第 9 卷 第 5 期
2020 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 9 No. 5
Sep. 2020

引文格式：
吕迪, 徐坤, 李慧云, 等. 融合类人驾驶行为的无人驾驶深度强化学习方法 [J]. 集成技术, 2020, 9(5): 34-47.
Lv D, Xu K, Li HY, et al. Human-like driving strategy based on deep reinforcement learning for autonomous vehicles [J]. Journal of Integration Technology, 2020, 9(5): 34-47.

融合类人驾驶行为的无人驾驶深度强化学习方法

吕 迪[1,2,3] 徐 坤[1,2] 李慧云[1,2] 潘仲鸣[1,2]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](深圳市电动汽车动力平台与安全技术重点实验室 深圳 518055)
[3](哈尔滨理工大学 哈尔滨 150000)

摘 要 现有无人车辆的驾驶策略过于依赖感知-控制映射过程的“正确性”，而忽视了人类驾驶汽车时所遵循的驾驶逻辑。该研究基于深度确定性策略梯度算法，提出了一种具备类人驾驶行为的端到端无人驾驶控制策略。通过施加规则约束对智能体连续行为的影响，建立了能够输出符合类人驾驶连续有序行为的类人驾驶端到端控制网络，对策略输出采用了后验反馈方式，降低了控制策略的危险行为输出率。针对训练过程中出现的稀疏灾难性事件，提出了一种更符合控制策略优化期望的连续奖励函数，提高了算法训练的稳定性。不同仿真环境下的实验结果表明，改进后的奖励塑造方式在评价稀疏灾难性事件时，对目标函数优化期望的近似程度提高了 85.57%，训练效率比传统深度确定性策略梯度算法提高了 21%，任务成功率提高了 19%，任务执行效率提高了 15.45%，验证了该方法在控制效率和平顺性方面具备明显优势，显著减少了碰撞事故。

关键词 深度强化学习；端到端控制；无人驾驶；类人驾驶；奖励塑造
中图分类号 TG 181 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20200515001

Human-Like Driving Strategy Based on Deep Reinforcement Learning for Autonomous Vehicles

LV Di[1,2,3] XU Kun[1,2] LI Huiyun[1,2] PAN Zhongming[1,2]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*Shenzhen Key Laboratory of Electric Vehicle Power Platform and Safety Technology*, *Shenzhen* 518055, *China*)
[3](*Harbin University of Science and Technology*, *Harbin* 150000, *China*)

Abstract The driving decisions of human drivers have the social intelligence to handle complex conditions in addition to the driving correctness. However, the existing autonomous driving strategies mainly focus on

收稿日期：2020-05-15 **修回日期**：2020-06-04
基金项目：国家重点研发计划项目(2016YFD0700602)；国家自然科学基金项目(61603377)
作者简介：吕迪，硕士研究生，研究方向为无人地面车辆智能控制；徐坤(通讯作者)，博士，高级工程师，研究方向为智能控制，E-mail：kun.xu@siat.ac.cn；李慧云，博士，研究员，研究方向为智能驾驶；潘仲鸣，硕士，工程师，研究方向为智能系统。

the correctness of the perception-control mapping, which deviates from the driving logic that human drivers follow. To solve this problem, this paper proposes a human-like autonomous driving strategy in an end-to-end control framework based on deep deterministic policy gradient (DDPG). By applying rule constraints to the continuous behavior of the agents, an unmanned end-to-end control strategy was established. This strategy can output continuous and reasonable driving behavior that is consistent with the human driving logic. To enhance the driving safety of the end-to-end decision-making scheme, it utilizes the posterior feedback of the policy output to reduce the output rate of dangerous behaviors. To deal with the catastrophic events in the training process, a continuous reward function is proposed to improve the stability of the training algorithm. The results validated in different simulation environments showed that, the proposed human-like autonomous driving strategy has better control performance than the traditional DDPG algorithm. And the improved reward shaping method is more in line with the control strategy to model the catastrophic events of sparse rewards. The optimization expectation of the objective function can be increased by 85.57%. The human-like DDPG autonomous driving strategy proposed in this paper improves the training efficiency of the traditional DDPG algorithm by 21%, the task success rate by 19%, and the task execution efficiency by 15.45%, which significantly reduces collision accidents.

Keywords　deep reinforcement learning; end-to-end control; autonomous driving; human-like driving; reward shaping

1　引　言

在人工智能所面临的诸多任务中，无人驾驶是一个极具挑战的场景。与图像处理、自然语言理解等应用场景相比，无人驾驶必须能应对车-路(环境)交互频繁、场景复杂多变、实时性要求高、容错率低等挑战[1]。

近年来，学术界提出了基于深度强化学习(Deep Reinforcement Learning)方法的端到端无人驾驶策略，将具有感知抽象特征能力的深度学习与能实现自适应决策的强化学习相结合。通过模拟人类认知映射行为，实现从感知输入到控制输出的端到端功能。深度 Q 网络(Deep Q Network)[2]、深度确定性策略梯度(Deep Deterministic Policy Gradient，DDPG)[3]、递归确定性策略梯度(Recurrent Deterministic Policy Gradients)[4]等深度强化学习方法逐渐被尝试应用到无人车辆的决策控制任务中。

端到端的无人驾驶控制思想最早起源于 1989 年，卡耐基梅隆大学首先使用名为 ALVINN 的反馈神经网络进行了端到端的车辆运动规划[5]。ALVINN 使用了一个 3 层反馈神经网络，以摄像头采集的信息作为输入，直接决策出车辆的转向角。2005 年，Lekhachev 等[6]提出了一个端到端的神经网络，进行车辆避障所需的轨迹规划，使用一个 6 层的卷积神经网络，以左右摄像头的信息作为输入，直接输出车辆的转向命令。2017 年，中国科学院深圳先进技术研究院夏伟与李慧云[7]提出了一种基于深度强化学习的自动驾驶策略学习方法。该方法采用在线交互式学习方法对深度网络模型进行训练，通过对状态空间进行聚类再采样，提高了算法的训练效

率。2018 年，伯克利学院的 Chowdhuri 等[8]学者提出了 MultiNet 的多模态网络架构，以解决直道、弯道、超车、停车等不同模式下的多参数设置问题，使端到端的规划方法更具鲁棒性。上述研究揭示了端到端控制在实现无人驾驶复杂决策控制方面的潜力。

理想的无人驾驶策略的输出应该由一系列符合类人(Human-Like)驾驶逻辑的连续有序行为组成。这些连续有序驾驶行为除了具备确保车辆行驶安全的“正确性”外，还需具有一定的社会智能属性[9]。然而，现有的无人驾驶策略过于强调感知-控制映射过程[10]的“正确性”，往往忽视了人类驾驶汽车时兼具一定社会智能的驾驶逻辑[9]，这可能会导致严重的安全事故。例如，谷歌无人驾驶汽车曾在 2016 年发生与市政公交车的碰撞事故，起因是无人驾驶策略错误判断实际没有让行意图的公交车会让行，这一判断明显不符合人类驾驶汽车时的驾驶逻辑，没有实现人类驾驶员作决策时的社会智能[11]。即当前的无人驾驶策略，并未考虑人类社会在长期驾驶经验中所积累的驾驶逻辑与决策智能。

针对无人驾驶策略过于注重感知-控制映射过程的“正确性”而忽略“社会智能”[12]属性的问题，有学者提出采用将端到端驾驶决策任务中的“决策-驾驶”任务分开处理的方式，来降低输出驾驶行为的不合理性。通过对无人驾驶策略在某些重要时刻的逻辑判断进行改进，使无人驾驶汽车在一些复杂的情况下具备类人决策和判断能力[9-11]。Codevilla 等[10]引入条件反射概念，将“决策-驾驶”任务分开处理，以人类驾驶经验对无人驾驶的决策输出进行先验优化。但该方法改进的仍然是无人车对关键驾驶行为的决策规划问题，输出依旧是对驾驶过程中某些重大决策事件做出的离散概率。Li 等[11]注意到现行无人驾驶决策算法中决策思维与人类思维不符合(AI 决策思维非类人化)的问题，并提出了一个类人驾驶系统，通过先验建立的车辆决策规则使无人驾驶汽车的决策思维更贴合人类思维。虽然该方法建立了类人化的先验数据集与仿真环境，并在训练过程中对算法的策略逻辑进行了前验约束。但是，这一研究改进的方面主要是在变道、超车等行为发生时，智能车所做出的决策判断。因此其输出的是离散的选取动作概率值，并未能在整个无人驾驶任务中形成连续的合理动作序列。相较于基于强化学习的端到端控制，基于规则的无人驾驶策略拥有更符合类人逻辑的驾驶规则。Montemerlo 等[13]对车辆行为进行细分，建立了一个拥有 13 个状态的有限状态机组成的决策系统。但是基于规则的无人驾驶决策方式更侧重于实现功能，而不是实现高驾驶性能[14]。而无人驾驶汽车数据来源的不确定性，使依赖精准环境判断且基于规则的无人驾驶策略无法拥有足以应对真实路面环境的决策鲁棒性[15]。

无人驾驶输出的连续策略应符合人类驾驶汽车时的驾驶逻辑，即无人驾驶策略应具备“类人”驾驶行为的逻辑。现实道路中的车辆驾驶行为是一个连续的过程，因此无人驾驶策略的输出应该由一系列符合类人逻辑的连续有序行为组成。此外，策略网络输出的应该是贯穿整个驾驶任务的连续规则，而并非只是在需要做出某些重大判断时的离散概率[16]。因此，在保证驾驶任务顺利完成的同时，连续的、类人化的控制规则对正确的驾驶决策至关重要。

为提高车辆在真实道路条件下的无人驾驶性能，本文主要针对现有算法缺乏类人社会智能的不足，结合深度确定性策略梯度算法，提出一种具备一定类人驾驶能力的无人驾驶策略。本研究的主要贡献是：(1)在端到端无人驾驶控制算法中引入了基于类人逻辑的规则约束，建立了能够输出符合类人逻辑且具有连续有序行为的无人驾驶端到端控制网络。同时通过基于环境与规则对网络策略的多维度后验反馈，成功降低了网络的

危险行为输出率。(2)通过将驾驶过程中出现的稀疏的、灾难性的离散事件视为一个有状态的连续过程，同时建立了在时序上连续的奖惩机制，从而避免了其产生的策略过拟合问题，并加速了训练策略向目标函数的收敛。

2　端到端控制的“感知–控制”映射问题

2.1　“感知–控制”映射模糊性

强化学习本质上是一个序贯决策问题[17]，即智能体(Agent)如何根据当前可观测到的状态(State)选择一个动作(Action)，使获得的累积回报(Reward)最大，并将状态映射为动作的函数即为策略(π)。目前主流的端到端控制算法遵循如图 1 所示的感知-控制的逻辑映射过程，其中控制器接收来自环境的观测值 o_t 和命令 c_t，并接收环境对当前动作的反馈信息进入下一步。控制策略的输出，取决于智能体在此刻对于环境的观测。这一状态可描述为典型的马尔可夫决策过程(Markov Decision Process)[18]。

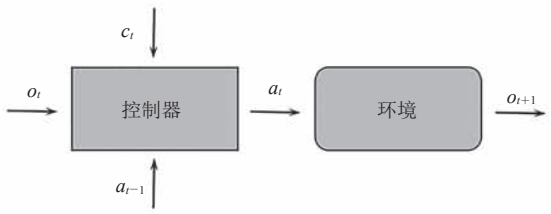


图 1　“感知-控制”映射过程概述

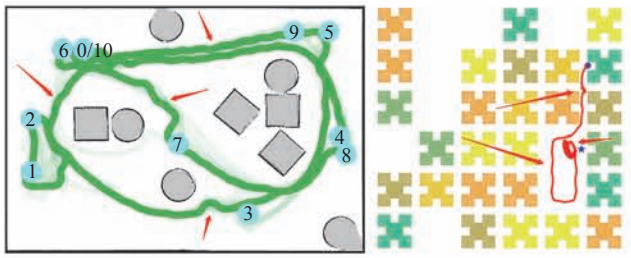
Fig. 1　Overview of the “perception-control” mapping process

在真实道路场景下，感知-控制的逻辑映射过程往往会具有模糊性，其主要原因是完成驾驶任务所需的正确决策往往无法仅凭感知输入单独推断获取。在这种情况下，从信息输入到控制命令输出的映射不再是一个明确的映射函数。因此，算法策略梯度的拟合函数逼近必然会遇到困难。由于训练者无法直接控制网络决策行为判断的方式，即使现有算法的训练网络可以解决部分映射中的模糊性问题，但仍无法推断出最优解。这种感知-控制映射过程中的模糊性，会造成控制网络输出不符合类人逻辑的危险动作。此时，应用马尔可夫决策过程的前提条件不再成立。

感知-控制映射过程中的模糊性也无法通过拓展端到端控制算法至部分可观察马尔可夫决策过程的方式[4]完全解决。规则空间中的类人逻辑是人类社会经过长期驾驶实践获取的经验法则，其不存在于无人驾驶系统此刻或任何之前时刻的观测中，也无法通过车辆对自身驾驶行为的观测统计得出。因此，通过记录时序上的历史状态也无法真正解决感知-控制映射过程中的这一模糊性。当前无人驾驶策略的输出仅依赖感知信息，由于信息输入的不完整，导致感知-控制的映射不再是一个明确的映射函数。而“感知-控制”映射具有模糊性，会使无人驾驶策略缺乏人类驾驶时所遵循的驾驶逻辑与社会智能。因此，如何实现无人驾驶的类人逻辑是无人驾驶端到端控制亟待解决的问题。

2.2　现有策略的不合理行为

在现有的无人驾驶控制算法中，输出策略不符合类人逻辑的情况普遍存在(图 2)。这种行为影响行车的安全性与舒适性，使无人车辆的驾驶行为像是“醉驾”。



(a)DDPG 算法[19](移动机器人导航)　(b)RDPG 算法[20](无人机导航)

图 2　当前端到端控制算法中存在的策略输出不合理问题

Fig. 2　The problem of unreasonable policy output in the current end-to-end control algorithm

在无障碍直线场景下的巡航任务是体现无

人驾驶策略是否拥有类人驾驶逻辑的典型验证场景。借助 Carla 无人驾驶仿真器，本文构造了一条如图 3 所示的直道，使用 DDPG 无人驾驶算法[10]控制无人车执行巡航任务。图 4 为无人车执行巡航任务时的运动轨迹与控制过程。

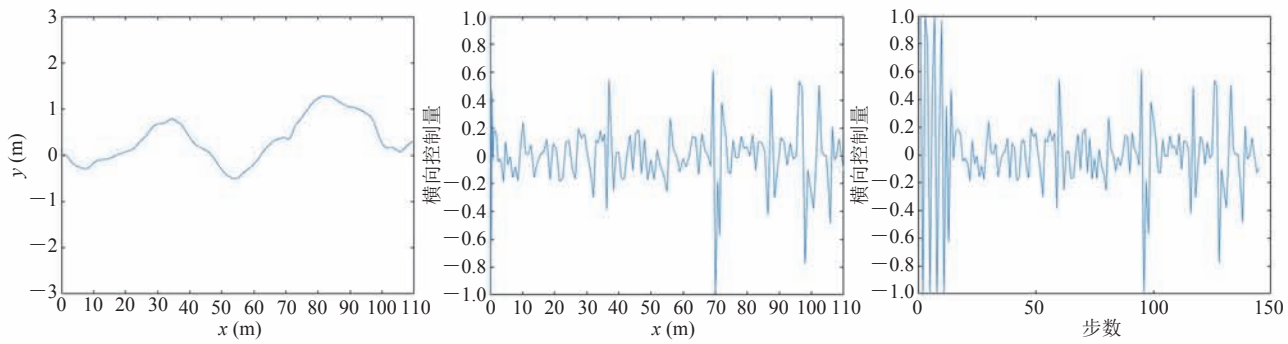
图 3 用以执行巡航任务的直线场景

Fig. 3 Straight line scene used to perform cruise tasks

从图 4 可看出，传统的 DDPG 算法无法输出连续合理的、具备类人逻辑的驾驶策略。但由于无人车执行任务过程中并没有出现碰撞、越出车道、压线等异常表征行为。即使无人车在一条平直的道路上表现出了左摇右摆的“醉驾”行为，却仍然获得了较高的奖励回报(图 5)。因此，如何实现无人驾驶任务中“贯穿整个驾驶过程且符合类人逻辑”的控制规则，使无人驾驶控制算法输出“符合类人逻辑的连续有序行为”，也是无人驾驶端到端控制领域亟待解决的问题。

2.3 对稀疏的灾难性事件的奖励塑造

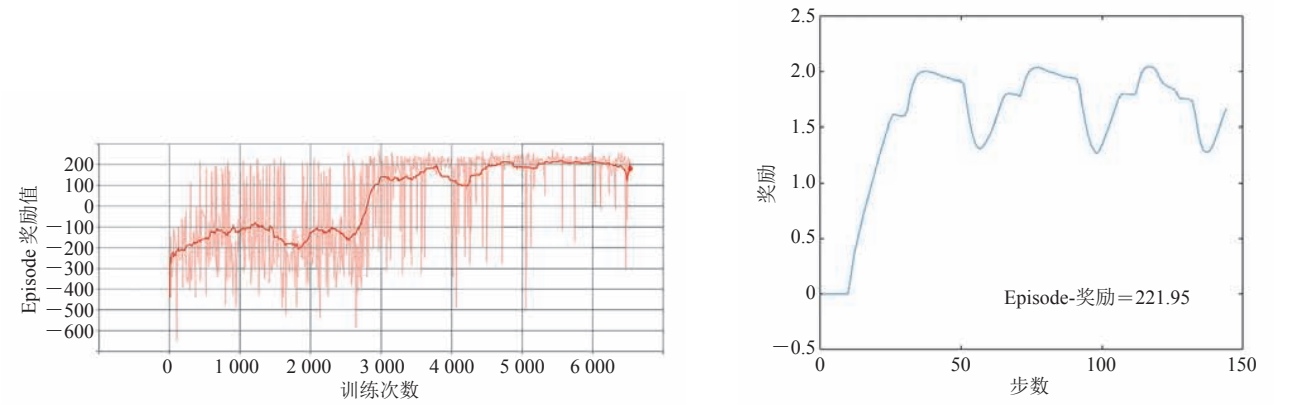
基于深度强化学习的无人驾驶端到端控制是带有奖励函数的智能体与环境的交互过程。奖励塑造(Reward Shaping)[21]方法为无人车驾驶策略的学习提供了确定性的解决方案。无人驾驶的驾驶场景中，危险性最高的意外情况是碰撞，同时碰



(a) 无人车的运动轨迹 (b) 无人车在不同运动位置的横向控制量输出情况 (c) 无人车在时序的横向控制量输出情况

图 4 使用现有 DDPG 算法在图 3 所示场景中进行无人车巡航任务的轨迹

Fig. 4 Results of the cruise task using existing DDPG algorithm



(a) DDPG 算法[8]在训练过程中的收敛情况 (b) DDPG 算法在图 3 所示场景中进行巡航任务的奖励回报

图 5 巡航实验获得的奖励回报

Fig. 5 Rewards from cruise experiments

撞事故也是失稳、甩尾、侧翻等所有意外工况的最终表征，且几乎每次碰撞事故都会带来不可估计的严重后果。

本研究对现有的算法进行调研发现，在当前的训练环境中，现有研究通常将碰撞事故设置为稀疏的瞬发行为，对于碰撞惩罚的构建都是在碰撞时刻所给出的一个较大的、稀疏的确定惩罚。但是这种稀疏的灾难性事件对策略网络更新的回报方式只局限于优化期望的控制目标，没有对于风险的具体建模或优化[22]。与驾驶行为一样，在现实世界中，碰撞事故的产生同样是一个带有状态的过程。车辆与碰撞对象之间的相互接近也是一个渐进而非瞬时的过程，危险存在于整个不符合逻辑的驾驶过程之中，而不是仅仅存在于产生碰撞的时刻。

现有端到端控制算法的研究尚缺乏能够有效处理灾难事件的概率(通常包括大的、负的回报函数)方法。这些分布在 Episode 中的稀疏、离散的奖惩设置无法使用简单的二次函数逼近器进行拟合。除了容易引发训练策略的过拟合外，还会引起策略向目标函数逼近方向的偏离，进而影响算法的训练效率。因此，针对这些稀疏的、灾难性的事件，探寻能够更贴合对目标函数优化期望，并能够结合驾驶任务更好地被智能体拟合并理解的奖励塑造方式，对无人驾驶端到端控制算法的训练效率与策略效果非常重要。

3　类人确定性策略梯度无人驾驶策略

3.1　类人深度确定性策略梯度的无人驾驶决策方法

在分析无人驾驶中的模糊性、不合理行为、以及稀疏灾难性事件奖励塑造的基础上，本文进一步提出了融合类人驾驶行为的无人驾驶深度强化学习方法(HL-DDPG)。算法整体流程如表 1 所示，该方法的核心包括类人逻辑驾驶约束和奖励塑造。

表 1　HL-DDPG 算法流程

Table 1　HL-DDPG algorithm

HL-DDPG algorithm
采用权重 θ^{Q} 和 θ^{μ} 随机初始化 Critic 网络 $Q(s,a\mid\theta^{Q})$ 和 Actor 网络 $\mu(s\mid\theta^{\mu})$
分别采用权重 $\theta^{Q'}\leftarrow\theta^{Q}$ 初始化目标网络 Q'，权重 $\theta^{\mu'}\leftarrow\theta^{\mu}$ 初始化目标网络 μ'
先验建立基于类人逻辑的规则约束
初始化回放缓冲区 R
全局训练开始，记初始 Episode＝1
在类人逻辑区间内初始化随机过程以进行动作探索
初始化智能体时序连续行为数据集 h_a
接收初始观察状态 s_1
开始探索 Episode，记初始动作序列 t＝1
综合当前策略、探索噪音与时序上的连续动作逻辑，在当前时刻的规范动作集 $a\in\mathrm{Std}_a$ 中选择动作
执行动作 a，观察奖励 r，并观察新状态 s_{t+1}
记录当前时刻动作 a 至时序连续行为数据集 h_a 中
记录 (s_t,a_t,r_t,s_{t+1}) 至初始化回放缓冲区 R 中
从 R 中抽取 N 个过渡点 (s_i,a_i,r_i,s_{i+1}) 的随机 Mini-Batch
设置 $y_i=r_i+\gamma Q'[s_{i+1},\mu'(s_{i+1}\mid\theta^{n})\mid\theta^{Q}]$
通过最小化损失来更新评论家： $L=\frac{1}{N}\sum_{i}\left[y_i-Q(s_i,a_i\mid\theta^{Q})\right]^2$
更新策略： $\nabla\eta(\mu)\approx\mathbb{E}_{s\sim d_{\mu}(s)}\left[\nabla_{\theta}\mu(s)Q_{\mu}(h_a,a)\mid_{a=\mu(h_a)}\right]$
探索结束
全局训练结束

3.2　符合类人逻辑的驾驶约束

针对目前无人驾驶端到端控制算法的连续策略缺乏时序上的逻辑关系问题，建立了符合类人逻辑的驾驶约束。同时，将无人驾驶策略所依赖的“感知-控制”映射过程拓展为如图 6 所示的“感知＋驾驶逻辑-控制”映射。控制器接收来自环境的观测值 o_t 和命令 c_t，同时接收自身在之前时序上已产生的行为状态 a_t。在考虑驾驶逻辑后，输出符合类人逻辑的有序行为，并接收环境对当前动作的反馈信息进入下一步。

在原始 DDPG 算法的基础上，HL-DDPG 算法在无人驾驶策略中添加了类人的驾驶逻辑约束，然后基于先验类人驾驶经验建立了类人驾驶

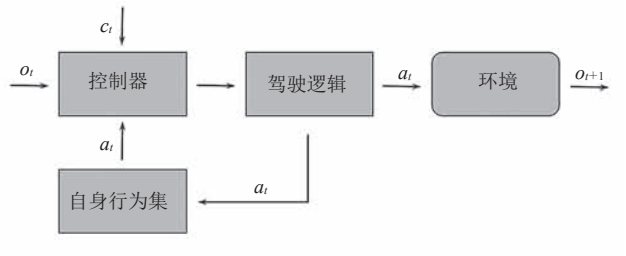


图 6 “感知＋驾驶逻辑-控制”映射过程概述

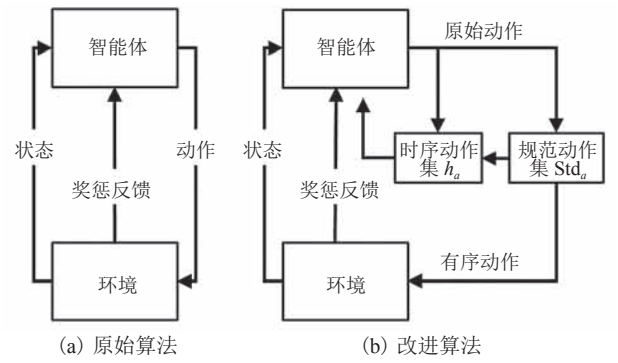
Fig. 6 Overview of “perception ＋ driving regulation-control” mapping process

规则。除原有的感知-控制映射机制外，约束无人驾驶策略的规范动作空间 $a \in \mathrm{Std}_a = [a_1, a_2, \cdots, a_n]$，约束智能体在符合规范动作空间的类人逻辑区间内初始化随机过程以进行动作探索。定义包含智能体历史时序上的连续行为数据集 $h_a = [a_1, a_2, \cdots, a_t]$，当智能体完成当前时序动作后，先将已完成动作记录至时序连续行为数据集 h_a 中，然后再进行下一步动作。本文将智能体在时序上输出的连续动作行为的逻辑性也纳入控制输出的考量范围，通过先验知识对策略网络建立基于类人逻辑的规则约束，以智能体自身已产生的行为状态为基准，对算法网络输出的策略进行后验约束。随后，将约束结果塑造为奖惩函数反馈给训练网络，并对网络的策略输出进行改善，以确保无人车驾驶行为是一个连续有逻辑的过程。图 7 为改进前后智能体与环境的交互模式示意图。这种改进将无人车的驾驶行为视为一个连续的、有状态的过程，使用符合类人逻辑的规则对策略输出进行约束，并构造了新的奖惩机制。

原始 DDPG 算法未将策略输出的连续动作视为有状态的连续过程，也并未将智能体多步连续动作间自身的合理性纳入策略更新的考虑范围。针对这一缺陷，拓展 HL-DDPG 的策略更新方式为：

$$\nabla \eta(\mu) \approx \mathbb{E}_{s \sim d_\mu(s)} \left[\nabla_\theta \mu(s) Q_\mu(h_a, a)|_{a=\mu(h_a)} \right] \quad (1)$$

其中，h_a 为包含智能体历史时序上的连续行为数据集，$h_a = [a_1, a_2, \cdots, a_t]$。



(a) 原始算法 (b) 改进算法

图 7 两种智能体与环境的交互模式

Fig. 7 Two modes of the interaction between agent and environment

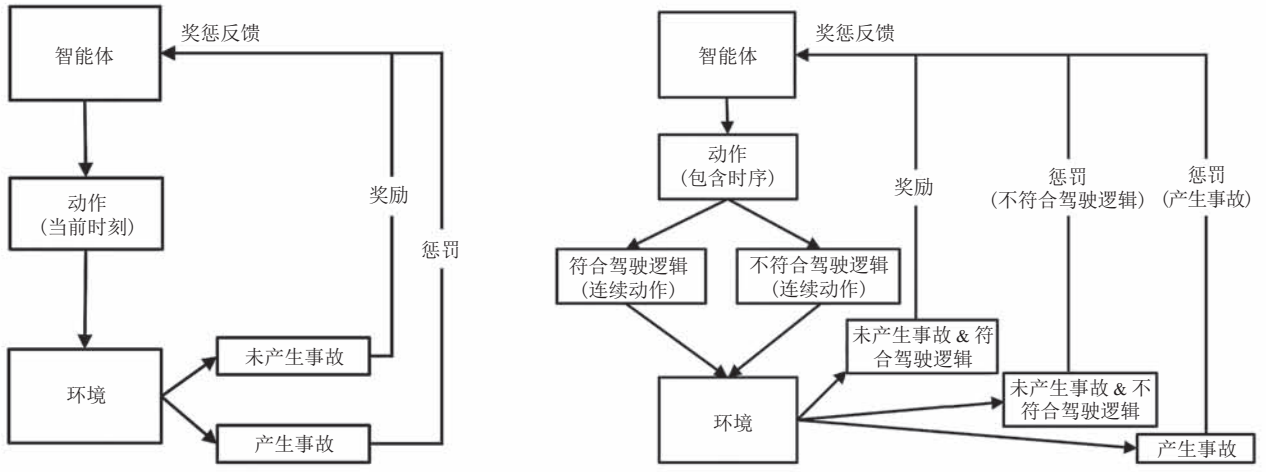
3.3 对稀疏的灾难性事件的奖励塑造

现有算法对稀疏的灾难性事件的奖励塑造难以符合对目标函数的优化期望。本文将无人车在驾驶任务中产生的一些稀疏灾难性事件构建为一系列具有状态函数的连续过程，以相对平滑且连续的奖励塑造替代原有环境设置中大的稀疏离散惩罚。对于未表征出灾难性事件的潜在危险驾驶行为，添加了无人车自身的行为规范评价，以对未产生环境反馈的潜在危险驾驶行为进行惩罚。改进算法与原始算法对比如图 8 所示。其中，图 8(a) 为原始评价方式，以对环境反馈的结果为基准，对智能体的行为构造一些大的、稀疏的奖惩条件；图 8(b) 为改进后的奖惩评价方式，将无人车的驾驶行为视为一个连续的、有状态的过程。改进方法通过使用驾驶逻辑去惩罚有安全隐患的危险行为，将灾难事件的概率构造成连续的奖惩函数，在碰撞风险产生前就对策略输出进行连续反馈。

4 实验结果

4.1 仿真实验设置

本文考虑到无人车在平直路面上的驾驶场景，因此无人车在行驶时的活动区域仅限于 *x*-*y* 平面，不会产生纵向偏移，故可将无人车状态简



(a) 原始评价方式　　(b) 改进后的奖惩评价方式

图 8　两种算法对驾驶行为的评价方法

Fig. 8　Two evaluation methods of driving behavior

化描述为[x, y, v, ω]。其中，(x, y)为无人车在 x-y 平面的位置坐标，v 为当前时刻无人车的速度，ω 为当前时刻无人车的横摆角速度。

本研究使用分支结构对改进算法进行训练，分别对两种基于 DDPG 端到端控制方法进行了对比，具体如图 9 所示。图 9(a)为原始 DDPG 算法；图 9(b)为 HL-DDPG 算法，HL-DDPG 算法在原始 DDPG 算法的基础上，使用驾驶逻辑对策略输出量进行了约束。在分支结构中，感知图像、自身状态、训练任务三个输入由 3 个模块独立处理，分别为感知模块 $I(i)$、状态模块 $M(m)$ 和任务模块 $T(t)$。感知模块由卷积网络实现，状态模块和任务模块为全连接网络。所有模块的输出联合表示为：

$$j=J\left(i,m,t\right)=\left\langle I\left(i\right),M\left(m\right),T\left(t\right)\right\rangle \tag{2}$$

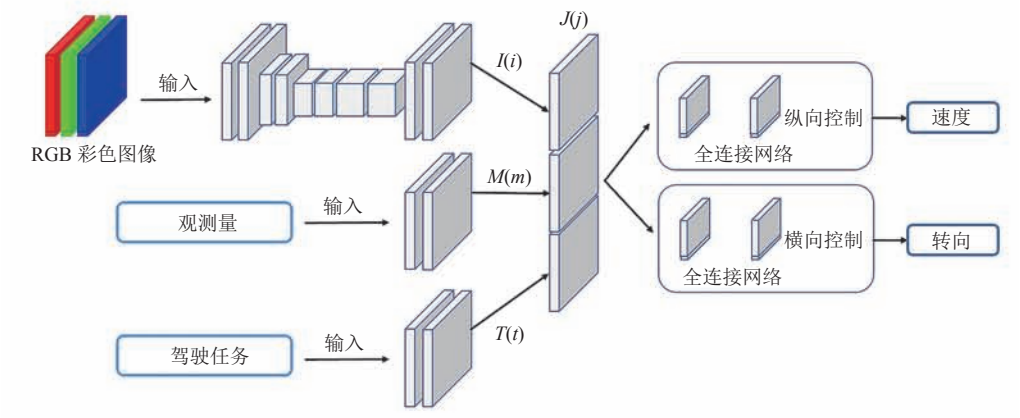
其中，m 为汽车的当前状态量；i 为输入的图象数据；t 为当前的驾驶任务。所有网络都由具有相同架构的模块组成[23]，不同之处在于模块和分支的配置。如图 9 所示，感知模块由 8 个卷积和 2 个全连接层组成。其中，第一层卷积核大小为 5，第二层为 3，第 1、3、5 个卷积层的步长为 2。通道的数量从首层 32 个递增至末层 256 个，在卷积层之后进行归一化处理。各全连接层包含 512 个单元。除感知模块外，状态模块和任务模块都实现为标准的多层感知器。本研究使用对所有的隐藏层进行 ReLU 非线性处理，其中在全连接层后设置下采样(Dropout)为 50%，在卷积层后设置下采样为 20%。

无人车辆输出的动作是一组二维矢量：[纵向控制量，横向控制量]，给定一个预测动作 a 和一个真实动作 a_{gt}，每个样本的损失函数定义为：

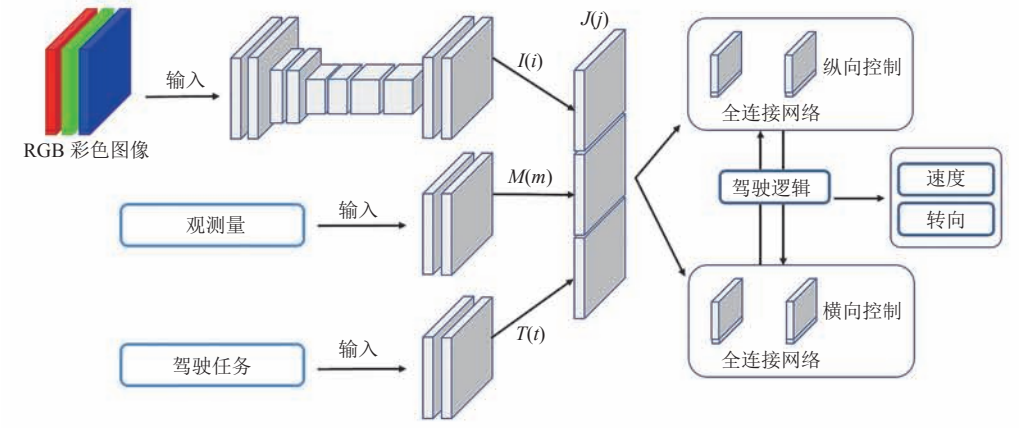
$$\begin{aligned}\ell\left(a,a_{gt}\right)&=\ell\left\langle\left(s,a\right),\left(s_{gt},a_{gt}\right)\right\rangle\\&=\left\|s-s_{gt}\right\|^2+\lambda_a\left\|a-a_{gt}\right\|^2\end{aligned} \tag{3}$$

4.2　改进前后算法控制输出对比

为验证本研究使用类人逻辑的规则对现有策略不合理行为的改进效果，对改进后算法的效果进行了实验验证，并与原始算法的实验效果进行了对比。图 10 为改进前后的算法执行巡航任务时的运动轨迹与控制过程。平顺的驾驶过程有利于行驶安全性与舒适性，并避免额外的能量损失，提高任务执行效率[24]。从图 10 数据曲线可直观看出，具备类人逻辑的控制算法的控制曲线明显比原始控制算法[10]更为平顺。这表明所提出的算法在



(a) 原始 DDPG 算法



(b) HL-DDPG 算法

图 9　本文进行对比的两种端到端控制方法

Fig. 9　Two end-to-end control methods for comparison

巡航任务中展现出了更合理的控制过程，并对任务表现出了更高的执行效率。

为更好地对实验结果进行量化展示，本研究统计了两种算法控制过程中横向控制量的方差与标准差，并对比了两种算法完成任务所需的控制步数，结果如表 2 所示。

从表 2 可知，在对相同实验任务的执行过程中，改进后的算法比原始算法拥有更为平顺的控制过程。具体体现为改进后算法的横向控制输出量的方差与标准差都远低于改进前的算法。此外，本研究所改进的算法完成相同任务所需时间为改进前算法的 93.1%，拥有更好的任务执行效率。

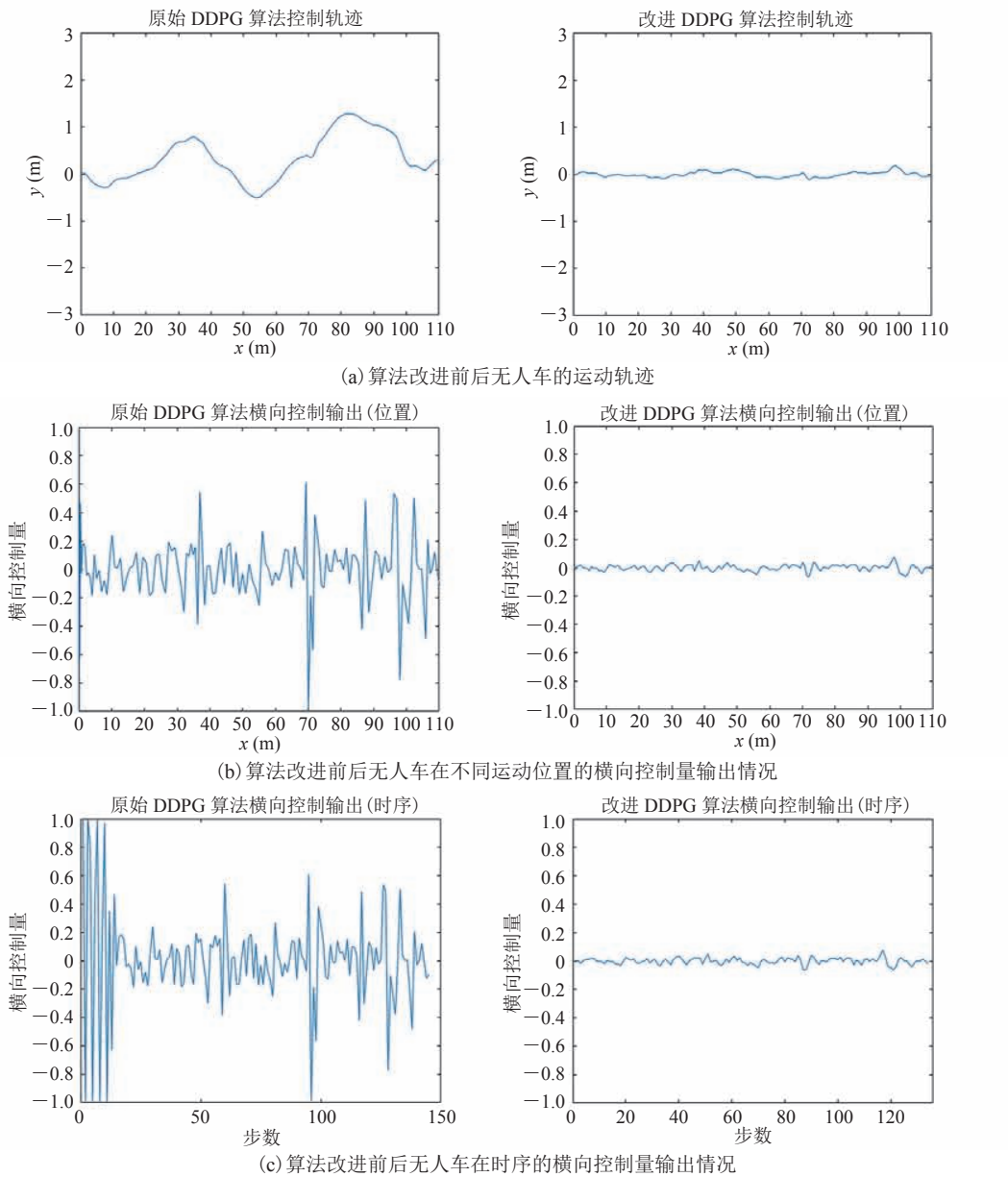
表 2　两种算法直线巡航实验的控制效果对比

Table 2　Comparison of the control performance of the two algorithms in the straight line cruise task

算法	横向控制量		所需时间
	方差	标准差	（比例）
原始算法	0.109 7	0.331 2	1
改进算法	0.000 5	0.022 7	0.931 0

为说明智能体的连续行为不合理的现象不是个例，本研究引入另一个仿真环境 Gazebo 对研究结果进一步说明，相应实验仍然基于已经被广泛验证过的 DDPG 算法[19]。在 Gazebo 中取消了道路约束，搭建了一个如图 11 所示的开放空间的移动机器人导航场景。

研究分别使用原始控制算法[19]与考虑类人驾驶逻辑的改进控制算法进行仿真实验的验证。实验为两个算法在完全相同的实验环境下设置了相



(a) 算法改进前后无人车的运动轨迹

(b) 算法改进前后无人车在不同运动位置的横向控制量输出情况

(c) 算法改进前后无人车在时序的横向控制量输出情况

图 10　改进后算法与原始算法[9]在图 3 所示场景中进行巡航任务的效果对比

Fig. 10　Comparison of the improved algorithm and the original algorithm for the cruise task in the scenario shown in figure 3

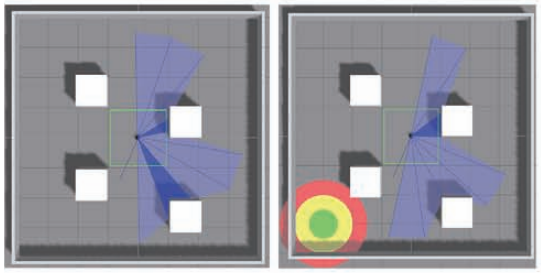
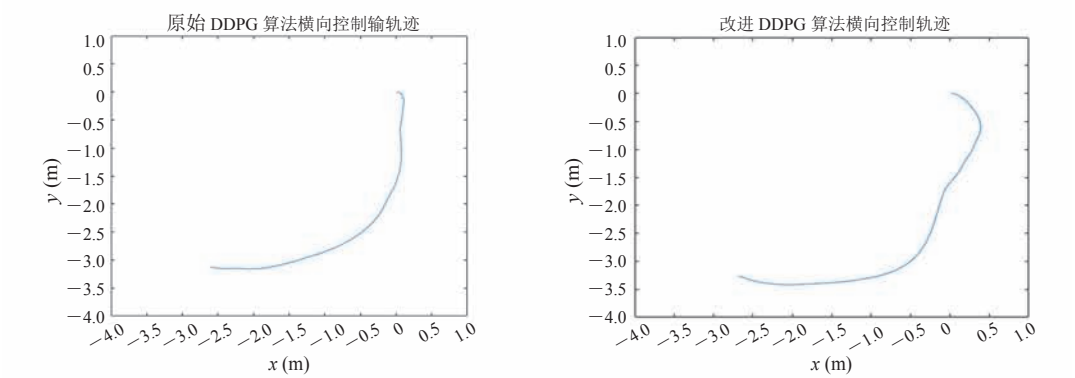


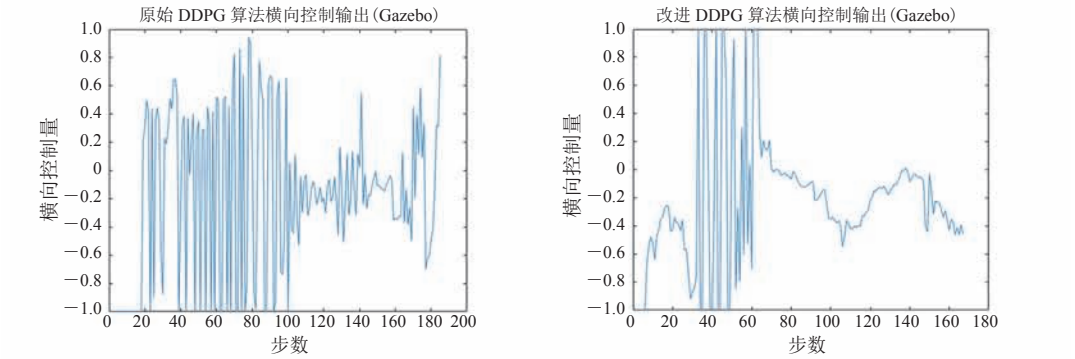
图 11　在 Gazebo 中搭建的仿真环境

Fig. 11　Simulation environment built in Gazebo

同的导航起点与终点，并详细记录了两次实验中移动机器人的运动轨迹与控制过程。在进行实验之前，两种算法都进行了训练并都已充分收敛。从图 12 可以直观地看出，具备类人逻辑的控制算法在实验过程中明显表现出比原始算法更合理的控制过程，并表现出更高的执行效率。为更好地对实验结果进行量化展示，本研究统计了两种算法控制过程中横向控制量的方差与标准差，并



(a) 原始控制算法[19]在实验过程中的运动轨迹对比



(b) 本研究改进驾驶逻辑的控制算法与原始控制算法[19]在实验过程中的控制过程对比

图 12 Gazebo 的仿真实验结果

Fig. 12 Gazebo's simulation experiment results

对比了两种算法完成任务所需的控制步数比，结果如表 3 所示。

表 3 两种算法避障导航任务的控制效果对比

Table 3 Comparison of the control performance of two algorithms for the obstacle avoidance navigation task

算法	横向控制量		所需时间
	方差	标准差	（比例）
原始算法	0.079 8	0.282 5	1
改进算法	0.050 6	0.225 0	0.902 7

结合表 2 和表 3 可以看出，在不同的仿真环境中，HL-DDPG 算法都比原始 DDPG 算法展示出了更平顺的控制过程和更高的任务执行效率。这不仅验证了本研究对于算法策略输出的改进效果，也说明了本研究的改进方式在不同仿真环境中都拥有良好的泛化性。

4.3 平滑且连续的奖励塑造的改进效果

汽车事故种类繁多，产生碰撞的原因也不尽相同，但所有的事故都可以归纳描述为：发生→发展→结束的状态过程，碰撞标志着这一过程的结束，而非开始。为描述该问题，本文仍使用图 3 所示的 Carla 环境模拟了汽车的碰撞过程，构造了无人车在直线巡航场景下的碰撞实验(图 13)。

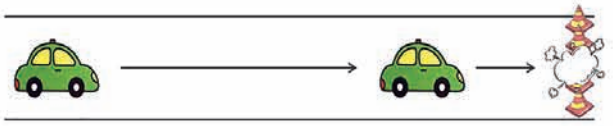


图 13 无人车碰撞实验

Fig. 13 Unmanned vehicle collision test

图 14 为在对不同碰撞过程的奖励塑造下，环境对智能体碰撞事故产生的反馈数据图。结果显示，在两种不同的奖励塑造条件下，改进的带有过程状态的碰撞建模对产生碰撞的无人车给出了更为平顺与密集的惩罚。值得一提的是，在目前的强化学习算法中，智能体策

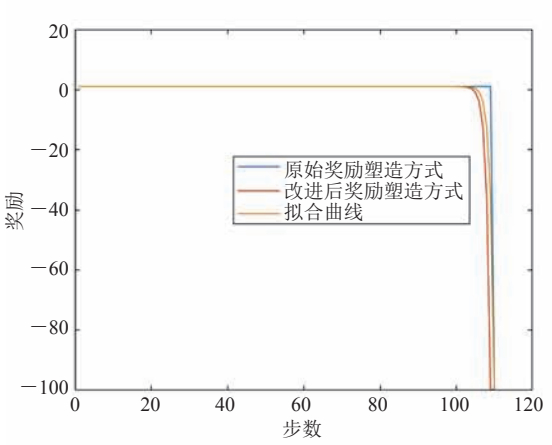


图 14　两种奖励塑造方式的拟合结果

Fig. 14　Fitting results of two reward shaping methods

略的更新完全依赖于环境所给予的奖励回馈：$\pi(a|s)=\arg\max E[R]$，策略梯度是得分函数和价值函数乘积的期望：

$$\nabla_\theta J(\theta)=\mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log\pi_\theta(s,a)Q^{\pi_\theta}(s,a)\right] \quad (4)$$

本研究所使用的 DDPG 算法中，确定性策略也包含两部分：Critic 估计行为价值函数，Actor 估计行为价值函数的梯度。Actor 根据策略梯度调整 $\mu_\theta(s)$的参数 θ，并且用 $Q^{\omega}(s,a)\approx\varepsilon Q^{\pi}(s,a)$ 来逼近真实值。对无人车碰撞事故所构建的稀疏的、严重离群的灾难性惩罚，除了易导致策略的过拟合问题外，对策略梯度的下降也会造成严重不良影响。

在无人驾驶任务中，单步奖励只有带入 Episode 中才有意义。本研究使用二次函数逼近器对图 14 中两种奖励塑造方式进行了拟合。从图 14 可以看出，改进后的奖励塑造方式得出的奖励曲线与拟合后的曲线更加贴近。

记 $Q^{\omega}(s,a)=\varepsilon Q^{\pi}(s,a)$，$\varepsilon$ 为 Episode_奖励拟合的相似度：

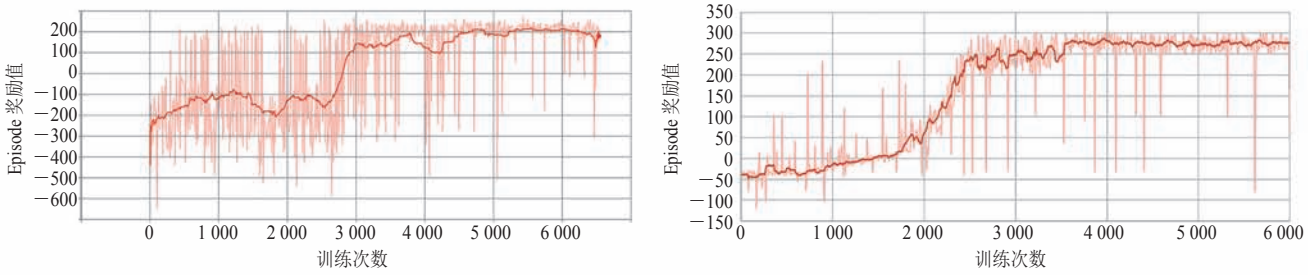
$$\varepsilon=\frac{\int_d\left\langle Q^{Fit.\omega}(s,a)\right\rangle}{\int_d\left\langle Q^{\pi}(s,a)\right\rangle}(d=process) \quad (5)$$

为进一步量化说明，本研究对两种奖励塑造方式的 ε 值进行计算，获得原始奖励 ε 值为 0.0918，改进奖励 ε 值为 1.0526。记 $|1-\varepsilon|$ 为实际奖励与期望的偏离度，可见，改进前的奖励塑造方式所带来的拟合结果是完全失真的。相较于改进前的方式，本研究改进后带有过程状态的碰撞奖励塑造方式对目标函数优化期望的近似程度提高了 85.57%。这种改进可以有效避免由稀疏奖励带来的过拟合现象，也使得智能体的优化目标更符合期望，为算法带来了更快的收敛速度。

4.4　算法的训练效率与控制效果

本研究记录了改进后的算法在训练中的收敛情况，并和原始算法[10]的收敛情况进行了对比，结果如图 15 所示。

从图 15 可以看出，相比于改进前的原始 DDPG 算法，HL-DDPG 算法明显展现出了更稳定的收敛过程，并在算法收敛后具有更稳定的策略表现。原始算法训练至基本收敛状态耗时约 2 900 步，HL-DDPG 算法训练至基本收敛状态耗时约 2 400 步，训练效率提高了 21%。为对本研究改进后算法的控制效果进行更精确的量化展



(a) 原始 DDPG 算法在训练中的收敛情况　　(b) 本研究改进后的 HL-DDPG 算法在训练中的收敛情况

图 15　两种算法在训练中的收敛情况对比

Fig. 15　Convergence of two algorithms in training

示，分别使用两种算法进行了无人车巡航实验，并对两种算法的实验结果进行了统计，具体如表 4 所示。从表 4 可以看出，相比于改进前的原始 DDPG 算法，HL-DDPG 算法任务成功率提高了 19%，任务执行效率提高了 15.45%，并减少了碰撞事故。

表 4 两种算法的控制效果对比

Table 4 Comparison of control performance of two algorithms

算法	任务成功率（%）	产生碰撞次	执行任务时间
原始 DDPG	76	11	1
HL-DDPG	95	2	0.866 1

5 讨论与分析

在目前对无人驾驶端到端控制领域的研究工作中，对无人驾驶策略缺乏类人逻辑这一问题，大部分改进都集中在无人车对关键驾驶行为的决策规划问题上[9-11]。因此其输出的是对驾驶过程中某些重大决策事件做出的离散选取动作概率值，因此，这些改进都并未能在整个无人驾驶任务中形成连续的合理动作序列。

本研究提出一种具备类人逻辑的无人驾驶策略，使策略网络能够输出符合类人驾驶逻辑的连续有序行为。但是无人驾驶规则空间中的类人逻辑是人类社会经过长期驾驶实践得来的经验法则，其引入到策略训练中的实际效果仍然取决于人为的先验设置是否足够精细。但是，过于严苛的驾驶规则设置又会影响无人车对环境与策略的探索，导致策略陷入局部极值。因此，如何在不影响智能体对环境自由探索的前提下，以尽可能小的工作量设置更合理的驾驶规则，仍需进一步的探索与研究。

在智能体自身对策略的探索方面，长短时记忆模型（Long Short Memory Network）[25]解决了传统的 RNN 模型在处理长时数据时，较远序列梯度消失的问题，使得策略网络对长时数据的理解成为可能。虽然由于规则空间中类人逻辑的不可观测性，智能体无法直接通过对环境与状态信息的连续观测获取驾驶逻辑。但是得益于长短时记忆模型对长时数据的理解能力，借助先验建立的驾驶规则，使得无人驾驶策略在较长时序上的“观测-逻辑”配对成为可能。因此，如何运用长短时记忆模型，对智能体在时序上的环境观测信息赋予逻辑语义，并研究逻辑语义与环境观测信息的配对网络策略与训练的影响，也是值得继续深入研究的问题。

6 结 论

现有无人驾驶策略过于依赖感知-控制映射过程的“正确性”，往往忽视了人类驾驶汽车时所遵循的驾驶逻辑。针对这一问题，本文研究了具备类人驾驶行为的无人驾驶策略（HL-DDPG）。本研究在基于深度强化学习的端到端无人驾驶控制网络中，施加类人驾驶的规则约束对智能体连续行为的影响，建立了能够输出符合类人驾驶逻辑的连续有序行为的无人驾驶端到端控制网络。为增强端到端决策行为的安全性，采用对策略输出进行后验反馈的方式，降低了控制策略的危险行为输出率。此外，针对训练过程中出现的难以被拟合的稀疏灾难性事件，提出了连续且更符合控制策略优化期望的奖励函数，提高了算法训练的稳定性。

多个不同仿真环境的结果表明，控制网络中添加规则约束改进的算法比原始算法控制性能更优。改进后的奖励塑造方式在评价稀疏的灾难性事件时，对目标函数优化期望的近似程度比改进前提高了 85.57%，训练效率比传统 DDPG 算法提高了 21%，任务成功率提高了 19%，任务执行效率提高了 15.45%。这表明，采用本文所提出的类人端到端驾驶控制策略，显著减少了碰撞事故，保证驾驶安全的同时提高了驾驶性能。

参 考 文 献

[1] Wang HJ, Yuan SH, Guo MY, et al. Tactical driving decisions of unmanned ground vehicles in complex highway environments: a deep reinforcement learning approach [J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2020: 0954407019898009.

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

[3] Lillicrap TP, Hunt JJ, Pritzel A, et al. Continuous control with deep reinforcement learning [Z/OL]. arXiv Preprint, arXiv:1509.02971, 2015.

[4] Heess N, Hunt JJ, Lillicrap TP, et al. Memory-based control with recurrent neural networks [Z/OL]. arXiv Preprint, arXiv:1512.04455, 2015.

[5] Nash A, Daniel K, Koenig S, et al. Theta*: any-angle path planning on grids [C] // AAAI, 2007, 7: 1177-1183.

[6] Likhachev M, Ferguson D, Gordon G, et al. Anytime search in dynamic graphs [J]. Artificial Intelligence, 2008, 172(14): 1613-1643.

[7] 夏伟, 李慧云. 基于深度强化学习的自动驾驶策略学习方法 [J]. 集成技术, 2017, 6(3): 29-40.

[8] Chowdhuri S, Pankaj T, Zipser K. MultiNet: multi-modal multi-task learning for autonomous driving [C] // 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 2019.

[9] Sheridan TB. Human–robot interaction: status and challenges [J]. Human Factors, 2016, 58(4): 525-532.

[10] Codevilla F, Miiller M, López A, et al. End-to-end driving via conditional imitation learning [C] // 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018: 1-9.

[11] Li LZ, Ota K, Dong MX. Humanlike driving: empirical decision-making system for autonomous vehicles [J]. IEEE Transactions on Vehicular Technology, 2018, 67(8): 6814-6823.

[12] Qureshi AH, Nakamura Y, Yoshikawa Y, et al. Intrinsically motivated reinforcement learning for human-robot interaction in the real-world [J]. Neural Networks, 2018, 107: 23-33.

[13] Montemerlo M, Becker J, Bhat S, et al. Junior: the stanford entry in the urban challenge [J]. Journal of Field Robotics, 2008, 25(9): 569-597.

[14] Guo M, Xu YC, Zhang YJ, et al. A decision-making method for unmanned cars based on drivable area cutting [C] // 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, 2012, 3: 1210-1215.

[15] Brechtel S, Gindele T, Dillmann R. Probabilistic decision-making under uncertainty for autonomous driving using continuous POMDPs [C] // 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2014: 392-399.

[16] Chen YP, Wang JK, Li JH, et al. Lidar-video driving dataset: learning driving policies effectively [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5870-5878.

[17] Xu HZ, Gao Y, Yu F, et al. End-to-end learning of driving models from large-scale video datasets [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2174-2182.

[18] Monahan GE. State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms [J]. Management Science, 1982, 28(1): 1-16.

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

[20] Wang C, Wang J, Shen Y, et al. Autonomous navigation of UAVs in large-scale complex environments: a deep reinforcement learning approach [J]. IEEE Transactions on Vehicular Technology, 2019, 68(3): 2124-2136.

[21] Sammut C, Webb GI. Encyclopedia of Machine Learning and Data Mining [M]. New York: Springer US, 2017.

[22] Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning [C] // Proceedings of the 34th International Conference on Machine Learning, 2017, 70: 2817-2826.

[23] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks [C] // European Conference on Computer Vision, 2014: 818-833.

[24] Chen G, Su SH. Driver-behavior-based robust steering control of unmanned driving robotic vehicle with modeling uncertainties and external disturbance [J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2020, 234(6): 1585-1596.

[25] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.