分阶段奖励优化的变曲率匝道深度强化学习车队协同控制

靳 双[1†]，刘安龙[1]，赵 杭[2,3]，吴仕勋[1]

(1. 重庆交通大学 信息科学与工程学院，重庆 400074；2. 重庆邮电大学 自动化学院，重庆 400065；
3. 重庆邮电大学 智能网联汽车与车路协同重庆市重点实验室，重庆 400065)

摘 要: 针对车队在变曲率匝道等复杂道路场景下行驶易出现队形不稳、控制迟滞及安全性下降等问题, 本文提出一种基于分阶段奖励优化的双延迟深度确定性策略梯度算法 (Stage Reward Shaping Twin Delayed DDPG, SR-TD3), 该算法结合车联网信息与多智能体强化学习框架, 通过在道路的进入、保持和驶出三个曲率阶段分别设计安全、平滑与效率导向的奖励函数, 实现车队在变曲率道路下的稳定协同控制. 在算法结构上采用集中训练与分布执行架构, 并融合参数共享与优先经验回放机制, 并在 Critic 网络中引入残差正则项以抑制价值波动. 基于 CARLA 仿真平台的实验结果表明, 与现有的五种算法相比, SR-TD3 算法在收敛速度、稳定性及跟车精度方面均有显著提升. 其中, 在道路曲率过渡阶段车速均方根误差较五种算法分别降低了 41.31%、86.53%、58.25%、73.08% 和 81.60%; 在整个变曲率匝道上, 车距控制较前五者分别提升了 27.30%、54.52%、36.66%、77.51% 和 76.37%, 同时奖励曲线收敛更快、波动更小, 表现出更高的控制稳定性与学习效率.

关键词: 车队协同控制；深度强化学习；分阶段奖励函数；变曲率道路；多智能体；TD3

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1145

引用格式: 靳双, 刘安龙, 赵杭, 等. 分阶段奖励优化的变曲率匝道深度强化学习车队协同控制 [J]. 控制与决策.

Phase-based reward optimization for deep reinforcement learning–based platoon cooperative control on variable-curvature ramps

JIN Shuang[1†]，*LIU An-long*[1]，*ZHAO Hang*[2,3]，*WU Shi-xun*[1]

(1. School of Information Science and Engineering，Chongqing Jiaotong University，Chongqing 400074，China；
2. School of Automation，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；
3. Chongqing Key Laboratory of Intelligent Connected Vehicles and Cooperative Vehicle-Infrastructure Systems，Chongqing University of Posts and Telecommunications，Chongqing 400065，China)

Abstract: To address the issues of unstable formation, control lag, and degraded safety when vehicle platoons drive in complex road scenarios such as varying-curvature ramps, this paper proposes a Stage Reward Shaping Twin Delayed Deep Deterministic Policy Gradient (SR-TD3) algorithm. By combining connected vehicle information with a multi-agent reinforcement learning framework, this algorithm achieves stable cooperative control of platoons on varying-curvature roads. It accomplishes this by designing safety-, smoothness-, and efficiency-oriented reward functions specifically for the three curvature stages: entering, maintaining, and exiting the road. Structurally, the algorithm adopts a Centralized Training with Decentralized Execution architecture, integrates parameter sharing and prioritized experience replay mechanisms, and introduces a residual regularization term into the Critic network to suppress value fluctuations. Experimental results based on the CARLA simulation platform demonstrate that, compared with five existing algorithms, the SR-TD3 algorithm achieves significant improvements in convergence speed, stability, and car-following precision. Specifically, during the road curvature transition stage, the root mean square error of vehicle speed is reduced by 41.31%, 86.53%, 58.25%, 73.08%, and 81.60% respectively compared to the five baseline algorithms. Throughout the entire varying-curvature ramp, inter-vehicle distance control is improved by 27.30%, 54.52%, 36.66%, 77.51%, and 76.37% respectively. Furthermore, the reward curve converges faster with smaller fluctuations, demonstrating superior control stability and learning efficiency.

收稿日期: 2025-11-04；**录用日期:** 2026-03-12.

基金项目: 国家自然科学基金项目 (62403090); 中国博士后科学基金面上项目 (2022M710546).

†通信作者. E-mail: jsfj@cqjtu.edu.cn.

本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

Keywords: platoon cooperative control； deep reinforcement learning； phase-based reward function； variable-curvature road；multi-Agent；TD3

0 引 言

近年来, 随着智能网联汽车 (Intelligent and Connected Vehicles, ICVs) 与智能交通系统的发展, 车队协同控制已成为智能驾驶领域的重要研究方向. 通过车辆间信息交互与协同决策, 车队能够在一定通信拓扑下保持稳定队形, 从而提高道路通行效率、降低能耗与排放并增强交通安全性[1]. 该技术在高速公路自动驾驶、城市物流运输及编队行驶等场景中具有广泛应用前景. 然而, 在变曲率匝道、弯道及坡道等复杂道路环境下, 弯道半径、转角和坡度等因素会显著影响车辆速度及稳定性[2]. 因此, 在复杂道路条件下实现稳定高效的车队协同控制仍是亟待解决的关键问题.

针对上述挑战, 学术界已开展了广泛探索. 早期研究多基于传统控制理论, 如自适应巡航控制 (Adaptive Cruise Control, ACC)、协同自适应巡航控制 (Cooperative Adaptive Cruise Control, CACC)、滑模控制 (Sliding Mode Control, SMC)、鲁棒控制 (Robust Control) 和模型预测控制 (Model Predictive Control, MPC) 等方法[3-9]. 此类方法在平直道路或简单工况下表现优异, 但在变曲率、强机动等复杂动态场景中, 往往难以兼顾控制性能与实时性. 近年来, 强化学习 (Reinforcement Learning, RL) 为车队控制提供了新的解决思路. 其通过与环境交互来学习控制策略, 在无需精确建模的情况下表现出良好的自适应能力[10]. 特别是深度强化学习 (Deep Reinforcement Learning, DRL) 能够有效处理高维连续状态与动作空间问题[11], 其中深度确定性策略梯度 (Deep Deterministic Policy Gradient, DDPG) 因擅长处理连续控制任务而被广泛应用于自动驾驶[12]. 文献 [13] 提出了一种提升智能体学习率的 DDPG 算法; 文献 [14] 则构建了一种融合 ACC 机制与 RL 优势的混合模型, 有效提升了算法的收敛速度与执行效率; 文献 [15] 基于强化学习算法, 提出了一种多目标车辆跟随决策算法, 提升了车辆跟随的舒适性需求; 文献 [16] 将竞争双重深度 Q 网络 (Dueling Double Deep Q-Network, D3QN) 算法与 DDPG 算法结合, 提出了一个双层决策模型, 提升了车辆在变道和跟车过程中的效率、安全和舒适性.

针对多智能体控制问题, 文献 [17] 针对车辆多目标优化问题提出了一种 RL 算法, 显著增强了车辆行驶稳定性; 文献 [18] 提出了一种 ADAC(actor-double-attention-critic) 算法, 该算法提高了多智能体在混合合作-竞争任务中的协作性能. 文献 [19] 提出了一种混合 DDPG 算法, 该算法使多智能体在收敛速度和任务完成度方面有明显提高. 文献 [20] 提出了一种基于 MADDPG (Multi-Agent Deterministic Policy Gradient) 算法上的单值网络对所有智能体的状态和行为进行评估的 PS-MACDDPG(Parameter Sharing Multi-Agent Cooperative DDPG) 算法. 然而,DDPG 在实际训练中往往存在高估偏差、收敛速度慢以及策略不稳定等问题, 限制了其在复杂车队控制场景中的应用. 为克服 DDPG 的不足, 研究者提出了双延迟深度确定性策略梯度 (Twin Delayed Deep Deterministic Policy Gradient,TD3) 算法. TD3 通过引入双 Critic 网络来缓解价值高估问题, 并采用延迟更新与目标策略平滑等机制, 有效提升了训练的稳定性和收敛性能. 基于这一改进,TD3 逐渐成为控制领域的主流方法之一.

基于上述车队在复杂道路场景难以控制、现有控制算法收敛慢以及不稳定性的问题, 本文提出一种基于分阶段奖励优化的双延迟深度确定性策略梯度算法 (SR-TD3), 并将其应用于变曲率匝道车队协同控制. 该方法按车辆行驶特性将道路划分为进入阶段、保持阶段和驶出阶段, 并针对不同阶段设计差异化奖励函数: 进入阶段侧重姿态调整与安全过渡, 保持阶段侧重车道跟踪稳定性与行驶舒适性, 驶出阶段则注重平滑加速与通行效率. 此外, 融合车联网技术实时交互周围车辆的状态与道路拓扑信息, 增强了控制系统的精确度与前瞻性. 实验结果表明, 相较于现有算法该方法能够在复杂变曲率道路条件下实现稳定、高效的车队协同控制.

本文基于 CARLA 仿真平台进行实验, 将所提 SR-TD3 方法与传统 MADDPG、改进型 PS-MACDDPG、CACC-LKA、CACC-TD3 和 MPC-TD3 进行对比分析, 从奖励收敛速度、控制性能和队形稳定性等方面验证了其优越性. 本文的主要研究贡献体现在以下三个方面: 1) 针对变曲率匝道伴随坡度变化的极端工况, 提出了一种能兼顾入弯安全减速与出弯高效加速的控制策略, 解决了传统单一奖励函数在复杂动态场景下难以平衡多维目标的难题; 2) 构建了基于道路几何特征驱动的动态奖励机制, 通过实时监测曲率变化率精准划分行驶阶段, 并据此动态切换奖励计算逻辑, 实现了车队在时变环

境下的自适应优化; 3) 在多智能体协同框架中融合TD3 算法, 并在 Critic 网络中引入残差正则项, 有效抑制了复杂路况下的 Q 值剧烈波动, 显著提升了算法在连续动作空间中的训练稳定性与收敛效率.

1　问题陈述

1.1　控制场景

如图 1 所示, 本文与以往大多数的强化学习车队控制场景不同, 针对的是变曲率道路下的车队协同行驶场景进行研究. 车队由一辆领航车与多辆跟驰车组成, 沿着变曲率匝道轨迹行驶, 道路曲率在不同阶段呈连续变化形式, 车队可以根据不同阶段曲率的连续性调整转向角和油门, 并在此过程中实现速度与间距的动态协调.

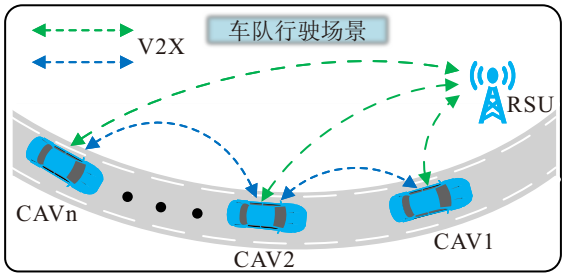


图1　车队行驶场景图

在该场景中, 领航车需依据车联网通信 (Vehicle-to-Everything,V2X) 与路基设施进行交互, 获得前方道路的曲率、坡度信息, 通过控制算法计算出适配的车速和航向角, 并进行相应的控制调整, 来适应不同道路状况的行驶. 而跟驰车需依据 V2X 通信获取的前车状态信息, 结合本地感知状态 (位置、速度、航向角等), 通过强化学习策略生成连续控制动作, 实现恒定车间距控制与纵向行驶平滑性. 同时, 控制算法需在曲率连续变换和道路坡度变化等非线性条件下保持车队稳定性与安全性, 避免车辆发生碰撞或脱离队形.

1.2　建　模

在控制场景基础上, 本文将车队协同行驶问题抽象为一个完全协作性的多智能体强化学习 (MARL) 模型. 车队中的每辆车被视为一个智能体 (Agent), 通过与环境的交互不断学习最优控制策略; 而环境信息由道路拓扑结构、前车行为、周围车辆状态及物理动力学约束共同构成. 智能体在每一时刻根据自身状态与环境反馈生成控制动作, 通过试错与奖励信号优化其策略, 以实现稳定、高效、安全的车队协同控制.

在完全合作场景下, 问题可视为多智能体马尔可夫决策过程 (Multi-Agent Markov Decision Process, MMDP), 可采用集中训练-分散执行 (Centralized Training with Decentralized Execution, CTDE) 范式[21-22], 常用形式为:

$$(\mathcal{N}, S, \{A_i\}_{i\in\mathcal{N}}, P, R, \gamma). \tag{1}$$

其中$\mathcal{N}=\{1,\ldots,N\}$为智能体集合, S为状态空间; 每个智能体在时刻t的状态s_t由自车的运动学信息 (位置、速度、加速度、航向角) 与前车的相对距离与相对速度、道路曲率及车道偏移等特征共同构成. 状态信息不仅反映个体自身的动态特性, 也包含对周围环境变化的感知, 从而为策略学习提供充分的环境. A_i为第i个智能体的动作空间, 智能体在连续动作空间内选择控制量$a_t=[\delta_t,\tau_t]$, 其中δ_t表示转向角控制量, τ_t表示油门输入. $P(s'|s,a)$为联合状态转移概率; $R(s,a)$为共享的全局奖励函数; γ为折扣因子.

在t时刻, 各智能体根据局部观测o_i^t执行动作$a_i^t\sim\pi_i(a_i|o_i^t)$, 联合动作为$a_t=(a_1^t,\ldots,a_N^t)$. 环境据此进行状态转移$s_{t+1}\sim P(s_{t+1}|s_t,a_t)$, 并产生全局奖励$r_t=R(s_t,a_t)$.

系统的联合策略期望回报函数定义为:

$$J(\pi)=\mathbb{E}_{s_0,a_t\sim\pi,s_{t+1}\sim P}[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)]. \tag{2}$$

其中$\pi=\{\pi_i\}_{i\in\mathcal{N}}$表示联合策略集合. 在此建模框架下, 整个车队系统通过集中训练与分布执行的机制实现多智能体协同. 训练阶段各智能体共享全局状态信息, 以提升学习效率; 执行阶段则依据本地观测独立决策, 从而具备较强的分布式执行能力与环境适应性. 通过这种对现有大多数控制算法的改进, SR-TD3 算法能够有效捕捉车辆间的动态耦合关系, 并实现变曲率道路下的稳定协同控制.

2　基于深度强化学习的车队协同控制算法设计

2.1　协同控制总体架构

本文基于 SR-TD3 算法构建的车队协同控制系统整体架构采用“感知–决策–控制”三部分, 实现多车在变曲率道路下的协同行驶控制, 其架构框图如图 2 所示.

感知层通过车载传感器和 V2X 通信模块实时感知车辆自身及邻车的运动状态, 并区别现有大多数车队架构, 本架构采用非链式的信息交互机制以增强车队的稳定性. 融合道路曲率、车速、位置及车距等多源信息, 作为智能体的状态输入. 决策层中采用改进的双延迟深度确定性策略梯度算法 (SR-

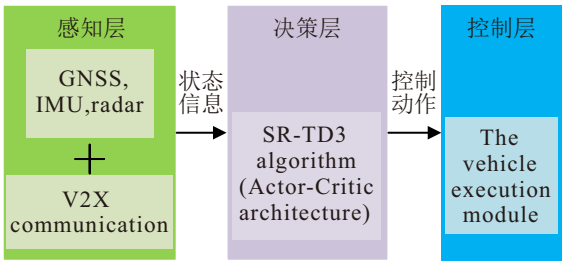


图2 控制系统架构框图

TD3), 通过集中式训练与分布式执行框架实现多智能体策略学习; 在训练阶段, 各车辆智能体共享参数以提升收敛效率与稳定性, 而在执行阶段, 各智能体根据本地感知与通信获取的状态信息独立输出控制动作, 从而保证系统的实时性与可扩展性. 控制层依据智能体输出的连续控制量驱动车辆, 实现对目标车距、速度和轨迹的精确控制.

2.2 观测空间与动作空间设计

本文所提出的 SR-TD3 算法采用集中训练分布执行架构, 其网络体系主要由策略网络 (Actor) 与价值网络 (Critic) 及其对应的目标网络构成. 车队的协同控制属于部分观测马尔可夫决策过程 (POMDP), Actor 网络的输入被设计为智能体的局部观测向量 O_i. 该观测向量依次经过两个节点数分别为 400 与 300 的全连接隐藏层, 层间采用 ReLU 激活函数进行非线性变换, 最终通过带有 Tanh 激活函数的输出层生成范围在 [−1,1] 的二维无量纲动作向量. 此动作向量随后经过式 (3) 定义的线性映射关系, 转化成为车辆执行模块所需的实际控制指令:

$$\begin{bmatrix} throttle \\ steer \end{bmatrix} = J \begin{bmatrix} \alpha 1 \\ \alpha 2 \end{bmatrix} = Ja. \tag{3}$$

式中u为真实的控制输入向量, 包含油门和转向两个分量; J为动作映射矩阵; α表示智能体 Actor 网络输出的无量纲动作向量, 范围在 [−1,1], $\alpha 1$对应油门通道的原始动作输出; $\alpha 2$对应转向通道的原始动作输出.

与 Actor 不同,Critic 网络在训练阶段利用全局信息, 其输入设定为包含所有智能体观测信息的全局状态 $S = [o_1, o_2, \cdots, o_N]$与联合动作的拼接向量. Critic 网络内部同样维持了 [400,300] 的隐藏层结构与 ReLU 激活配置, 并采用线性输出层以精确估计 Q 值. 为了克服传统 DDPG 算法中的 Q 值高估问题, 本算法融合了 TD3 算法的双 Q 网络 (Twin Critic) 机制以及目标网络延迟更新策略, 通过构建包含主网络与目标网络在内的“双 Actor+四 Critic”拓扑结构, 有效提升了策略在复杂变曲率道路场景下学习的稳定性与收敛速度.

2.3 奖励函数设计

鉴于变曲率道路协同控制的复杂需求本文确立了“密集奖励为主, 稀疏奖励为辅”的设计原则. 该机制利用基于速度与车距的密集奖励加速初期策略收敛, 同时结合基于避障与队形保持的稀疏奖励以强化特定任务表现. 为了使车队更加地适应连续变曲率和坡度的道路场景, 在奖励函数的设计上结合了动力学约束的动态阈值. 针对变曲率道路的具体分阶段奖励设计如下:

在车队即将进入变曲率道路时, 车辆会通过地图实时获取本车道的路基点, 从而得到当前道路曲率k.

1) 在曲率道路的进入阶段行驶时, 我们的首要目标是通过速度控制和航向角调整实现安全过渡, 其触发条件是:

$$k > k_{thresh} and |\dot{k}| > \epsilon. \tag{4}$$

式中k表示当前道路曲率, k_{thresh}表示进入弯道的阈值曲率; 式中$\dot{k}$为曲率变化率, ϵ为曲率变化阈值.

其中k_{thresh}可由车辆动力学理论进行计算得出, 推导得到公式如 (5) 所示.

阈值曲率:

$$k_{thresh} = \frac{\mu g}{v^2}. \tag{5}$$

式中μ为轮胎与路面的摩擦系数; g为重力加速度 (9.81 $\mathrm{m/s^2}$); v为当前车辆行驶速度.

其中ϵ可由动力学约束理论进行计算得出, 推导公式如 (6) 和 (7) 所示.

在短时间内忽略速度变化, 可以简单近似推到:

$$a_{lateral} = v^2 k \Rightarrow \frac{d_{a_{lateral}}}{dt} \approx v^2 \frac{dk}{dt}. \tag{6}$$

由上述推导可得:

$$\epsilon = \frac{j_{max}}{v^2}. \tag{7}$$

奖励函数设计为:

$$R_{enter} = w1 \cdot (v_{target} - |v - v_{target}|) + w2 \cdot e^{-|d_{center}|} + w3 \cdot (-|\delta|) + w4 \cdot (-a_{lateral}). \tag{8}$$

变量说明:

v_{target}: 基于摩擦系数 μ、重力加速度 g 和当前曲率k的安全速度; d_{center}: 车辆与车道中心线的横向偏移 (单位: 米); δ: 方向盘转角 (单位: 弧度); $a_{lateral}$: 横向加速度 (单位: 米/二次方秒).

2) 在曲率道路的保持阶段行驶时, 我们的首要目标是稳定跟踪车道中心线并维持安全车速保证乘客的舒适性, 其触发条件是:

$$k > k_{thresh} and |\dot{k}| \leqslant \epsilon. \tag{9}$$

奖励函数:

$$R_{cruise} = w5 \cdot e^{-|d_{center}|} + w6 \cdot (\frac{v}{v_{safe_max}}) + w7 \cdot (-|\dot{\delta}|) + w8 \cdot (-|\theta_{error}|). \tag{10}$$

变量说明:

$v_{safe\_max} = (\sqrt{\mu g/\kappa}, v_{road\_limit})$; $\dot{\delta}$: 转角变化率 (方向盘抖动惩罚); θ_{error}: 车身朝向和道路切线方向的偏差.

3) 在曲率道路的驶出阶段行驶时, 我们以安全和效率为首要目标使车队平滑加速驶出曲率道路, 其触发条件是:

$$k < k_{thresh} and \dot{k} < 0. \tag{11}$$

奖励函数:

$$R_{exit} = w9 \cdot (v_{target_exit} - |v - v_{target_exit}|) + w10 \cdot e^{-|d_{center}|} + w11 \cdot (-|\theta_{error}|) + w12 \cdot (-j). \tag{12}$$

变量说明:

v_{target_exit}: 从当前速度线性加速到直道目标速度; j: 加加速度, 加速度变化率 (单位: 米/三次方秒).

2.4　算法实现

本文提出了改进型的 SR-TD3 算法, 如图 3 所示, 该算法在保留参数共享与集中式训练框架的基础上, 从网络结构、目标值计算以及样本利用等多个层面对算法进行了优化设计.

在网络结构方面, SR-TD3 引入了双重 Critic 网

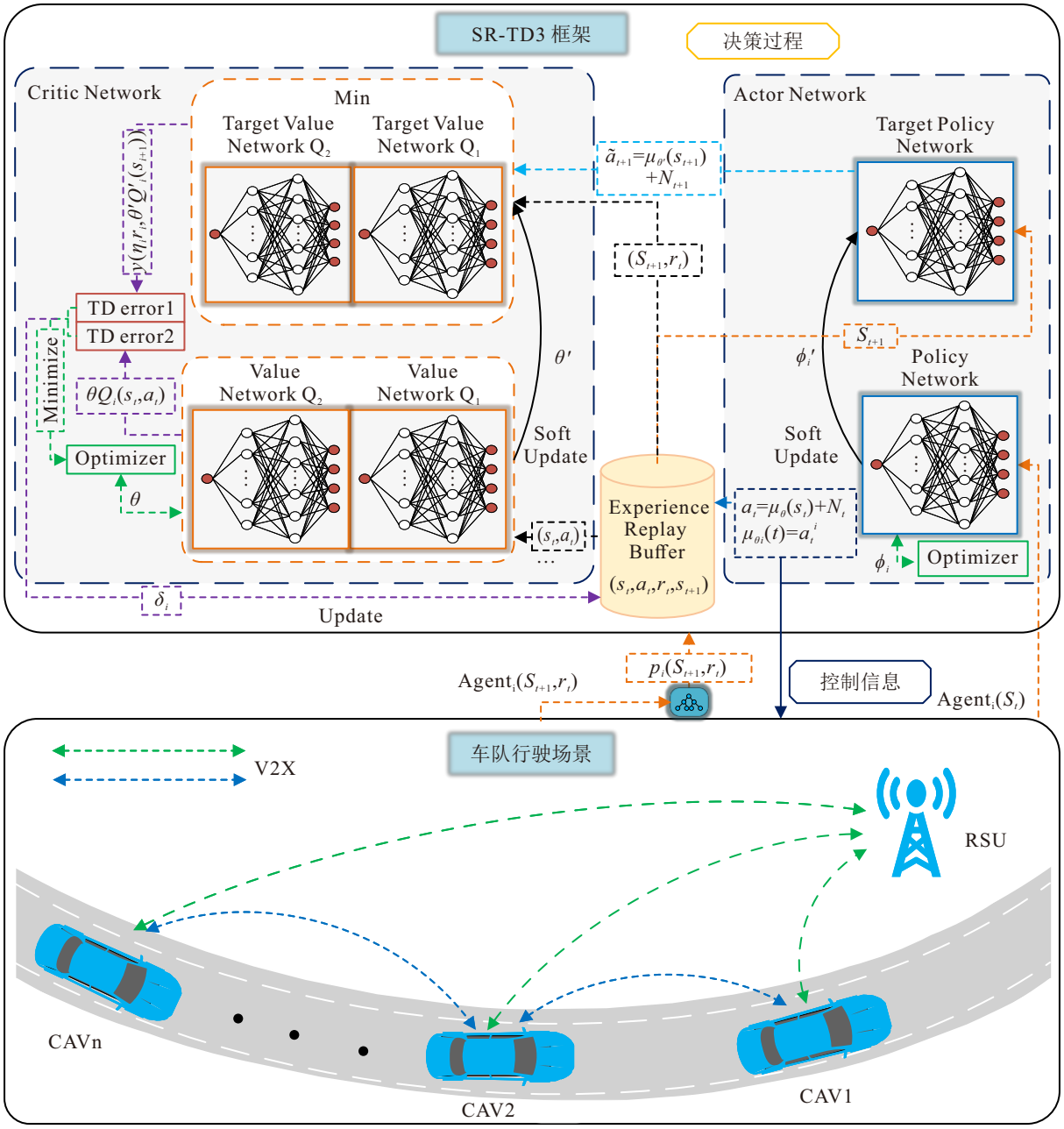


图3　SR-TD3 算法框架图

络与延迟策略更新机制, 有效的缓解了传统 DDPG 算法在单 Critic 模型中存在Q值过估计的问题. 设两个价值网络分别为$Q_{\phi_1}(s_t,a_t)$与$Q_{\phi_2}(s_t,a_t)$, 在原有参数共享模式上引入加权的 Min–Max 目标 Q 值融合, 本文针对变曲率匝道这一具体表达式如下:

$$\mathcal{Y}_t=\eta_t r_t+\gamma[\lambda \underset{i=1,2}{min} Q_{\phi_i'}(s_{t+1},\widetilde{a}_{t+1})+ (1-\lambda)\underset{i=1,2}{max} Q_{\phi_i'}(s_{t+1},\widetilde{a}_{t+1})]. \tag{13}$$

其中, $\widetilde{a}_{t+1}=\mu_{\theta'}(s_{t+1})+\epsilon,\epsilon\sim\mathcal{N}(0,\sigma^2)$表示平滑噪声项, $\lambda\in[0,1]$控制最小值与最大值的加权比例. 该设计在确保策略更新稳定的同时, 提升了目标值估计的精确性. Critic 网络的优化目标为最小化加权平方误差:

$$L_Q=\frac{1}{B}\sum_{I=1}^{B}w_i(Q(s_i,a_i)-\mathcal{Y}_i)^2. \tag{14}$$

其中w_i为重要性采样权重. Actor 网络采用延迟更新策略, 仅在每d次 Critic 更新后执行一次参数更新, 其梯度计算形式为:

$$\nabla_\theta J(\theta)=\frac{1}{B}\sum_{I=1}^{B}\nabla_a Q_{\phi_1}(s_i,a)|_{a=\mu_\theta(s_i)}\nabla_\theta\pi_\theta(s_i). \tag{15}$$

从而实现策略的稳定优化. 通过这种延迟策略更新机制, SR-TD3 算法在训练早期能够保持稳定性, 避免了由于策略的频繁调整而导致的梯度振荡, 从而使训练时收敛更快.

在经验回放机制上, SR-TD3 算法引入了优先经验回放 (Prioritized Experience Replay, PER) 方法, 以提高关键样本的利用效率. 对于经验池$\mathcal{D}$中的每个样本, 其采样概率由时间差分误差$\delta_i=|\mathcal{Y}_i-Q_\phi(s_i,a_i)|$决定, 定义为:

$$P(i)=\frac{(|\delta_i|+\varepsilon)^\alpha}{\sum_k(|\delta_k|+\varepsilon)^\alpha}. \tag{16}$$

其中a控制优先采样程度, ε为防止零概率的平滑项. 为补偿非均匀采样带来的估计偏差, 引入重要性采样权重修正:

$$w_i=(\frac{1}{N\cdot P(i)})^\beta. \tag{17}$$

其中β在在训练过程中逐渐从 0.4 增加至 1.0, 以平衡样本重要性与训练稳定性. 此外, 在原有算法的框架上, SR-TD3 算法在 Critic 损失函数中加入了残差正则项:

$$L_{res}=\eta(\mathcal{Y}_t-Q_{\phi_1}(s_t,a_t))^2. \tag{18}$$

用于限制目标 Q 值的波动范围, 从而进一步提高了训练的平滑性与稳定性.

在多智能体结构设计上, SR-TD3 延续了 PS-MACDDPG 的参数共享机制 (Parameter Sharing) 与集中式 Critic 设计. 假设车队中共有N个智能体, 其全局状态与动作可分别表示为$S=[o_1,o_2,\cdots,o_N]$与$A=[a_1,a_2,\ldots,a_N]$. 共享的策略网络$\mu_\theta(s_i)$基于每辆车的局部状态输出相应动作a_i, 而集中式 Critic $Q_\phi(S,A)$则利用全局状态与联合动作进行统一评估. 该架构在保证训练阶段信息完整性的同时, 使执行阶段仍具分布式独立性, 从而兼顾算法的协同性与扩展性.

在算法设计中, 折扣因子设定为$\gamma=0.99$, 以平衡短期与长期回报. Actor 网络与 Critic 网络分别采用 Adam 优化器, 学习率设置为1×10^{-4}与1×10^{-3}. 在策略执行阶段, 向动作添加均值为零、标准差为$\sigma=0.2$的高斯噪声, 并对其进行 [−0.4,0.4] 截断, 同时噪声强度随训练逐步衰减, 最低不小于 0.06, 以保证探索性. 每次参数更新均从经验回放缓冲区中随机采样 256 条交互样本, 经验回放缓冲区容量在训练脚本中设为 30 000, 并结合优先经验回放机制以提升采样效率. 为了提高稳定性, 采用软更新策略, 参数τ=0.005, 并设置延迟更新步长为 3, 即每进行三次 Critic 更新才对 Actor 进行一次更新.Actor 学习率和 Critic 学习率分别是 1e-4 和 1e-3.

2.5 约束条件

由于仿真环境是连续的, 为了进行合理的仿真训练, 在部分条件判定时, 当$done=1$时终止当前训练的回合. 这些条件分别是:

1、碰撞条件

$$done=1,if C_i=1\ (i\in\{leader,follower1,follower2\}). \tag{19}$$

2、车道偏离条件

$$done=1,if|e_{lane}(i)|>\delta_{lane}. \tag{20}$$

3、速度异常条件

$$done=1,if(v_i<v_{min})\vee(v_i>v_{max}). \tag{21}$$

4、车队队形约束条件

$$done=1,if d_{follow}(i)>d_{max},\ i\in\{follower1,follower2\}. \tag{22}$$

5、路径限制条件

$$done=1,if d_{leader}=d_{route}. \tag{23}$$

6、时间限制条件

$$done = 1, if\ step \geq N_{max}. \tag{24}$$

3 实验与结果分析

3.1 仿真环境与参数

实验是结合 RoadRunner 和 CARLA 软件进行的, 分别承担场景生成与仿真控制的任务, 图 4 为地图场景示意图. 在实验中车辆的动力学由 CARLA 仿真环境提供, 其输入控制量为油门 throttle 与转向角 steer, 经过无量纲缩放油门和转向的动作范围取值分别是 [0,1] 和 [−0.4,0.4]. 该控制量作为输入传递给 CARLA 内部的车辆物理引擎, 从而实现算法对车队的控制.

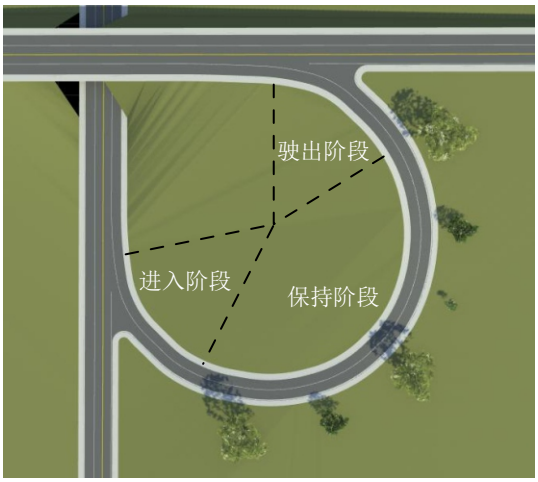


图4　实验道路场景示意图

在训练过程中, 对一个由三辆车组成的车队进行训练, 车辆检测距离设置为 280 m. 车道宽度为 3.2 m.在车辆通过曲率道路时, 我们按照曲率道路车辆的速度特性分成三个部分, 分别是进入曲率路段、曲率保持路段和驶出曲率路段. 我们设定生成车队中每辆车的间距为 5 m, 车辆在变曲率道路中行驶的理想速度为 5 m/s.

3.2 算法模型训练对比

本小节训练的目的是通过比较各种算法在车队控制中的性能, 来评估各种算法在训练的收敛性能. 表 1 详细说明了实验设置.

表1　实验算法配置表

实验序号	领航车数量	跟驰车数量	算法名称	学习结构
1	1	2	SR-TD3	PF
2	1	2	CACC-TD3	PF
3	1	2	MPC-TD3	PF
4	1	2	PS-MACDDPG	PF
5	1	2	MADDPG	PLF

五组实验都采用领队-跟驰者结构来进行队列控制训练, 学习结构包括参数共享架构 (Parameter Sharing Framework, PF) 和独立参数学习架构 (Parameter Learning Framework, PLF). 在 DRL 中, 通常采用回合奖励 (episode reward, ER) 与平均奖励 (average reward, AR) 反映模型训练的收敛水平和学习效果. 设计的训练回合数为 3 000, 分别观察每种算法的奖励曲线图 5 到图 7 为模型训练过程中奖励值变化对比. 由于 CACC-TD3 算法, MPC-TD3 算法和 SR-TD3 算法框架结构与参数设计一致, 且训练效果一致, 所以不再展示 CACC-TD3 算法和 MPC-TD3 算法的奖励函数图.

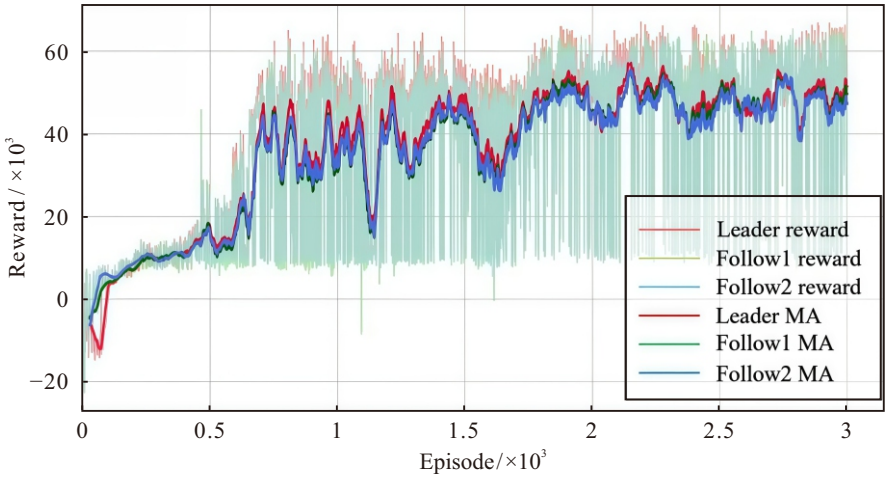


图5　SR-TD3 奖励函数图

从图 5 中可以看出引入 PER 的 SR-TD3 算法在训练的过程中, 奖励值在 600 回合左右开始增加, 在 1 800 回合左右基本收敛, 最终在 2000 回合后奖励值均值基本稳定且相差不大, 表明训练基本已收敛至全局最优, 智能体已学习到有效策略. 而 PS-MACDDPG 算法虽然在训练的开始奖励值就增加, 奖励值在 600 回合左右开始收敛, 但在 600 回合过后奖励值波动大, 不具有稳定性. 而 MADDPG 算法收敛过晚, 在 800 回合左右奖励值才开始增加,1 250 回合左右奖励值开始收敛, 且收敛后奖励值同样有

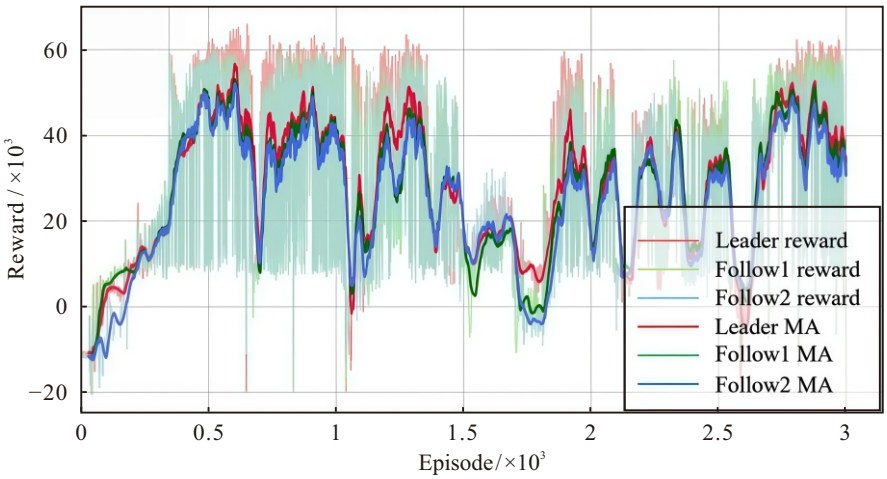


图6 PS-MACDDPG 奖励函数图

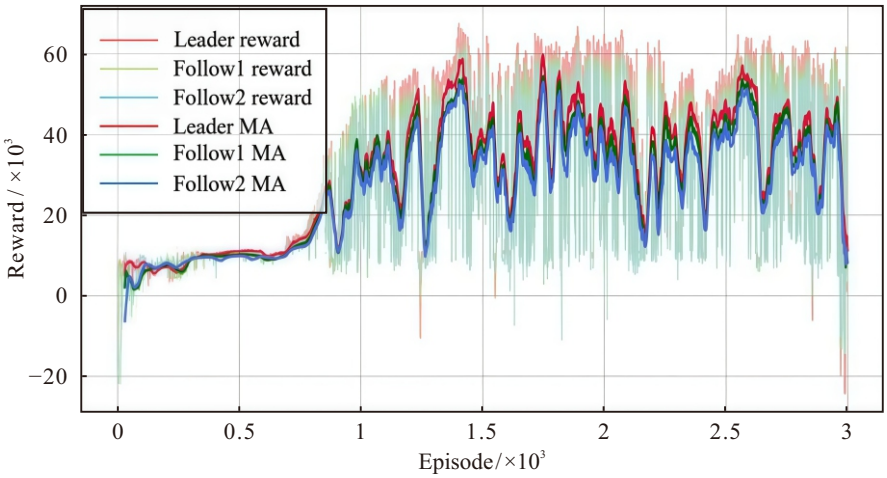


图7 MADDPG 奖励函数图

很大的波动. 虽然三种算法再训练过程中都获得了较高奖励值, 但是从奖励函数曲线图中可以看到除了 SR- TD3 算法外, 其他两种算法在收敛后奖励值波动大, 由此模型训练效果可以得出 SR-TD3 算法智能体在整个探索过程中, 学习得更加稳定.

3.3 算法模型测试对比

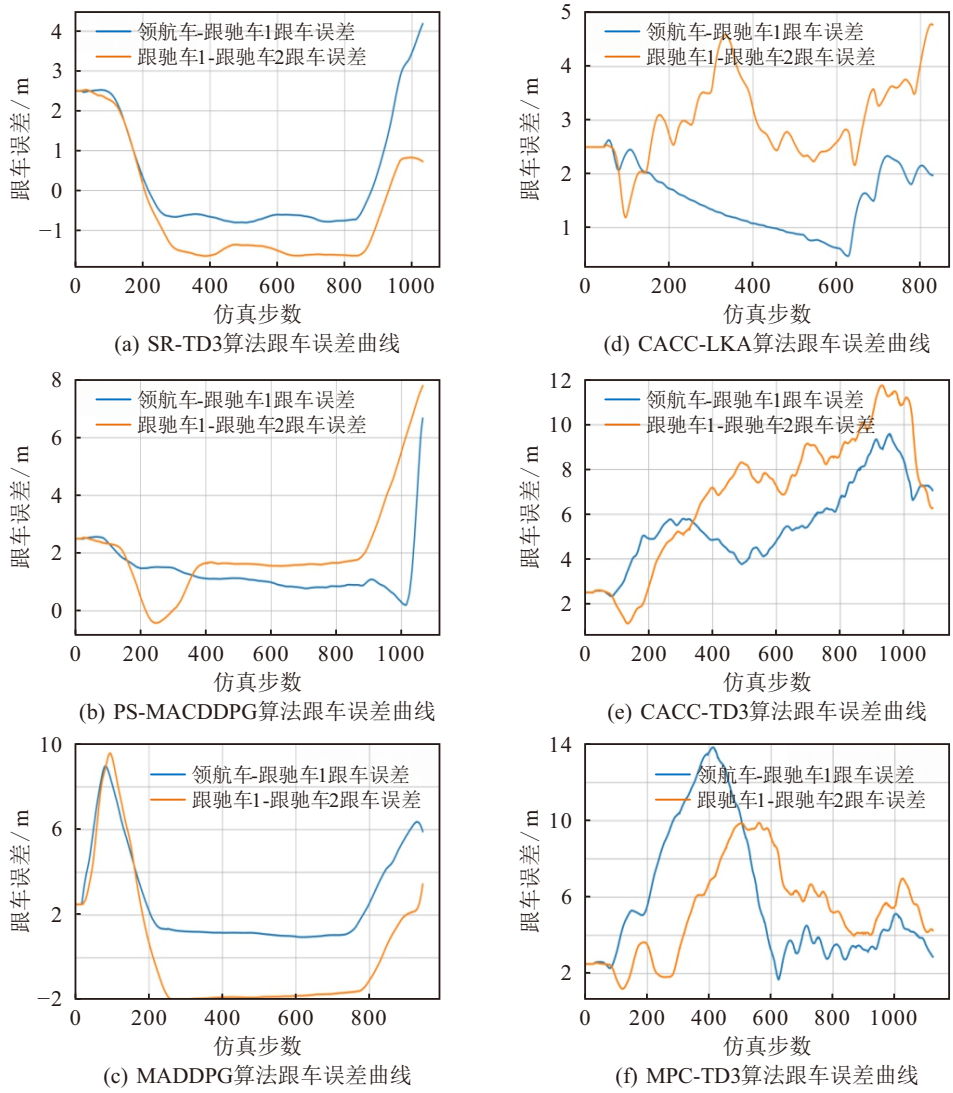
本小节的目的是评估车队在连续变曲率道路场景中的行驶性能. 实验开始时, 车辆间初始间距为 5 m.

车队行驶过车中车间距为 2.5m 时奖励值最高, 预期效果车队行驶过程中会随着变曲率匝道的不同阶段进行速度的调整.

在变曲率匝道场景中, 车队在不同道路阶段 SR-TD3、PS-MACDDPG、MADDPG、CACC-LKA、CACC-TD3 和 MPC-TD3 算法的跟车误差曲线和跟车误差对比分析分别如图 8 和图 9 所示.

如图 8 所示, 车队在整个行驶过程中从初始车间距不断向目标车间距进行收敛. 其中前三种算法在控制车队通过变曲率道路过程的车间距都出现了先收敛稳定后车距扩大的过程, 这是由于车队在从变曲率道路的进入阶段到保持阶段, 又从保持阶段到驶出阶段的速度变化所引起的. 相比之下, MPC-TD3 的失效归因于物理模型失配与残差修正权限的冲突模型偏差导致基础控制量错误, 而受限的 RL 输出无法有效补偿. CACC-TD3 则受困于预热数据的分布偏移, 导致策略收敛至“远距避险”的局部最优. 传统 CACC-LKA 受限于线性反馈原理, 缺乏对动态工况的前馈补偿, 不可避免地产生了响应滞后与稳态误差. 如图 9 所示, 在使用 PS-MACDDPG 和 MADDPG 算法对车队跟车距离的控制中, 可观察到车队在从变曲率道路的进入阶段到保持阶段过程中, 实际跟车距离与目标车距之间的波动较大, 且非常的不稳定; 而 SR-TD3 算法在控制车队在整个变曲率道路行驶过程中, 实际跟车距离与目标车距相贴合, 且在整个过程中波动最小.

通过数据计算出在整个变曲率道路上, 六种算法所控制的车队领航车与跟驰车 1 车间距均方根误差值, 跟驰车 1 和跟驰车 2 车间距均方根误差值. 可以得到 SR-TD3 算法在控制车队整个变曲率道路的行驶过程中, 在领航车与跟驰车 1 和跟驰车 1 与跟

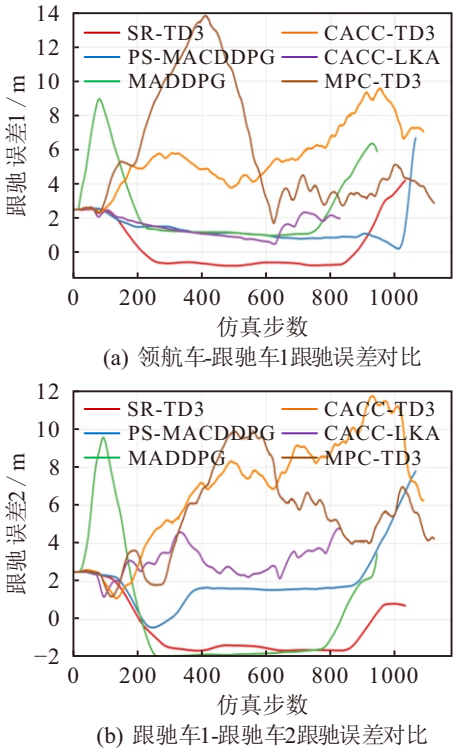


(a) SR-TD3算法跟车误差曲线　(d) CACC-LKA算法跟车误差曲线

(b) PS-MACDDPG算法跟车误差曲线　(e) CACC-TD3算法跟车误差曲线

(c) MADDPG算法跟车误差曲线　(f) MPC-TD3算法跟车误差曲线

图8　不同算法下跟车误差曲线图



(a) 领航车-跟驰车1跟驰误差对比

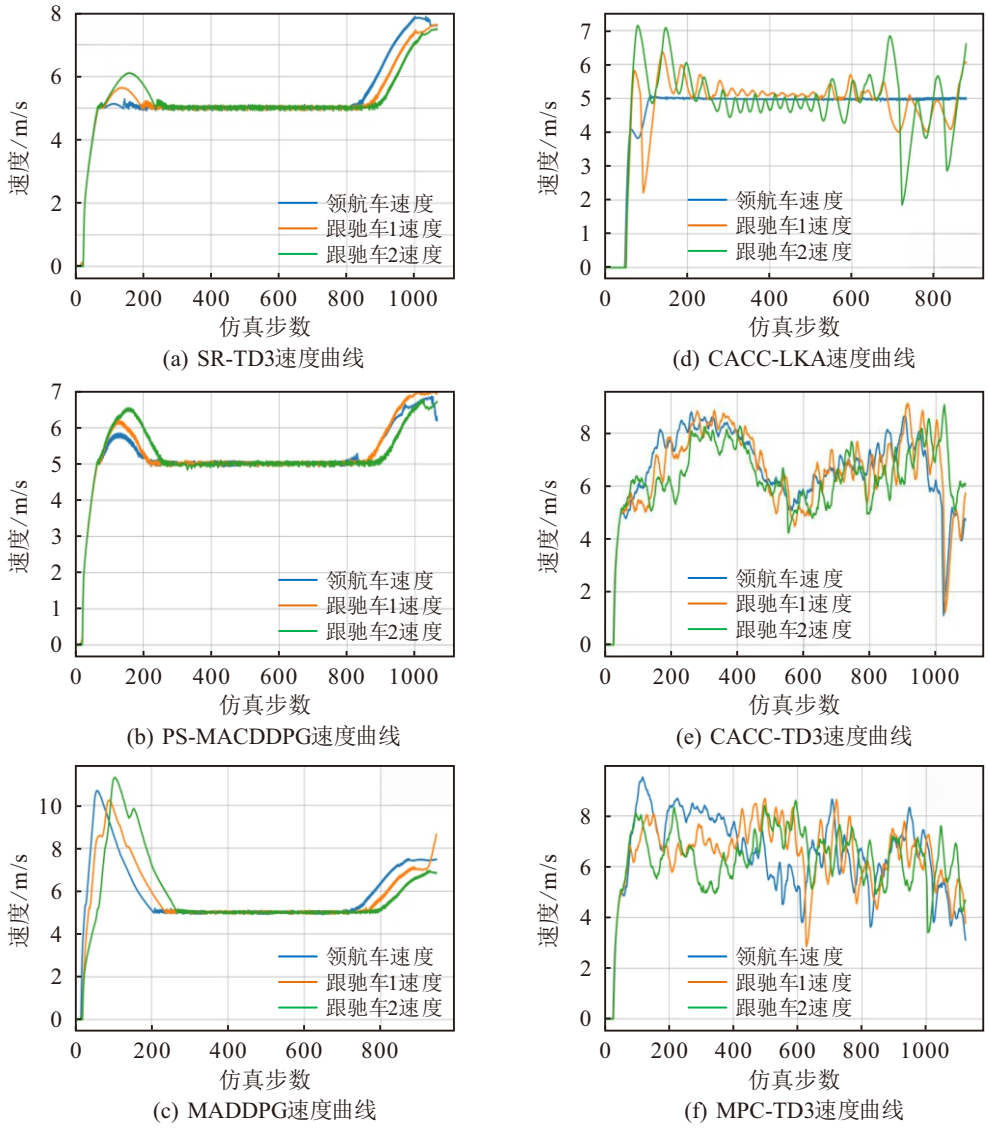
(b) 跟驰车1-跟驰车2跟驰误差对比

图9　跟车误差对比图

驰车 2 的跟车稳定性上, 相较于 PS-MACDDPG 算法分别提升了 5.49% 和 40.57%; 相较于 MADDPG 算法分别提升了 57.49% 和 51.22%; 相较于 CACC-LKA 算法分别提升了 11.36% 和 50.37%; 相较于 CACC-TD3 算法分别提升了 74.90% 和 79.57%; 相较于 MPC-TD3 算法分别提升了 78.43% 和 73.95%. 对于整个车队的跟车稳定性上分别提升了 27.30%、54.52%、36.66%、77.51% 和 76.37%, 对比结果表明, SR-TD3 算法在车队的间距控制性能上更加的稳定, 跟车效果最优.

如图 10 所示, 展示了车队在变曲率匝道的进入阶段到驶出阶段整个过程中, 六种算法控制下各车速度变化. 其中在从变曲率道路的进入阶段到保持阶段 (仿真步数 50-250) 中, 不同控制算法下各车的速度变化对比如图 11 所示.

通过计算该过程中各车速度均方根误差值, 可以得出 SR-TD3 算法在过渡阶段领航车的速度稳定性相较与 PS-MACDDPG 算法、MADDPG 算法、



(a) SR-TD3速度曲线　(d) CACC-LKA速度曲线

(b) PS-MACDDPG速度曲线　(e) CACC-TD3速度曲线

(c) MADDPG速度曲线　(f) MPC-TD3速度曲线

图10　不同算法下行驶速度变化曲线图

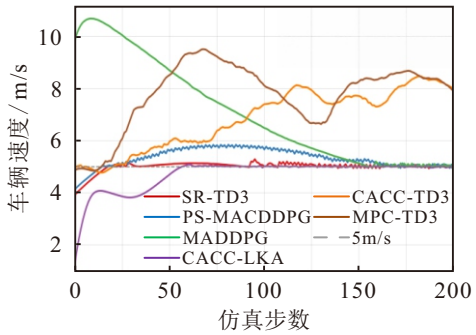
CACC-LKA 算法、CACC-TD3 算法和 MPC-TD3 算法分别提升了 59.96%、93.02%、70.19%、90.87% 和 93.53%; 跟驰车 1 的速度稳定性分别提升了 43.82%、86.93%、66.52%、75.59% 和 79.54%; 跟驰车 2 速度稳定性分别提升了 30.23%、81.26%、44.26%、31.12% 和 65.51%. 对于整个车队的整体速度稳定性上分别提升了 41.21%、86.53%、58.25%、73.08% 和 81.60%. 安全性指标和组队稳定性也是常用于衡量车队控制性能的两个重要指标, 安全性指标最小碰撞时间 (Minimum Time-to-Collision,Min TTC) 反映了车辆在当前相对速度下的碰撞风险; 组队稳定性用于衡量车队在行驶过程中抑制交通扰动传播的能力[23]. 表 2 分别从车队的安全指标和组队稳定性上, 对六种算法在该阶段对车队的控制性能进行了比较, 基于车辆动力学和人类反应特性的综合考量 TTC<1.5s 视为危险; 组队稳定性由加速度均方根比值得到, 在 Ratio>1.0 时表示车队不具有稳定性. 从表 2 中可以看出由 SR-TD3 算法所控制的车队在行驶时, 性能不但同时满足安全指标和组队稳定性, 且在组队稳定性上相较于其他五种算法有明显的领先效果. 从各种实验结果的对比分析表明在连续曲率道路中,SR-TD3 算法对车队的控制性能优于其他五种算法.

表2　车队性能比较

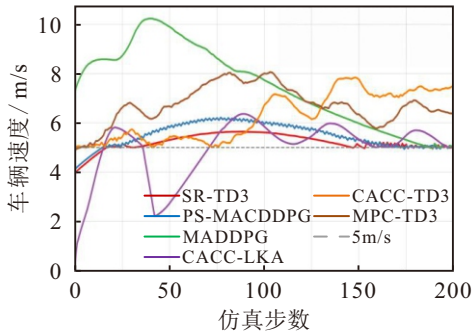
控制方法	安全指标 (Min TTC)(秒)		是否安全	组队稳定性 (加速度RMS比率)		是否稳定
	跟驰车1	跟驰车2		跟驰车1/领航车	跟驰车2/跟驰车1	
SR-TD3	6.03	3.32	√	0.88	0.67	√
PS-MACDDPG	>10	2.85	√	0.96	1.01	—
MADDPG	5.80	0.45	—	1.02	1.47	—
CACC-LKA	>10	1.91	√	3.36	1.13	—
CACC-TD3	9.62	2.86	√	1.42	0.87	√
MPC-TD3	7.04	2.27	√	0.91	1.06	—

4　结　论

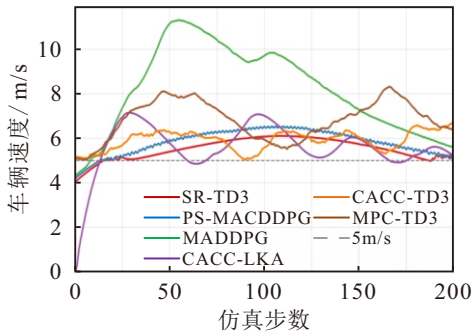
本文围绕变曲率匝道这一复杂场景下的车队协



(a) 领航车速度变化曲线对比



(b) 跟驰车1速度变化曲线对比



(c) 跟驰车2速度变化曲线对比

图11　阶段切换过程中各车速度对比图

同控制问题, 提出并验证了一种基于分阶段奖励优化的改进型 TD3 算法 (SR-TD3). 通过在进入、保持与驶出三个阶段分别构建差异化奖励函数, 并结合优先经验回放与动作平滑约束, 本文方法有效提升了车队在高曲率道路中的控制稳定性和收敛效率. 实验结果表明,SR-TD3 相较于 PS-MACDDPG、MADDPG、CACC-LKA、CACC-TD3 和 MPC-TD3 在奖励收敛速度、速度变化平稳性以及车间距控制方面均表现出更优的性能, 能够在较少训练回合内收敛至全局最优策略, 并显著降低速度和车距波动, 保证了车队的安全性与舒适性.

参考文献 (References)

[1] Liu W, Hua M, Deng Z Y, et al. A systematic survey of control techniques and applications in connected and automated vehicles[J]. IEEE Internet of Things Journal, 2023, 10(24): 21892-21916.

[2] 徐进, 崔强, 林伟, 等. 螺旋匝道和螺旋桥的小客车行驶速度特性[J]. 中国公路学报., 2019, 32(7): 158-171.
(Xu J, Cui Q, Lin W, et al. Speed behavior of passenger cars on helical ramps and helical bridges[J]. China Journal of Highway and Transport, 2019, 32(7): 158-171.)

[3] Zhao J, Wang Z G, Lv Y F, et al. Robust optimal prescribed performance control of adaptive cruise control systems with unknown dynamics[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(4): 4757-4769.

[4] Ma G Q, Pagilla P R, Darbha S. Robust cooperative adaptive cruise control system design: Trade-off between parasitic actuation lag and communication delay[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(6): 7980-7989.

[5] 宋秀兰, 陈雨, 陈新, 等. 联合通信资源分配的网联车协同自适应巡航时滞反馈控制[J]. 控制与决策, 2023, 38(10): 2888-2896.
(Song X L, Chen Y, Chen X, et al. Joint communication resource allocation and cooperative adaptive cruise delay-feedback control of connected vehicles[J]. Control and Decision, 2023, 38(10): 2888-2896.)

[6] Gao Z B, Wu Z Z, Hao W, et al. Optimal trajectory planning of connected and automated vehicles at on-ramp merging area[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 12675-12687.

[7] Sun K X, Chen Y Q, Nur A M, et al. Fuzzy sliding mode control for improving the handling stability of intelligent vehicles with 4WS[C]. 2025 44th Chinese Control Conference. Chongqing: IEEE, 2025: 355-360.

[8] Du G D, Zou Y, Zhang X D, et al. Efficient motion control for heterogeneous autonomous vehicle platoon using multilayer predictive control framework[J]. IEEE Internet of Things Journal, 2024, 11(23): 38273-38290.

[9] Bao H Q, Kang Q, Shi X D, et al. Robust learning-based model predictive control for intelligent vehicles with unknown dynamics and unbounded disturbances[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(2): 3409-3421.

[10] Du H X, Chen J. Reinforcement learning-based autonomous driving path planning for smart connected vehicles[C]. 2024 International Conference on Power, Electrical Engineering, Electronics and Control. Athens: IEEE, 2024: 981-985.

[11] Wang W B, Hui F, Zhang J F, et al. Deep reinforcement learning method for trajectory planning of connected and autonomous vehicles in the round about lane-changing scenario[C]. 2024 4th International Symposium on Computer Technology and Information Science. Xi'an: IEEE, 2024: 168-173.

[12] 王云泽, 孙宇, 骆中斌, 等. 基于深度强化学习的自动驾驶行为决策研究综述[J]. 控制与决策, 2026, 41(2): 305-328.
(Wang Y Z, Sun Y, Luo Z B, et al. Review of autonomous driving behavior decision-making based on deep reinforcement learning[J]. Control and Decision, 2026, 41(2): 305-328.)

[13] Tiong T, Saad I, Teo K T K, et al. Autonomous vehicle driving path control with deep reinforcement learning[C]. 2023 IEEE 13th Annual Computing and Communication Workshop and Conference. Las Vegas: IEEE, 2023: 0084-0092.

[14] Chen J Z, Wu X B, Lv Z K, et al. Collaborative control of vehicle platoon based on deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2024, 73(10): 14399-14414.

[15] 邓小豪, 侯进, 谭光鸿, 等. 基于强化学习的多目标车辆跟随决策算法[J]. 控制与决策, 2021, 36(10): 2497-2503.
(Deng X H, Hou J, Tan G H, et al. Multi-objective vehicle following decision algorithm based on reinforcement learning[J]. Control and Decision, 2021, 36(10): 2497-2503.)

[16] Peng J K, Zhang S Y, Zhou Y, et al. An integrated model for autonomous speed and lane change decision-making based on deep reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 21848-21860.

[17] 李永福, 周发涛, 黄龙旺, 等. 基于深度强化学习的网联车辆队列纵向控制[J]. 控制与决策, 2024, 39(6): 1879-1887.
(Li Y F, Zhou F T, Huang L W, et al. Longitudinal control of connected vehicle platoons based on deep reinforcement learning[J]. Control and Decision, 2024, 39(6): 1879-1887.)

[18] Kong H, Xing Q L, Wang Q, et al. ADAC: Actor-double-attention-critic for multi-agent cooperation in mixed cooperative-competitive environments[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(7): 9579-9592.

[19] 陈亮, 梁宸, 张景异, 等. Actor-Critic 框架下一种基于改进 DDPG 的多智能体强化学习算法[J]. 控制与决策, 2021, 36(1): 75-82.
(Chen L, Liang C, Zhang J Y, et al. A multi-agent reinforcement learning algorithm based on improved DDPG in Actor-Critic framework[J]. Control and Decision, 2021, 36(1): 75-82.)

[20] Ji H N, Du J, Chen R. Research on collaborative control algorithms for vehicle platoons based on PS-MACDDPG[C]. 2024 5th International Conference on Artificial Intelligence and Electromechanical Automation. Shenzhen: IEEE, 2024: 1089-1094.

[21] Hao J Y, Yang T P, Tang H Y, et al. Exploration in deep reinforcement learning: From single-agent to multiagent domain[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 8762-8782.

[22] Farag W. Multi-agent reinforcement learning using the deep distributed distributional deterministic policy gradients algorithm[C]. 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies. Sakheer: IEEE, 2020: 1-6.

[23] Martínez-Díaz M, Al-Haddad C, Soriguera F, et al. Impacts of platooning of connected automated vehicles on highways[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(7): 6366-6396.

作者简介

靳双 (1992−), 男, 博士, 硕士生导师, 主要研究方向为智能网联汽车协同控制与智慧交通, E-mail: jsfj@cqjtu.edu.cn;

刘安龙 (2000−), 男, 硕士生, 主要研究方向为车路协同/车车协同与自动驾驶、交通信息与控制, E-mail: m17823575216@163.com;;

赵杭 (1993−), 男, 副教授, 硕士生导师, 主要研究方向为带状态约束的车辆编队控制、网联环境下的车辆跟驰模型、基于人工智能的智能网联汽车控制, E-mail: zhaohang@cqupt.edu.cn;;

吴仕勋 (1983−), 男, 副教授, 硕士生导师, 主要研究方向为无线定位技术、无线通信、人工智能, E-mail: wushixun333@163.com.