a_{\rm g}$
    选择执行动作 $a_t=Q_{\rm p}>Q_{\rm g}?\,a_{\rm p}:a_{\rm g}$
    执行动作 a_t，获得奖励 r_t，获得新观测 o_{t+1}
    存储 (o_t,a_t,r_t,o_{t+1}) 到经验缓冲区 β
    更新历史信息 $h_i\leftarrow h_i,o_t,a_t,r_t$
    β 随机采样 N 个样本 (o_t,a_t,r_t,o_{t+1})
    $a'_t\leftarrow\pi_\omega(o_{t+1},h_i)+\varepsilon$
    $y\leftarrow r_t+\gamma\min\limits_{i=1,2}Q'_{\theta_i}(o_{t+1},h_i,a'_t)$
    更新双时序评价网络：
    $\theta_i=\dfrac{1}{NT}\arg\min\limits_{\theta_i}\sum\limits_n\sum\limits_t(y-Q_{\theta_i}(o_t,h_i,a_t))^2$
    更新时序策略网络：
    $\nabla_\omega J(\omega)\approx\dfrac{1}{NT}\sum\limits_n\sum\limits_t\nabla_\omega\ln\pi_\omega\nabla_{a_t}(Q_\theta-u(h_i))$
    更新目标网络：
    $\theta'\leftarrow\sigma\theta+(1-\sigma)\theta'$, $\omega'\leftarrow\sigma\omega+(1-\sigma)\omega'$
  end for
end for

4 实验和结果（Experiment and results）

4.1 实验设置

在不同复杂程度的动态场景中进行实验，以验证 TS-TD3 算法。模型训练是在一台搭载 NVIDIA GTX 3060 显卡、32G 内存的计算机中进行，操作系统为 Ubuntu 18.04LTS，采用 1.10 版本的 PyTorch 训练网络模型，模型的超参数设置如表 1 所示。使用 Gazebo9 软件搭建动态仿真环境，移动机器人选用二轮差速运动模型，最大线速度为 0.5 m/s，最大角速度为 1 rad/s，搭载 2D 激光雷达，扫描范围为 $[-150^\circ,\ 150^\circ]$，扫描频率为 10 Hz，测量半径为 [0.2 m, 8 m]。

表 1 模型超参数设置

Tab.1 Model hyperparameter setting

参数	取值
折扣率 γ	0.99
学习率 μ	5×10^{-4}
经验池尺寸	10^6
历史经验池尺寸	5×10^6
经验池批量大小	50
历史经验池批量大小	1
探索噪声参数 τ	2×10^{-1}
目标网络更新率 σ	5×10^{-3}

在图 5 和图 6 所示的动态仿真环境中对 TS-TD3 模型进行训练和测试，仿真环境分为训练环境和测试环境，训练环境的基本尺寸为 11 m×11 m，测试环境的基本尺寸是 21 m×21 m，环境中移动的行人在每轮训练中运动的轨迹相同，行人不会主动识别、避让机器人，以提供一个标准的训练、测试环境。

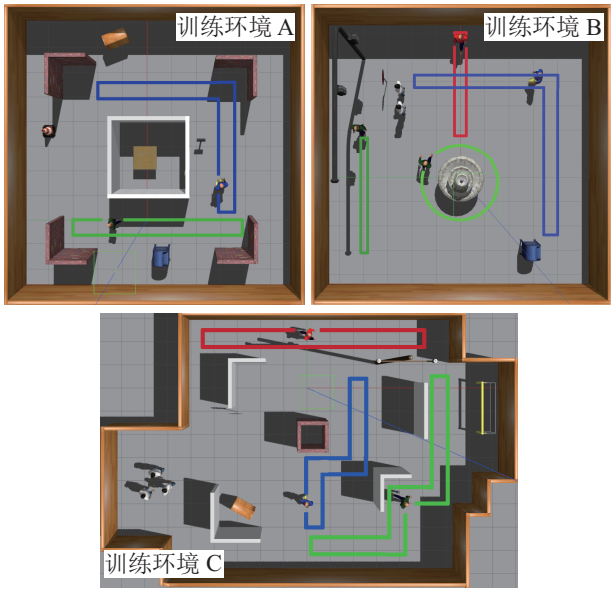

图 5　训练环境

Fig.5　Training environment

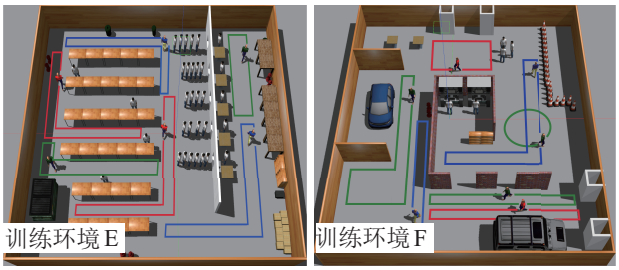

图 6　测试环境

Fig.6　Testing environment

在训练环境中训练 1500 轮，每轮训练开始时机器人的初始位置和目标位置在场景内随机分配，当机器人到达目标点、发生碰撞，或是步数达到 2000 步时，结束本轮训练。每进行 5000 步训练，将训练环境替换为测试环境，进行 10 轮性能测试，并记录训练曲线。为了提高模型的泛化性和策略探索能力，初始训练环境为场景 A；训练达到 500 轮时，训练环境改为场景 A 和场景 B 交替；训练达到 1000 轮时，训练环境改为场景 A、B、C 交替。

4.2　训练阶段

4.2.1　对比实验设置

为在训练阶段评估 TS-TD3 算法，选择目前主流的 DRL 算法进行对比实验，包括 MDP 算法和 POMDP 算法，使用与 TS-TD3 算法相同的训练方式，每种算法训练 1500 轮，每轮训练的最大步数为 2000 步，算法训练最大步数 10^6 步，对比算法的介绍如下：

(1) DDPG[9]：基于 DPG 以及演员－评论家框架，从经验池中抽样得出训练模式，能够处理高维输入，适用于需要连续控制的任务。

(2) PPO（近端策略优化算法）[29]：基于演员－评论家框架，采用重要性采样技术，实行多阶段批量更新，是 OpenAI 的基准方法之一，适用于连续控制任务。

(3) TD3[10]：DDPG 的有效改进模型，采用一对延迟更新的评价网络限制对动作的过高评价，采用目标网络技术减缓误差的累积，适用于连续控制。

(4) RDPG（循环确定性策略梯度算法）[16]：上述 3 种方法均基于 MDP，而 RDPG 算法则是基于 POMDP，使用时间反向传播（BPTT）法总结历史信息，是 DPG 在环境部分可观测场景中的延伸。

(5) TD3-LSTM[30]：同样是基于 POMDP，使用 LSTM 处理历史信息后将其输入模型，策略网络采用多重复合策略进行导航，评价网络采用多重评价体系指导导航策略，是 TD3 模型在环境部分可观测场景中的延伸。TD3-LSTM 模型与 TS-TD3 模型均对历史信息进行预处理，区别之处在于 TS-TD3 模型内置 LSTM 组件对经过预处理的历史信息、动作和观测再次整合，对有效片段进行筛选与记忆。同时，TD3-LSTM 是针对不同任务的模型框架，并非专门针对机器人自主导航，其多重策略的同时更新、多重评价体系的设置易造成导航轨迹混乱的情况。

除此之外，引入基于地图的导航框架进行对比，将 ROS 内置的依靠地图的导航框架 move_base[2] 引入对比实验，其框架明确地划分为先验地图建立、全局路径规划、局部避障规划 3 个子模块，具有代表性。实验环境设置与基于 DRL 的模型训练相同。采用 Gazebo 仿真环境的自动回溯机制收集样本数据，move_base 数据集包括从 Gazebo 的训练环境中自动收集的 10^6 步样本，单步的样本包括某一时刻机器人执行的动作和该时刻环境给予的奖励。

4.2.2　专家经验网络训练

在 Gazebo9 仿真软件的训练场景 A、B、C 中人工建立 1000 轮训练，每轮训练不超过 2000 步的依赖人工操作的专家经验数据集（见 3.2 节），由熟悉二轮差速自主移动机器人的技术人员根据平面激光雷达的扫描数据（环境观测）控制机器人运动（执行动作），采集单步数据集（环境观测和执行动作）。每轮训练的最大步数为 500 步。将专家经验数据集输入进专家经验网络进行训练，训练工作参照 3.2 节。使用训练完成的专家经验网络将专家经验数据集扩充至 10^6 步，将其用于对比实验。

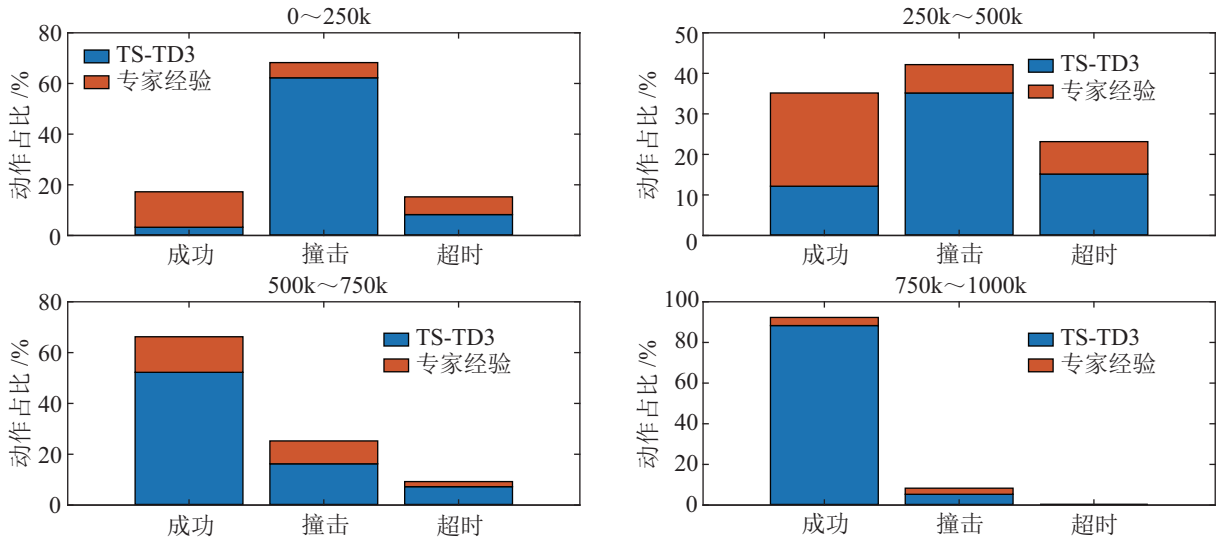

图 7 不同训练阶段的专家经验网络使用率

Fig.7 Usage percentage of the expert experience network at different stages of training

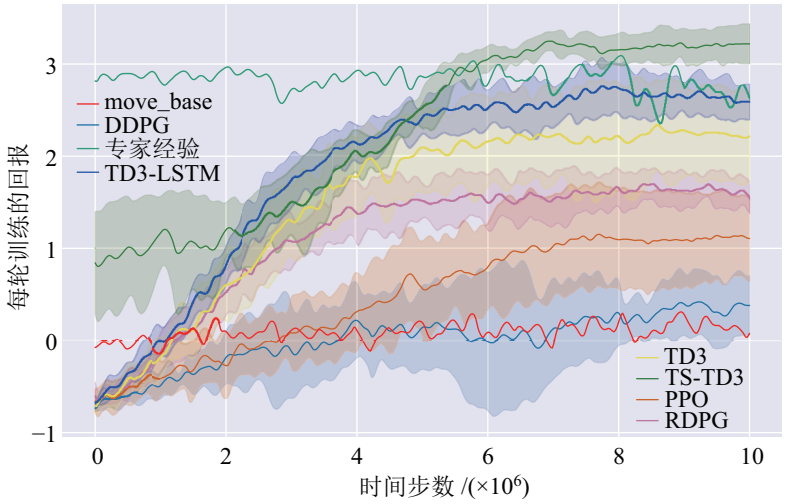

图 8 回报曲线图

Fig.8 Reward curve diagram

训练前期（步数 0～250k）机器人大多借助专家经验的指导才能到达目标地点，随着 TS-TD3 模型训练量的增加，专家经验网络的影响逐渐减弱。图 7 是 TS-TD3 模型在训练期间，机器人处于每轮训练的最终状态时（最终状态包括成功到达、发生碰撞、超时），专家经验网络输出的指导动作与策略网络输出的导航动作在训练集中的分时使用率。

4.2.3 训练曲线分析

实验每进行 5000 步，将训练环境替换为测试环境进行 10 轮的测试，取累积回报的平均值作为迭代步的累积回报进行记录，曲线如图 8 所示。

依赖地图的传统导航框架 move_base [2] 无法满足动态环境的导航要求，其原因在于机器人将行人连续运动轨迹判断为障碍物，并将其雷达扫描数据映射到了代价地图当中，形成了不存在的障碍物，造成了“机器人冻结”问题 [3]。

DDPG 和 PPO 算法假设当前传感器信息中包含了周围环境的所有信息。动态环境下，传感器的输入仅包含部分的环境信息，这导致训练曲线震荡，累积回报低，DDPG 算法出现模型崩溃的现象，无法承担动态环境下的导航任务。PPO 算法采用 N 步更新策略，包含一定时序信息，导航效果优于 DDPG 算法。同样基于 MDP 的 TD3 算法在训练到 7×10^5 步后能够在动态环境下进行导航，但同样存在训练曲线剧烈震荡的问题，在训练后期也存在一定波动。

基于 POMDP 的 RDPG 算法通过 BPTT 遍历与总结历史信息，训练过程中曲线抖动较小，但算法

需要遍历完整历史信息，训练稳定性较差，在急速变化的动态环境下导航效果不稳定。TD3-LSTM 模型累积回报低于 TS-TD3 模型，但训练时间更短，模型收敛更快。其多重导航策略与多重评价体系的设置使 LSTM 总结历史信息时没有一定标准可循，不同批次的模型训练效果差别较大，训练后期仍然存在曲线抖动的情况，稳定性不如 RDPG 算法与 TS-TD3 算法。

TS-TD3 算法在训练前期，专家经验网络使得迭代累积回报明显高于其他算法，但此时评价网络输出不准确，无法准确鉴别价值更高的动作，因而训练曲线抖动，导航效果不稳定。在 5.4×10^5 步后，算法的导航策略逐渐优于专家经验的指导。历史缓冲区不断更新，将训练前期存入的不稳定训练轨迹弹出，存入优于专家经验的历史信息，历史信息基准引导 DPG 进行有益探索，需要考虑环境的部分可观测性。时序评价网络在评价策略时，需要总结前后时刻动作的时序性。模型收敛速度加快，收敛后训练曲线稳定不抖动，训练曲线累积回报在所选算法中属于最优。

4.3　消融实验

设计消融实验进一步评估 TS-TD3 模型各组件的影响，引入 TD3-LSTM [30]（见 4.2.1 节）作为基线模型进行对比，移除部分组件参与消融实验的模型如图 9 所示。

在固定场景（场景 A）和完整训练场景（场景 A、B、C 交替）分别开展消融实验，实验设置与对比实验相同，获得的累积回报曲线如图 10 所示。(1) 移除专家经验网络的 TS-TD3-Rex 模型结构可视作基线模型 TD3-LSTM 增加内置 LSTM 的版本（模型更新方式则完全不同）。训练前期，TS-TD3-Rex 模型的累积回报明显低于引入专家经验网络的其他

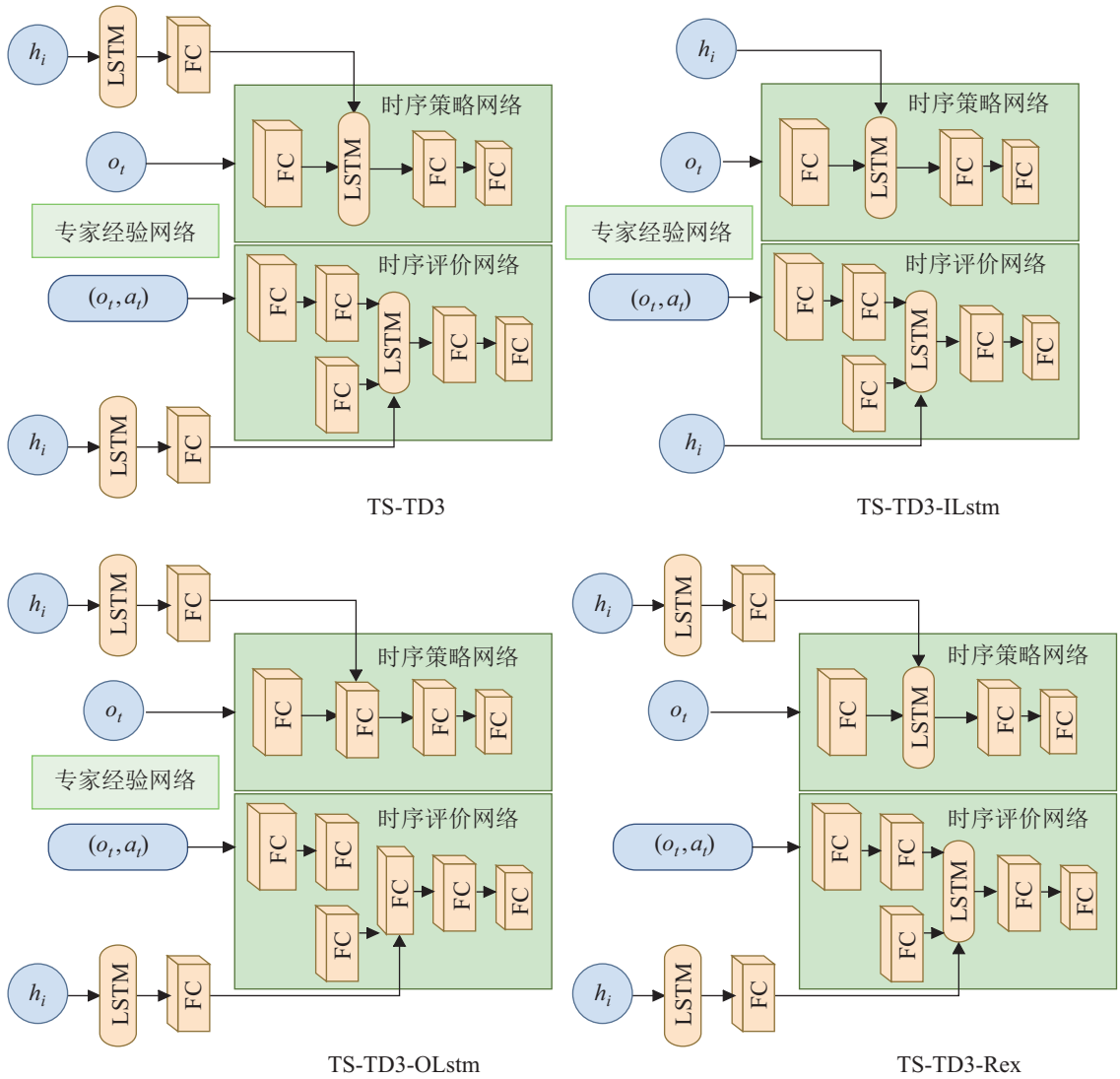

图 9　消融实验模型结构图

Fig.9　Model structure diagram in ablation experiment

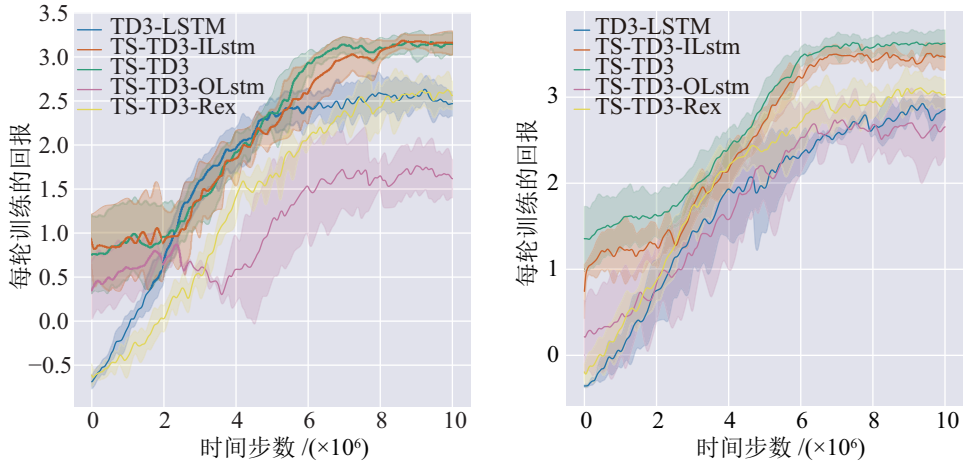

图 10 完整训练场景（左）和单一场景（右）回报曲线

Fig.10 Reward curve diagram in complete training scenario (left) and single scenario (right)

版本，训练时间在所有版本中最长。模型收敛后，累积回报仍然较低，但高于基线模型 TD3-LSTM。(2) 内置 LSTM 的 TS-TD3-ILstm 模型收敛后累积回报略低于完整 TS-TD3 模型，训练时间延长，但性能差别不大。(3) 外置 LSTM 的 TS-TD3-OLstm 模型性能出乎意料地最低，累积回报低于基线模型 TD3-LSTM，稳定性较差、训练时间较长，专家经验与历史信息的引入反而降低了模型训练效果，可能的原因是移除内置 LSTM 的模型不具备整合、总结历史信息的能力，造成了曲线震荡、模型发散。(4) 基线模型 TD3-LSTM 累积回报不如完整 TS-TD3 模型，但在完整训练场景中训练时间更短，模型稳定性劣于 TS-TD3 模型，4.4 节将证明完整 TS-TD3 模型应对复杂情况的泛化性强于 TD3-LSTM。

移除特定组件（见表 2）对模型训练时间、累积回报、模型稳定性等性能参数造成不同程度的影响，但将所有组件整合的完整 TS-TD3 模型具备良好的性能。

表 2 消融实验模型设置

Tab.2 Setting of the model for ablation experiment

名称	描述
TS-TD3	完整模型
TS-TD3-Rex	移除专家经验网络
TS-TD3-ILstm	内置 LSTM 模型：移除模型外对历史信息进行预处理的记忆组件（包括 LSTM 和全连接层）
TS-TD3-OLstm	外置 LSTM 模型：移除模型内对观测、动作、历史信息进行整合总结的 LSTM 组件，使用全连接层进行替代
TD3-LSTM	对比模型，使用长短期记忆网络对历史信息进行预处理，使用多重评价标准

4.4 测试阶段

将机器人部署在测试环境 E、F 下分别进行 100 轮测试（图 11），定量评估 TS-TD3 算法，机器人的初始位置和目标位置均固定。记录的数据结果如表 3、表 4 所示。

测试环境 E 模拟人流量较大的餐厅取餐环境，排队的人群呈明显的队列模式。TD3 算法执着于路径的最短性，沿着墙边反复尝试离开墙壁，导航策略表现出的“机器人摆头”行为在行人场景中造成安全隐患，多次进行不必要的探索，延长了导航时间。RDPG 算法的导航动作的连续性明显优于 PPO 和 TD3 算法，但 RDPG 算法面对突然出现的行人时，规避动作显得僵硬不流畅。采用 TD3-LSTM 算法导航时机器人能与墙边保持一定的安全距离，但面对行人时，避障行为展现出一条不自然的曲线，可能会在人群中造成困扰。TS-TD3 算法的导航轨迹流畅，选择的路径较短，面对障碍物时的避障行为自然。

测试环境 F 模拟行人随意行走、聚集与交谈的动态场景。PPO 算法明显偏离了最优轨迹，不断尝试向场景的开阔地带行驶。TD3 算法选择了场景中拐角较多的方向，能够到达目标地点，但是该算法在拐角处不断地调整行驶角度，这一行为会造成安全隐患。TD3-LSTM 算法在 F 场景中的导航数据较好，但是该算法使机器人沿着墙边行驶，机器人高速且突然的出现在多行人场景中是难以被接受的。TS-TD3 算法在某些导航指标上不如 TD3-LSTM 算法，但在实际表现上，其导航动作符合一定连续性，面对动态障碍物的避障行为自然流畅，对部分可观测的场景具有较强的反应能力，成功率在所有算法中最高。

表 3　训练环境 E 导航测试

Tab.3　Navigation test in training environment E

	行驶距离平均值 /m	行驶距离最大值 /m	行驶距离最小值 /m	行驶时间平均值 /s	行驶时间最大值 /s	行驶时间最小值 /s	行驶距离标准差	行驶时间标准差	成功率
PPO	16.52	26.74	14.04	93.39	100.93	66.20	6.43	15.13	69%
RDPG	15.34	18.88	12.25	79.14	90.48	69.22	3.42	9.57	77%
专家经验	15.09	16.31	14.07	75.02	78.29	71.56	0.67	2.73	97%
TD3	15.79	21.60	12.83	96.40	104.38	73.45	4.67	14.28	74%
TD3-LSTM	14.62	17.38	11.26	76.76	100.87	66.16	3.67	10.23	82%
TS-TD3	14.21	17.34	13.59	72.20	85.57	65.98	1.24	8.54	90%

表 4　训练环境 F 导航测试

Tab.4　Navigation test in training environment F

	行驶距离平均值 /m	行驶距离最大值 /m	行驶距离最小值 /m	行驶时间平均值 /s	行驶时间最大值 /s	行驶时间最小值 /s	行驶距离标准差	行驶时间标准差	成功率
PPO	18.06	32.19	15.62	82.61	111.82	73.54	16.82	22.05	65%
RDPG	17.08	26.97	13.94	63.46	80.84	39.81	11.47	18.69	82%
专家经验	16.25	18.23	11.45	55.04	62.26	46.61	2.13	4.61	96%
TD3	16.89	26.58	10.41	53.28	71.44	37.52	13.87	21.26	71%
TD3-LSTM	15.65	25.67	11.35	47.74	69.55	36.56	8.45	19.85	79%
TS-TD3	15.97	24.69	12.78	54.42	68.50	43.91	9.67	15.23	88%

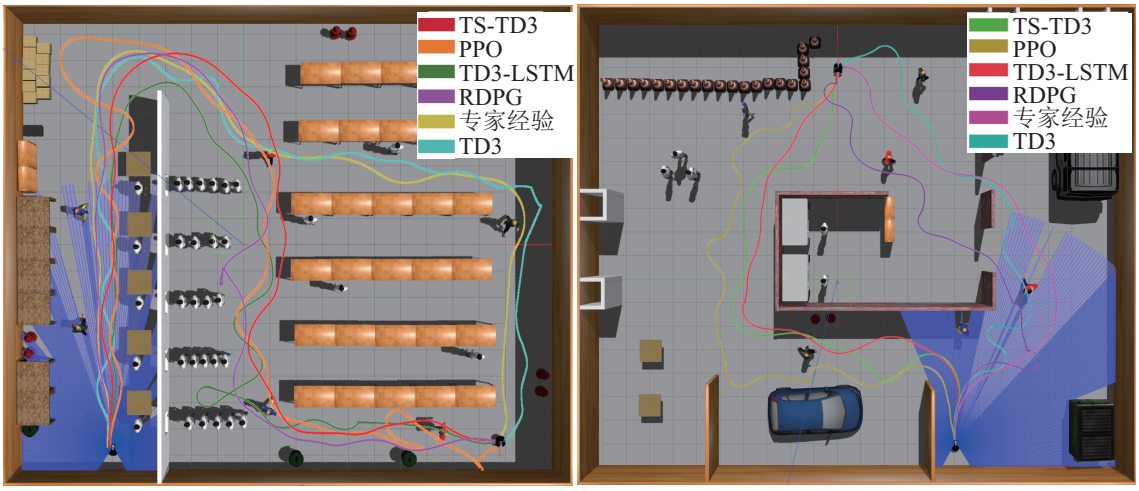

图 11　测试环境 E（左）和测试环境 F（右）

Fig.11　Testing environment E (left) and F (right)

4.5　真机测试

将仿真训练环境的模型移植到实际环境中（sim2real）是深度强化学习领域具有挑战性的方向。部分仿真实验算法无法完成从仿真环境到实际环境的迁移任务，本文将 TS-TD3 模型直接部署在真实机器人当中，模型参数不作调整。真实机器人采用二轮差速运动模型，最大线速度和最大角速度与仿真实验相同，传感器使用思岚 A2 平面激光雷达，机器人搭载 i7-7700 处理器，内置 Ubuntu 18.04 系统的计算平台。实际机器人定位使用 AMCL（自适应蒙特卡洛定位）算法，通过里程计估计机器人的运动速度，软件部署细节参考文 [31]，需要指出的是，全局地图是真实机器人定位算法 AMCL 所必需的输入，导航策略将不会参考全局地图。

为了定量分析 TS-TD3 算法在实际环境下的表现，在图 12 所示的实际动态场景中进行 50 轮实机测试，实际环境的基本尺寸为 8 m×9 m，环境中安排 3 台按照固定轨迹移动的机器人充当动态障碍物。并引入 TD3 算法进行比较，导航测试效果如表 5 所示。

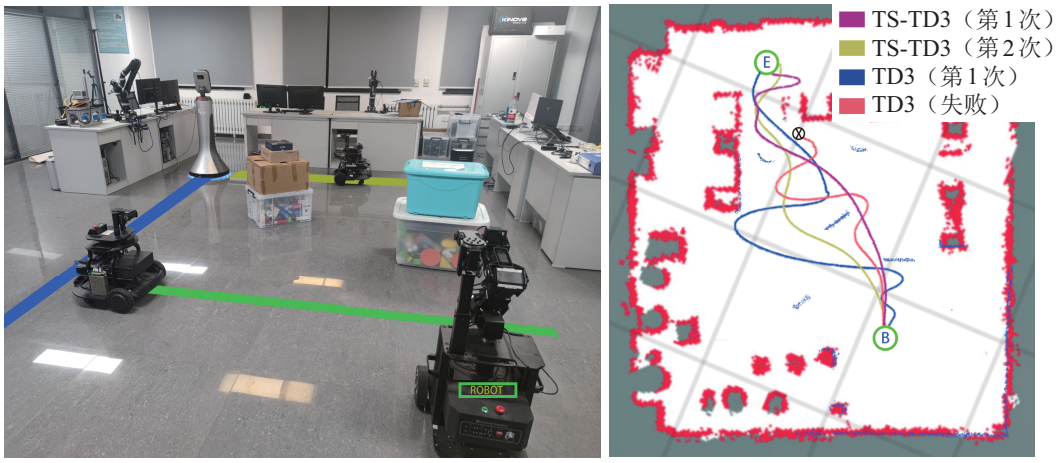

图 12 真实环境测试

Fig.12 Test in the real environment

表 5 真实机器人导航测试

Tab.5 Navigation test with real robots

	行驶距离平均值 /m	行驶距离最大值 /m	行驶距离最小值 /m	行驶时间平均值 /s	行驶时间最大值 /s	行驶时间最小值 /s	行驶距离标准差	行驶时间标准差	成功率
TS-TD3	7.73	12.48	7.16	30.25	37.58	26.69	1.92	6.05	84%
TD3	10.41	20.97	7.94	34.46	40.84	32.81	6.47	8.39	66%

基于 MDP 的 TD3 算法在面对动态障碍物时，算法的实时避障策略更接近随机选择一个方向进行转向以避开动态障碍物，不关注时序关联性的导航动作导致机器人面对单动态障碍物时表现一般，有时需要多次转向以离开动态障碍物运动的区域，更难以处理多动态障碍物交汇的情况。

TS-TD3 算法在面对单一动态障碍物时，关注动态障碍物的运动学信息，实时避障自然，导航轨迹前后连续，能够在不中断导航进行等待的情况下处理多动态障碍物交汇情况，在令机器人不断接近目标地点的同时导航动作符合一定的时序关联性。

5 结论（Conclusion）

动态复杂环境下的导航任务是移动机器人真正进入大众生活无法绕开的一个问题。TS-TD3 无地图导航方法实现了一种在动态场景中，机器人既关注隐藏的环境信息，同时又关注自身导航动作时序关联性的导航策略，并采用专家经验网络在训练前期引导机器人导航动作。主要创新有：(1) 将导航任务定义为 POMDP，引入历史信息来解决动态环境的部分可观测问题。(2) 为 DPG 引入历史信息基准，引导策略梯度探索环境的隐藏状态。在模型中引入 LSTM 网络处理历史信息，总结历史信息中的时序关联性，并映射到机器人导航动作中。(3) 引入专家经验网络指导模型训练，使机器人导航动作符合人工操作的特定规范，避免机器人为追求环境的高分回报，而在动态环境中反复循环、摆头。

同时必须指出，由于缺少全局地图，本文模型难以保证全局路径的最优性 [32]。这在狭窄、动态的环境中影响不大，因为机器人需要随着动态障碍物随时改变运动轨迹。但在大范围场景中，缺少全局地图可能导致机器人进行不必要的探索，并对机器人定位系统提出挑战 [33]。因此，未来的工作包括将 TS-TD3 算法与大范围场景全局地图相融合，允许全局地图保留黑盒地区，允许全局地图只对重要环境建筑进行建模。

参考文献（References）

[1] Li J, Qin H, Wang J Z, et al. OpenStreetMap-based autonomous navigation for the four wheel-legged robot via 3D-Lidar and CCD camera[J]. IEEE Transactions on Industrial Electronics, 2021, 69(3): 2708-2717.

[2] Pütz S, Simón J S, Hertzberg J. Move base flex a highly flexible navigation framework for mobile robots[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 3416-3421.

[3] Sathyamoorthy A J, Patel U, Guan T, et al. Frozone: Freezing-free, pedestrian-friendly navigation in human crowds[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4352-4359.

[4] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[5] Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications[J]. IEEE Transactions on Cybernetics, 2020, 50(9): 3826-3839.

[6] Quan H, Li Y S, Zhang Y. A novel mobile robot navigation method based on deep reinforcement learning[J]. International Journal of Advanced Robotic Systems, 2020, 17(3). DOI: 10.1177/1729881420921672.

[7] Chen Y F, Liu M, Everett M, et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2017: 285-292.

[8] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms[C]//International Conference on Machine Learning. PMLR, 2014: 387-395.

[9] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[DB/OL]. (2019-07-05) [2022-12-01]. https://arxiv.org/abs/1509.02971.

[10] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//International Conference on Machine Learning. PMLR, 2018: 1587-1596.

[11] Tai L, Paolo G, Liu M. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 31-36.

[12] Xie L H, Wang S, Rosa S, et al. Learning with training wheels: Speeding up training with a simple controller for deep reinforcement learning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 6276-6283.

[13] Everett M, Chen Y F, How J P. Motion planning among dynamic, decision-making agents with deep reinforcement learning [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 3052-3059.

[14] Devo A, Mezzetti G, Costante G, et al. Towards generalization in target-driven visual navigation by using deep reinforcement learning[J]. IEEE Transactions on Robotics, 2020, 36(5): 1546-1561.

[15] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[16] Heess N, Hunt J J, Lillicrap T P, et al. Memory-based control with recurrent neural networks[DB/OL]. (2015-12-14) [2022-12-01]. https://arxiv.org/abs/1512.04455.

[17] Malinowski M, Doersch C, Santoro A, et al. Learning visual question answering by bootstrapping hard attention[C]// European Conference on Computer Vision. Cham, Switzerland: Springer, 2018: 3-20.

[18] Liu H C, Huang Z Y, Wu J D, et al. Improved deep reinforcement learning with expert demonstrations for urban autonomous driving[C]//IEEE Intelligent Vehicles Symposium. Piscataway, USA: IEEE, 2022: 921-928.

[19] Xie L H, Miao Y S, Wang S, et al. Learning with stochastic guidance for robot navigation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 166-176.

[20] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. PMLR, 2015: 448-456.

[21] Weaver L, Tao N. The optimal reward baseline for gradient-based reinforcement learning[DB/OL]. (2013-01-10) [2022-12-01]. https://arxiv.org/abs/1301.2315.

[22] Sutton R S, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation [C]//Advances in Neural Information Processing Systems 12. Red Hook, USA: Curran Associates Inc., 1999: 1057-1063.

[23] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8(3): 229-256.

[24] Silva A, Gombolay M. Encoding human domain knowledge to warm start reinforcement learning[C]//AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI, 2021: 5042-5050.

[25] Xie L H, Markham A, Trigoni N. SnapNav: Learning mapless visual navigation with sparse directional guidance and visual reference[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 1682-1688.

[26] Hermann K M, Malinowski M, Mirowski P, et al. Learning to follow directions in street view[C]//AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI, 2020: 11773-11781.

[27] Wawrzynski P. Control policy with autocorrelated noise in reinforcement learning for robotics[J]. International Journal of Machine Learning and Computing, 2015, 5(2): 91-95.

[28] Jin C, Krishnamurthy A, Simchowitz M, et al. Reward-free exploration for reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2020: 4870-4879.

[29] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[DB/OL]. (2017-08-28) [2022-12-01]. https://arxiv.org/abs/1707.06347.

[30] Meng L, Gorbet R, Kulić D. Memory-based deep reinforcement learning for POMDPs[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2021: 5619-5626.

[31] Everett M F. Robot designed for socially acceptable navigation [D]. Cambridge, USA: MIT, 2017.

[32] 宁宇铭，李团结，姚聪，等．基于快速扩展随机树－贪婪边界搜索的多机器人协同空间探索方法[J]．机器人，2022，44(6)：708-719.
Ning Y M, Li T J, Yao C, et al. Multi-robot cooperative space exploration method based on rapidly-exploring random trees and greedy frontier-based exploration[J]. Robot, 2022, 44(6): 708-719.

[33] 王浩，卢德玖，方宝富．动态环境下基于增强分割的 RGB-D SLAM 方法[J]．机器人，2022，44(4)：418-430.
Wang H, Lu D J, Fang B F. RGB-D SLAM method based on enhanced segmentation in dynamic environment[J]. Robot, 2022, 44(4): 418-430.

作者简介：

姜　杨（1982 –），男，博士，副教授。研究领域：移动机器人自主导航与路径规划。

曾铁文（1999 –），男，硕士生。研究领域：移动机器人无地图导航。

万东东（1998 –），男，硕士生。研究领域：强化学习，移动机器人自主导航与路径规划。