控制与决策
Control and Decision

基于深度强化学习的自动驾驶行为决策研究综述

王云泽, 孙宇, 骆中斌, 张春波

引用本文:
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.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0441

基于深度强化学习的自动驾驶行为决策研究综述

王云泽[1,2†]，孙 宇[1,2]，骆中斌[3,4,5]，张春波[1,2]

(1. 石家庄铁道大学 交通运输学院，石家庄 050043；
2. 河北省交通安全与控制重点实验室，石家庄 050043；3. 重庆大学 计算机学院，重庆 400044；
4. 招商局重庆交通科研设计院有限公司，重庆 400067；5. 自动驾驶技术交通运输行业研发中心, 重庆 400067)

摘 要: 自动驾驶行为决策是车辆实现智能化的核心技术, 深度强化学习 (DRL) 因其环境交互特性和端到端决策优势在该领域展现出显著潜力. 鉴于此, 通过多维度分析, 系统梳理基于 DRL 的自动驾驶行为决策的研究内容和发展趋势: 首先, 回顾行为决策的发展历程, 并分析 DRL 在自动驾驶中的应用趋势; 然后, 提出“状态-动作-奖励-策略-评价”五维框架, 分析算法要素与跟驰、换道等驾驶任务的映射关系; 接着, 结合匝道合流、交叉口和施工区等典型场景, 剖析 DRL 在不确定性环境中的应用方案; 最后, 指出多车协同、长尾事件及可解释性等挑战, 并提出未来研究方向. 研究表明: 技术上, DRL 算法选择与优化日趋多元化, 模型向多模态、轻量化发展; 应用上, 决策范式正从单车智能向车路云协同升级, 从功能实现向人性化交互进化, 突破现有技术“算法创新-硬件加速-法规适配”的协同演进路径.

关键词: 汽车工程；自动驾驶；深度强化学习；行为决策；综述；典型交通场景；多车协同

中图分类号: U471.1 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0441

引用格式: 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.

Review of autonomous driving behavior decision-making based on deep reinforcement learning

WANG Yun-ze[1,2†]，*SUN Yu*[1,2]，*LUO Zhong-bin*[3,4,5]，*ZHANG Chun-bo*[1,2]

(1. College of Traffic and Transportation，Shijiazhuang Tiedao University，Shijiazhuang 050043，China；2. Hebei Key Laboratory of Traffic Safety and Control，Shijiazhuang 050043，China；3. College of Computer Science，Chongqing University，Chongqing 400044，China；4. China Merchants Chongqing Communications Technology Research & Design Institute Co.，Ltd.，Chongqing 400067，China；5. Transportation Industry R&D Center of Autonomous Driving Technology，Chongqing 400067，China)

Abstract: Behavior decision-making is a core technology for vehicle intelligence. Deep reinforcement learning (DRL), with its environment-interactive capability and end-to-end decision-making advantages, has shown great potential in this field. This paper conducts a multidimensional analysis and systematically reviews the core content and development trends of DRL-based autonomous driving behavior decision-making research. First, the development of behavioral decision-making is reviewed, and the application trends of DRL in autonomous driving is analyzed. Second, a five-dimensional framework “state-action-reward-policy-evaluation” is proposed to analyze the mapping between algorithmic components and driving tasks such as car-following and lane-changing. Third, application schemes of DRL in uncertain environments are examined through typical traffic scenarios including ramp merging, intersections, and construction zones. Finally, we identify key challenges such as multi-vehicle coordination, long-tail event handling, and algorithm interpretability, and suggest future research directions. The study shows that, technically, DRL algorithm selection and optimization are becoming more diverse, with models evolving toward multi-modal and lightweight designs. In terms of application paradigms, behavior decision-making is transitioning from single-vehicle intelligence to vehicle-road-cloud collaboration, and from function-driven implementation to human-centric interaction. Overcoming current technical bottlenecks requires a co-evolution path of algorithm innovation, hardware acceleration, and

收稿日期: 2025-04-25；**录用日期:** 2025-08-14.

基金项目: 国家重点研发计划项目 (2023YFB2504704)；2024 年度河北省社会科学发展研究课题 (202402302)；河北省省级科技计划软科学研究专项资助项目 (25350801D)；河北省自然科学基金项目 (E2024210032).

†通信作者. E-mail: wangyunze@stdu.edu.cn.

regulatory adaptation.

Keywords: automotive engineering; autonomous driving technology; deep reinforcement learning; behavioral decision-making; review; typical traffic scenarios; multi-vehicle collaboration

0 引 言

在自动驾驶系统中, 行为决策系统扮演着“大脑中枢”的核心角色, 负责整合来自环境感知层收集的精确数据, 依据驾驶任务和控制目标生成决策指令, 并将其传递至控制执行层, 从而确保智能化操作的高效实现[1]. 一个可靠的行为决策系统不仅能够提升自动驾驶车辆的智能性、安全性、经济性和舒适性, 还可以增强乘驾人员的信任度、接受度和满意度, 同时促进交通系统的合法性、协调性与高效性[2].

现阶段研究中, 自动驾驶行为决策可大致分为4种类型: 基于规则 (rules-based)[3]、基于优化理论 (optimization theory-based)[4]、基于控制理论 (control theory-based)[5] 和基于学习 (learning-based)[6] 的行为决策方法. 基于规则的方法通过预设规则库, 根据交通环境、法律法规和驾驶经验推理决策, 具有可解释性强和安全性高的优势, 但其对复杂、动态驾驶环境的适应性较差, 难以扩展至新场景[7]. 基于优化理论的方法将行为决策建模为数学优化问题, 通过构造目标函数和约束条件进行全局规划与路径优化, 表现出较高的效率和鲁棒性[8]. 然而, 该方法通常依赖于精确的环境建模, 在处理高维、非线性问题时计算复杂度较高, 且难以保证实时性. 基于控制理论的方法侧重于实时反馈控制和稳定性优化, 适用于短期动态调整场景, 如跟车和车道保持等. 尽管其能够提供快速的局部优化, 但缺乏全局视角和决策能力, 难以应对复杂的长时任务[9]. 相比之下, 基于学习的方法能够从大量驾驶数据中自动提取特征并优化驾驶策略, 高效地处理复杂任务, 快速响应并适应高维环境变化, 表现出极强的灵活性和适应性.

在基于学习的方法中, 强化学习 (reinforcement learning, RL) 因其通过环境交互和试错机制优化策略的特点, 成为近年来自动驾驶行为决策研究的热点. 然而, 传统强化学习在应对高维复杂动态环境时仍受到学习效率和表征能力的限制[10]. 为克服这些挑战, 深度强化学习 (deep reinforcement learning, DRL) 通过结合深度学习的感知能力和强化学习的决策能力, 为自动驾驶行为决策提供了更加高效和灵活的解决方案[11]. 例如基于深度 Q 网络 (deep Q-network, DQN) 的高速公路换道模型[12], 基于深度确定性策略梯度 (deep deterministic policy gradient, DDPG) 的多信息融合行为决策方法[13], 以及结合循环近端策略优化算法 (recurrent proximal policy optimization, RPPO) 的智能决策系统[14] 等, 这些方法都在特定场景下解决了行为决策的实际问题.

尽管已有文献对基于RL的自动驾驶研究进行了综述, 但仍存在一定的局限性. 例如, 文献 [15] 系统地总结了不同类型的DRL算法在自动驾驶领域的应用及实际部署, 但在解析关键设计问题 (如奖励函数、评价指标等) 和算法多样化应用方面缺乏深入讨论, 未能全面覆盖复杂交通环境中的挑战; 文献 [16] 聚焦于最新RL算法在行为规划领域的进展和趋势, 但未能结合自动驾驶行为决策的核心任务进行系统分析; 文献 [17] 重点阐述了不同层次RL算法的应用技术路线, 却未深入分析具体交通场景中行为决策的挑战及优化方法; 文献 [18] 梳理了DRL在自动驾驶系统环境感知、决策规划和控制执行等关键技术领域的应用现状, 但对于决策问题的具体实现方法探讨不足. 综合来看, 现有综述文献多聚焦于RL算法、自动驾驶系统整体框架以及运动规划等领域, 较少从基于DRL的自动驾驶行为决策任务的关键维度, 如状态空间、动作空间、奖励函数、策略更新及评价指标等角度进行深入总结. 同时, 这些研究在理解、优化与应用等方面的系统性思考和对典型交通场景中实际应用指导作用也有所不足.

基于上述研究背景, 本文聚焦于基于DRL的自动驾驶行为决策问题, 针对现有综述的不足进行系统梳理与扩展. 主要贡献如下: 1) 全面详细地介绍DRL算法的基本原理及其在自动驾驶行为决策中的应用分类; 2) 揭示DRL在自动驾驶行为决策中的“离散动作决策→连续决策优化→多智能体协同”3阶段技术跃迁规律, 阐明算法演进与交通场景复杂度提升的协同关系; 3) 突破现有综述侧重算法和系统框架的局限, 从“状态空间-动作空间-奖励函数-策略更新-评价指标”五位一体的决策维度构建分析框架, 结合跟驰、换道、转弯等驾驶行为, 建立DRL算法与驾驶任务的映射关系; 4) 结合主线通行、匝道合流、交叉口及施工区等4类典型场景, 针对性地探讨不同交通场景下所面临的关键挑战及其对应的解决方案, 为不确定环境下的自主决策提供了技术路径; 5) 从多车协同控制、长尾事件处理、算法可解释性及伦理问题等角度系统梳理基于DRL自动驾驶行为决策所面临的核心挑战, 并展望未来研究的重点,

为理论创新与实际工程应用的结合提供重要参考.

1　自动驾驶行为决策概述

自动驾驶行为决策技术经历了从规则驱动到数据驱动, 再到大模型阶段的演进过程[19]. 随着计算能力增强、数据资源积累以及机器学习算法的发展, 系统决策能力不断提升, 实现了从封闭环境中的逻辑验证向开放环境下的自主学习和智能决策的跨越.

1.1　规则驱动与基础架构奠定期

在自动驾驶技术发展的早期阶段, 行为决策主要依赖于基于规则的算法和模型预测控制 (model predictive control, MPC) 方法[20]. 这些基于编码规则的系统虽能在已知场景中稳定运行, 但面对复杂和未知环境时, 缺乏灵活性与泛化能力. 早期典型架构由感知、规划和控制 3 大模块组成, 决策过程依赖预设逻辑判断. 例如, Waymo 初期结合高精地图和传感器融合技术, 通过传统方法完成基础驾驶任务[21].

波士顿动力公司团队在 DARPA 挑战赛中展示了集成机器学习与感知系统的自动驾驶原型, 尽管仍以规则为主, 但为数据驱动方法的兴起奠定了基础[22].

在此阶段, 自动驾驶系统多部署于封闭测试环境, 如 DARPA 赛道、Google 的仿真平台、CARLA 等, 为后续研究提供了验证平台[23-24].

1.2　机器学习与决策模型探索期

随着数据和计算能力的积累, 研究者开始应用机器学习方法优化决策策略. 此时, 隐马尔可夫模型、支持向量机和动态贝叶斯网络, 以及马尔可夫决策过程 (Markov decision process, MDP) 等浅层机器学习算法实现了从数据中自动学习有效策略, 逐步替代人工规则[25-27].

深度学习技术的应用显著提升了感知精度, 为高质量的决策输入奠定了基础. DRL 在该阶段初步应用于行为决策, 展现出较强的策略学习能力.

与此同时, 测试环境开始从封闭道路向真实城市道路拓展. 中国的百度 Apollo 测试场与上海示范区等引入复杂交通情景, 推动了该技术在真实环境中的落地[28-29].

1.3　大模型与综合系统集成期

近年来, 自动驾驶决策进入“大模型”阶段. 深度神经网络和 RL 模型广泛应用于提升系统全局理解能力与泛化性能. 离线强化学习 (offline RL) 通过历史数据训练, 显著降低了成本与风险[30-31]; 分层 RL、安全 RL 和联邦 RL 等方法进一步提升了系统鲁棒性与安全性[32-34].

尽管大模型具备强大能力, 但其高算力需求与训练成本仍是实际部署的障碍. 车载计算单元面临功耗、空间和实时性等限制, 制约了大模型的直接应用. 因此, 研究者正积极探索轻量化模型 (如网络剪枝、量化和知识蒸馏) 与边缘计算协同机制, 以提升资源受限条件下的部署效率[35]. 如何实现低延迟、高性能的快速推理, 成为后续发展的关键.

在仿真训练方面, CARLA、SUMO 和 PreScan 等平台广泛应用, 支持大规模、多样化场景测试. 自动驾驶测试区也持续引入更具挑战性的环境, 以提升算法的泛化与实际适应能力[36].

2　基于 DRL 的自动驾驶行为决策应用

2.1　DRL 算法概述

DRL 结合了 RL 的试错机制与深度神经网络的高维特征提取能力, 使智能体在与环境的交互中不断优化策略, 以实现复杂任务下的高效决策.

DRL 的基本框架基于 RL 中的 MDP, 主要包括环境 (environment, E)、状态空间 (state space, S)、动作空间 (action space, A) 和奖励函数 (reward function, R). 其中: S用于表征智能体所感知的环境信息, A定义可采取的行为范围, R用于衡量动作的优劣, E决定状态转移与反馈机制.

智能体依据当前状态选择动作, 环境返回新的状态与奖励, 目标是最大化长期累积奖励的期望值, 即

$$J(\pi)=\arg\max_{\pi}\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma r_t\right]. \tag{1}$$

其中: π为策略, $J(\pi)$为策略的目标函数; γ为折扣因子, 表示未来奖励的重要性; r_t为奖励函数在第t步提供的及时反馈.

DRL 的核心机制包括价值函数 (value function)、策略 (policy) 和贝尔曼方程 (Bellman equation).

价值函数用来评估状态或状态-动作对的长期收益, 常用形式包括状态价值函数和状态-动作价值函数, 计算公式为

$$V^{\pi}(s)=\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma r_t|s_0=s\right], \tag{2}$$

$$Q^{\pi}(s,a)=\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma r_t|s_0=s,a_0=a\right]. \tag{3}$$

其中: $V^{\pi}(s)$为状态价值函数, 表示从状态s开始, 智能体遵循策略π所能获得的期望长期奖励; $Q^{\pi}(s,a)$为状态-动作价值函数, 表示在状态s下执行动作a后, 遵循策略π所能获得的期望长期奖励; $s_0=s$表

示从状态s开始计算价值; $a_0 = a$表示从状态s开始并采取动作a后计算价值.

策略是从状态s到动作a的映射, 分为确定性策略和随机性策略. 确定性策略直接输出具体动作$a = \mu(s)$; 随机性策略输出一个动作分布$\pi(a|s)$, 通过采样决定最终动作. DRL 通过策略优化使智能体决策能力不断增强, 并最终实现最优行为决策.

贝尔曼方程是价值函数的递归定义, 也是 DRL 算法的理论基础, 计算公式为

$$V(s_t) = \mathbb{E}\Big[R(s_t, a_t) + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t)V(s_{t+1})\Big]. \tag{4}$$

其中: $V(s_t)$为当前状态价值函数; s_t、a_t、s_{t+1}、a_{t+1}分别为当前状态、当前动作、下一状态、下一动作; P为状态转移概率分布; $V(s_{t+1})$为下一状态的价值.

进一步地, 贝尔曼最优方程描述了最优策略下的价值函数特性, 计算公式为

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{s' \sim P}[r_t + \gamma \cdot \max_{a'} Q^{\pi}(s_{t+1}, a_{t+1})]. \tag{5}$$

DRL 算法可根据智能体数量和交互模式分为单智能体深度强化学习 (single-agent deep reinforcement learning, SADRL) 算法和多智能体深度强化学习 (multi-agent deep reinforcement learning, MADRL) 算法, 分类总体框架如图 1 所示.

1) SADRL.

SADRL 可以根据是否使用模型分为有模型、无模型和模仿学习 3 类[37].

基于模型的 DRL 算法通过学习一个模型来描述状态转换和奖励评估. 虽然这种方法显著提高了样本效率, 但由于无法预先建模所有未知情况, 构建有模型的控制过程仍然相对困难. 因此, 通常采用神经网络近似环境模型, 这导致模型对网络的依赖性较强. 代表算法包括 World Models、AlphaZero 和基于模型的策略优化 (model-based policy optimization, MBPO) 等.

无模型的 DRL 算法主要包括基于值函数、基于策略和基于动作-评价 (actor-critic, AC) 框架 3 类. 基于值函数的 DRL 算法首先评估值函数, 然后利用该值函数改善当前策略, 适用于离散的状态和动作空间. 代表算法有 Q-Learning 和 SARSA 等. 基于策略的方法无需显式估计每个状态-动作对的 Q 值, 而是通过估计策略函数的参数, 利用训练好的策略模型进行决策. 代表算法有策略梯度算法 (policy gradient, PG)、蒙特卡洛策略梯度算法 (Monte Carlo policy gradient) 等. 基于值和策略相结合的 DRL 算法通常称为 AC 方法, 该方法引入动作网络 (actor network) 和评判网络 (critic network), 结合了值函数与策略函数的特点, 既能学习最优策略, 又能估计最优值函数. AC 方法的优点在于能够同时学习值函数和策略函数, 从而实现更高效的学习和决策. 此外, AC 方法还能解决连续动作空间问题和处理噪声环

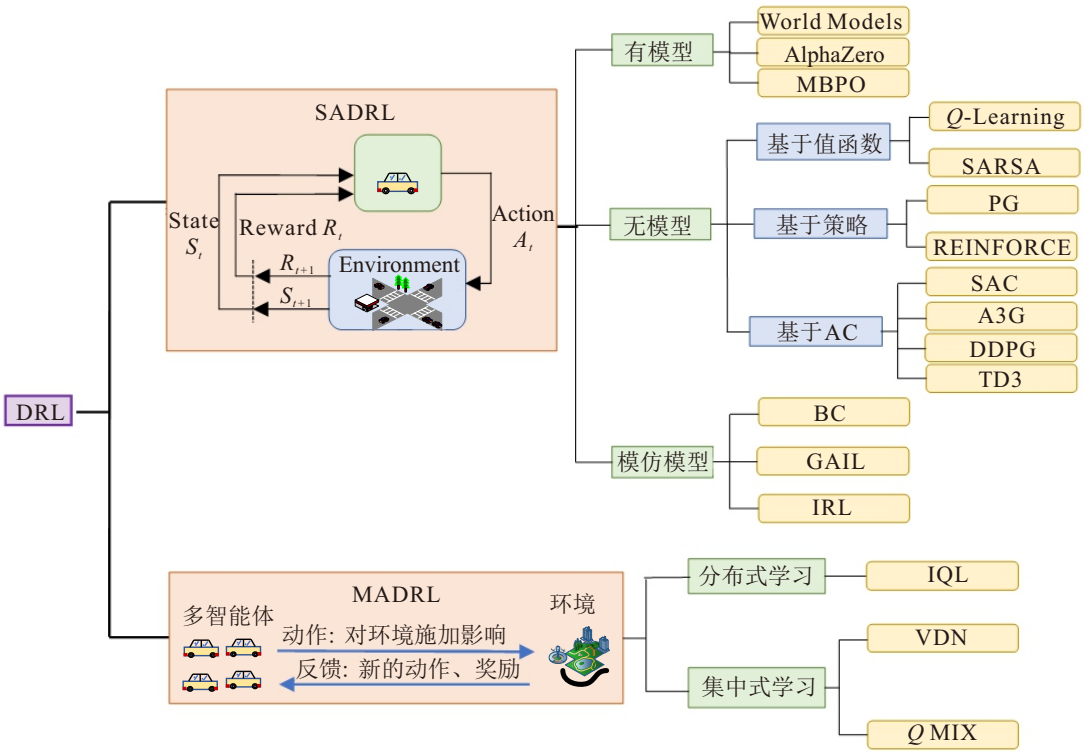

图1 DRL 算法的分类

境, 与传统 PG 方法相比, 其效率更高. 代表算法包括优势演员-评论家算法 (advantage actor-critic, A2C)、异步优势演员-评论家算法 (asynchronous advantage actor-critic, A3C)、DDPG、双延迟深度确定性策略梯度算法 (twin delayed deep deterministic policy gradient, TD3) 和软演员-评论家算法 (soft actor-critic, SAC) 等.

模仿学习 (imitation learning) 是一种重要的学习范式, 旨在通过模仿专家行为加速智能体的学习过程. 在模仿学习框架下, 专家提供一系列状态-动作对, 反映其在特定环境下的行为. 智能体 (模仿者) 利用这些数据进行训练, 以接近专家的策略水平, 无需依赖环境的奖励信号. 代表算法包括行为克隆 (behavioral cloning, BC)、生成对抗模仿学习 (generative adversarial imitation learning, GAIL) 和逆强化学习 (inverse reinforcement learning, IRL) 等.

2) MADRL.

MADRL 系统遵循马尔可夫博弈 (随机博弈) 过程. 在马尔可夫博弈中, 所有智能体根据当前环境状态同时选择动作. 这些动作构成的联合行为不仅会影响环境状态的转移和更新, 还会进一步影响智能体获得的奖励, 如图 2 所示. 这种通过多个智能体相互作用共同完成复杂任务的能力, 使 MADRL 成为解决自动驾驶中多车协同决策问题的有效方法.

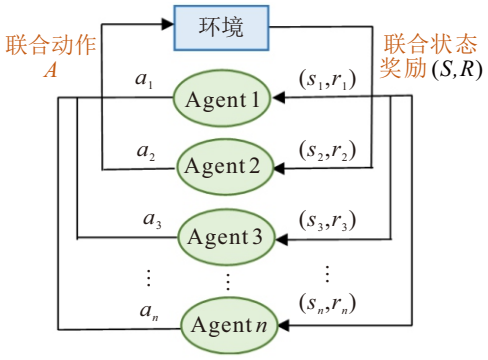

图2　多智能体的马尔可夫决策过程

在学习方式上, MADRL 可分为集中学习与分布式学习, 适用于不同的任务需求. 集中学习依赖中心控制器处理所有智能体的决策, 适合在训练阶段进行全局优化. 典型的集中学习算法包括 Q-mix 和值分解网络 (value decomposition networks, VDN). 这种方法可以从全局视角优化系统目标, 更容易获得整体最优解. 然而, 集中学习需要大量通信资源, 难以满足实时性要求, 在实际应用场景中扩展性也较差. 分布式学习则让每个智能体独立学习和决策, 特别适合实时任务场景. 这种方法无需中心控制器, 减少了通信需求并提高了系统的实时性和扩展性. 以独立 Q 学习 (independent Q-learning, IQL) 为代表的分布式算法, 能够在动态多变的交通环境中快速调整智能体的决策策略. 分布式学习的缺点在于可能导致局部最优解, 难以保证全局最优.

2.2　基于 DRL 的自动驾驶行为决策发展历程

自动驾驶行为决策技术的智能化进程与 DRL 的发展呈现出高度耦合的态势. DRL 在自动驾驶中的应用经历了从离散决策问题到连续动作决策, 再到多智能体协同决策的逐步演进, 如图 3 所示.

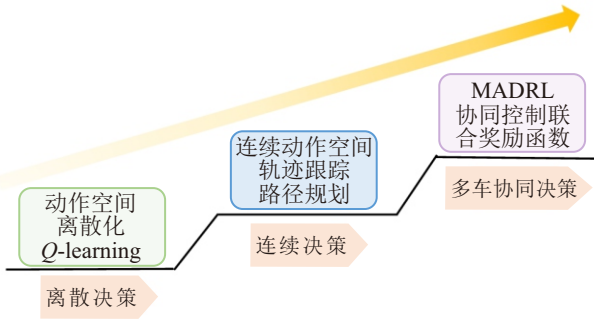

图3　基于 DRL 的自动驾驶行为决策发展历程

2.2.1　离散决策

2015 年, Google DeepMind 提出了 DQN, 通过神经网络逼近状态-动作值函数 $Q(s,a)$, 在 Atari 游戏环境中成功实现了从高维感知输入到动作输出的端到端学习[38]. 随后, DQN 被应用于 TORCS、CarSim 等仿真平台, 在车道保持和静态障碍物规避等结构化场景中验证了其有效性[39]. 这些成果展示了 DRL 在自动驾驶决策中的潜力.

然而, DQN 算法存在两个主要问题: 一是只适用于离散动作空间, 难以解决实际场景中的连续控制问题; 二是样本效率较低, 对训练数据依赖性强, 且在泛化性和适应性方面存在不足.

2.2.2　连续决策

为了克服离散算法的局限性, 研究者们提出了基于 AC 框架的算法, 推动了 DRL 在自动驾驶连续控制任务中的应用. AC 框架结合了值函数和策略函数的优点, 采用两个独立网络: Actor 网络输出连续的动作策略 $\pi_\theta(a|s)$; Critic 网络评估状态-动作对的价值 $Q(s,a)$, 提供策略的更新反馈信号. 在该框架的基础上, DDPG、近端策略优化算法 (proximal policy optimization, PPO)、TD3 和 SAC 等经典算法被相继提出, 其特性总结如表 1 所示. 这些算法显著提升了 DRL 在自动驾驶连续决策任务中的表现, 广泛应用于变道、转弯、避障和路径规划等连续决策任务[44-46].

表1 典型算法总结

算法名称	特点	优点	不足	发表年份
DDPG	采用AC网络; 双层网络更新; 引入噪声增加随机性	可以处理连续动作空间问题; 收敛速度快	收敛稳定性不足; 对神经网络的参数敏感; 会产生Q值过估计问题	2015[40]
PPO	使用随机梯度上升方法优化	样本利用效率高; 训练稳定性强; 适用性广泛; 简单易实现	对样本要求高, 容易陷入局部最优解	2017[41]
TD3	利用双层目标函数; 对策略进行延迟更新	解决了Q值过估计问题; 降低了高方差的影响	目标网络较复杂; 训练过程较长	2018[42]
SAC	结合了策略梯度和价值函数; 支持多个代理同时进行; 同步更新	学习过程快; 泛化能力强; 稳定性和效率高; 适用于连续动作和复杂状态的任务	梯度估计可能存在较大方差	2018[43]

2.2.3 向多智能体协同决策的演进

在真实交通环境中, 单车智能已难以满足多车交互与协同需求, 因此, MADRL 成为研究热点. 该类方法支持多车间策略博弈与合作, 通过建模复杂交互关系, 实现系统级最优.

典型算法如 MADDPG、MAPPO 等, 广泛应用于交叉路口通行、车队编组等场景, 表现出优于单智能体的协调能力[47-48]. 在架构上, “集中训练、分布执行”成为主流范式, 有效解决了环境非平稳性带来的策略震荡问题.

此外, 车路协同感知与通信能力不断增强, 为多智能体之间的信息共享提供了支撑, 进一步推动了从“单车智能”向“群体智能”的跃迁[49].

尽管 MADRL 为自动驾驶中的多车交互提供了有效的解决框架, 但仍面临通信成本高、通信延迟、全局最优解难以保证以及系统鲁棒性不足等挑战. 当前的研究主要通过引入延迟扰动和延迟补偿等策略来减轻延迟带来的负面影响[50].

2.3 DRL 在自动驾驶中的广泛应用与研究趋势

DRL 凭借其在复杂高维环境中的信息处理和自主学习能力, 已在 Waymo、特斯拉等企业的自动驾驶系统中得到实际应用, 展现出良好的环境适应能力和策略优化能力[51-52].

基于 DRL 的自动驾驶行为决策也在学术领域引起了广泛关注. 通过在中国知网 (CNKI) 和 Web of Science 等学术平台上以“深度强化学习”和“自动驾驶”为关键词进行检索, 提取了自 2019 年以来的相关研究文献, 并生成词云图, 如图 4 所示.

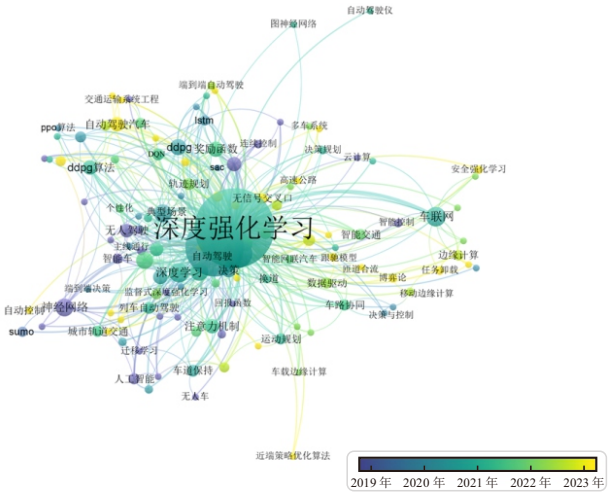

图4 DRL 在自动驾驶决策领域的研究范畴

从图 4 中可以看出, 自动驾驶领域的研究主要集中在以下几个方面:

1) 算法融合与演化.

DQN、PPO、DDPG、SAC 等基础 DRL 算法仍然是自动驾驶行为决策研究的重点. 但近年来, 研究更加关注不同范式的融合与演化. 例如, 引入 Transformer 机制构建序列决策模型, 提升多车交互处理能力; 使用联邦深度强化学习 (federated DRL) 在保护隐私前提下提升泛化性, 更好地适应车路协同场景的需求[53-54].

2) 任务复杂化与系统协同.

在传统任务方面, 像“跟驰模型”“转弯”和“换道”等任务的优化仍然在持续进行. 同时, 研究的重点逐步向系统协同决策拓展. 端到端决策框架的研究逐渐深化, 通过直接将多传感器感知信息映射到控制指令上, 从而简化了传统“感知-规划-控制”分层架构中的冗余性. 这不仅提高了决策的效率, 而且加强了决策的实时性. 此外, 车路协同数据的融合使得自动驾驶系统能够在全局轨迹规划与局部动态避障之间实现协同决策, 从而显著提升复杂场景 (如无信号交叉口、匝道合流等) 的通行效率与安全性[55].

3) 交通场景多样化应用.

随着算法的不断优化, 研究的重心逐步从结构化交通场景 (如高速公路) 拓展到城市复杂路口和长尾异常场景等非结构化场景. 现实交通环境中普遍存在着“长尾事件”, 即虽在统计上出现概率极低、但一旦发生将对安全构成严重威胁的极端情况, 例如行人突然横穿马路、车辆违规驶入施工区域、极端天气下传感器失效等. 由于此类事件在训练数据中极为稀缺, 导致基于 DRL 的决策模型在应对此类情景时存在泛化能力不足、鲁棒性差等问题. 为提升模型应对能力, 研究者已尝试引入极端情境模拟、风险敏感型奖励函数设计及模仿学习等手段. 但如何系

统提升 DRL 模型在长尾事件下的响应能力, 仍是当前和未来的关键研究方向之一.

4) 智能体集成化发展.

自动驾驶的研究也从单一智能体决策发展到多智能体协同和车路云协同的集成. 在这一过程中, MADRL 成为研究的热点. 通过优化车与车、车与路侧设施之间的策略交互, 有效提升了交通流的全局效率 (如高速公路车队行驶、路口冲突消解等)[56].

3　基于 DRL 的自动驾驶行为决策方法

基于 DRL 的自动驾驶行为决策, 其本质是将高维动态交通环境信息 (如自车状态、周围车辆参数、障碍物分布及道路结构) 映射为高效策略, 并通过奖励函数与策略搜索机制, 实现在复杂场景下的实时最优行为决策. 其核心理念围绕“状态-动作-奖励-策略-评价”五元设计框架的闭环优化展开, 通过端到端的交互学习, 突破传统规则驱动方法的泛化性瓶颈.

在自动驾驶任务中, 跟驰、换道和转弯是最基本且广泛适用的驾驶行为, 涉及纵向控制、横向决策及动态交互的综合处理, 几乎覆盖了大多数基础驾驶场景. 在高速公路和城市道路上, 车辆需频繁地完成跟驰、换道和转弯操作, 以实现高效和平稳的驾驶体验[57].

跟驰行为可进一步扩展为紧急刹车、加速追赶等任务, 旨在通过精准的纵向控制提升安全性和响应能力[58-60]. 换道行为则需应对并道、横向跟踪及紧急避障等场景, 要求系统在动态多目标交互中实现准确高效的横纵向联合决策[61-62]. 而转弯行为广泛应用于急弯、匝道和交叉路口等高风险场景, 其决策质量直接影响车辆运行效率和行车安全[63]. 因此, 本文从这 3 类行为出发, 系统探讨 DRL 在状态空间、动作空间、奖励函数、策略优化和评价指标等方面的设计方法.

3.1　状态空间

状态空间是自动驾驶车辆感知环境、做出决策的基础, 其建构直接影响系统的泛化能力和精度表现. 理想的状态空间应全面描述车辆状态、道路环境与任务要素, 遵循“动态表征-交互建模-场景解耦”的设计原则.

3.1.1　跟驰行为

跟驰行为决策常见于交通量较大的单车道纵向控制之中, 目标是通过调整车辆速度, 使之与前车保持安全距离, 从而优化行驶效率. 因此, 在跟驰场景中, 传统状态空间通常关注车辆绝对状态信息, 包括速度、加速度及前后车距离[64]. 这一方法虽能保障基础控制稳定性, 却难以应对复杂交通流的动态扰动.

为了提高建模能力, 状态空间架构经历了纵向控制到多目标协同优化的过程: 最初, 通过引入相对运动特征 (速度差、车身占比等), 建立车辆间交互的微分几何描述[65]; 之后, 开始整合道路摩擦系数、坡度倾角等环境耦合参数, 构建基于物理约束的状态空间[66]; 现在, 许多研究将能源状态 (电池荷电量、等效氢耗等) 与驾驶舒适度 (加速度、震动频谱) 纳入特征空间, 逐渐形成“能量-效率-舒适性”的多目标协同优化范式[67-68], 为高效自动驾驶决策奠定了坚实基础. 跟驰行为状态表示如图 5 所示.

3.1.2　换道行为

换道行为不仅涉及纵向控制, 还需要精确处理横向轨迹控制. 因此, 其状态空间的设计必须全面刻画车辆与周围环境的动态交互关系, 涵盖自车、前车、目标车道及相邻车道的车辆动态信息.

传统方法多采用相对位置、相对速度以及车道密度等变量作为核心状态特征, 同时引入最小安全距离和车头时距等指标以强化对安全性的描述[69]. 这些方法计算效率高, 在一定程度上满足了自动驾驶的基本需求, 但由于其主要聚焦于静态变量及简单动态关系, 难以精准表达复杂换道场景中的横向运动及多目标优化需求.

为弥补这一缺陷, 相关研究引入了轨迹预测方法, 通过将横摆角速度偏差、横向误差、路径曲率等信息纳入状态空间, 以增强对换道过程中车辆运动轨迹的精准描述[70]. 这一改进不仅提升了换道决策的鲁棒性, 还能使自动驾驶车辆在高动态环境下做出更加灵活且安全的换道决策. 状态空间表示如图 6 所示.

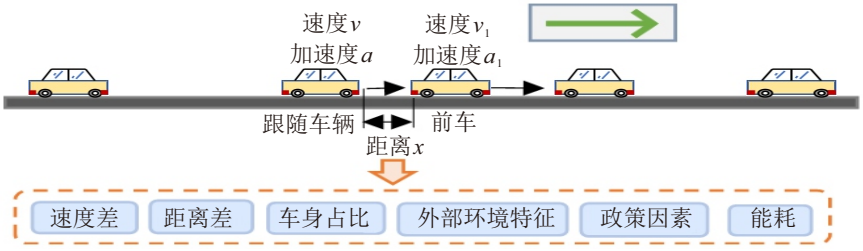

图5　跟驰行为状态表示

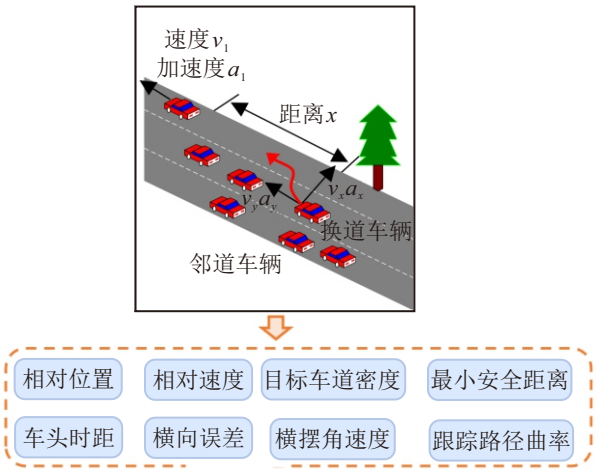

图6 换道行为状态表示

3.1.3 转弯行为

转弯行为需要在单车道范围内综合考虑纵向和横向运动, 以确保车辆能够安全、平稳地通过弯道. 由于转弯场景涉及复杂的道路几何特性, 状态空间的构建不仅需要关注车辆自身的动态信息, 还需精准表征弯道路况、曲率及坡度等因素.

传统状态空间如图 7 所示, 主要包括横向位置、纵向位置、线速度、加速度、横摆角速度及方向盘角度, 这一设计物理意义明确, 计算量较小, 适用于大部分弯道路况[71]. 然而, 这种设计对道路几何结构等静态特征的描述能力较弱, 限制了自动驾驶车辆在复杂场景中的适应性和鲁棒性.

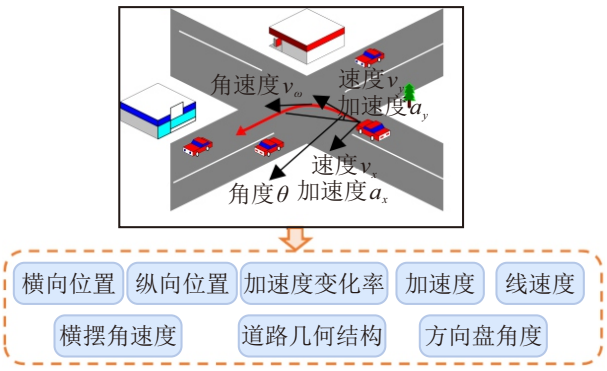

图7 转弯行为状态表示

为应对这一挑战, 相关研究提出了基于场景向量化的状态表示方法, 通过高效向量化表达所有转弯车辆的状态, 提升对复杂场景的描述能力[72]. 同时, 针对道路几何特性描述能力的不足, 一些研究采用了网格化交通状态编码的方法, 将转弯场景划分为固定长度的网格, 并利用布尔值标注网格内的车辆存在情况[73]. 这种网格化方法不仅能够精确捕捉转弯车辆的高分辨率动态状态, 还能有效降低模型的计算复杂度, 使自动驾驶系统能够更高效地处理非结构化道路环境.

3.2 动作空间

动作空间是车辆在特定任务中可执行的操作集合, 其设计直接影响驾驶策略的精度、车辆控制的平滑性及算法的学习效率. 根据动作空间的离散化程度, 通常分为离散型和连续型两类, 分别适用于不同的任务场景.

3.2.1 跟驰行为

传统跟驰决策采用离散型动作空间, 通过固定加速度级别调整车速, 计算简单, 但控制精度较低, 易导致行驶不平稳. 相比之下, 连续型动作空间允许车辆输出在特定范围内的任意加速度值, 能够实现更平滑的加速和减速控制, 显著提升驾驶体验和跟驰行为的稳定性[74]. 然而, 连续动作空间具有无限维性, 增加了动作选择的复杂性与学习难度. 为综合离散与连续型动作空间的优势, 相关研究提出了一种混合动作设计的方法, 该方法结合了离散目标速度选择与连续控制微调, 能够兼顾决策稳定性和精细化控制[75-76].

3.2.2 换道行为

换道行为对动作空间的灵活性和可控性提出了更高的要求. 其动作空间的设计不仅需要同时满足横向与纵向动作的需求, 还需综合考虑车道边界、目标车道的可行性以及周围车辆的动态信息[77].

离散型动作空间通常包含有限的换道选项, 并将纵向加速度离散为固定值, 作为附加维度与换道选项组合形成有限动作集合[78]. 这一设计具有实现简单、符合驾驶逻辑以及便于策略学习等优点, 但适应性较弱, 难以处理复杂的交通事件. 连续型动作空间通过直接输出横向与纵向加速度, 实现更平滑的换道轨迹, 适用于高精度运动规划与轨迹预测[78]. 然而, 单纯解耦横纵向加速度可能导致控制协同性不足, 影响换道稳定性.

为进一步优化换道过程, 当前研究通过 3 个方面进行改进: 1) 使用加速度值与方向角结合的方法, 使轨迹更流畅, 同时提高换道成功率; 2) 考虑车辆的动力学特性, 结合运动规划模型, 使换道过程更加贴近真实驾驶行为; 3) 进一步融合油门、刹车与转向控制, 以提升换道过程中对车速与轨迹的精细化调节能力[79-80].

3.2.3 转弯行为

转弯行为的核心目标是实现横向和纵向的精准路径跟踪. 传统方法采用离散型动作空间, 通常将角速度或方向盘转角离散化, 使车辆能够以固定步长

调整转向角度. 这种方法计算量低, 易于收敛, 但由于无法动态调整角速度, 容易导致转弯轨迹生硬或不够平滑, 在高速转弯或复杂弯道环境下表现不佳[81]. 为提高轨迹平滑性, 连续型动作空间逐渐成为主流, 例如直接输出方向盘转角或前轮转角, 以精确控制横向轨迹[82]. 这种设计不仅贴近实际驾驶行为, 还能有效提高转弯轨迹的平滑性, 但同样会增加计算复杂性.

3.3 奖励函数

奖励函数是 DRL 算法中衡量智能体行为优劣的核心机制, 对指导策略学习和优化智能体行为起着至关重要的作用. 由于奖励函数的设计原则和策略具有通用性, 可广泛适用于跟驰、换道和转弯等多种场景, 因此本节未对不同行为进行单独讨论.

3.3.1 设计原则

奖励函数的设计不仅需要优化单一驾驶行为, 还需考虑系统整体性能的全面性. 理想的奖励函数应能够帮助自动驾驶系统在多变的交通环境中实现安全、平稳、高效的驾驶体验. 其设计原则包括以下几点:

1) 同时满足可行性和最优性要求[83]. 可行性保证了算法能够在复杂的现实环境中找到可行解, 而最优性则旨在寻求综合约束条件下性能最优的方案.

2) 综合考虑自动驾驶的具体任务目标与环境约束. 无论是在跟驰、换道还是转弯等行为, 奖励函数的设计都需兼顾安全性、效率、舒适性、节能以及法规遵循等多项目标需求.

3) 综合考虑局部最优和长期收益. 由于长期收益会强化对全局最优解的探索能力[84], 奖励函数的设计需要在短期收益与长期目标之间实现平衡, 避免因过度优化局部指标而导致整体目标偏离或智能体陷入局部最优解.

4) 考虑伦理与法规合规性. 奖励函数不仅需要满足技术性能上的要求, 还应符合相关法律法规和社会伦理. 例如, 德国的自动驾驶法案规定了自动驾驶车辆必须具备高度自动化水平, 并且在出现紧急情况时必须能够保护乘客和其他道路使用者的安全[85]. 中国的智能网联汽车准入政策则要求自动驾驶系统在道路测试时必须符合严格的安全标准, 并进行实时监控与数据记录[86].

3.3.2 设计策略

奖励函数的策略设计直接影响智能体的学习效率和策略质量. 在自动驾驶场景中, 不同的设计策略能够针对性地提升行为决策系统的性能和鲁棒性. 以下是常见的奖励函数设计策略.

1) 基于目标的奖励函数设计.

此类设计旨在通过明确的任务目标来指导智能体的学习过程. 在自动驾驶行为决策任务中, 常见的目标包括实现最短时间内到达目的地、最大化交通流量和最小化燃油消耗等[87]. 通过设计与目标相关的奖励信号, 车辆能够在不断的交互中逐步逼近最优解. 这种设计方式直接推动了智能体向特定目标优化, 并能够在一定程度上提高学习效率.

2) 基于约束的奖励设计.

基于约束的奖励设计注重对安全性、稳定性和法律法规等方面的约束条件的遵循. 通过引入相关约束条件作为奖励函数的组成部分, 可以有效引导智能体避免不合规或危险的行为. 在这种设计中, 自动驾驶车辆会因符合约束条件的行为获得正奖励, 而在违反约束时则会受到负奖励[88]. 例如, 超速、碰撞、急刹车等行为将被惩罚, 从而确保自动驾驶系统不会做出过于激进或不安全的决策.

3) 自适应奖励设计.

自适应奖励设计是一种根据智能体学习过程与环境动态变化, 实时调整奖励函数结构与强度的方法. 该策略通过评估智能体当前的学习状态、策略表现或任务难度, 动态调节奖励信号, 从而提升训练收敛速度与策略优化效果. 在自动驾驶任务中, 自适应奖励机制可根据交通流量、道路类型及风险等级等环境特征灵活调整奖励项权重, 使系统在不同驾驶情境下保持高效且稳定的行为策略. 该方法能够降低过拟合风险, 提高对复杂场景的适应能力与策略的泛化性能, 是当前自动驾驶复杂决策场景中广泛采用的重要手段之一[89].

4) 先验奖励函数.

先验奖励函数是基于领域知识或专家经验设计的初始奖励函数. 在缺乏大量实验数据或示范的情况下, 设计者可以利用已有的经验和规则来构建奖励函数, 帮助智能体启动学习过程[90]. 在自动驾驶行为决策过程中, 可以基于交通规则、车辆物理学、行驶安全要求等领域知识设计奖励函数. 通过引导智能体规避危险行为 (如碰撞、违规变道等), 先验奖励函数为智能体提供了一个相对安全和稳定的学习起点[91]. 尽管先验奖励函数在早期学习阶段非常有效, 但需要依赖于人工设计, 且难以捕捉复杂的环境和任务特征. 因此, 先验奖励函数通常需要结合其他学习方法 (如逆强化学习) 进一步优化.

5) 逆强化学习.

逆强化学习 (inverse reinforcement learning, IRL) 是一种通过学习专家行为轨迹反推奖励函数的算法,

其核心目标是重构专家在特定情境下所遵循的潜在奖励机制, 而非直接优化策略本身. 在自动驾驶行为决策中, IRL 通过分析人类驾驶员在复杂动态交通环境中的操作, 提取其行为的隐性偏好和目标, 从而构建能够模拟专家驾驶行为的奖励函数[92]. 该方法特别适用于任务复杂、环境非结构化且难以人工设计奖励函数的场景, 具有较强的泛化性和解释能力.

6) 模仿学习.

模仿学习 (imitation learning) 是一种通过学习专家示范行为来训练智能体策略的方法, 常作为 IRL 的补充或替代路径, 尤其适用于缺乏明确奖励信号的任务场景[93]. 在自动驾驶任务中, 模仿学习广泛应用于行驶、变道、超车等典型操作的策略学习, 车辆可以利用专家驾驶数据进行监督训练, 以近似人类驾驶行为[94]. 其主要优势是规避了复杂的奖励函数设计过程, 显著降低建模难度. 然而, 由于专家数据可能存在非最优或不一致的情况, IL 模型在泛化能力和鲁棒性方面可能受到限制, 容易导致策略误导或性能退化, 需通过数据筛选与策略微调进行补强.

3.3.3 设计指标

随着自动驾驶技术的发展, 奖励函数设计趋于精细化和多维化, 以适应多样化的场景和任务需求. 以下是一些常见的设计指标:

1) 通行安全.

安全性是自动驾驶系统的首要目标. 通过引入碰撞时间 (time-to-collision, TTC)、避免碰撞减速率 (deceleration rate to avoid collision, DRAC) 和碰撞时间导数 (time-to-collision derivative, TTCD) 等[95], 能够精确评估车辆在复杂环境下的安全性.

2) 通行效率.

在通行效率方面, 速度是评估通行效率的核心指标. 奖励函数通常通过设定车辆当前速度与期望速度之间的差值来反映通行效率. 为避免正负值相互抵消, 常对该差值进行平方处理, 以更精确地监测速度偏离程度[96]. 此外, 为提升通行效率的稳定性, 奖励函数可进一步引入跟车时距等指标, 以弥补仅依靠单一速度指标可能带来的局限性[97].

3) 节能.

低碳驾驶是现代自动驾驶系统的重要目标. 通过优化车辆的加速和减速模式, 减少急加速、急刹车以及不必要的速度波动[98], 从而有效降低能量消耗, 提升燃油经济性或电池续航能力. 为此, 奖励函数通常设定为加速度变化率、单位时间电能能耗或者微观的油耗排放模型[99].

4) 舒适性.

在舒适性方面, 奖励函数通过优化加速度和角速度的变化来降低乘客的不适感. 除了传统的加速度变化率约束外, 还可引入横向加速度变化的评估, 减少急转弯和横向摆动引起的不适感[100]. 部分研究还在奖励函数中加入了驾驶人特征指标, 使得自动驾驶系统能够模仿不同驾驶风格, 以满足个性化需求[101].

5) 法规遵循.

在法规遵循方面, 奖励函数主要关注车辆在不同场景下的合规性行为, 如速度限制、车道线保持等[102]. 对于违反限速、车道偏移以及碰撞等行为, 通常会通过设置相应的惩罚函数加以约束[103]. 例如, 为防止车辆超速, 常采用车辆速度与最大限速的比值作为限制条件[104]. 在转弯场景中, 还可考虑车辆与道路中心的偏移距离, 以确保其不会压线行驶. 针对换道行为, 如果测试中未将碰撞作为终止条件, 则换道的惩罚函数可以直接基于碰撞次数进行设置[105]. 这种设计有助于强化对法规遵循的约束, 从而提升自动驾驶系统的安全性与合法性.

3.3.4 权重设置

在 DRL 中, 奖励函数中的每个指标通常通过参数权重进行量化, 这些权重直接影响智能体对各目标的偏好和决策行为, 决定了自动驾驶车辆如何在安全性、流畅性和效率等多个目标之间进行平衡. 以下是几种常见的权重设置方法:

1) 手动调试法.

手动调试法是传统的权重设置方法, 依赖于专家知识和任务需求来设定权重. 这种方法适用于目标明确且场景特定的任务. 例如, 在复杂交通环境中, 安全性通常是最重要的考虑因素, 因此可以为安全性设置较高的权重; 而在交通较为顺畅的环境下, 效率可能成为优先考虑的目标, 此时可以适当提高效率指标的权重. 手动调试法的优点是简单易懂, 可以基于专家经验迅速调整, 但其缺点是灵活性较差, 且无法充分适应复杂动态的驾驶环境.

2) 动态权重调整法.

由于驾驶场景的动态性和复杂性, 固定权重往往难以适应不同情境下目标优先级的变化. 因此, 动态权重调整法被广泛应用于自动驾驶系统中, 以提高车辆的适应能力. 常见的动态权重调整方法包括贝叶斯优化算法、遗传算法、粒子群优化算法和梯度下降法等[106-107]. 贝叶斯优化算法通过模型预测权重对奖励的影响, 逐步更新权重, 以最大化长期回报. 遗传算法和粒子群优化算法基于进化策略对权重参数进行全局搜索, 适用于复杂多目标的决策场景. 梯

度下降法通过优化权重的梯度, 动态调整各目标的权重, 使其更加符合当前环境的变化. 动态权重计算方法为

$$w_i(t)=\frac{\text{priority}_i(t)}{\sum_{j=1}^{n}\text{priority}_j(t)}. \tag{6}$$

其中: $w_i(t)$为第i个元素在时间t的权重; $\text{priority}_i(t)$为在时间t时刻目标i的优先级, 动态归一化处理确保权重总和为 1.

3) 参数归一化处理.

奖励函数中不同目标的量纲差异可能会导致权重分配不均. 例如, 安全性通常基于距离 (单位为m), 而流畅性可能基于加速度 (单位为m/s^2). 这种尺度差异可能影响不同目标间的相对重要性. 因此, 归一化技术被广泛应用于奖励值的处理, 以确保各目标之间具有可比性, 从而提高学习效率[108]. 常见的归一化方法包括最大最小归一化和 Z-Score 标准化, 计算公式分别为

$$R_i^{\text{norm}}=\frac{R_i-R_i^{\min}}{R_i^{\max}-R_i^{\min}}, \tag{7}$$

$$R_i^{\text{norm}}=\frac{R_i-\mu_i}{\sigma_i}. \tag{8}$$

其中: R_i^{norm}为归一化后的奖励值, 经过公式处理后, 被映射到$[0,1]$的范围内; R_i为原始奖励; $R_i^{\min}$为目标i奖励值的最小可能值 (通常通过数据分析得出); $R_i^{\max}$为目标i奖励值的最大可能值 (通常通过数据分析得出); μ_i和σ_i分别为奖励值的均值和标准差.

这两种归一化方法的主要作用是将各目标的尺度统一, 使它们在优化过程中具有同等的影响力, 从而提高学习算法的效率. 通过归一化处理, 智能体能够更加有效地权衡和优化多目标任务中的不同指标, 避免某一目标因尺度差异过大而主导整个优化过程.

3.4　策略与更新

在自动驾驶行为决策中, DQN、DDPG 和 PPO 是典型的 DRL 算法, 它们分别代表了基于值函数、确定性策略梯度和概率策略优化 3 种典型的策略更新范式, 广泛应用于不同控制精度与任务复杂度的场景中.

1) DQN.

DQN 是基于值函数的策略更新代表算法, 通过更新 Q 值函数来选择最优动作. 该方法的策略更新流程如图 8 所示. 核心公式为 Q 值更新, 更新方法为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha(r_{t+1}+\gamma\max_{a'}Q(s_{t+1},a')-Q(s_t,a_t)). \tag{9}$$

其中: $Q(s_t,a_t)$为当前状态s_t和动作a_t对应的 Q 值; a'为下一步可能的所有动作中最大值的动作; α为学习率, 决定每次更新的幅度.

也可以采用贝尔曼公式进行 Q 值更新, 即

$$Q(s_t,a_t)=r_t+\gamma\max_{a'}Q_{\text{target}}(s_{t+1},a';\theta^-). \tag{10}$$

其中: Q_{target}为目标 Q 网络的值函数, 使用固定参数θ^-.

2) DDPG.

DDPG 是一种基于确定性策略梯度更新的代表算法, 通过分别训练 Actor 和 Critic 网络来优化策略与估值函数, 能够精细调整控制参数, 如车辆的加速、刹车力度等. 该方法的策略更新流程如图 9 所示, Actor 和 Critic 网络更新的计算公式为

$$Q_{\text{target}}=r_{t+1}+\gamma Q_{\text{crictic}}(s_{t+1},a_{t+1}), \tag{11}$$

$$\nabla_{\theta_\pi}J(\theta_\pi)\approx\mathbb{E}[\nabla_{\theta_\pi}\log\pi_{\theta_\pi}(s_t,a_t)\cdot Q_{\text{target}}(s_t,a_t)]. \tag{12}$$

其中: θ_π为参数; $Q_{\text{crictic}}(s_{t+1},a_{t+1})$为下一个状态-动作对的估值; Q_{target}为智能体在当前策略下可能获得的总回报; $J(\theta_\pi)$为策略的目标函数; $\pi_{\theta_\pi}(s_t,a_t)$为给定状态$s_t$时, 选择动作$a_t$的概率.

3) PPO.

PPO 是一种基于概率策略优化更新的代表算法, 具有较好的训练稳定性和样本训练效率. 其核心思想是通过计算新旧策略的概率比, 结合优势函数来优化策略, 并且在更新过程中通过裁剪机制防止过大的更新幅度, 如图 10 所示, 策略更新计算公式为

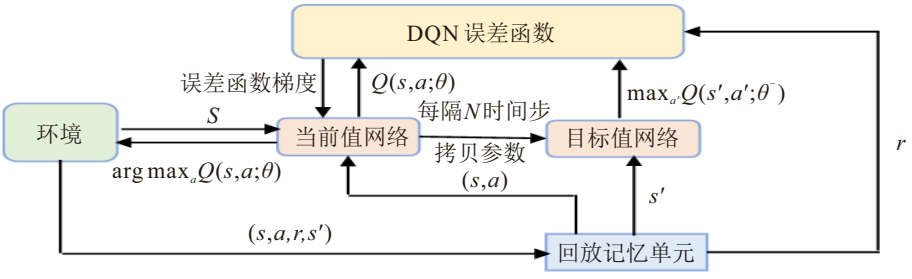

图8　DQN 流程

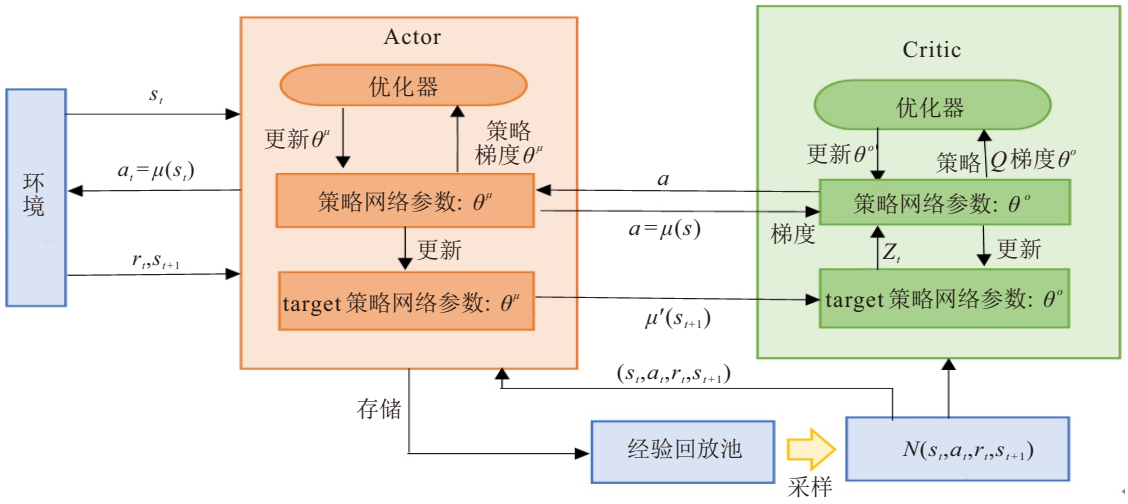

图9 DDPG 算法流程

$$L_{\mathrm{CLIP}}(\theta)=\mathrm{E}_t[\min(r_t(\theta)A_t,\mathrm{clip}(r_t(\theta),1-\epsilon,1+\epsilon)A_t)],\quad(13)$$

$$r_t(\theta)=\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t\mid s_t)}.\quad(14)$$

其中: $L_{\mathrm{CLIP}}(\theta)$为 PPO 的目标函数; E_t为对时间步t的期望值; $r_t(\theta)$为新旧策略的概率比; π_θ为当前策略; $\pi_{\theta_{\mathrm{old}}}$为旧策略; A_t为优势函数; ϵ为控制策略更新幅度的裁剪阈值; clip为限制$r_t(\theta)$的取值范围, 避免策略更新过大; $\pi_\theta(a_t|s_t)$为新策略在状态s_t下选择动作a_t的概率; $\pi_{\theta_{\mathrm{old}}}(a_t|s_t)$为旧策略在状态$s_t$下选择动作$a_t$的概率.

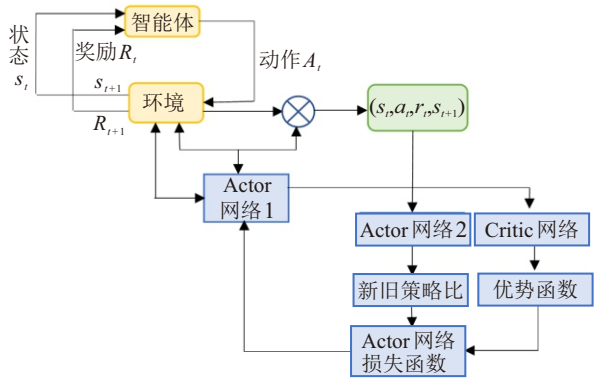

图10 PPO 算法流程

3 种典型 DRL 算法的优缺点和适用场景总结如表 2 所示.

表2 DQN、DDPG 和 PPO 算法对比

特性	DQN	DDPG	PPO
动作空间类型	离散	连续	离散或连续
使用场景	简单、低动态	高动态、需精细控制	动态复杂、多样化
稳定性	中等	中等	较高
学习效率	较低	较高	较高
实现复杂度	较低	较高	中等

3.5 评价指标

在自动驾驶行为决策系统中, DRL 模型的评价至关重要. 有效的评价不仅需要衡量模型的学习性能, 还应综合考虑计算效率与收敛速度、稳定性、安全性以及系统的整体性能等. 以下是常见的评价指标.

1) 学习性能.

学习性能反映了 DRL 模型在自动驾驶环境中的训练效果与策略优化能力. 常见的学习性能评价指标包括平均奖励、学习曲线和奖励方差. 其中: 平均奖励是多个训练回合中的平均值, 体现了模型在多样化环境中的表现[109]; 学习曲线用于直观展示模型的学习进度和在复杂环境中持续优化决策的能力; 奖励方差用于评估模型训练中的波动性与稳定性, 较低的方差表明智能体行为较为一致和可靠.

2) 训练效率和推理效率.

训练效率和推理效率是自动驾驶系统中两项至关重要的评价指标, 分别对应于训练过程和训练后的决策过程. 训练效率通常通过收敛速度和训练时间来衡量[110]. 推理效率侧重于实时性和计算效率, 即在训练完成后的策略推理过程中, 智能体做出决策的能力[111].

3) 稳定性.

DRL 模型的稳定性是保障自动驾驶系统可靠性和可预测性的关键[112]. 训练稳定性可以通过奖励的方差或损失函数的波动情况评估. 稳定性测试可以通过开环测试和闭环测试两种方法进行, 这两种方法各自侧重于不同的测试角度和实验环境.

开环测试通常使用如 KITTI、Waymo Open Dataset、Apollo Scape 等真实驾驶环境数据集进行评价[113-114], 主要测试模型在已知条件下的稳定性, 确保其在没有实时反馈的情况下依然能提供一致的决

策. 闭环测试可以利用 CARLA、SUMO 等仿真平台进行动态环境下的评价, 模型需在实时环境中接收反馈并调整决策[115-116]. 此类测试能够更真实地反映智能体的动态适应能力及其在复杂交互情境中的表现, 特别是在多车协同与复杂交通流下的表现.

4) 安全性.

安全性是自动驾驶系统最核心的评估指标, 因为其直接影响驾驶员、乘客和行人的生命安全. 在 DRL 模型中, 安全性指标主要衡量模型在决策过程中是否能够有效避免碰撞、遵守交通规则, 并规避其他潜在的危险行为. 安全性指标通常与奖励函数设计紧密相关, 通过引导模型在复杂环境中选择最安全的行为策略, 确保系统的鲁棒性和可靠性. 在评价时, 通常会使用如 NGSIM、InD 和 High D 等数据集, 测试模型在不同情境下 (如跟车、换道、交叉路口等) 是否能够有效地避免碰撞[117]. 此外, 违反交通规则所带来的安全分析也可通过 Nu Scenes 和 Lyft 数据集进行评价[118], 以确保模型在实际驾驶中遵循交通法规和道路规则.

5) 系统整体性能.

系统的整体性能决定了自动驾驶系统在实际应用中的表现, 主要评价内容包括通行效率和节能效果. 通行效率用于衡量系统在不增加交通拥堵的情况下, 如何高效地完成交通任务. 常见的评价指标包括通行能力、通过速度、平均延误和总延误等. 通行效率的优化效果可以通过 High D 和 Argoverse 等数据集进行测试, 以评价系统在速度和流畅性方面的表现[119]. 节能效果则评价系统是否能够优化行驶路径和驾驶行为, 减少能量浪费, 常用的指标包括碳排放量和油耗等. 节能效果通过 Waymo 开放数据集和 Apollo Scape 数据集验证加速度波动优化程度[120], 并测试如何通过优化驾驶行为和路径规划来减少不必要的能量消耗, 从而提升整体的能源使用效率.

4　典型场景中的应用

传统以跟驰、换道和转弯等基本行为为核心的建模方法, 难以应对高复杂性、高风险性驾驶场景的需求. 因此, 本节对驾驶场景进行更为精细地分类, 提炼具有代表性的典型应用场景, 从而为基于 DRL 的行为决策系统设计提供坚实的研究基础.

按照道路类型, 交通场景可划分为高等级公路、城市道路、乡村道路和特殊路段. 高等级公路主要解决车辆长期保持安全稳定运行以及交通高效流转的问题, 其典型场景包括主线运行、匝道合流和匝道分流. 相比之下, 城市道路具有更复杂和多样的特性, 需要应对行人、机动车、非机动车等混合交通流的挑战, 典型场景包括信控交叉口和非信控交叉口. 而在乡村道路中, 横向突发闯入是主要挑战, 这种行为不仅可能发生在交叉口, 甚至出现在普通路段上. 特殊路段 (如急弯、陡坡和施工区) 由于环境的不确定性和动态变化, 对自动驾驶系统的感知、预测和决策能力提出了更高的要求.

通过文献梳理可以发现, 现有研究主要集中于特定的道路场景, 而对于乡村道路等“长尾问题”场景的关注相对较少. 在研究重点方面, 除了跟车、变道、转弯等基本行为外, 现有研究主要聚焦于持续运行场景 (如主线长距离驾驶)、交汇场景 (如匝道合流、交叉口) 和并线场景 (如施工区车道变更). 基于此, 本文选择主线运行、匝道合流、交叉口和施工区四类典型场景, 通过分析其核心挑战与关键技术问题, 探讨基于 DRL 的优化策略与应用方法, 为复杂场景下的自动驾驶行为决策实现提供参考. 本节内容的结构如图 11 所示.

4.1　主线通行场景

主线通行场景对自动驾驶车辆在持续运行中的综合能力提出了考验, 要求车辆在长时间、高动态环境下平衡车速控制、路径规划之间的关系. 基于 DRL 的自动驾驶行为决策方法在主线通行场景中的应用, 主要集中于不良天气应对与多车编队.

4.1.1　不良天气应对

不良天气 (如大雨、大雾、积雪或强风) 是主线通行场景中对自动驾驶行为决策影响最大的因素之一. 它不仅会显著降低传感器的感知性能, 还会改变车辆与道路之间的物理交互特性 (如轮胎抓地力降低), 增加了交通事故的风险.

在不良天气场景下, DRL 算法通过多模态感知融合 (如激光雷达、毫米波雷达和摄像头数据结合) 及鲁棒性优化, 增强自动驾驶的决策能力, 从而提高其对复杂环境的适应性. 通过构建高维状态空间, DRL 框架能够实时捕捉天气条件对道路摩擦系数的动态影响, 进而优化车辆的速度调节、加速控制和轨迹规划策略[121].

针对湿滑路面或低能见度等天气条件, 相关研究在奖励函数中加入了针对刹车距离、车速偏差和车道偏移的惩罚项, 以提升决策的安全性[122]. 结合动态驾驶风险感知和运动规划, DRL 模型能够在高风险天气条件下优先保障安全性, 进一步降低事故发生率[123].

在实际应用方面, 索尼公司开创了基于视觉的

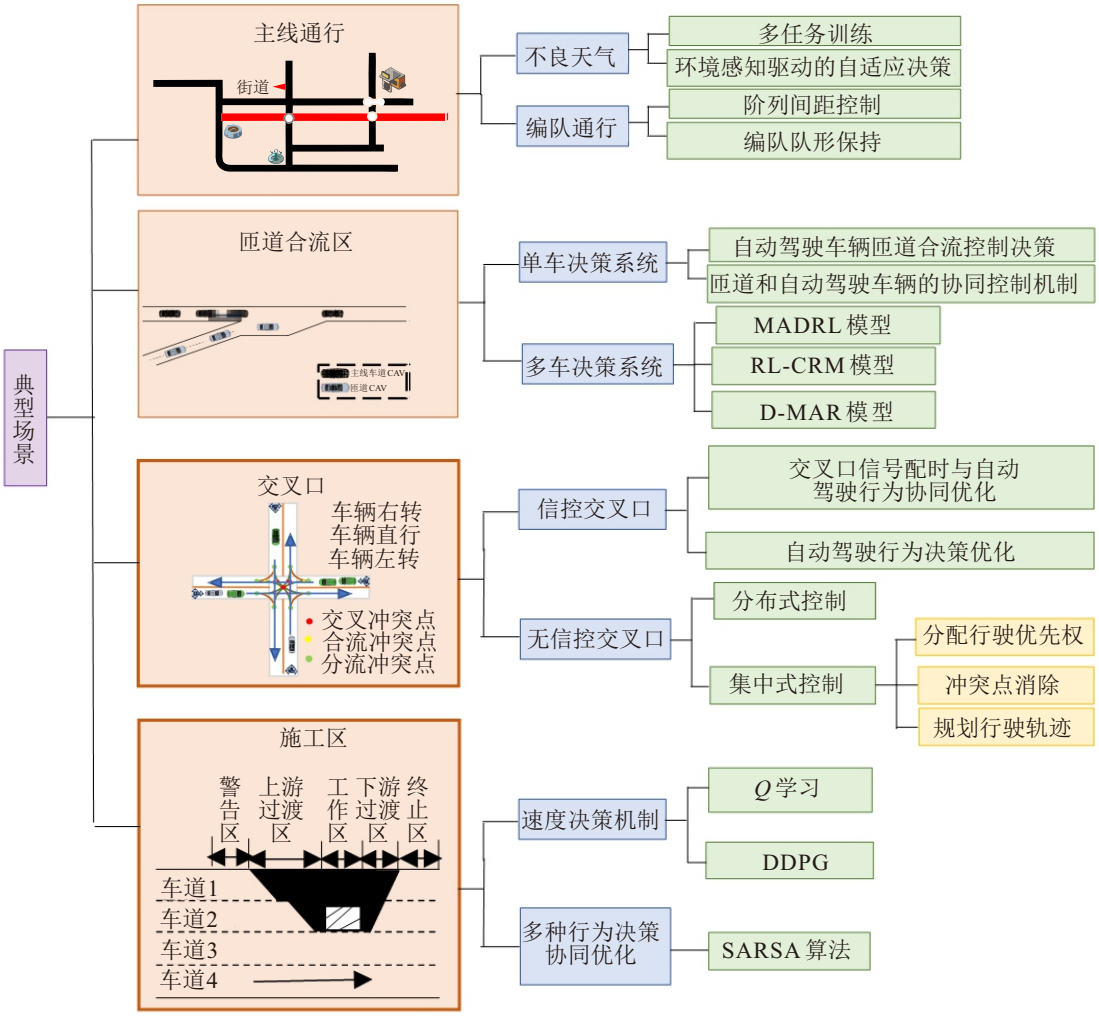

图11 典型场景结构

自动驾驶强化学习框架, 为在复杂环境中实现高效决策提供了重要支持[124]. 同时, 沃尔沃公司则专注于解决强化学习中的随机性和认知不确定性问题, 这为应对不良天气等复杂场景下的自动驾驶挑战提供了有效的解决方案[125].

4.1.2 多车编队

多车编队通行是主线道路中提升交通效率和节能减排的关键技术之一. 通过车间通信实现的信息共享, 车辆编队能够协同控制速度和加速度, 保持队形并减少空气阻力. 编队通行需要在多车协作与全局优化之间找到平衡, 增加了行为决策的复杂性.

基于 DRL 的多车编队研究主要集中于队列间距控制、编队队形保持和动态队列调整. 通过 SAC、DQN、DDPG 等算法, 自动驾驶车辆能够通过协作学习制定加入或退出编队的最优策略[126-127]. 研究者通常在奖励函数的设计中加入多目标优化指标, 如空气阻力节省、队形稳定性和交通效率等[128], 以引导车辆保持稳定的队形, 减少车间距偏差; 也可以在奖励函数中加入对车间距偏差的惩罚项, 同时在编队解散时最小化对整体交通流的干扰.

此外, 一些研究还探索了基于“集中式训练、分散式执行”框架的编队通行方法. 在训练阶段, 自动驾驶车辆通过全局信息优化编队协作策略, 而在执行阶段, 各车根据局部信息独立决策, 形成稳定队列, 从而有效降低实时性要求[129-130].

4.2 匝道合流区域通行场景

匝道合流区是高速公路与支路的交汇场景, 其核心难点是车辆动态汇入主路、车速匹配和车辆协作优化. 现有的研究主要分为单车行为决策和多车协同行为决策[131].

4.2.1 单车行为决策

单车行为决策系统以匝道车辆为控制核心, 充分考虑主线车辆的运动特征, 通过优化车辆的速度调整与路径选择, 实现安全且高效的合流行为. 该方法在设计上简洁高效, 尤其适用于低交通流密度场景, 能够在保证决策效果的同时显著降低计算开销. 目前, 该类方法多采用如 DDPG、TD3 和 PPO 等先进的深度强化学习算法, 以提升合流策略的智能化

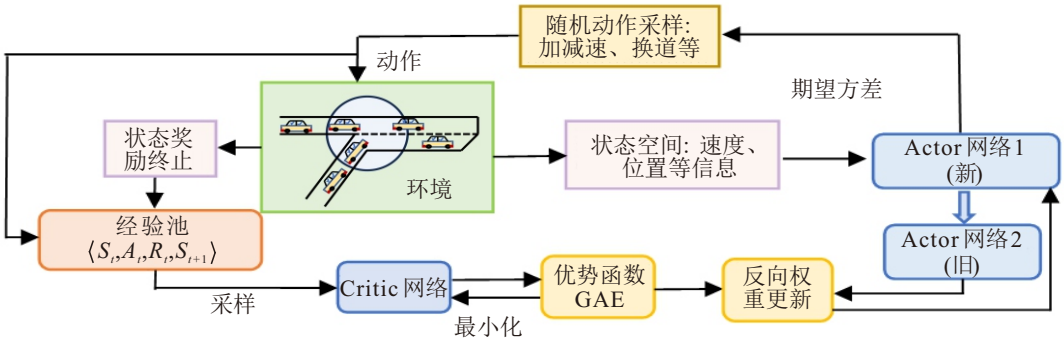

图12　基于 PPO 的匝道合流控制流程

水平和泛化能力[132-134]. 其中, PPO 算法在处理匝道问题时表现尤为出色, 能够通过实时优化车辆速度和相对距离来制定合流策略[135]. 与传统的基于规则的合流方法相比, PPO 能够显著减少主路车辆与匝道车辆间的干扰, 提高交通流畅性. 该算法在解决匝道合流问题时的主要流程如图 12 所示.

4.2.2　多车协同决策

多车协同决策系统能够同时控制主路和匝道上的多辆车辆, 强调车辆间的协作与竞争, 旨在通过全局优化提升整体交通效率与安全性[136]. 此类方法适用于高交通流密度或混合交通场景, 算法复杂度较高, 通常采用 MADRL 模型, 例如 MADDPG 或 Q-MIX[137-138].

一种常用的框架是在训练阶段获取合流区全局信息以优化策略, 而在执行阶段, 各车根据局部环境独立决策[139]. 该框架使主路车辆与匝道车辆能够进行交互学习, 优化了行为策略. 主线车辆通过调整速度为匝道车辆创造合流机会, 而匝道车辆则根据主路交通流状态选择最佳时机汇入. 这种协同学习机制显著减少了合流车辆与主线车辆间的冲突[140-141].

尽管多车协同决策模型具有显著优势, 但其在高维状态空间下, 计算复杂度和实时性仍然是主要挑战. 为此, 研究者引入状态空间压缩技术和分布式计算方法, 以降低计算负担并提升实时响应能力. 这些改进措施显著增强了多车协同决策模型的实际应用价值[142].

4.3　交叉口通行场景

交叉口场景涉及多方向、多流交汇, 尤其在通行优先权分配、多目标路径规划和信号控制优化等方面具有较高复杂性. 作为交通网络的关键节点, 交叉口的多方向交汇特点易导致通行效率下降和事故风险增加[143]. DRL 算法通过实时感知和决策, 能够快速应对其他车辆及行人的动态行为, 在交叉口通行场景中展现出显著优势. 随着 5G 通信和车路协同技术的发展, 基于 DRL 的自动驾驶决策在信控交叉口和无信控交叉口中都有广泛应用[144]. 交叉口行为控制流程如图 13 所示.

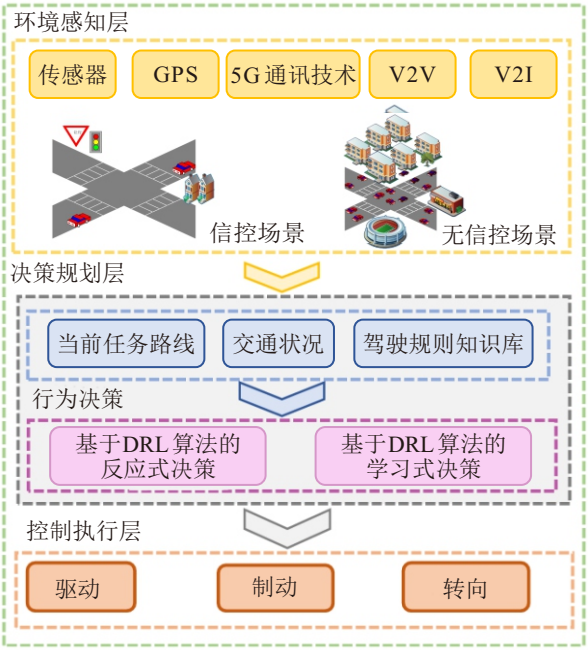

图13　交叉口行为决策流程

4.3.1　信控交叉口

在信控交叉口中, 自动驾驶车辆面临交通信号随机性、多车交互复杂性及不同车辆决策策略的不确定性, 基于 DRL 的研究主要聚焦于信号配时与车辆行为协同优化、自动驾驶行为决策优化.

1) 信号配时与自动驾驶行为协同优化.

基于 DRL 的自动驾驶行为决策方法通过设计包含车辆位置和速度的状态空间, 并结合动态奖励函数, 实现信号配时与车辆行为决策的协同优化. 例如, 结合 PPO 算法动态调整信号配时和可变车道设置, 可减少车辆排队长度, 提高通行效率, 如图 14 所示[145].

2) 自动驾驶行为决策优化.

在信控交叉口的上游, 自动驾驶车辆需要提前完成换道操作, 以确保高效通行. 基于 DRL 的自动驾驶行为决策方法能够通过优化换道决策和路径规

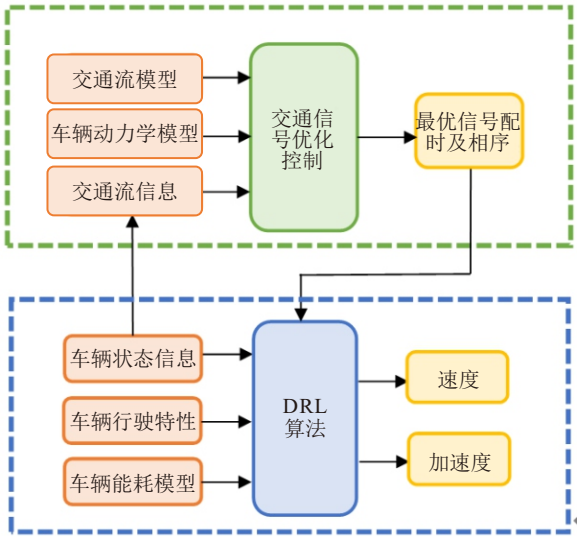

图14 交叉口信号配时与自动驾驶行为协同优化

划, 提升车辆在复杂交通环境中的适应能力. 例如, 采用群体控制方法协调多辆车辆的行为策略, 从系统最优角度提高交叉口的通行效率[146].

4.3.2 无信控交叉口

无信控交叉口缺乏交通信号灯指引, 车辆需自主决策通行顺序以避免碰撞, 同时兼顾通行效率. 主要挑战包括多方向交互、冲突点规划、感知与决策实施, 以及多目标权衡. 常见的决策方法包括集中式控制[147] 和分布式控制[148].

1) 集中式控制.

集中式控制是优化无信控交叉口的常用方法, 通过控制中心管理交叉口通行顺序, 以提高通行效率. 其优势在于全局优化, 适用于交通流密度高或多方向交互复杂的场景, 主要通过通行优先权分配和轨迹规划来解决冲突问题.

在无信控交叉口, 集中式控制通过 DRL 算法动态调整车辆速度和路线, 解决多辆车辆同时到达冲突点的问题. 利用多智能体 MDP 模型, 中央控制器根据实时交通流量和车辆需求输出通行优先权动作, 实现车流有序通行[149].

在复杂动态环境中, 集中式控制通过轨迹规划优化车辆通行路径. 通过将实时感知交叉口状态传入状态空间, 并以最优速度或路线为输出动作, 避免潜在冲突, 同时动态调整到达时间窗以提升流畅性[150].

2) 分布式控制.

分布式控制通过车辆自主学习与决策, 依靠车联网技术实现多车协同交互, 以应对实时性和复杂性挑战. 其优势在于实时响应和灵活适配, 适合交通流分布不均或场景动态变化的情况, 但依赖于车联网环境的可靠性, 可能因局部信息不足或策略冲突而受限.

在无信控交叉口的多车交互场景中, 车辆需动态避让彼此以避免冲突. 利用分布式模型预测控制或基于蒙特卡洛树搜索与 DRL 相结合的策略, 能够实现多车协同轨迹规划[151-152]. 每辆车独立计算行驶轨迹, 并通过多轮交互学习最优的通行序列和速度控制策略, 选择最优路径以最大化安全性和通行效率[153]. 此方法在低延迟通信环境下表现优异, 具备更高的灵活性与实时响应能力.

4.4 施工区域通行场景

施工区域作为动态约束场景, 复杂的道路条件 (如变窄车道和临时交通标志) 以及多变的环境因素 (如施工车辆与人员的频繁移动) 使传统自动驾驶算法难以精准应对, 增加了交通事故的风险. 针对这些挑战, 基于 DRL 的自动驾驶行为决策方法通过融合环境建模和实时策略调整, 可实现动态避障与路径优化, 提升通行效率与安全性. 同时, 研究更先进的速度决策机制和优化施工区内车辆行为的协同控制也成为解决问题的关键.

4.4.1 速度决策机制

施工区内速度变化模式具有显著特性, 受制于施工区域布局、施工设备停放位置以及施工人员活动范围等多重因素. 基于 DRL 的速度决策机制, 通过对施工区环境进行实时建模与感知, 结合动态优化策略, 可以有效降低风险并提升车辆通行效率. 例如, 深度 Q 学习算法能够检测施工区内潜在的拥堵趋势, 并通过调整车速引导车辆有序流动, 减少交通冲突, 决策流程如图 15 所示[154]. 此外, DDPG 算法凭借对连续动作空间的处理能力, 可在施工区域内实现精细化的限速优化和动态速度调整[155]. 这种基于速度决策的行为优化, 不仅能够降低施工区域内的交通事故率, 而且可以缓解由复杂施工环境引发的交通震荡, 确保通行效率和安全性的平衡.

4.4.2 多种行为决策协同优化

施工区域的动态复杂性对自动驾驶车辆提出了多目标协同优化的挑战, 包括避障、路径规划和速度控制等行为决策的同步优化. 针对这一问题, 研究者提出了基于分布式多智能体 DRL 算法的解决方案, 通过动态调整跟驰距离和换道策略, 实现车辆在施工区域内的平稳通行[156].

分布式在线 SARSA 算法通过车联网环境中的信息共享机制, 有效协调车辆间的行为选择, 可显著提高施工区域的交通流畅性与安全性[157]. 该方法能够应对施工场景中复杂的道路环境、动态变化的障

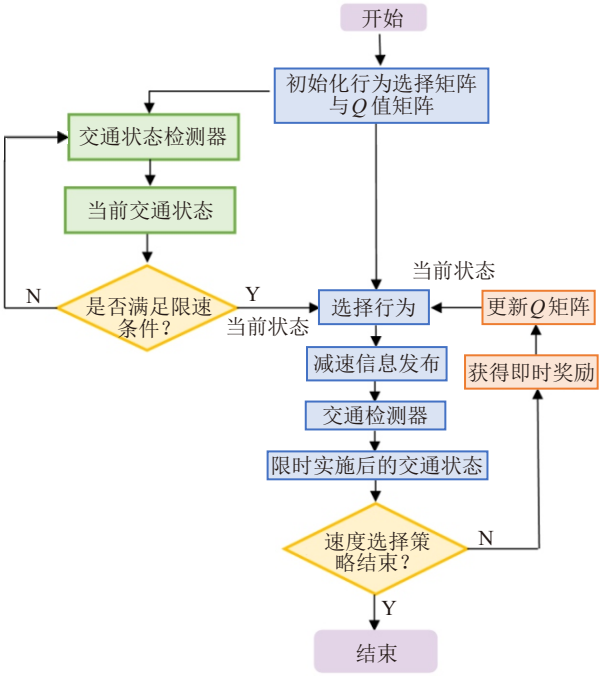

图15　速度决策机制流程

碍物以及行人等不可预测因素, 增强了系统在处理实时变化和潜在冲突时的灵活性.

通过多种行为的协同优化, 自动驾驶系统可以更加精准地平衡施工场景中多目标需求. 例如, 结合避障与路径规划的动态权衡, 系统能够同时满足通行效率与安全性的要求, 从而在复杂施工环境中实现更加平稳的车辆决策与控制.

5　总结与展望

基于 DRL 的自动驾驶行为决策方法已从理论探索逐步迈向工程化应用, 其发展路径可概括为“单体决策 → 协同优化 → 系统可信”三阶段跃迁. 本文通过对 DRL 在行为决策中的发展、设计、优化与应用的全链条分析, 揭示了其核心挑战与突破方向. 未来, 如何推动自动驾驶行为决策技术向更智能化、精准化、个性化发展, 并提升其在实际场景中的社会接受度, 仍是亟待解决的重要课题. 未来研究还需重点关注以下几个方面:

1) 多智能体协同决策的演进路径.

在复杂交通交互场景中, 多智能体 DRL 已成为提升系统整体决策效率的关键手段. 通过车辆间的协同策略, 可实现局部最优与全局最优之间的平衡. 然而, 随着参与智能体数量增加, 系统将面临策略一致性、通信负载与计算复杂性等挑战. 未来研究应围绕高效通信协议设计、低延迟分布式架构与博弈理论融合展开, 确保多车协同在大规模动态环境下的实时性与稳定性.

2) 多目标动态权衡与奖励函数创新.

自动驾驶行为决策需同时优化安全性、通行效率、乘坐舒适性与环保性等多个目标, 但传统静态加权的奖励设计难以适应不同情境的目标动态变化, 易造成行为失衡. 未来应探索动态自适应权重机制, 结合场景感知实现多目标之间的实时调节. 同时引入驾驶经验、交通规则与用户偏好, 构建更具人性化与实际适应性的奖励函数体系, 提升系统综合表现.

3) 长尾场景的解决与边缘鲁棒性增强.

当前基于 DRL 的自动驾驶行为决策系统在罕见或极端场景 (即长尾事件) 下表现不稳定, 严重制约了其实用性与安全性. 未来应通过多样化仿真环境构建、数据增强与场景重建提升模型泛化能力; 并引入对抗训练、分布式强化学习与异常检测机制, 增强系统在非典型环境下的鲁棒性与恢复能力, 实现对边缘场景的有效应对.

4) 可解释性增强与可信决策体系构建.

作为“黑箱”模型, DRL 算法缺乏透明的决策过程解释, 难以满足监管与用户对系统可信度的要求. 未来应结合因果推理、注意力机制等方法, 建立从输入状态到行为输出的可解释路径图谱. 同时辅以可视化决策分析工具与用户交互界面, 提升策略输出的透明性与可追溯性, 助力构建可信、可验证的自动驾驶系统.

5) 法规与伦理约束的量化研究及治理体系构建.

自动驾驶在面临安全事故、隐私保护和伦理抉择时尚缺乏统一的法律与技术应对体系, 未来需推动伦理规则与法规体系的量化建模, 结合伦理强化学习与多准则决策分析方法, 使算法决策过程能够符合社会价值观. 同时, 应推动跨学科协同, 构建覆盖“开发-部署-监管”全流程的治理框架, 为政策制定与产业标准落地提供技术支撑.

综上所述, DRL 在自动驾驶行为决策中的应用正迈向协同化、人性化、可信化的新阶段. 突破现有瓶颈需以算法创新为核心, 同时兼顾硬件加速与法规约束, 方能构建“人-车-路-云”全域协同的智能驾驶生态系统.

参考文献 (References)

[1] Urmson C, Anhalt J, Bagnell D, et al. Autonomous driving in urban environments: Boss and the urban challenge[J]. Journal of Field Robotics, 2008, 25(8): 425-466.

[2] 朱冰, 贾士政, 赵健, 等. 自动驾驶车辆决策与规划研究综述[J]. 中国公路学报, 2024, 37(1): 215-240.
(Zhu B, Jia S Z, Zhao J, et al. Review of research on decision-making and planning for automated vehicles[J]. China Journal of Highway and Transport,

2024, 37(1): 215-240.)

[3] Li N, Yao Y, Kolmanovsky I, et al. Game-theoretic modeling of multi-vehicle interactions at uncontrolled intersections[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(2): 1428-1442.

[4] Li S Y, Liu C Y, Chen W H. An integrated MPC decision-making method based on MDP for autonomous driving in urban traffic[C]. 2024 6th International Conference on Industrial Artificial Intelligence. Shenyang, 2024: 1-6.

[5] Jeong Y. Probabilistic game theory and stochastic model predictive control-based decision making and motion planning in uncontrolled intersections for autonomous driving[J]. IEEE Transactions on Vehicular Technology, 2023, 72(12): 15254-15267.

[6] Al-Sharman M, Dempster R, Daoud M A, et al. Self-learned autonomous driving at unsignalized intersections: A hierarchical reinforced learning approach for feasible decision-making[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(11): 12345-12356.

[7] Bae S H, Joo S H, Pyo J W, et al. Finite state machine based vehicle system for autonomous driving in urban environments[C]. 2020 20th International Conference on Control, Automation and Systems. Busan, 2020: 1181-1186.

[8] Yin T Q, Li Y, Fan J H, et al. A novel gated recurrent unit network based on SVM and moth-flame optimization algorithm for behavior decision-making of autonomous vehicles[J]. IEEE Access, 2021, 9: 20410-20422.

[9] Liao Y P, Yu G Z, Chen P, et al. Integration of decision-making and motion planning for autonomous driving based on double-layer reinforcement learning framework[J]. IEEE Transactions on Vehicular Technology, 2024, 73(3): 3142-3158.

[10] Li Q, Peng H, Li J X, et al. A survey on text classification: From traditional to deep learning[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 13(2): 1-41.

[11] Dongfeng Yuexiang Technology Co Ltd. Intelligent driving vehicle autonomous parking system based on deep reinforcement learning[P]. CN117698695A. 2024-03-15.

[12] Wang J J, Zhang Q C, Zhao D B, et al. Lane change decision-making through deep reinforcement learning with rule-based constraints[C]. 2019 International Joint Conference on Neural Networks. Budapest, 2019: 1-6.

[13] 周润发. 融合动态场景信息和 DDPG 算法的智能车决策规划方法研究与应用[D]. 成都: 电子科技大学, 2021.
(Zhou R F. Research and application of intelligent car decision planning method combining dynamic scene information and DDPG algorithm[D]. Chengdu: University of Electronic Science and Technology of China, 2021.)

[14] Wu Y Q, Liao S Q, Liu X, et al. Deep reinforcement learning on autonomous driving policy with auxiliary critic network[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(7): 3680-3690.

[15] Kiran B R, Sobh I, Talpaert V, et al. Deep reinforcement learning for autonomous driving: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 4909-4926.

[16] Wu J D, Huang C, Huang H L, et al. Recent advances in reinforcement learning-based autonomous driving behavior planning: A survey[J]. Transportation Research Part C: Emerging Technologies, 2024, 164: 104654.

[17] 金立生, 韩广德, 谢宪毅, 等. 基于强化学习的自动驾驶决策研究综述[J]. 汽车工程, 2023, 45(4): 527-540.
(Jin L S, Han G D, Xie X Y, et al. Review of autonomous driving decision-making research based on reinforcement learning[J]. Automotive Engineering, 2023, 45(4): 527-540.)

[18] 韩胜明, 肖芳, 程纬森. 深度强化学习在自动驾驶系统中的应用综述[J]. 西华大学学报: 自然科学版, 2023, 42(4): 25-31.
(Han S M, Xiao F, Cheng W S. Overview of the application of deep reinforcement learning in autonomous driving systems[J]. Journal of Xihua University: Natural Science Edition, 2023, 42(4): 25-31.)

[19] Liu Q, Li X Y, Tang Y J, et al. Graph reinforcement learning-based decision-making technology for connected and autonomous vehicles: Framework, review, and future trends[J]. Sensors, 2023, 23(19): 8229.

[20] Guo H Y, Cao D P, Chen H, et al. Model predictive path following control for autonomous cars considering a measurable disturbance: Implementation, testing, and verification[J]. Mechanical Systems and Signal Processing, 2019, 118: 41-60.

[21] Gomes L. When will Google's self-driving car really be ready?[J]. IEEE Spectrum, 2016, 53(5): 13-14.

[22] de Donato M, Polido F, Knoedler K, et al. Team WPI-CMU: Achieving reliable humanoid behavior in the DARPA robotics challenge[J]. Journal of Field Robotics, 2017, 34(2): 381-399.

[23] Hu Y N, Zhao D, Wang Y, et al. DAnoScenE: A driving anomaly scenario extraction framework for autonomous vehicles in urban streets[J]. Journal of Intelligent Transportation Systems, 2025, 29(1): 32-52.

[24] Malik S, Khan M A, Aadam, et al. CARLA+: An evolution of the CARLA simulator for complex environment using a probabilistic graphical model[J]. Drones, 2023, 7(2): 111.

[25] Liu S W, Zheng K, Zhao L, et al. A driving intention prediction method based on hidden Markov model for autonomous driving[J]. Computer Communications, 2020, 157: 143-149.

[26] Zhao D Y, Zhao S E. Sparse least squares support vector machine based methods for vehicle driving

behavior recognition[J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2024, 238(6): 1392-1404.

[27] Ma J C, Xie H, Song K, et al. A Bayesian driver agent model for autonomous vehicles system based on knowledge-aware and real-time data[J]. Sensors, 2021, 21(2): 331.

[28] Han I, Park D H, Kim K J. A new open-source off-road environment for benchmark generalization of autonomous driving[J]. IEEE Access, 2021, 9: 136071-136082.

[29] 吴琼. 北京自动驾驶车辆道路测试报告 (2019)[J]. 智能网联汽车, 2020(2): 46-55.
(Wu Q. Beijing autonomous vehicle road test report (2019)[J]. Intelligent Connected Vehicles, 2020(2): 46-55.)

[30] Lin H H, Ding W H, Liu Z X, et al. Safety-aware causal representation for trustworthy offline reinforcement learning in autonomous driving[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4639-4646.

[31] 王雪松, 杨露, 程玉虎. 基于温和泛化的不确定性离线强化学习[J]. 控制与决策, 2025, 40(11): 3329-3339.
(Wang X S, Yang L, Cheng Y H. Uncertainty-aware offline reinforcement learning with mild generalization[J]. Control and Decision, 2025, 40(11): 3329-3339.)

[32] Gu Z Q, Gao L P, Ma H T, et al. Safe-state enhancement method for autonomous driving via direct hierarchical reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(9): 9966-9983.

[33] Nguyen H D, Han K. Safe reinforcement learning-based driving policy design for autonomous vehicles on highways[J]. International Journal of Control, Automation and Systems, 2023, 21(12): 4098-4110.

[34] Wu L Z, Lin H, Wang X D. Federated training generative adversarial networks for heterogeneous vehicle scheduling in IoV[J]. IEEE Internet of Things Journal, 2025, 12(5): 4888-4898.

[35] 张浩然, 李君, 邢立宁, 等. 大模型与智能优化算法集成研究综述[J]. 控制与决策, DOI: 10.13195/j.kzyjc.2025.0121.
(Zhang H R, Li J, Xing L N, et al. A research review on the integration of large models and intelligent optimization algorithms[J]. Control and Decision, DOI: 10.13195/j.kzyjc.2025.0121.)

[36] 孙闻鹏, 崔昀宽, 倪心睿, 等. 车路云一体化自动驾驶推进高速公路新质生产力[J]. 中国公路, 2025(9): 114-117.
(Sun W P, Cui Y K, Ni X R, et al. Vehicle-road-cloud integrated autonomous driving promotes the new quality productivity of highways[J]. China Highway, 2025(9): 114-117.)

[37] Gläscher J, Daw N, Dayan P, et al. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning[J]. Neuron, 2010, 66(4): 585-595.

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

[39] Sallab A E, Abdou M, Perot E, et al. End-to-end deep reinforcement learning for lane keeping assist[J/OL]. 2016, arXiv: 1612.04340.

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

[41] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J/OL]. 2017, arXiv: 1707.06347.

[42] Fujimoto S, Hoof H V, Meger D. Addressing function approximation error in actor-critic methods[J/OL]. 2018, arXiv: 1802. 09477.

[43] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[J/OL]. 2018, arXiv: 1801.01290.

[44] Tan H N. Reinforcement learning with deep deterministic policy gradient[C]. 2021 International Conference on Artificial Intelligence, Big Data and Algorithms. Xi'an, 2021: 82-85.

[45] Paczolay G, Harmati I. A new advantage actor-critic algorithm for multi-agent environments[C]. 2020 23rd International Symposium on Measurement and Control in Robotics. Budapest, 2020: 1-6.

[46] Hadi B, Khosravi A, Sarhadi P. Cooperative motion planning and control of a group of autonomous underwater vehicles using twin-delayed deep deterministic policy gradient[J]. Applied Ocean Research, 2024, 147: 103977.

[47] 李佩璋, 费庆, 陈振, 等. 具备可解释性的决策依据自编码多智能体强化学习方法[J]. 控制与决策, 2025, 40(9): 2748-2758.
(Li P Z, Fei Q, Chen Z, et al. Interpretable decision-basis autoencoder for multi-agent reinforcement learning[J]. Control and Decision, 2025, 40(9): 2748-2758.)

[48] 程龙. 基于深度强化学习的强制换道场景下智能网联车驾驶决策研究[D]. 北京: 北京交通大学, 2022.
(Cheng L. Research on driving decision-making of intelligent networked vehicles in forced lane change scenario based on deep reinforcement learning[D]. Beijing: Beijing Jiaotong University, 2022.)

[49] Beijing Institute Tech. Automatic driving vehicle control method based on multi-agent reinforcement learning[P]. China: CN116394968A. 2023-07-07.

[50] Li Z Z, Wang S, Zhang S Y, et al. Edge-assisted V2X motion planning and power control under channel uncertainty[J]. IEEE Transactions on Vehicular Technology, 2023, 72(7): 9641-9646.

[51] Waymo LLC. Interactive autonomous vehicle agent[P]. United States: US11067988B1. 2021-07-20.

[52] An H, Ying C, Chen Y, et al. Deep-learning-based T1-enhanced selection of linear coefficients (DL-TESLA) for PET/MR attenuation correction[P]. United States:

US2022044399A1. 2022-02-10.

[53] 陈维兴, 李晨辉, 李业波. 基于 Transformer-DRL 的机坪特种车群调度策略研究[J]. 控制与决策, 2025, 40(6): 1939-1949.
(Chen W X, Li C H, Li Y B. Research on scheduling strategy of special vehicle cluster on apron based on Transformer-DRL[J]. Control and Decision, 2025, 40(6): 1939-1949.)

[54] University Jiangsu. Complex network cognition-based federated reinforcement learning end-to-end autonomous driving control system, method, and vehicular device[P]. United State: US2025128720A1. 2023-10-03.

[55] University Tsinghua. Collaborative mode diversification-oriented unsupervised multi-agent reinforcement learning method[P]. China: CN115496208A(B). 2022-12-20.

[56] University Lanzhou. Collaborative automatic driving method for risk-sensitive multi-agent reinforcement learning [P]. China: CN118569298A. 2024-08-30.

[57] 张志勇, 黄大洋, 黄彩霞, 等. TD3 算法改进与自动驾驶汽车并道策略学习[J]. 机械工程学报, 2023, 59(8): 224-234.
(Zhang Z Y, Huang D Y, Huang C X, et al. TD3 algorithm improving and lane-merging strategy learning for autonomous vehicles[J]. Journal of Mechanical Engineering, 2023, 59(8): 224-234.)

[58] 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.)

[59] Puccetti L, Yasser A, Rathgeber C, et al. Speed tracking control using model-based reinforcement learning in a real vehicle[C]. 2021 IEEE Intelligent Vehicles Symposium. Nagoya, 2021: 1213-1219.

[60] 张辰, 徐云雯, 李德伟. 车路协同环境下数据驱动的混合交通流速度调控方法[J]. 控制与决策, 2024, 39(9): 2950-2958.
(Zhang C, Xu Y W, Li D W. Data-driven speed control method for mixed traffic flow in vehicle-road cooperative environment[J]. Control and Decision, 2024, 39(9): 2950-2958.)

[61] 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, Piscataway: IEEE, 2024: 168-173.

[62] 张健, 李青扬, 李丹, 等. 基于深度强化学习的自动驾驶车辆专用道汇入引导[J]. 吉林大学学报: 工学版, 2023, 53(9): 2508-2518.
(Zhang J, Li Q Y, Li D, et al. Merging guidance of exclusive lanes for connected and autonomous vehicles based on deep reinforcement learning[J]. Journal of Jilin University: Engineering and Technology Edition, 2023, 53(9): 2508-2518.)

[63] Zhang X F, Wu L, Liu H, et al. High-speed ramp merging behavior decision for autonomous vehicles based on multiagent reinforcement learning[J]. IEEE Internet of Things Journal, 2023, 10(24): 22664-22672.

[64] 韩卓呈. 自适应巡航控制系统拟人化算法研究[D]. 长春: 吉林大学, 2024.
(Han Z C. Research on human-like adaptive cruise control algorithm[D]. Changchun: Jilin University, 2024.)

[65] 柳鹏, 赵克刚, 梁志豪, 等. 基于深度强化学习 CLPER-DDPG 的车辆纵向速度规划[J]. 汽车安全与节能学报, 2024, 15(5): 702-710.
(Liu P, Zhao K G, Liang Z H, et al. Vehicle longitudinal speed planning based on deep reinforcement learning CLPER-DDPG[J]. Journal of Automotive Safety and Energy, 2024, 15(5): 702-710.)

[66] 陈菁, 赵聪, 马裕城, 等. 车路云一体化架构下面向舒适性提升的自动驾驶速度智能决策方法[J]. 中国公路学报, 2025, 38(2): 243-257.
(Chen J, Zhao C, Ma Y C, et al. Intelligent decision-making method for autonomous driving speed for comfort improvement under vehicle-road-cloud integrated architecture[J]. China Journal of Highway and Transport, 2025, 38(2): 243-257.)

[67] Zheng K, Yang H J, Liu S W, et al. A behavior decision method based on reinforcement learning for autonomous driving[J]. IEEE Internet of Things Journal, 2022, 9(24): 25386-25394.

[68] Lee H, Kim K, Kim N, et al. Energy efficient speed planning of electric vehicles for car-following scenario using model-based reinforcement learning[J]. Applied Energy, 2022, 313: 118460.

[69] 王新凯, 王树凤, 王世皓. 基于规则约束的深度强化学习智能车辆高速路场景下行驶决策[J]. 汽车技术, 2023(9): 18-26.
(Wang X K, Wang S F, Wang S H. Rule-based constrained deep reinforcement learning for intelligent vehicle driving decisions in highway scenarios[J]. Automobile Technology, 2023(9): 18-26.)

[70] 冯耀, 景首才, 惠飞, 等. 基于深度强化学习的智能网联车辆换道轨迹规划方法[J]. 汽车安全与节能学报, 2022, 13(4): 705-717.
(Feng Y, Jing S C, Hui F, et al. Deep reinforcement learning-based lane-changing trajectory planning method of intelligent and connected vehicles[J]. Automotive Safety and Energy, 2022, 13(4): 705-717.)

[71] 耿玺钧, 崔立堃, 熊高, 等. 子目标驱动 DQN 算法的无人车狭窄转弯环境导航[J]. 控制与决策, 2024, 39(11): 3637-3644.
(Geng X J, Cui L K, Xiong G, et al. Navigation in narrow turning environment of unmanned vehicle based on subgoal-driven DQN algorithm[J]. Control and Decision, 2024, 39(11): 3637-3644.)

[72] 王曙燕, 万顷田. 自动驾驶车辆在无信号交叉口右转

驾驶决策技术研究[J]. 计算机应用研究, 2023, 40(5): 1468-1472.
(Wang S Y, Wan Q T. Right-turn driving decisions of autonomous vehicles at signal-free intersections[J]. Application Research of Computers, 2023, 40(5): 1468-1472.)

[73] Zhang R S, Ishikawa A, Wang W L, et al. Using reinforcement learning with partial vehicle detection for intelligent traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(1): 404-415.

[74] Li Z, Yuan S H, Yin X F, et al. Research into autonomous vehicles following and obstacle avoidance based on deep reinforcement learning method under map constraints[J]. Sensors, 2023, 23(2): 844.

[75] Savari M, Choe Y. Utilizing human feedback in autonomous driving: Discrete vs. continuous[J]. Machines, 2022, 10(8): 609.

[76] Ni H Y, Yu G Z, Chen P, et al. An integrated framework of lateral and longitudinal behavior decision-making for autonomous driving using reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2024, 73(7): 9706-9720.

[77] Xu R C, Xu J M, Liu X, et al. Safe hybrid-action reinforcement learning-based decision and control for discretionary lane change[J]. Machines, 2024, 12(4): 252.

[78] Yao X, Du Z C, Sun Z B, et al. Cooperative lane-changing in mixed traffic: A deep reinforcement learning approach[J]. Transportmetrica A: Transport Science, 2024: 1-23.

[79] Wang Z, Huang H L, Tang J J, et al. A deep reinforcement learning-based approach for autonomous lane-changing velocity control in mixed flow of vehicle group level[J]. Expert Systems with Applications, 2024, 238: 122158.

[80] Shi Y Y, Liu J H, Liu C Q, et al. DeepAD: An integrated decision-making framework for intelligent autonomous driving[J]. Transportation Research Part A: Policy and Practice, 2024, 183: 104069.

[81] Yang Z, Zhang R S, Pandey G, et al. A hierarchical vehicle behavior prediction framework with traffic signals and interactive agents[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(10): 11066-11079.

[82] Xiao W X, Yang Y Y, Mu X Y, et al. Decision-making for autonomous vehicles in random task scenarios at unsignalized intersection using deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2024, 73(6): 7812-7825.

[83] Lavalle S M. Planning algorithms[M]. Cambridge: Cambridge University Press, 2006.

[84] Yang H, Feng Z Y, Wei Z Q, et al. Intelligent computation offloading for joint communication and sensing-based vehicular networks[J]. IEEE Transactions on Wireless Communications, 2024, 23(4): 3600-3616.

[85] Park K S. A study for improving driving safety assurance for fully autonomous vehicles-focusing on amendments of the German road traffic act and the Japanese road traffic act[J]. Journal of Auto-Vehicle Safety Association, 2019, 15(1): 45-54.

[86] 郑飞. 中国自动驾驶汽车法律规制的基本构想[J]. 社会科学辑刊, 2025(4): 186-200.
(Zheng F. The basic structure of China's legal regulation of autonomous vehicles[J]. Social Science Journal, 2025(4): 186-200.)

[87] Chen C, Jiang J G, Lv N, et al. An intelligent path planning scheme of autonomous vehicles platoon using deep reinforcement learning on network edge[J]. IEEE Access, 2020, 8: 99059-99069.

[88] Talamini J, Bartoli A, de Lorenzo A, et al. On the impact of the rules on autonomous drive learning[J]. Applied Sciences, 2020, 10(7): 2394.

[89] Bin Issa R, Das M, Rahman M S, et al. Double deep *Q*-learning and faster R-CNN-based autonomous vehicle navigation and obstacle avoidance in dynamic environment[J]. Sensors, 2021, 21(4): 1468.

[90] 傅明建, 郭福强. 基于深度强化学习的无信号灯路口决策研究[J]. 计算机工程, 2024, 50(5): 91-99.
(Fu M J, Guo F Q. Research on decision-making at intersection without traffic lights based on deep reinforcement learning[J]. Computer Engineering, 2024, 50(5): 91-99.)

[91] 刘一鸣. 基于奖励设计的深度强化学习算法研究与应用[D]. 北京: 北京邮电大学, 2020.
(Liu Y M. Research and application of deep reinforcement learning algorithm based on reward design[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.)

[92] Huang Z Y, Liu H C, Wu J D, et al. Conditional predictive behavior planning with inverse reinforcement learning for human-like autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(7): 7244-7258.

[93] 郑川, 杜煜, 刘子健. 基于模糊收敛和模仿强化学习的自动驾驶横向控制方法[J]. 汽车技术, 2024(7): 29-36.
(Zheng C, Du Y, Liu Z J. A lateral control method of autonomous driving based on fuzzy convergence and imitative reinforcement learning[J]. Automobile Technology, 2024(7): 29-36.)

[94] Ok J H, Woo H. Development of personalized autonomous driving agents using imitation learning[J]. Journal of KIISE, 2024, 51(6): 558-566.

[95] 李光泽. 混合交通环境下网联自动驾驶车辆行驶决策与控制方法研究[D]. 西安: 长安大学, 2022.
(Li G Z. Research on decision-making and control method of connected and automated vehicles in mixed traffic[D]. Xi'an: Chang'an University, 2022.)

[96] 任玥, 邹博文, 尹旭, 等. 考虑驾驶员特性的个性化跟驰控制策略研究[J]. 西南大学学报: 自然科学版, 2022, 44(3): 12-19.
(Ren Y, Zou B W, Yin X, et al. Study of personalized

car-following control strategy by considering the driver characteristics[J]. Journal of Southwest University: Natural Science Edition, 2022, 44(3): 12-19.)

[97] Zhu M X, Wang Y H, Pu Z Y, et al. Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving[J]. Transportation Research Part C: Emerging Technologies, 2020, 117: 102662.

[98] 陈越, 焦朋朋, 白如玉, 等. 基于深度强化学习的自动驾驶车辆跟驰行为建模[J]. 交通信息与安全, 2023, 41(2): 67-75.
(Chen Y, Jiao P P, Bai R Y, et al. Modeling car following behavior of autonomous driving vehicles based on deep reinforcement learning[J]. Journal of Transport Information and Safety, 2023, 41(2): 67-75.)

[99] Liu X, Liu Y W, Chen Y, et al. Enhancing the fuel-economy of V2I-assisted autonomous driving: A reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2020, 69(8): 8329-8342.

[100] Liao J D, Liu T, Tang X L, et al. Decision-making strategy on highway for autonomous vehicles using deep reinforcement learning[J]. IEEE Access, 2020, 8: 177804-177814.

[101] 刘学高. 基于深度强化学习的自动驾驶拟人化跟驰决策算法研究[D]. 重庆: 西南大学, 2023.
(Liu X G. Research on autonomous driving human-like car-following decision algorithm based on deep reinforcement learning[D]. Chongqing: Southwest University, 2023.)

[102] Al-Qizwini M, Bulan O, Qi X W, et al. A lightweight simulation framework for learning control policies for autonomous vehicles in real-world traffic condition[J]. IEEE Sensors Journal, 2021, 21(14): 15762-15774.

[103] Lin J Q, Zhang P, Li C G, et al. APF-DPPO: An automatic driving policy learning method based on the artificial potential field method to optimize the reward function[J]. Machines, 2022, 10(7): 533.

[104] Du D F, Shen M Y, Guo X R, et al. Hierarchical driving strategy for connected and autonomous vehicles making a protected left turn at signalized intersections[J]. Journal of Transportation Engineering, Part A: Systems, 2023, 149(3): 04022154.

[105] 牟浪. 基于深度强化学习的单智能体高速路段自动驾驶算法研究[D]. 成都: 西南财经大学, 2022.
(Mu L. Research on autonomous driving algorithm based on deep reinforcement learning for single agent on high speed road sections[D]. Chengdu: Southwestern University of Finance and Economics, 2022.)

[106] Prathiba S B, Raja G, Dev K, et al. A hybrid deep reinforcement learning for autonomous vehicles smart-platooning[J]. IEEE Transactions on Vehicular Technology, 2021, 70(12): 13340-13350.

[107] Huang H S, Li Y, Song G, et al. Deep reinforcement learning-driven UAV data collection path planning: A study on minimizing AoI[J]. Electronics, 2024, 13(10): 1871.

[108] He X K, Lv C. Toward personalized decision making for autonomous vehicles: A constrained multi-objective reinforcement learning technique[J]. Transportation Research Part C: Emerging Technologies, 2023, 156: 104352.

[109] Huang Y J, Yang S, Wang L W, et al. An efficient self-evolution method of autonomous driving for any given algorithm[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(1): 602-612.

[110] Chen L Q, He Y, Pan W K, et al. A novel generalized meta hierarchical reinforcement learning method for autonomous vehicles[J]. IEEE Network, 2023, 37(4): 230-236.

[111] Guan J Y, Chen G, Huang J, et al. A discrete soft actor-critic decision-making strategy with sample filter for freeway autonomous driving[J]. IEEE Transactions on Vehicular Technology, 2023, 72(2): 2593-2598.

[112] Lee D S, Kwon M. Stability analysis in mixed-autonomous traffic with deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2023, 72(3): 2848-2862.

[113] Liu C, Wei W, Liang B F, et al. ConvMLP-mixer based real-time stereo matching network towards autonomous driving[J]. IEEE Transactions on Vehicular Technology, 2023, 72(2): 2581-2586.

[114] Lee H, Jee J, Oh C, et al. Derivation of driving stability indicators for autonomous vehicles based on analyzing waymo open dataset[J]. The Journal of the Korea Institute of Intelligent Transport Systems, 2024, 23(4): 94-109.

[115] Rosero L A, Gomes I P, da Silva J A R, et al. Integrating modular pipelines with end-to-end learning: A hybrid approach for robust and reliable autonomous driving systems[J]. Sensors, 2024, 24(7): 2097.

[116] Zhou Y, Chen Y X. Learning to drive in the NGSIM simulator using proximal policy optimization[J]. Journal of Advanced Transportation, 2023, 2023: 4127486.

[117] Geng M S, Cai Z E, Zhu Y Z, et al. Multimodal vehicular trajectory prediction with inverse reinforcement learning and risk aversion at urban unsignalized intersections[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(11): 12227-12240.

[118] 李弘扬, 李阳, 王晖杰, 等. 自动驾驶开源数据体系: 现状与未来[J]. 中国科学: 信息科学, 2024, 54(6): 1283-1318.
(Li H Y, Li Y, Wang H J, et al. Open-sourced data ecosystem in autonomous driving: The present and future[J]. Science in China: Information Sciences, 2024, 54(6): 1283-1318.)

[119] Chandra R. Towards autonomous driving in dense, heterogeneous, and unstructured traffic[D]. Maryland: University of Maryland, 2022.

[120] Wen X, Jian S S, He D B. Modeling the effects of

autonomous vehicles on human driver car-following behaviors using inverse reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 13903-13915.

[121] Wang J, Sun H Y, Zhu C. Vision-based autonomous driving: A hierarchical reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2023, 72(9): 11213-11226.

[122] Ben Elallid B, Bagaa M, Benamar N, et al. A reinforcement learning based autonomous vehicle control in diverse daytime and weather scenarios[J]. Journal of Intelligent Transportation Systems, 2024: 1-14.

[123] Khalil Y H, Mouftah H T. Exploiting multi-modal fusion for urban autonomous driving using latent deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2023, 72(3): 2921-2935.

[124] Sony Corp. Vision-based sample-efficient reinforcement learning framework for autonomous driving[P]. United States: US11106211B2(A1). 2019-10-03.

[125] Hoel Carl Johan, Laine Leo, VOLVO Autonomous Solutions AB. Managing aleatoric and epistemic uncertainty in reinforcement learning, with applications to autonomous vehicle control[P]. United States: US2022374705A1. 2022-11-09.

[126] 邹沅江. 多智能体强化学习协同编队算法与实验研究[D]. 成都: 电子科技大学, 2024.
(Zou Y J. Research on multi-agent reinforcement learning collaborative formation algorithm and experiment[D]. Chengdu: University of Electronic Science and Technology of China, 2024.)

[127] 同济大学. 一种基于强化学习的自动驾驶编队队形重组方法[P]. 中国: 202410857114.1. 2024-11-01.
(Tongji University. A formation reorganization method for autonomous driving formation based on reinforcement learning[P]. China: 202410857114.1. 2024-11-01.)

[128] 周舒雅. 基于强化学习的货车编队系统自适应资源优化分配研究[D]. 成都: 西南交通大学, 2021.
(Zhou S Y. Research on adaptive resource allocation of truck formation system based on reinforcement learning[D]. Chengdu: Southwest Jiaotong University, 2021.)

[129] Liu D Y, Liu H, Lv J H, et al. Time-varying formation of heterogeneous multiagent systems *via* reinforcement learning subject to switching topologies[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2023, 70(6): 2550-2560.

[130] 北京捷升通达信息技术有限公司. 一种无人车编队轨迹规划控制系统及方法 [P]. 中国: 202411771964.6. 2025-03-07.
(Beijing Jiesheng Tongda Information Technology Co., Ltd. An unmanned vehicle formation trajectory planning control system and method[P]. China: 202411771964.6. 2025-03-07.)

[131] 李春, 吴志周, 曾广, 等. 合流区智能网联汽车协同控制方法综述[J]. 计算机工程与应用, 2024, 60(12): 1-17.
(Li C, Wu Z Z, Zeng G, et al. Review of connected autonomous vehicle cooperative control at on-ramp merging areas[J]. Computer Engineering and Applications, 2024, 60(12): 1-17.)

[132] Wang C, Xu Y, Zhang J, et al. Integrated traffic control for freeway recurrent bottleneck based on deep reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(9): 15522-15535.

[133] 鲁子洋. 基于深度强化学习的自动驾驶匝道汇入决策研究[D]. 长春: 吉林大学, 2023.
(Lu Z Y. Research on ramp merging decision-making for autonomous driving based on deep reinforcement learning [D]. Changchun: Jilin University, 2023.)

[134] Chen D, Hajidavalloo M R, Li Z J, et al. Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(11): 11623-11638.

[135] 王子洋. 基于深度强化学习的车辆匝道合流控制方法研究[D]. 北京: 北京交通大学, 2022.
(Wang Z Y. Research on vehicle on-ramp merging control method based on deep reinforcement learning[D]. Beijing: Beijing Jiaotong University, 2022.)

[136] Le N T T. Multi-agent reinforcement learning for traffic congestion on one-way multi-lane highways[J]. Journal of Information and Telecommunication, 2023, 7(3): 255-269.

[137] Zhan G, Zhang X M, Li Z C, et al. Multiple-UAV reinforcement learning algorithm based on improved PPO in ray framework[J]. Drones, 2022, 6(7): 166.

[138] 同济大学. 一种基于 MADDPG 的自动驾驶车辆匝道合流协同控制方法及系统 [P]. 中国: 202210889860.X. 2023-08-29.
(Tongji University. A cooperative control method and system for ramp merging of autonomous vehicles based on MADDPG[P]. China: 202210889860.X. 2023-08-29.)

[139] 焦岩. 基于多智能体强化学习的智能网联车辆协同决策控制研究[D]. 重庆: 重庆理工大学, 2024.
(Jiao Y. Research on cooperative decision control of intelligent connected vehicles based on multi-agent reinforcement learning[D]. Chongqing: Chongqing University of Technology, 2024.)

[140] Liu B, Ding Z T. A distributed deep reinforcement learning method for traffic light control[J]. Neurocomputing, 2022, 490: 390-399.

[141] Wang H C, Huang S C, Huang P J, et al. Curriculum reinforcement learning from avoiding collisions to navigating among movable obstacles in diverse environments[J]. IEEE Robotics and Automation Letters, 2023, 8(5): 2740-2747.

[142] Harris A, Valade T, Teil T, et al. Generation of spacecraft operations procedures using deep reinforcement learning[J]. Journal of Spacecraft and Rockets, 2022, 59(2): 611-626.

[143] 马万经, 李金珏, 俞春辉. 智能网联混合交通流交叉口控制: 研究进展与前沿[J]. 中国公路学报, 2023, 36(2): 23-40.
(Ma W J, Li J J, Yu C H. Intersection control in mixed traffic with connected automated vehicles: A review of recent developments and research frontiers[J]. China Journal of Highway and Transport, 2023, 36(2): 23-40.)

[144] 杨晓光, 赖金涛, 张振, 等. 车路协同环境下的轨迹级交通控制研究综述[J]. 中国公路学报, 2023, 36(9): 225-243.
(Yang X G, Lai J T, Zhang Z, et al. Review of trajectory based traffic control in a vehicle-infrastructure cooperative environment[J]. China Journal of Highway and Transport, 2023, 36(9): 225-243.)

[145] 许润南. 车路协同环境下交叉口信号与车道协同控制方法研究[D]. 赣州: 江西理工大学, 2023.
(Xu R N. Cooperative control method of intersection signal and lane in vehicle-road cooperative environment [D]. Ganzhou: Jiangxi University of Science and Technology, 2023.)

[146] 徐泽洲, 曲大义, 洪家乐, 等. 智能网联汽车自动驾驶行为决策方法研究[J]. 复杂系统与复杂性科学, 2021, 18(3): 88-94.
(Xu Z Z, Qu D Y, Hong J L, et al. Research on decision-making method for autonomous driving behavior of connected and automated vehicle[J]. Complex Systems and Complexity Science, 2021, 18(3): 88-94.)

[147] Cakija D, Assirati L, Ivanjko E, et al. Autonomous intersection management: A short review[C]. 2019 International Symposium ELMAR. Zadar, 2019: 21-26.

[148] Pan X, Chen B L, Dai L, et al. A hierarchical robust control strategy for decentralized signal-free intersection management[J]. IEEE Transactions on Control Systems Technology, 2023, 31(5): 2011-2026.

[149] Xu B, Li S E, Bian Y G, et al. Distributed conflict-free cooperation for multiple connected vehicles at unsignalized intersections[J]. Transportation Research Part C: Emerging Technologies, 2018, 93: 322-334.

[150] 吴伟豪. 无信号灯交叉口智能车路径规划算法研究[D]. 南京: 南京林业大学, 2023.
(Wu W H. Research on intelligent vehicle path planning algorithm for unsignalized intersections [D]. Nanjing: Nanjing Forestry University, 2023.)

[151] 金立生, 魏青嵩, 谢宪毅, 等. 基于 DMPC 的无信控交叉口智能网联车辆多车协同轨迹规划[J]. 汽车安全与节能学报, 2024, 15(2): 235-241.
(Jin L S, Wei Q S, Xie X Y, et al. Multi-vehicle cooperative path planning at untrusted intersections based on DMPC[J]. Journal of Automotive Safety and Energy, 2024, 15(2): 235-241.)

[152] University Jilin. Non-signalized intersection complete autono-mous traffic flow traffic control method[P]. China: CN117636661A (B). 2024-03-01.

[153] 董玮, 李岩, 郭宏伟, 等. 无信号交叉口自动驾驶汽车协同驾驶策略研究[J]. 交通科技与管理, 2024, 5(4): 8-12.
(Dong W, Li Y, Guo H W, et al. Research on cooperative driving strategy of autonomous vehicles at unsignalized intersections[J]. Transportation Technology and Management, 2024, 5(4): 8-12.)

[154] 季文韬. 基于强化学习的高速公路施工区可变限速控制方法[D]. 南京: 东南大学, 2020.
(Ji W T. Variable speed limit control method for highway construction area based on reinforcement learning [D]. Nanjing: Southeast University, 2020.)

[155] 肖哲. 车联网条件下八车道高速公路施工区可变限速控制方法[D]. 南京: 东南大学, 2022.
(Xiao Z. Variable speed limit control in the work zone area of two-way eight-lane highway under the existence of connected automated vehicles[D]. Nanjing: Southeast University, 2022.)

[156] 梁志康. 高速公路施工区智能网联车辆控制策略仿真研究[D]. 长春: 吉林大学, 2019.
(Liang Z K. Modeling control strategies of intelligent connected vehicle in expressway work zone[D]. Changchun: Jilin University, 2019.)

[157] 李晓虎. 基于强化学习的高速公路施工区智能车辆运行控制研究[D]. 西安: 长安大学, 2021.
(Li X H. Research on connected and autonomous vehicle operation control in freeway work zone area based on reinforcement learning[D]. Xi'an: Chang'an University, 2021.)

作者简介

王云泽 (1988−), 男, 副教授, 硕士生导师, 主要研究方向为智能交通系统、车路协同控制、交通大数据分析, E-mail: wangyunze@stdu.edu.cn;

孙宇 (2001−), 女, 硕士生, 主要研究方向为自动驾驶决策与路径规划、交通行为建模, E-mail: Sunyuuu0622@163.com;

骆中斌 (1987−), 男, 高级工程师, 博士, 主要研究方向为交通仿真与建模、城市交通管理、交通基础设施智能化, E-mail: luozhongbin@stu.cqu.edu.cn;

张春波 (1988−), 男, 博士, 硕士生导师, 主要研究方向为交通系统优化、智慧高速与出行服务、交通安全评估, E-mail: zhangchunbochn@yeah.net.