基于 TD3 算法的无人驾驶消防车匝道汇入模型

陈泽宇

中国消防救援学院，北京 102202

摘　　要： 在匝道汇入时，无人驾驶消防车应能清晰感知周围交通状况，以确保车辆适应不同交通场景，并做出相应决策，避免碰撞、优化路径。为解决现有基础无人驾驶消防车匝道汇入模型（以下简称基础模型）中深度强化学习算法存在的收敛速度慢，汇入成功率较低等问题，提出建立将 TD3(Twin Delayed Deep Deterministic Policy Gradient) 算法与长短期记忆网络相结合的新型智能车匝道汇入（LSTM-TD3）模型，用长短期记忆网络分别替换 Actor 网络的输入层和 Critic 网络的输入层，以有效预测输出结果。试验结果表明，LSTM-TD3 模型比基础模型收敛速度和汇入成功率更高，同时能够帮助无人驾驶消防车做出最佳汇入决策。

关 键 词： 无人驾驶；消防车；匝道汇入；深度强化学习；长短期记忆网络

Intelligent Fire Truck Ramp Merge Model Based on TD3 Algorithm

Chen Zeyu

China Fire and Rescue Academy, Department of Emergency Communications and Information Engineering, Beijing 102202

Abstract: During on-ramp merging, autonomous firefighting vehicles should be able to clearly perceive the surrounding traffic conditions to ensure that the vehicle can adapt to different traffic scenarios, make appropriate decisions, avoid collisions, and optimize routes. To address issues such as slow convergence and lower merging success rates present in the existing baseline autonomous firefighting vehicle merging model (hereafter referred to as the baseline model), this paper proposes a novel intelligent vehicle merging model (LSTM-TD3) that combines the TD3 (Twin Delayed Deep Deterministic Policy Gradient) algorithm with Long Short-Term Memory (LSTM) networks. In this new model, LSTM networks replace the input layers of both the Actor network and the Critic network to effectively predict output results. Experimental results show that the LSTM-TD3 model achieves higher convergence speed and merging success rate compared to the baseline model, and it helps the autonomous firefighting vehicle make optimal merging decisions.

Keywords: Smart Car; Ramp Entry; Deep Reinforcement Learning; Long and Short Term Memory Network

引言

近年来，随着我国社会高速发展，公交已经无法完全满足人民的生活出行需要，因此社会车辆保有量日益增长，这一定程度上增加了车祸的发生概率[1]。同时，社会车辆的不断增多，也容易导致消防车在出警时遇到堵塞等紧急情况，间接降低了救援时效性和成功率。所以推出无人驾驶消防车刻不容缓。自动驾驶领域在过去几年中取得了飞速发展，无论是在学术研究还是在工业发展方面。尽管已经取得了很大进展，但完全自动驾驶尚未上市，只有部分组件能够帮助驾驶员执行自动换道和停车等有限的操作。目前，从匝道并入主路车道这一重要问题尚未完全解决。匝道汇入区域通常具有复杂的道路结构，包括多个车道、交叉口和匝道。研究无人驾驶消防车，使其具备准确理解和处理这些复杂道路情境的能力，能够确保安全导航和路径规划。同时，在匝道汇入区域，可能有交通事故或火灾等紧急情况发生。如果无人驾驶消防车具备快速而可靠的应急响应机制，能够有效地处理紧急情况。

智能车匝道汇入算法最初基于传统学习方法。随着机器学习的逐渐兴起，该方法也逐渐用于智能车匝道汇入场景中。但无论是传统学习方法还是机器学习方法都有一定的局限性，无法准确预测车辆周围环境，不能有效处理突发情况。近年来，随着人工智能和强化学习的兴起与发展，基于深度强化学习的智能车算法不断涌现。乔良[2]、鲍实[3]、Wang H[4] 等人都基于深度强化学习算法 DQN(Deep Q-Network)[5] 建立模型，对比使用传统强化学习，汇入成功率和汇入精度有所提升，但其对复杂环境的适应性较差，难以应用到真实场景。从 DQN 算法衍生出的 DDPG(Deep Deterministic Policy Gradient)[6] 算法，改进了这一缺点，可以更好地解决汇入车流这种车辆连续控制的问题。吴思凡[7] 采用一种基于 DDPG 算法的智能车汇入模型，将汇流问题转化为序列决策问题求解，结果表明该模型的收敛速度、稳定性和成功率都高于基于 DQN 算法的模型，但是仍然存在收敛速度慢、稳定性不高的缺点。Lin[8] 等人使用 DDPG 算法来研究匝

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

道汇入，并通过调整奖励函数的权重以降低碰撞率。尽管如此，他们在实际计算过程中发现 Q 值被高估，这对算法的性能和学习效率产生了负面影响。。Yu Chen[9]、Mingli Zhang[10]、Sina Alighanbari[11] 等人均基于 DDPG 算法建立模型，其中 Yu Chen 通过引入注意力机制优化 DDPG 性能，Mingli Zhang 采用基于 Armijo Goldstein 准则的 BFGS 方法作为值函数网络的训练算法，Sina Alighanbari 使用非线性模型预测控制（NMPC）控制器训练基于 DDPG 算法的智能车辆模型，通过对比发现在成功率等性能上均有所提高。

目前使用深度强化学习研究匝道汇入的算法大多是基于 DDPG 算法，但是 DDPG 算法存在算法稳定性不足、动作价值函数估计过高等缺点，导致汇入策略模型精度不高、收敛困难。而 TD3[12] 改善了 Q 值函数高估的缺点，降低了近似 Q 值的方差，提高了 DDPG 算法的性能；不过，精度提高的同时也增加了策略学习的复杂性，造成算法的收敛困难，鲁棒性差。因此本文提出一种基于 LSTM-TD3 算法的匝道汇入模型，通过与 DPPG、TD3 匝道汇入模型对比，验证本其高效性和有效性。

一、基于 LSTM-TD3 算法的无人驾驶消防车匝道汇入模型

（一）TD3 算法

TD3 算法基于 Actor-Critic 框架，是 Fujimoto S 等人针对 DDPG 算法缺点做出改进后的算法。专门用于处理连续动作空间中的强化学习任务。它是对 DDPG（Deep Deterministic Policy Gradient）算法的改进，提出了多个创新措施来提升算法的稳定性和性能。主要包括：

1. 双重 Critic 网络（Twin Critic Networks）

减少 Q 值的过度估计。TD3 使用两个独立的 Critic 网络（Q1 和 Q2），通过计算两个网络的 Q 值，并取其最小值作为实际的 Q 值，这可以减少由于 Q 值过高估计带来的偏差，从而提高训练的稳定性。

2. 策略延迟更新（Delayed Policy Updates）

减少策略更新对 Q 值估计的影响。TD3 采用延迟更新策略，即 Actor 网络（策略网络）的更新频率低于 Critic 网络。具体来说，Critic 网络每次更新时，Actor 网络只更新一次，通常每次 Critic 更新后，Actor 网络更新次数为 1/2 或 1/4。这种策略减少了策略更新频繁导致的训练不稳定。

3. 目标策略平滑（Target Policy Smoothing）

减少目标 Q 值计算中的方差。TD3 在计算目标 Q 值时，为目标策略网络生成的动作添加噪声。这样可以避免在训练过程中由于目标动作的微小变化导致目标 Q 值的不稳定，从而提高训练过程的稳定性。噪声通常是高斯噪声或均匀噪声。

尽管在许多方面有了改进，但是 TD3 算法仍然存在一些不足：例如 TD3 使用多个网络和额外的噪声引入，增加了计算和存储的开销；TD3 有多个超参数（如延迟更新的步数、噪声的强度等），需要仔细调节以获得最佳性能；尽管 TD3 在中等维度的动作空间中表现优异，但在处理非常高维的动作空间时，训练过程可能变得更加复杂且不稳定等。基于这些缺点本文提出了引入 LSTM 网络，在许多实际应用中，智能体的观测可能是部分可观测的，传统的 TD3 算法通常假设环境的状态是完全可观测的。然而，LSTM 网络能够处理这种部分可观测的问题，因为它们具备记忆能力，可以从序列数据中学习时间依赖性，并从历史信息中提取有用的状态信息；同时在一些复杂环境中，当前的决策可能受到很久以前的事件的影响。LSTM 网络可以捕捉和记住这些长时序依赖。

（二）LSTM-TD3 算法

本文提出的 LSTM-TD3 算法又在 TD3 算法的基础上加入了 LSTM[13] 神经网络，使其能够基于历史信息学习到准确的策略行为，从而提高算法的鲁棒性和智能车的汇入成功率。LSTM 是一种用作处理时序数据的循环神经网络，其结构如图 1 所示。本文用两层 LSTM 神经网络代替了 Actor 网络的输入层和 Critic 网络的输入层，以有效地预测输出结果；同时，为了减少顺序更新对神经网络和强化学习的影响，只对 Actor 网络中当前动作的输出进行顺序更新。

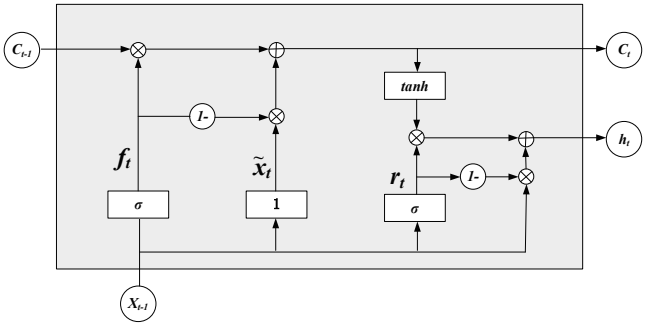


图 1 LSTM 结构示意图

每个 LSTM 单元中包含输入门、输出门和遗忘门，其中遗忘门主要负责对前一单元中不重要的信息进行选择性遗忘处理，如式 (1) 所示。

$$f_t = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right) \quad (1)$$

输入门主要控制输入到当前单元的信息比例，如式 (2) (3) 所示。

$$i_t = \sigma\left(X_t W_{xi} + H_{t-1} W_{hi} + b_i\right) \quad (2)$$

$$\tilde{C}_t = \tan h\left(X_t W_{xc} + H_{t-1} W_{hc} + b_c\right) \quad (3)$$

式 (4) 表示了将上个单元保留的历史信息与本单元输入的信息进行加权处理

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \quad (4)$$

输出门主要控制当前单元有多少信息被输出给下一个单元，

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

如式 (5) (6) 所示。

$$o_t = \sigma\left(X_t W_{xo} + H_{t-1} W_{ho} + b_o\right) \quad (5)$$

$$H_t = o_t \otimes \tan h\left(C_t\right) \quad (6)$$

加入 LSTM 网络后，LSTM-TD3算法的总结构图如图2所示。

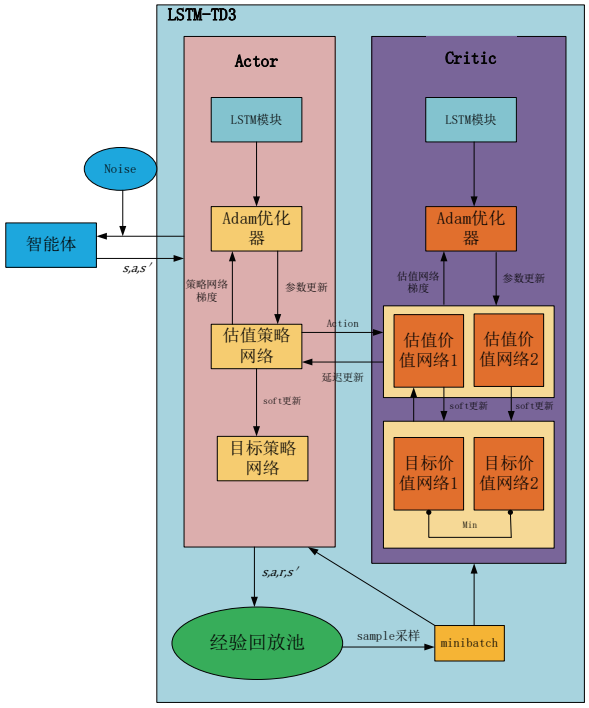


图2 LSTM-TD3算法结构示意图

代理车辆首先从仿真环境中获得传感器信息，包含 (s_t, a_t, s_{t+1})，其中 s_t 表示 t 时刻代理车的状态，a_t 表示 t 时刻代理车所做动作，s_{t+1} 表示 $t+1$ 时刻代理车做完 a_t 后的状态；然后将信息传递LSTM-TD3算法中的策略网络 Actor；之后，Actor 根据代理车辆周围环境信息对代理车辆发出指令，代理车辆在做出相应的动作后得到从环境反馈回来的奖励值，Critic 网络根据奖励函数，并基于环境信息对该动作进行价值评价，输出动作值函数 $Q=(s,a)$；代理车辆通过优化器调整网络参数，以探索最优动作值函数，从而得到最优策略。其中，LSTM 网络用于提取代理车辆信息的特征，TD3框架用于做出控制代理车辆的决策。模型中的 LSTM 网络和 TD3算法是联合训练的。在训练过程中，当最近 epochs 的训练集奖励趋于稳定时，可以认为模型收敛，停止训练。

（三）状态空间

本文涉及的状态空间用一个5*9的 Tensor 表示，包含代理车辆（无人驾驶消防车）的自身信息和周围环境车辆信息等匝道汇入所需要的信息，如式 (7) 所示。

$$S = \left(x_i\right)\left(i = 1,2,3,4,5\right) \quad (7)$$

i 表示车辆编号，代理车辆编号为1，具体状态如图3所示，其中：

$$x_i = \left(V_i, X_{ij}, Y_{ij}\right)\left(i = 1,2,3,4,5; j = 1,2,3,4,5; i \neq j\right) \quad (8)$$

式中 X_{ij} 表示车辆 i 与车辆 j 之间的横向距离，Y_{ij} 表示车辆 i 与车辆 j 之间的纵向距离。

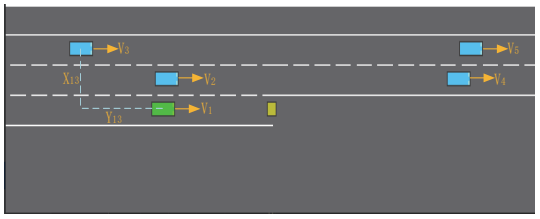


图3 自动驾驶状态描述

（四）动作空间

代理车辆动作空间主要包含五个元素，分别为左转、右转、匀速直行、加速和减速，其中加速和减速主要负责代理车辆的纵向控制，左转、右转和匀速直行主要负责代理车辆的横向控制。考虑到乘客舒适度，设定如下取值范围：

0—左转（固定角度 θ_{left} 为30度）、1—匀速直行（加速度为0，固定角度为0）、2—右转（固定角度 θ_{right} 为30度）、3—加速 $\left(accel\left(m \cdot s^{-2}\right) \in (0,2]\right)$、4—减速 $\left(decel\left(m \cdot s^{-2}\right) \in [-2,0)\right)$。

（五）奖励函数

奖励函数的作用是通过代理车辆与周围环境的交互，获得最大期望的策略，故其设置在强化学习中尤为重要。为使代理车辆快速、安全地汇入到主路，本文结合代理车辆本身的状态空间、动作空间和汇入场景下的目标，将其匀速直行和加速的奖励设置为 $r_{straight} = 0.1$ 和 $r_{accel} = 0.4$，将减速惩罚设置为 $r_{decel} = -0.1$。

待代理车辆安全汇入到主路后，为了避免与主路上的环境车辆相撞，应保持与前后车的安全距离，其计算公式为：

$$S_d = \left|V_{代理} - V_{环境}\right| \times 3.6 + D \quad (9)$$

其中，D 为司机的最小反应距离，取 $D = 5m$。将代理车辆的左转和右转设置为 $r_{change} = 0.1$，将代理车与环境车相撞惩罚设置为 $r_{die} = -10$。当超出最大步数时，给予惩罚数值 $r_{over} = -10$。奖励函数值大于10，认为汇入匝道成功；小于10，则认为汇入匝道失败。

二、仿真实验

本文基于 gym 库，采用 python 语言来搭建仿真环境，神经网络框架为 pytorch，计算机配置为：Windows 10系统，i7-10750 HQ 处理器（主频为 2.6 GHz），8GB 运行内存。最大训练回合为5000回合，在每个回合中，由于高速路上主路最低速度为80km/h，所以将环境车辆的初速度定为22 m/s- 25m/s 内的随机数值；考虑到乘客舒适度的问题，将加速度设置为 -2m/s^2—2m/s^2 内的随机数值，代理车辆的初速度设为12m/s。设置每回合最大限定步数为200步，如超过最大限定步数则视为发生碰撞，模型训练和模型测试均在 Python 搭建的仿真环境中运行。

具体实验过程如下（如图4所示）：利用仿真环境中传感器获得数据作为输入；代理车辆通过执行动作空间中的动作与环境进行交互，从而得到环境给予的奖励反馈；通过奖励值的高低来评判策略的优劣，利用深度强化学习算法更新策略，代理车辆面对新环境时会采取更优的汇入决策。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

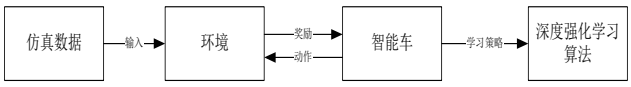


图4 代理车辆与环境交互的过程

（一）参数设定

本文设置最大仿真时长为20s，步长设定为0.1s，深度强化学习训练过程中主要超参数如表1所示

表1 LSTM-TD3算法主要超参数说明

参数名	参数值
LSTM 细胞单元数量 / 个	3
隐藏层层数 / 层	3
折扣系数	1e-4
动作网络学习率	0.01
评判网络学习率	0.01
优化算法类型	Adam
批量大小	m*n2
学习率衰减间隔	0.5
经验池尺寸	100
软更新速率	10
策略频率	2
策略噪声	0.05
动作噪声	0.3
动作噪声衰减步长	0.1

（二）环境搭建

本文基于 gym 库搭建仿真环境，有两条主路和一条匝道，主路最高限速为120km/h，匝道最高限速为45km/h，仿真道路长度约为1000m，起点在距匝道入口约100m 处。仿真环境如图5、图6所示，图中黄色箭头所指方向为车流方向，图5表示代理车辆即将进入匝道，图6表示代理车辆准备从匝道汇入主路。

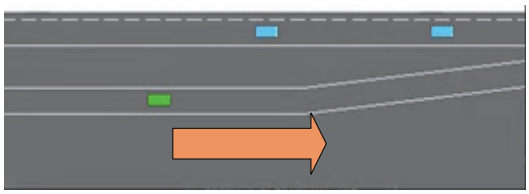


图 5 仿真环境示意图 (1)

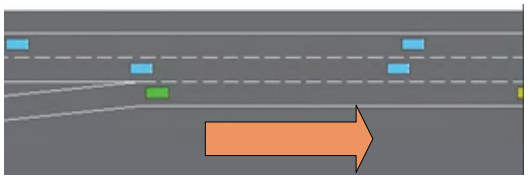


图 6 仿真环境示意图 (2)

代理车辆采用以中国消防救援学院教学使用的重汽8吨水罐消防车，环境车模型以市面上主流小轿车的尺寸为标准， 主要参数如表2所示。

表2 环境参数说明

环境参数 / 单位	参数数值
代理车车身长度 /m	8.2
代理车车身宽度 /m	2.5
环境车车身长度 /m	4.8
环境车车身宽度 /m	1.8
主路最高限速 /km ・h-1	120
匝道最高限速 /km ・h-1	45
仿真道路长度 /m	1000
环境刷新率 /s	0.1
实际道路宽度 /m	3.75

三、仿真结果及分析

（一）模型训练

本文主要目标是控制匝道上的代理车辆，根据车辆周围环境的变化，采用不同的策略进行汇入，并通过记录汇入成功率、训练模型的时间和奖励来描述模型学习的结果（如图7、图8、图9所示）。

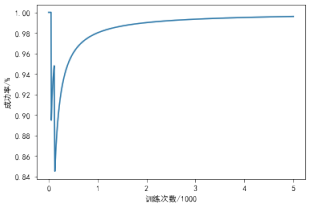


图7 LSTM-TD3模型汇入成功率

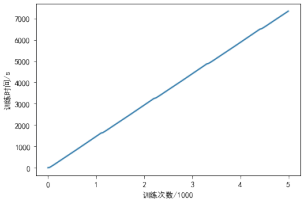


图8 LSTM-TD3模型训练时间

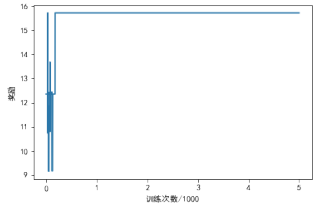


图9 LSTM-TD3模型单回合奖励

由图7可知，由于 LSTM-TD3算法为随机策略，起初汇入偶尔成功，但是随着训练次数的不断增加，成功率有所下跌；但是在大约200回合后，汇入成功率随着训练次数的增多而不断上升，训练结束时可达98.5%。

由图8可知，LSTM-TD3汇入模型的训练次数与训练时间呈线性正相关关系，模型在单位回合内的时间并不会随着训练次数的增加而增加，整个训练时间约为7105s。

由图9可知，LSTM-TD3汇入模型的单回合奖励随着训练次数的增加而逐渐呈现出上升的趋势，在大约200回合后，奖励值趋于稳定，表示模型已收敛。

3.2模型比较

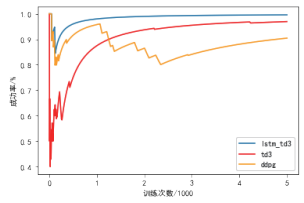


图10 模型汇入成功率对比

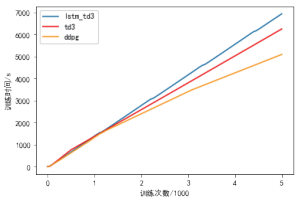


图11 模型训练时间对比

 | Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

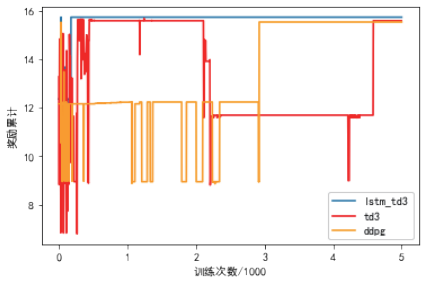


图 12 模型单回合奖励值对比

通过对比 LSTM-TD3、 TD3、DDPG 三种算法所得结果，从验证 LSTM-TD3算法在成功率、算法收敛时间方面的性能。通过在相同的仿真环境下进行仿真实验得到图10、图11、图12和表3，分析可知，在模型收敛速度方面，本文所提出的 LSTM-TD3模型在200回合左右就达到收敛，而 TD3模型和 DDPG 模型分别与1200回合左右和3000回合左右达到收敛；在成功率方面，LSTM-TD3模型在5000回合训练后成功率达到98.5%，而 TD3模型和 DDPG 模型成功率低于 LSTM-TD3模型，分别为93.7%和83.5%；在单回合奖励值方面，训练收敛后，LSTM-TD3算法的单回合奖励值略高于 TD3和 DDPG 模型，而 TD3模型略高于 DDPG 模型，说明经过训练后的 LSTM-TD3模型能够做出最佳汇入决策，且优于 TD3和 DDPG 模型；模型训练时间方面，由于 DDPG 模型网络结构相较于 TD3模型网络结构和本文所提出的 LSTM-TD3网络结构更为简单，所以在一定程度上花费较少的训练时间，但是成功率低和收敛速度较慢，而 LSTM-TD3模型尽管在模型训练时间上稍微高于 TD3模型和 DDPG 模型，但是成功率、收敛速度和决策优劣等方面均优于 TD3模型和 DDPG 模型。

表3 模型对比结果

模型	成功率 /%	训练时间 /s
DDPG	83.5	5160
TD3	93.7	6366
LSTM-TD3	98.5	7105

四、结语

本文提出了一种基于深度强化学习算法的无人驾驶消防车匝道汇入模型。具体地，选取 LSTM-TD3算法作为主要框架，通过分析 TD3模型在匝道合流场景下的不足进行改进。实验结果表明，本文提出的 LSTM-TD3汇入模型能够帮助无人驾驶消防车在面对复杂动态的环境时做出合理的决策，同时，对比 DDPG 模型和 TD3模型汇入成功率分别提升了15% 和4.8%，收敛速度方面也有明显优势。

客观来说，本文研究仍存在一定不足：

一是构建的匝道合流场景仍不能完全满足现实需要，与真实驾驶场景有差异。如在现实世界中，存在雨、雪、雾等恶劣天气，此时传感器的灵敏度和检测范围会受到较大影响，可能造成数据采集不准确。又如真实驾驶场景中车辆轮胎的摩擦也会影响制动距离（车辆行驶安全距离）。

二是只对匝道上的无人驾驶消防车辆进行控制，未涉及到多智能体共同决策。在真实的驾驶场景中，多智能体可以更接近真实场景，而多智能体的共同决策也是一项非常关键的技术。同时，智能网联汽车系统也是当前技术的发展趋势。它考虑了场景中智能车辆、交通信号灯和行人的博弈问题，较好地解决了虚拟与现实的转化问题。

综上，未来的工作将致力于以下几个方面：

1. 使用更加接近于真实世界的仿真软件，在恶劣天气条件下（如雨、雪、雾等）采集驾驶数据，建立包含各种天气条件的测试集。在匝道合流场景中引入天气模型，以模拟不同天气条件对传感器的影响。考虑传感器灵敏度和检测范围的变化，如雨水或雪对摄像头和雷达的干扰；

2. 在匝道合流场景中引入多个智能体（如不同类型的车辆、交通信号灯和行人），并定义它们的行为模型和交互规则。同时引入博弈论模型来解决智能体之间的资源分配、优先级和合作问题。利用强化学习或其他优化算法来优化策略。。

此外，还可以针对匝道合流场景尝试 PPO、A3C 等更多深度强化学习算法，以便更好地帮助自动驾驶消防车做出最佳汇入决策。

参考文献

[1] 刘传贺 . 关于增强汽车驾驶员安全驾驶意识的探讨 [J]. 时代汽车 ,2021(12):197-198.

[2] 乔良，鲍泓，玄祖兴，等 . 基于强化学习的无人驾驶匝道汇入模型 [J]. 计算机工程 ,2018,44(07):20-24+31.

[3] 鲍实，方勇 . 基于深度增强学习的无人车汇流决策方法 [J]. 电子测量技术 ,2018,41(19):44-49.

[4]Wang H，Yuan S，Guo M，et al. A deep reinforcement learning-based approach for autonomous driving in highway on-ramp merge[J]. Proceedings of the Institution of Mechanical Engineers Part D Journal of Automobile Engineering，2021，235(6):095440702199948.

[5]Mnih V，Kavukcuoglu K，Silver D，et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602，2013.

[6]Lillicrap T P，Hunt J J，Pritzel A，et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971，2015.

[7] 吴思凡，杜煜，徐世杰，等 . 基于深度确定性策略梯度的智能车汇流模型 [J]. 计算机工程，2020,46(01):87-92.

[8]Lin Y，McPhee J，Azad N L. Anti-jerk on-ramp merging using deep reinforcement learning[C]//2020 IEEE Intelligent Vehicles Symposium (IV). IEEE，2020：7-14.

[9]Chen Y，Dong C，Palanisamy P，et al. Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019：0-0.

[10]Zhang M，Zhang Y，Gao Z，et al. An improved DDPG and its application based onthe double-layer BP neural network[J]. IEEE Access，2020，8：177734-177744.

[11]Alighanbari S，Azad N L. Deep Reinforcement Learning with NMPC Assistance Nash Switching for Urban Autonomous Driving[J]. IEEE Transactions on Intelligent Vehicles，doi：10.1109/TIV.2022.3167616.

[12]Fujimoto S，Hoof H，Meger D. Addressing function approximation error in actor-critic methods[C]//International conference on machine learning. PMLR，2018：1587-1596.

[13]Hochreiter S，Schmidhuber J. Long short-term memory[J]. Neural computation，1997，9(8)：1735-1780.

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License