文章编号:1671-4598(2023)01-0153-07 DOI:10.16526/j.cnki.11—4762/tp.2023.01.023 中图分类号:TP311 文献标识码:A

基于深度强化学习的移动机器人动态路径规划算法

张柏鑫，杨毅镔，朱华中，刘安东，倪洪杰
（浙江工业大学 信息工程学院，杭州 310012）

摘要：为了在复杂舞台环境下使用移动机器人实现物品搬运或者载人演出，提出了一种基于深度强化学习的动态路径规划算法；首先通过构建全局地图获取移动机器人周围的障碍物信息，将演员和舞台道具分别分类成动态障碍物和静态障碍物；然后建立局部地图，通过LSTM网络编码动态障碍物信息，使用社会注意力机制计算每个动态障碍物的重要性来实现更好的避障效果；通过构建新的奖励函数来实现对动静态障碍物的不同躲避动作；最后通过模仿学习和优先级经验回放技术来提高网络的收敛速度，从而实现在舞台复杂环境下的移动机器人的动态路径规划；实验结果表明，该网络的收敛速度明显提高，在不同障碍物环境下都能够表现出好的动态避障效果。

关键词：移动机器人；LSTM；深度强化学习；动态路径规划；实时避障

Dynamic Path Planning Algorithm of Mobile Robot Based on Deep Reinforcement Learning

ZHANG Baixin, YANG Yibin, ZHU Huazhong, LIU Andong, NI Hongjie
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310012, China)

Abstract: In order to realize that mobile robot carries goods or performs manned performances in complex stage environment, a dynamic path planning algorithm based on deep reinforcement learning is proposed. Firstly, the obstacle information around the mobile robot is obtained by constructing a global map, and the actors and stage props are classified into dynamic obstacles and static obstacles respectively. Then a local map is established to acquire the dynamic obstacle information through LSTM network, and the importance of each dynamic obstacle is calculated to achieve better obstacle avoidance effect through social attention mechanism. Different avoidance situations of dynamic and static obstacles are realized by constructing a new reward function. Finally, the simulation learning and priority experience playback technology are used to improve the convergence speed of the network, so as to realize the dynamic path planning of the mobile robot in the complex stage environment. The experimental results show that the convergence speed of the network is significantly improved, and it can show the good dynamic effect in different obstacle environments.

Keywords: mobile robot; LSTM; deep reinforcement learning; dynamic path planning; real time obstacle avoidance

0 引言

为了丰富全民公共文化服务，特别是满足基层文化多样化服务需求，需要在基层小型文化服务综合体[1]活动空间中开展文化演出、会议议事、展览阅览以及民俗活动等文化服务功能。因服务空间不同，其功能空间内的配置设施和使用要求也不同。为了达到小型文化综合体“一厅多用”要求，往往要通过移动机器人协助完成多种功能空间相互快速组合以及切换，实现小型综合体空间内拥挤环境下的动态路径规划与快速搬运服务等，从而满足单一空间多种文化服务需求。

路径规划[2-3]是移动机器人实现各种功能的基础，分为全局静态路径规划[4]和局部动态路径规划[5]。在舞台环境这种充满动态障碍物的环境下对移动机器人的动态路径规划算法要求很高。目前的动态路径规划算法可以分成两类：基于反应的避障算法[6]和基于预测的避障算法[7]。基于反应的避障算法有人工势场法[8-9]，通过假设机器人在一种虚拟立场下运动受到障碍物的斥力和目标点的引力，但是在复杂环境下会陷入局部最优解或震荡。最佳反向碰撞（ORCA）[10-11]通过引入一个时间窗口，将相对位置转化为速度，给每个动态障碍物都设定速度区间，最后对所有线性空间用线性规划求出最优解，从而计算出最优路径。但是会存在抖动问题以及在复杂环境下规划不流畅的问题。基于反应的算法不能预测其他障碍物的运动趋势，从整个规划上来看，有时会产生不自然的轨迹，由于只考虑当前状态，

收稿日期:2022-06-11； **修回日期**:2022-07-14。
基金项目:国家自然科学基金项目（61973275）;浙江省省属高校基本科研业务（RF-A2020004）。
作者简介:张柏鑫(1998-),男,江苏兴化人,硕士生,主要从事移动机器人路径规划方向的研究。
引用格式:张柏鑫,杨毅镔,朱华中,等.基于深度强化学习的移动机器人动态路径规划算法[J].计算机测量与控制,2023,31(1):153-159,166.

依赖传感器快速更新速率对障碍物作出快速反应，所以规划出来的路径往往不是最优的。

基于预测的算法有动态窗口法[12-14]和基于深度强化学习的方法[15-18]，动态窗口法是在速度空间中采样多组速度，并模拟机器人在这些速度下一定时间内的轨迹，得到多组轨迹之后，选取最优轨迹对应的速度来驱动机器人移动。基于深度强化学习的方法将深度学习的感知能力和强化学习的决策能力相结合，可以实现端到端的控制方法。通过传感器直接获得周围的状态信息，经过网络处理，输出动作指令，具有非常好的自适应性，成为路径规划领域新的研究热点。

文献［19］提出了一种基于深度强化学习的分布式多机器人避障策略（CADRL），假设动态障碍物会主动避让机器人，通过值函数网络对周围动态障碍物的状态进行编码，通过预测障碍物的运动趋势，来规划出动态路径。文献［20］在文献［19］的基础上通过引入双经验池来提高算法的收敛速度。文献［21］提出了 LSTM 策略，通过引入长短期记忆神经网络来将动态障碍物的可变大小状态转换为固定长度向量，并以与机器人距离相反的顺序输入人类状态。上述的文献只考虑单个动态障碍物对机器人的影响，却没有考虑多个动态障碍物同时对机器人的影响。文献［22］提出了社会注意力机制，通过对动态障碍物的运动趋势分析，来捕获环境中动态障碍物的相对重要性。文献［23］提出了 SARL 策略，使用了一种图形结构来表示人群，并预测导航任务中的行人注意力得分。文献［24］直接将原始传感器数据和目标信息映射到控制命令上。文献［25］采用渐进式的由易到难的训练策略，将 DQN 与迁移学习相结合应用在导航中，提高了收敛网络收敛速度。文献［26］利用 Delaunay 三角剖分对障碍物进行编码，并结合扩展的混合 A * 方法，在时间－状态空间中有效地搜索最优解。

在现实环境下或拥挤的环境中，障碍物类型复杂，不仅存在动态障碍物，还存在静态障碍物。机器人对不同类型的障碍物处理应该也不相同。针对上述现有技术存在的问题，本文提出一种新的基于深度强化学习的动态路径规划算法。主要工作有：

1）将 LSTM 网络和社会注意力机制融合建立的新的神经网络模型来处理动态路径规划问题。

2）建立局部地图和全局地图来分别处理动静态障碍物信息，设计新的奖励函数来针对不同障碍物的情况。

3）针对训练前期收敛慢的问题，通过模仿学习对网络参数进行预训练，引入优先级经验回放技术来提高网络的收敛速度。

4）设计不同的障碍物环境来对比不同策略的动态避障效果。

1 基于深度强化学习的动态路径规划

如图 1 所示，在强化学习框架下，机器人通过与环境的交互来学习动作能够使其在给定环境中所获得的累计奖励最大化。然而机器人仅仅依靠与环境的交互以及动作的奖励来学习，在复杂环境下的表现往往不是很好。由于深度学习具有强大的感知能力，不仅能够为强化学习带来端到端优化的优势，而且使得强化学习不再受限于低维的空间中，为此引入深度学习来提高系统的感知能力，极大地拓展了强化学习的使用范围。

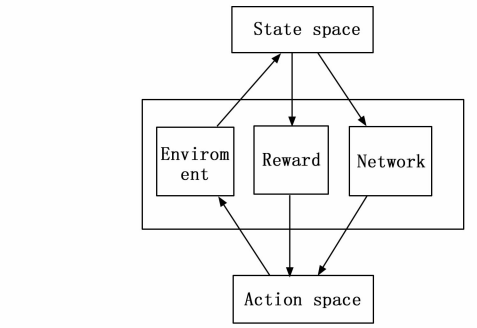

图 1 基于深度强化学习的动态路径规划流程

1.1 马尔可夫决策过程

基于强化学习的移动机器人的路径规划问题首先需要通过利用马尔可夫决策过程（MDP，markov decision process）来实现问题的形式化。MDP 可以描述为：机器人在某一时刻 t 下的状态为 S_t，根据最优策略 π^* 选择动作 a_t，机器人根据状态转移概率 P，从当前状态转移到下一个状态 $S_(t+1)$，时间间隔为 $\triangle t$ 然后根据当前状态得到奖励 r_t，再更新状态值函数：

$$V^*(S_t) = \sum \gamma^{\triangle t} P(S_t, a_t) \tag{1}$$

其中：状态值函数 $V^*(S_t)$ 是从状态 s_t 出发，按照最优策略 π^* 选取动作后得到的期望回报，机器人不仅要考虑当前的奖励还要考虑未来的奖励，所以设置折扣率 $\gamma \in (0, 1)$。

最优策略则通过最大化累计回报获得：

$$\pi^*(u_t) = \operatorname{argmax} R(u_t, a_t) + \gamma^{\triangle t} \cdot \int_{u_t+\triangle t} P(u_{u_t+\triangle t} \mid u_t, a_t) \cdot V^*(u_{u_t+\triangle t}) \mathrm{d}u_t + \triangle t \tag{2}$$

其中：u_t 表示当前移动机器人和障碍物的联合状态，a_t 表示 t 时刻的动作，γ 表示折扣率，$\triangle t$ 表示两个动作之间的时间间隔，V^* 表示最优值函数，P 表示状态转移函数，R 表示为奖励函数；模型的好坏主要在于状态空间 S、动作空间 A 和奖励空间 R 的设计。

1.2 状态空间

动态路径规划的状态空间包含机器人的状态和动静态障碍物的状态，机器人、动态障碍物和静态障碍物的状态分别定义为：

$$S_r = [P_x, P_y, G_x, G_y, V_x, V_y, \theta, r] \tag{3}$$

$$S_D = [P_x, P_y, V_x, V_y, r] \tag{4}$$

$$S_S = [P_x, P_y, r] \tag{5}$$

其中：$[P_x, P_y]$ 是物体的当前坐标，$[G_x, G_y]$ 是机器人的目标点，$[V_x, V_y]$ 是机器人或者动态障碍物的当前

时刻的速度，θ 是机器人当前时刻的航向角，r 是物体膨胀后的半径。

机器人周围的障碍物在 T 时刻内的联合状态即为网络的输入层：

$$u_t = [S_r, S_D, S_S] \tag{6}$$

1.3 奖励

机器人在移动过程中需要尽量避免与动态障碍物发生碰撞。由于动态障碍物的不确定性，随着与动态障碍物距离的减少，发生碰撞的概率就越大，而面对静态障碍物，随着距离的靠近并不会增加发生碰撞的概率。所以针对以上情况，设计了如下的奖励函数。

$$R(u_t, a_t) = \begin{cases} 1 & G_{x,y} = P_{x,y} \\ -0.25 & D_d \leq 0 \\ -0.25 + D_d \cdot 0.5 & 0 < D_d < 0.5 \\ -0.2 & D_s \leq 0 \\ 0 & others \end{cases} \tag{7}$$

其中：D_d，D_s 分别为为机器人和动态障碍物和静态障碍物之间的距离，为了减了机器人和动态障碍物发生碰撞的概率，当机器人靠近动态障碍物 0.5 米时，即 $D_d < 0.5$ 时，则认为机器人可能会和动态障碍物发生碰撞，随着距离的减少逐渐增加负奖励的值，从而降低移动机器人与动态障碍物发生碰撞的概率。而机器人靠近静态障碍物时，由于静态障碍物没有运动趋势，只要机器人不接触静态障碍物就不会发生碰撞，当机器人与静态障碍物发生碰撞时，即 $D_s \leqslant 0$ 时，给予负的奖励。从而在动态规划过程中，实现对动静态障碍物的不同应对策略。

1.4 动作空间

机器人的动作空间由线速度和角速度组成 $A = [w, v]$，v 为线速度 w 为角速度。为了符合动力学约束，本文将角速度在［0，90］区间内分成 15 等分，线速度按照函数 $y = 1/x$，x 取 1，2，3，4，5 可获得 5 个变化平滑的线速度，动作空间共有 75 种动作组合，其中不同的颜色代表那个区间动作选择的可能性，如图 2 所示。

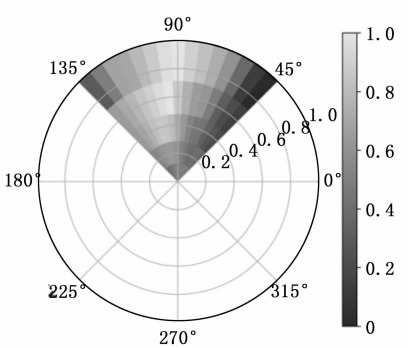

图 2　动作空间

2　基于值函数的深度强化学习方法

2.1 网络结构

文献［19］使用一个浅层的网络编码了障碍物信息，缺乏对障碍物信息的处理，本文此基础上通过引入 LSTM 网络[21]来编码障碍物信息。通过局部地图和社会注意力机制[22]来提取动态障碍物的特征，预测动态障碍物的运动趋势，通过全局地图来避开静态障碍物，网络模型具有更强的自适应性以及鲁棒性。

如图 3 所示，网络处理流程主要有：首先建立全局地图获取机器人和周围障碍物的状态，然后通过建立局部地图，通过 LSTM 网络编码动态障碍物状态和机器人状态输入社会注意力机制网络，来确定周围每个动态障碍物的注意力分数，然后通过 LSTM 网络处理静态障碍物的信息，将处理后的状态输入全连接层，最后通过激活函数对其进行归一化处理来得到最优值函数。

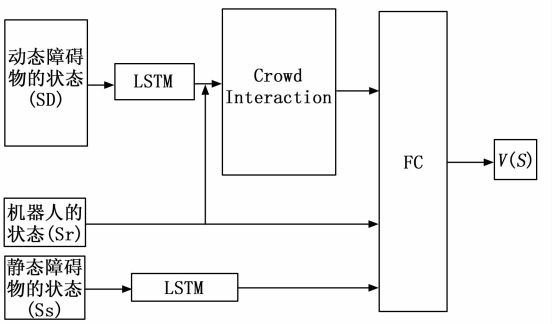

图 3　深度神经网络结构

2.2 网络输入模块

网络的输入参数由三部分组成：移动机器人的状态、动态障碍物的状态和静态障碍物的状态。静态障碍物的状态可视为已知信息存储，通过预先建立环境地图。动态障碍物的状态通过局部地图获得。本文以机器人为中心将地图参数化建立二维栅格局部地图。如图 4 所示，机器人位于原点，构造一个 $L \times L \times 3$ 映射张量 M_i，来编码动态障碍物 D_i 的位置和速度信息，张量 M_i 表示为：

$$M_i(a, b) = \sum\nolimits_{j \in N_i} \partial\, D_i \tag{8}$$

其中：D_i 是第 i 个动态障碍物的状态信息，(a, b) 用来存放坐标信息，∂ 是一个特征函数来确保障碍物在范围之内，N_i 是周围所有动态障碍物的状态信息的集合。

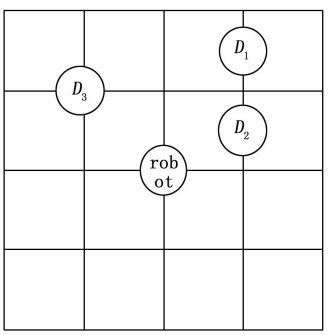

图 4　维栅格局部地图

2.3 社会注意力机制模块

社会注意力机制由多个多层感知器（MLP）组成，首

先将移动机器人的状态和动态障碍物的状态以及映射张量 M_i 通过 MLP 编码到固定长度的向量 e_i 中，然后再将 e_i 送入到下一层 MLP 中来获得机器人与动态障碍物之间的交互特征 h_i，然后再送入下一层 MLP 中来获得不同动态障碍物的注意力分数 α_i，注意力分数代表着动态障碍物与移动机器人发生碰撞的可能性。然后将所有的交互特征通过加权组合来表示所有动态障碍物对移动机器人的下一步动作的整体影响，将联合状态和交互特征输入到 MLP 来作为最优值函数的估计值。

$$e_i = \varnothing_e(s, M_i; w_e) \tag{9}$$

$$h_i = \Phi_h(e_i; w_h) \tag{10}$$

$$e_m = \frac{1}{n}\sum_{k=1}^{n} e_k \tag{11}$$

$$\alpha_i = \Phi_a(e_i, e_m; w_a) \tag{12}$$

$$C = \sum_{i=1}^{n} \mathrm{softmax}(\alpha_i) h_i \tag{13}$$

$$V = f_v(s, s_s, c; w_v) \tag{14}$$

其中：$\varnothing_e$ 是激活函数，w_e 是权重，Φ_h 是一个具有 ReLU 非线性的全连通层，w_h 是网络权值。e_m 是所有 e_i 向量的总和，Φ_a 是具有 ReLU 激活的 MLP，w_a 是权重。f_v 是具有 ReLU 激活的 MLP，权重用 w_v 表示。s 是机器人的状态，s_s 是静态障碍物的状态。

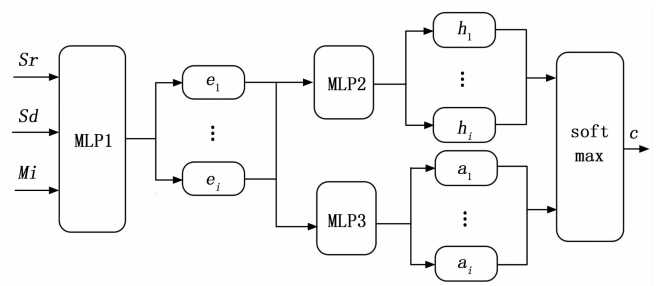

图 5 社会注意力机制模型

2.4 LSTM 模块

前馈神经网络需要固定长度的输入。然而移动机器人在移动的过程中遇到的障碍物的数量不固定，从而导致网络输入层的输入的参数也会发生变化，目前的方法通过设定固定的输入层参数，当障碍物数量过多或者很少的时候都会影响网络的性能。长短期记忆网络可以接收任意长度的序列参数，产生固定长度的输出，因此本文引入 LSTM 网络优化深度强化网络模型的性能。

LSTM 网络由其权重 $\{w_i, w_f, w_o\}$ 和偏差 $\{b_i, b_f, b_o\}$ 参数化，其中 $\{i, f, o\}$ 对应于输入门、遗忘门和输出门。如图 6 所示，机器人周围的障碍物按照距离依次送入 LSTM 单元，每个 LSTM 单元具有 3 个输入：障碍物的状态 S、先前的隐藏状态 h_i 和先前的单元状态 c_i。

在每个决策步骤中，每个障碍物的状态依次输入到 LSTM 单元中。LSTM 单元的初始状态 h_0，c_0 为 0，然后通过输入第一个状态 s_1 生成 $\{h_1, c_1\}$，然后运送到第二个 LSTM 单元生成 $\{h_2, c_2\}$，从而编码所有障碍物的状态信息。在输出障碍物信息时，可以通过 LSTM 网络调整障碍物信息的顺序，通过训练网络权重参数，通过遗忘门将远离机器人的障碍物信息遗忘，将靠近机器人的障碍物信息优先放置在最后编码的向量中。最后通过隐藏状态 h_n 将机器人周围障碍物的状态信息编码成一个固定长度的向量，并输送到前馈网络处。不仅可以解决前馈网络参数不固定的问题，还可以对障碍物状态信息进行排序，使移动机器人具有更好的性能。

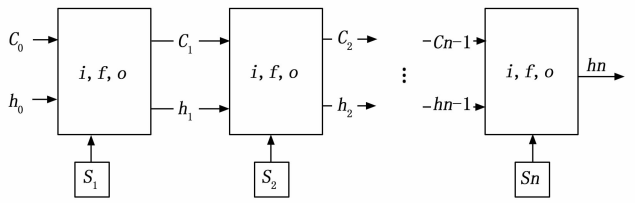

图 6 LSTM 网络模型

2.5 模仿学习

网络训练初期，由于环境复杂，障碍物多，发生碰撞的概率很大，网络需要很长时间才能度过探索阶段，网络的初始权重很差，机器人需要很久才能到达目标点。如图 7 所示。

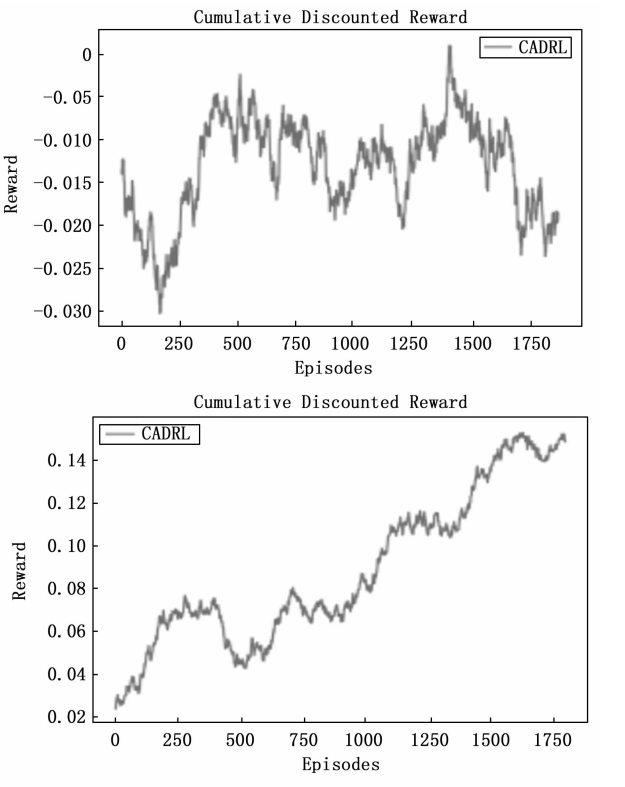

图 7 未引入模仿学习和引入模仿学习的回报曲线

左图是未引入模仿学习的回报函数，总奖励在很长一段时间内处于负值状态，而引入模仿学习的总奖励一直处于正值状态。所以通过使用模仿学习，对网络进行一定次数的预训练来优化网络权重。由于仿真环境一样，只是改

变机器人使用的避障策略，可以使策略度过探索期，所以通过预训练的网络可以提高收敛的速度。

2.6 优先级经验回放

DQN 算法采用了经验回放机制来解决经验数据的相关性和非平稳分布的问题，但存在一个问题是其采用均匀采样和批次更新，导致特别少但价值特别高的经验没有被高效的利用。本文采用一种“优先级经验回放”技术解决采样问题，从而提高算法的收敛速度。在优先级经验回放中，通过 TD-error 来给每一条经验添加重要性，TD-error 是某个时刻动作的值函数和当前网络的最优值函数的一个差值，差值越大则说明当前的经验比较差。所以我们定义：

$$P_t = (|\delta_t| + \varepsilon)^{\alpha} \quad (15)$$

其中：P_t 是选择当前经验的概率，α、ε 为常数，δ_t 为 TD-error。

3 实验验证

3.1 算法仿真与分析

本节主要对改进后的策略和当前主流策略的动态避障效果进行对比分析。本文的网络模型在 Python 中使用 Pytorch 实现，在 GeForce RTX 970 GPU 和 i7-4790K 的电脑上训练并测试。仿真环境实验是基于 python 中的 gym 库搭建的，动态障碍物在 6X6 的正方形边上随机生成起点，目标点为对应边上的一点，半径大小为 0.3 的圆，静态障碍物随机在环境下随机位置生成，为半径为 0.5～1.3 的圆。动态障碍物通过 ORCA 策略控制移动，可以避免动态障碍物之间以及和静态障碍物发生碰撞，其仿真参数如表 1 所示，通过时间差分法来对网络模型进行训练，通过 Adam 优化器[27]来优化最优值函数，网络的训练流程如下所示。

表 1　仿真参数

参数	含义	大小
rl_learning_rate	学习率	0.001
gamma	折扣因子	0.9
train_batches	训练批次	100
train_episodes	训练次数	10 000
sample_episodes	采样次数	1
target_update_interval	更新目标网络的间隔	50
evaluation_interval	策略评估的间隔	1 000
capacity	经验池大小	50 000

算法 1：网络训练流程。

步骤 1：使用模仿学习对网络进行预训练；

步骤 2：预训练结果初始化经验池；

步骤 3：初始化目标网络参数；

步骤 4：开始训练；

步骤 5：随机生成障碍物位置；

步骤 6：是否发生碰撞、到达目标或者超出限制时间；

步骤 7：步骤 6 成立结束当前 episode 否则执行步骤 8；

步骤 8：将联合状态输入网络；

步骤 9：根据当前状态使用最大化奖励选择动作；

步骤 10：获得奖励，更新下一刻状态；

步骤 11：将状态、动作、奖励和下一刻状态存储到经验池；

步骤 12：使用优先级经验回放从经验池里选择 batch；

步骤 13：根据 batch 优化价值网络；

步骤 14：通过 Adam 优化器优化网络参数；

步骤 15：更新目标网络参数；

步骤 16：返回最优值函数。

3.2 仿真结果

本节主要将本文的策略模型与 CADRL[19]、LSTM-RL[21]和 SARL[23]3 种策略模型进行比较，来验证网络模型的性能。

图 8（a）比较了不同策略的累计回报曲线，从图中可以看出我们的模型由于引入优先经验回放技术，在预训练的时候就获得了很好的效果，初始经验很高，所以收敛速度也是最快的。图 8（b）、（c）和（d）中比较了不同模型到达目标点所花费的时间，以及在仿真过程中的成功率和碰撞率，从图中可以看出我们的模型在相同起点和终点的条件下到达终点所花费的时间最短、碰撞率较低以及成功率很高。说明引入率 LSTM 网络和社会注意力机制后的模型效果表现很好，结合模仿学习以及最优经验回放后的网络训练效果更好，很好的验证了本文的网络模型的有效性。

3.3 不同策略的动态路径规划图

由于我们的模型对动态障碍物和静态障碍物都做了训练，因此设计了两组实验用来做对比实验。实验测试了 4 种策略在 500 次的实验下的动态规划结果。

3.3.1 不存在静态障碍物

首先在没有静态障碍物的环境下对比了不同策略的路径规划效果，设定了 5 个动态障碍物，动态障碍物采用 ORCA 策略，不会主动避让移动机器人。

本文的策略通过注意力机制，可以预测动态障碍物的运动趋势，避免与动态障碍物发生碰撞，当面对多个动态障碍物的时候，通过对障碍物排序可以实现更好的动作抉择，成功的避开障碍物。LSTM _ RL 虽然可以对障碍物进行排序，优先躲避最近的障碍物，但是却忽略了即将到来的障碍物对将来动作的影响，会出现即使躲避开当前障碍物，又在移动过程中与下一个障碍物相遇，导致规划的时间过长。SARL 虽然能够预测动态障碍物的趋势，但是太注重未来的障碍物而忽视了靠近的动态障碍物。从表 2 中和图 9 可以看到本文的策略，到达目标点的时间最短且路径更加平滑。

表 2　不存在静态障碍物的 500 次测试结果

策略	成功率/%	失败率/%	平均时间/s	总奖励
CADRL	0.94	0.06	11.49	0.252 5
LSTM-RL	0.89	0.04	12.65	0.214 7
SARL	0.99	0.01	11.08	0.317 4
Ours	1.00	0.00	10.38	0.341 6

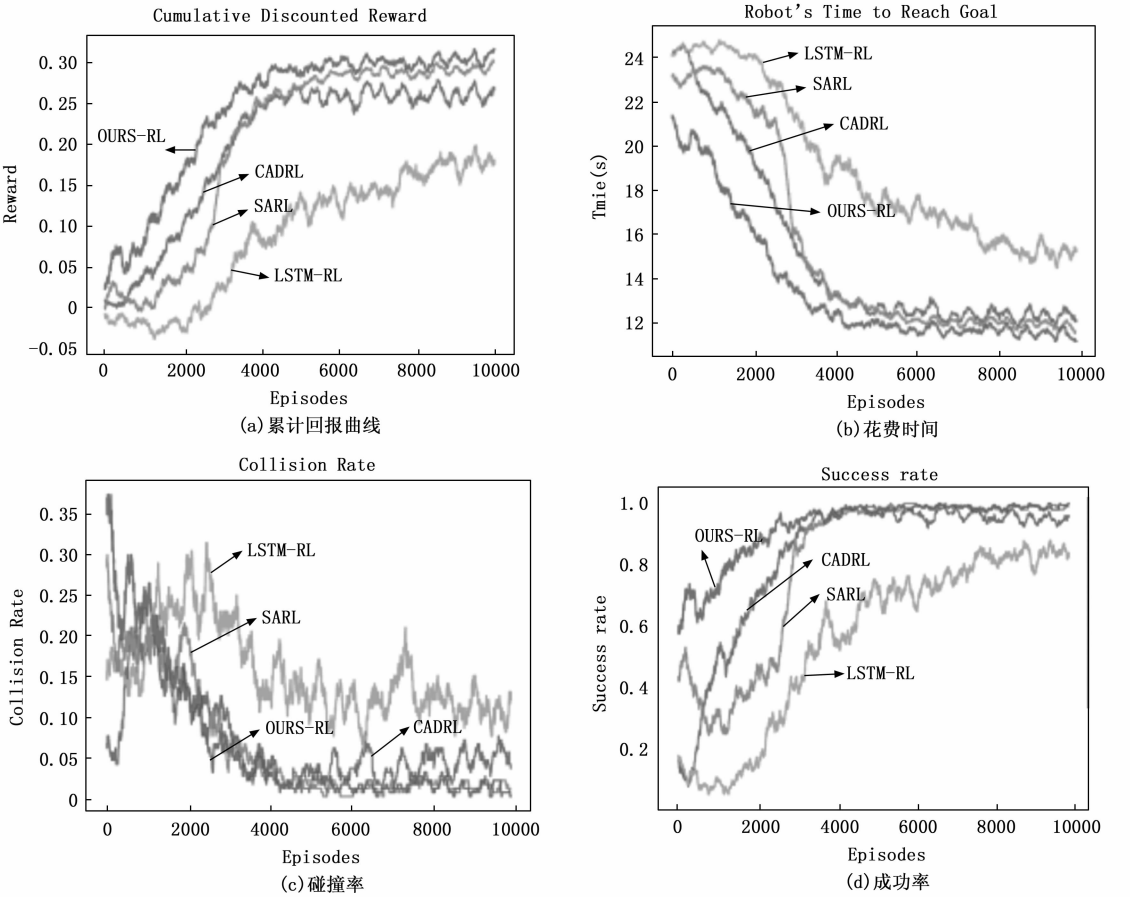

图 8 10 000 次训练结果

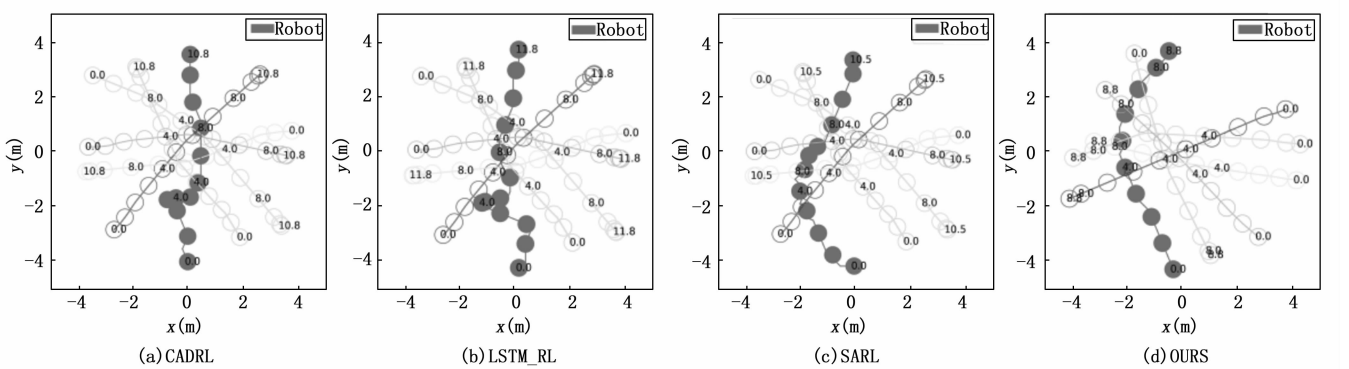

图 9 不同策略的一次测试案例

3.3.2 存在静态障碍物

由于障碍物的增加，为了更好的体现避障效果扩大了地图环境，由于静态障碍物的随机生成，本文主要分析了不同策略的避障效果，设定了 4 个静态障碍物和 3 个动态障碍物。

如图 10 所示 CADRL 仅使用一个浅层的网络不能很好的提取复杂动态环境中的信息，当环境复杂时很容易发生碰撞，LSTM _ RL 策略根据离障碍物的距离进行排序，但是容易陷死在拥挤的环境下，SARL 策略提取了行人的交互特征的有较好的避障效果，但是当被障碍物包围时，会出现" 冻结" 的问题，只有当障碍物离开时才能继续移动。而我们的策略不仅提取了行人的交互特征，而且排序了障碍物，针对不同的障碍物设置了不同的奖励策略，使机器人有更好的动作选择，更好的躲避不同的障碍物，即使在拥挤的环境下，也能规划出平滑的路径。如表 3 所示本文的模型成功率最高，花费的时间最短，均优于其他策略。

表 3 存在静态障碍物的 500 次测试结果

策略	成功率/%	失败率/%	平均时间/s	总奖励
CADRL	0.44	0.01	26.87	0.004 9
LSTM-RL	0.16	0.04	21.79	−0.030 5
SARL	0.74	0.03	19.02	0.046 6
Ours	0.80	0.03	16.78	0.129 9

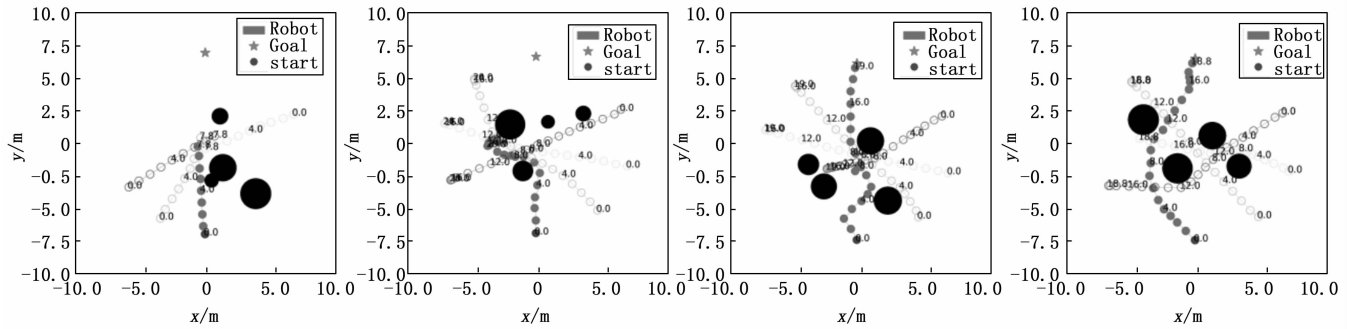

图 10　不同策略的一次测试案例

4　结束语

本文提出了一种基于深度强化学习的动态路径规划算法。首先通过构建全局地图获取移动机器人周围的障碍物信息，将障碍物分类成动态障碍物和静态障碍物。然后建立局部地图通过 LSTM 网络编码动态障碍物信息，通过社会注意力机制计算每个动态障碍物的重要性来实现更好的避障。通过构建新的奖励函数来应对动静态障碍物的不同躲避情况。最后通过模仿学习和优先级经验回放技术来提高网络的收敛速度，实验结果验证了该模型的准确性和有效性，表明了该模型能够实现更好的动态避障效果，但是没有现实机器人上实现，因此下一步研究如何在现实机器人上实现动态路径规划。

参考文献：

[1] 刘　榛，陈伟忠，王彦库，等. 多功能小型文化服务综合体内通与监督系统的要求分析和系统构型 [J]. 演艺科技，2021 (6)：50－55.

[2] 谭　辉. 移动机器人控制方法综述 [J]. 科技创新与应用，2021，11 (20)：125－127.

[3] 崔　奇，夏　浩，滕　游，等. 移动机器人自主导航系统及上位机软件设计与实现 [J]. 计算机测量与控制，2022，30 (1)：141－146.

[4] 王梓强，胡晓光，李晓筱，等. 移动机器人全局路径规划算法综述 [J]. 计算机科学，2021，48 (10)：19－29.

[5] 鲍庆勇，李舜酩，沈　峘，等. 自主移动机器人局部路径规划综述 [J]. 传感器与微系统，2009，28 (9)：1－4，11.

[6] 杨月全，韩　飞，曹志强，等. 基于激光传感器的动态拟合避障控制与仿真 [J]. 系统仿真学报，2013，25 (4)：704－708.

[7] 朱坤财，徐郑攀，赵自奇，等. 基于航迹预测的水面无人艇动态避障方法 [J]. 中国测试，2021，47 (11)：28－33，46.

[8] 石志刚，梅　松，邵毅帆，等. 基于人工势场法的移动机器人路径规划研究现状与展望 [J]. 中国农机化学报，2021，42 (12)：182－188.

[9] 王翔昌，吴训成，张伟伟，等. 基于改进人工势场算法的自主车辆局部路径规划方法研究 [J]. 计算机与数字工程，2022，50 (3)：554－558，630.

[10] BERG J，GUG S J，LIN M，et al. Reciprocal n-body collision avoidance [M]. Robotic sresearch. Springer，Berlin，Heidelberg，2011：3－19.

[11] 税　斌. 基于 ORCA 碰撞避免的人群疏散模拟 [J]. 现代计算机（专业版），2016 (2)：73－75.

[12] FOX D，BURGARD W，THRUN S. The dynamic window approach to collision avoidance [J]. IEEE Robotics & Automation Magazine，1997，4 (1)：23－33.

[13] 劳彩莲，李　鹏，冯　宇. 基于改进 A～＊与 DWA 算法融合的温室机器人路径规划 [J]. 农业机械学报，2021，52 (1)：14－22.

[14] 彭育强，黄泽龙，李少伟. 基于动态窗口法的移动机器人自动避障导航研究 [J]. 自动化仪表，2020，41 (10)：26－29，33.

[15] 赵玉新，杜登辉，成小会，等. 基于强化学习的海洋移动观测网络观测路径规划方法 [J]. 智能系统学报，2022，17 (1)：192－200.

[16] JIA Q，YANG M，MIAO Y，et al. Path planning using deep reinforcement learning based on potential field in complex environment [C] //Journal of Physics：Conference Series. IOP Publishing，2021，1748 (2)：022016.

[17] HU J，NIU H，CARRASCO J，et al. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning [J]. IEEE Transactions on Vehicular Technology，2020，69 (12)：14413－14423.

[18] 卢东来，郑战光. 基于深度学习的多机械手轨迹规划系统设计 [J]. 计算机测量与控制，2020，28 (11)：247－250.

[19] CHEN Y F，LIU M，EVERETT M，et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning [C] //2017 IEEE international conference on robotics and automation (ICRA). IEEE，2017：285－292.

[20] CHEN Y F，EVERETT M，LIU M，et al. Socially aware motion planning with deep reinforcement learning [C] //2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE，2017：1343－1350.

[21] EVERETT M，CHEN Y F，HOW J P. Motion planning among dynamic，decision-making agents with deep reinforcement learning [C] //2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE，2018：3052－3059.

（下转第 166 页）