**文章编号**：0253-374X(2024)04-0520-12
**DOI**：10.11908/j.issn.0253-374x.23397

# 强化学习在自动驾驶技术中的应用与挑战

何逸煦[1,2]，林泓熠[3]，刘　洋[3]，杨　澜[2]，曲小波[3]
(1. 华南理工大学 土木与交通学院，广东 广州 510640；2. 长安大学 信息工程学院，陕西 西安 710064；
3. 清华大学 车辆与运载学院，北京 100084)

**摘要**：围绕强化学习在自动驾驶领域的应用进行了多方面的概括和总结。对强化学习原理及发展历程进行了介绍；系统介绍了自动驾驶技术体系以及强化学习在自动驾驶领域的应用所需的基础；按不同的应用方向分别介绍了强化学习在自动驾驶领域中的应用案例；深入分析了现阶段强化学习在自动驾驶领域存在的挑战，并提出若干展望。

**关键词**：强化学习；自动驾驶；人工智能
**中图分类号**：U461　　**文献标志码**：A

## Applications and Challenges of Reinforcement Learning in Autonomous Driving Technology

*HE Yixu*[1,2], *LIN Hongyi*[3], *LIU Yang*[3], *YANG Lan*[2], *QU Xiaobo*[3]
(1. School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510640, China; 2. School of Information Engineering, Chang'an University, Xi'an 710064, China; 3. School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China)

**Abstract**: This paper provides a comprehensive overview and summary of the application of reinforcement learning in the field of autonomous driving. First, an introduction to the principles and development of reinforcement learning is presented. Following that, the autonomous driving technology system and the fundamentals required for the application of reinforcement learning in this field are systematically introduced. Subsequently, application cases of reinforcement learning in autonomous driving are described according to different directions of use. Finally, the current challenges of applying reinforcement learning in the field of autonomous driving are deeply analyzed, and several prospects are proposed.

**Keywords**: reinforcement learning; autonomous driving; artificial intelligence

近年来，人工智能在各个领域得到了广泛应用。其快速发展为智能交通系统注入了蓬勃动力，进一步推动了自动驾驶、车路协同等领域的发展，也助推了汽车工业的革新[1]。此外，人工智能技术的发展还推动了5G通信技术、车联网技术等相关科技的迅速崛起和创新。在此背景下，我国正大力推动互联网、大数据、人工智能同产业深度融合[2]。

赋予人工智能“思想”的机器学习算法，尤其是深度学习算法以其卓越的表现和广泛的适用性，被应用于解决各个领域的实际问题[3]。而在自动驾驶相关领域，现有的监督学习算法和无监督学习算法同样解决了许多关键问题[4-5]。例如，谷歌开发的SurfelGAN网络可以重建真实的自动驾驶汽车感知数据供自动驾驶系统学习[6]。特斯拉部分车型采用纯摄像头与深度学习算法即可实现辅助驾驶和主动安全功能。随着大模型(foundation models)技术的兴起，特斯拉还将基于Transformer的端到端大模型应用于FSD(full self-driving)完全自动驾驶功能的开发，并且在实际测试中达到了接近人类驾驶员的水平。上海人工智能实验室、武汉大学、商汤科技团队联合提出的感知决策一体化的自动驾驶通用大模型UniAD，首次将检测、跟踪、建图、轨迹预测、规划等任务整合到一个基于Transformer的端到端网络框架下，进一步论证了大模型与自动驾驶产业结合的潜力[7]。

收稿日期：2023-08-07
基金项目：国家重点研发计划(2021YFB2501205)；国家自然科学基金(52220105001, 52221005, 72322002, 72361137001, 72331001, 52131204)；长安大学研究生科研创新实践项目(300103723040)
第一作者：何逸煦，博士生，主要研究方向为智慧车辆与智能交通系统。E-mail：heyixu2023@163.com
通信作者：刘　洋，助理研究员，工学博士，主要研究方向为机器学习与智能交通系统。
E-mail：thu_ets_lab@tsinghua.edu.cn

论文拓展介绍

自从 DeepMind 的强化学习模型 AlphaGo 在 2016 年战胜多名人类围棋大师之后，强化学习在公众面前直观地展示了其强大的实力，引发了人们对人工智能潜力的无限遐想。强化学习的概念于上世纪 50 年代被提出，最早在最优控制等领域得到应用并逐步发展为机器学习的一大分支[8]，其本质是通过模拟人类学习新事物时不断试错并做出新的尝试的过程来学习最优的策略。在经历不断迭代与优化后，强化学习在目前的研究中已经得到了十分广泛的应用。OpenAI 于 2020 年开发出大型语言模型 GPT-3[9]，随后推出的 ChatGPT 聊天机器人一经面世就引起巨大轰动，作为可以“独立思考”的对话机器人，令公众对人工智能的能力有了全新的认知，而 ChatGPT 的诞生也得益于强化学习方法的巧妙运用，这也反映了强化学习已经成为人工智能新兴技术发展中不可或缺的一部分。在智能交通这个大量研究需要数据驱动的领域，学者们也在不断发掘强化学习用于解决实际问题的应用价值，以“reinforcement learning”和“autonomous driving”为关键词在 Web of Science 上按年份搜索得到的论文发表数量显示，自 2016 年起有越来越多的研究将强化学习应用于自动驾驶技术，并且相关论文数量呈逐年增加的趋势，不仅表明强化学习在自动驾驶领域的应用已经逐渐成为学术研究的热点，也体现了自动驾驶汽车产业的蓬勃发展与对自动驾驶技术关键解决方案的迫切追求。近期，清华大学等单位科研人员合作研发出基于密集强化学习的自动驾驶汽车加速测试方法，大大加速了自动驾驶技术的落地，相关研究成果发表在期刊《自然》上，将越来越多自动驾驶技术相关研究人员的目光吸引到强化学习这一机器学习方法上[10]。

目前关于强化学习的综述多着眼于强化学习的原理与发展以及其在多学科领域的整体应用，缺少针对强化学习在自动驾驶领域最新进展的系统性综述。本研究旨在填补这一空白，首先介绍强化学习的原理与发展，其次根据现有文献总结强化学习在自动驾驶领域的最新应用和发展趋势，分析该方向目前存在的主要挑战，并讨论其未来可能的发展方向。

# 1　强化学习原理与发展

## 1.1　基本原理

强化学习的基本原理是模仿人类学习的方式，让智能体在交互环境中不断试错，并根据反馈不断调整自身策略，从而学习到最优策略。强化学习主要包括状态 $s$(state)、策略 $\pi$(policy)、动作 $a$(action)、奖励 $r$(reward)等 4 个要素。强化学习的理论基础是马尔可夫决策过程(Markov decision process，MDP)，主要包括基于价值的强化学习方法和基于策略的强化学习方法两种代表性学习策略[11]。

在马尔可夫决策过程中，考虑一个智能体与一个随机且完全可观察的环境进行交互的情境。在这个情境中，智能体需要在离散的时间步骤中选择行动，目标是最大化累积奖励。马尔可夫决策过程是由 5 个元素组成的元组，记作 $(S, A, P, R, \gamma)$。其中，$S$ 代表状态的集合，$A$ 代表智能体可以选择的行动的集合，$R$ 代表奖励的集合，$\gamma \in (0, 1)$ 是一个折扣因子，用于调整未来的奖励，$P$ 代表满足马尔可夫性质的转移概率，具体表达式如公式(1)所示。

$$P(s_{t+1}|s_1, a_1, s_2, a_2, \cdots, s_t, a_t) = P(s_{t+1}|s_t, a_t) \tag{1}$$

式中：$s_t$，$a_t$ 分别为时间步 $t$ 的状态和智能体选择的动作，表示未来的状态 $s_{t+1}$ 只依赖于当前的状态 $s_t$ 和动作 $a_t$，而与过去的状态和动作无关。

智能体在马尔可夫决策过程中的学习步骤如图 1 所示，在每个时间步 $t$，智能体从环境中接收到状态 $s_t$，并根据策略 $\pi(a_t|s_t)$ 选择最佳可能的动作 $a_t$。这个策略是一个从状态 $s_t$ 到动作 $a_t$ 的映射。在采取动作 $a_t$ 后，状态会相应地更新，同时智能体会从环境中获得一个奖励 $r_t$，用来评估智能体采取动作 $a_t$ 的好坏。

在以上过程中，智能体的目标是最大化从每个状态 $s_t$ 开始的期望折扣奖励，记作 $G_t = \sum_{k=0}^{\infty} \gamma^k r_{k+t}$。这意味着智能体需要在考虑即时奖励的同时，也要考虑未来可能获得的奖励，而折扣因子 $\gamma$ 则决定了智能体对未来奖励的重视程度。

## 1.2　发展历程

### 1.2.1　传统强化学习

强化学习的发展历程可以追溯至 20 世纪 50 年代，当时的研究主要聚焦于试错学习的早期形式。1954 年，著名的认知科学家 Minsky 首次提出了强化学习的概念，奠定了强化学习的基础[8]。此后的研究工作继续深化和扩展这一领域。1957 年，Bellman 引入了马尔可夫决策过程的概念，并提出了动态规划(dynamic programming，DP)方法，这种方法可以解决一类具有已知环境模型的优化问题[11]。在接下来的几年里，Howard 于 1960 年提出了策略迭代方

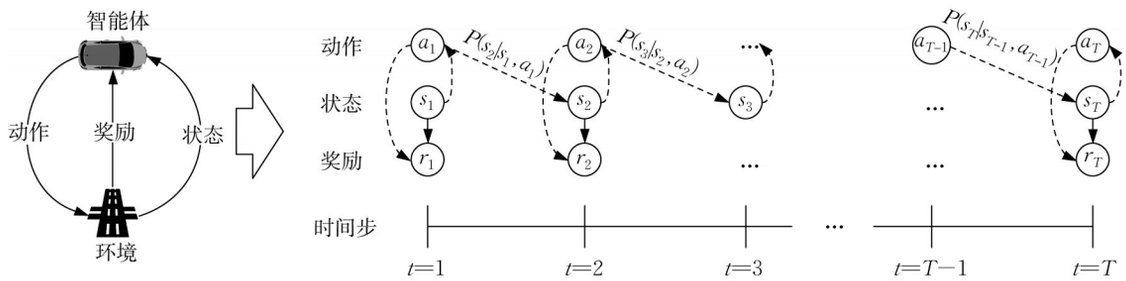



**图1 马尔可夫决策过程**

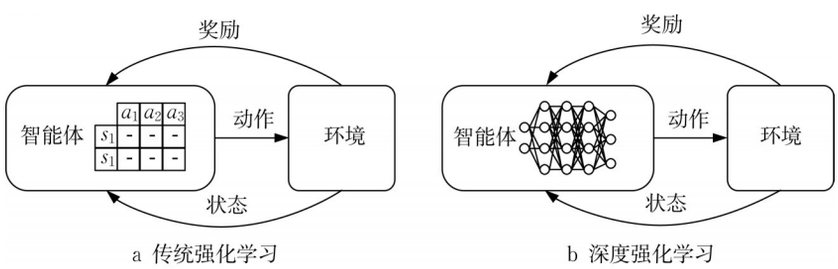
**Fig.1 Graphic representation of Markov decision process**

法,进一步丰富了马尔可夫决策过程的求解方法[12]。

20世纪80年代末,基于样本的学习方法开始崭露头角。1988年,Sutton[13]提出了时间差分学习(temporal difference learning,TD Learning)方法,该方法结合了蒙特卡罗方法和动态规划方法的优点。基于此,Watkins[14]在1989年提出了Q-learning算法,Rummery等[15]在1994年进一步提出了SARSA(state-action-reward-state-action)算法,这两种方法都是重要的基于价值的强化学习算法,用于处理有限的马尔可夫决策过程。

Q-learning算法和SARSA算法通常都使用一种叫做Q表的结构来存储并更新各个状态-动作对应的价值信息,如图2a所示,Q表中的行代表可能的状态,列代表潜在的动作,单元格代表预期的总奖励。然而,在处理具有大规模状态空间或连续状态空间的问题时,Q表的规模会庞大至无法处理,在信息爆炸、人工智能需求日益增长的背景下,这类传统强化学习方法的应用受到了很多限制。

### 1.2.2 深度强化学习




a 传统强化学习 b 深度强化学习

**图2 传统强化学习和深度强化学习原理图**

**Fig.2 Diagram of traditional reinforcement learning and deep reinforcement learning principles**

在现实世界应用中,强化学习通常需要处理高维连续的状态空间,为了克服传统强化学习在这方面的限制,Mnih等[16]在2015年引入了深度强化学习(deep reinforcement learning,DRL)的概念,提出DQN(deep Q-network)算法用于优化基于价值的强化学习方法。DQN采用深度神经网络作为函数近似器,该网络采用Q-learning算法的变体进行训练,使用随机梯度下降法更新权重。DQN算法的稳定表现还得益于另一关键技术:经验回放(experience replay)机制的引入。在经验回放机制中,用元组$(s_t, a_t, r_t, s_{t+1})$存储智能体与环境交互的历史经验,批量存放于经验回放池(experience replay buffer)中。

尽管DQN算法的提出为处理高维连续的状态空间提供了有效工具,但是基于价值的深度强化学习方法难以用于处理连续的动作空间。因此,为了解决连续动作空间的问题,研究者们开始转向基于策略的深度强化学习方法,通过深度神经网络近似策略,并借助策略梯度手段实现最优策略的求解,于是先后提出了信任区域策略优化(trust region policy optimization,TRPO)算法[17]、近端策略优化(proximal policy optimization,PPO)算法[18]、soft actor-critic(SAC)算法[19]等随机性策略方法

除了随机性策略方法外,深度强化学习中还存在学习确定性策略的方法。其中,深度确定性策略梯度(deep deterministic policy gradient,DDPG)算法作为一种应用广泛的确定性策略方法,由Lillicrap等[20]于2015年提出。DDPG算法采用行动者-评论家(actor-critic,AC)架构,包含两种用途不同的深度神经网络,用于导出确定性策略$\mu_\phi: S \to A$。其中,

Actor网络$\mu_\phi$以状态$s$作为输入，根据估计的$Q$值函数更新策略，并输出连续动作$a=\mu_\phi(s)$。Critic网络$Q_\theta$以状态$s$和动作$a$作为输入，并输出$Q$值函数$Q_\theta(s,a)$的标量估计值。与DQN算法类似，DDPG算法也使用经验回放池存储历史经验，并从中采样小批量的经验来更新网络。此外，DDPG算法引入了Actor网络和Critic网络的目标网络，分别表示为$\mu_\phi(s_t)$和$Q_\theta(s_t, a_t)$，并采用限制目标值变化速度的软更新方法来稳定学习过程。在DDPG算法的基础上，衍生出一系列确定性策略深度强化学习方法，进一步提升了算法的稳定性和效率。其中，Fujimoto等[21]在2018年提出的twin delayed deep deterministic policy gradient（TD3）算法被广泛应用，在处理高维连续动作空间和稀疏奖励等问题上表现出优越的性能。

近年来，学者在强化学习领域不断发掘新的研究方向，其中包括逆强化学习的提出和发展，这种新方法利用观察到的优化行为来推断相应的奖励函数。此外，多智能体强化学习也在不断探索和发展，该方法考虑在一个环境中存在多个学习和决策的主体，带来了全新的挑战和机遇。这些新领域的进展不仅拓宽了强化学习的应用范围，也为解决更复杂的问题提供了可能性。

## 2　强化学习在自动驾驶中的应用与挑战

### 2.1　自动驾驶简介

自动驾驶系统由感知、规划与决策、控制三个主要模块组成，如图3所示[2]。感知模块通过车辆周围的传感器数据获取环境信息，如道路状况、障碍物、交通信号灯和行人等，并将这些信息转化为车辆可以理解和处理的格式。规划与决策模块接收感知模块的数据，基于车辆当前位置、目标位置和环境动态，制定车辆行驶策略和路线规划。此模块还实时做出决策，如加速、减速、转弯、超车等，以确保车辆的安全性和高效性。控制模块负责执行决策模块的指令，通过控制发动机、刹车、转向等执行器来精确驾驶车辆。这三个模块协同工作，确保了自动驾驶系统的自主安全行驶。

在自动驾驶系统中，强化学习主要应用于规划与决策和控制模块。在规划与决策中，强化学习基于感知信息进行路径规划和适应不同道路状况的策略学习，使车辆能自主做出合适的决策。控制模块则利用强化学习算法根据驾驶操作和环境反馈优化控制策略，实现精准稳定的车辆控制。


**图3　自动驾驶系统主要组成模块**

**Fig.3　Main components of autonomous driving system**

### 2.2　数据集与训练环境

作为一种数据驱动的方法，强化学习依赖于与环境的交互以学习最优策略。在自动驾驶系统中，智能体通常需要面临一系列复杂多变的实际驾驶场景。为了能够将强化学习算法应用于自动驾驶系统并有效地训练模型，通常需要具备海量的真实数据和精确的仿真环境。真实数据可以帮助强化学习模型更准确地理解实际驾驶情况，而仿真环境则能够提供一个安全、有效、可控的平台来进行模型训练和策略优化。

自动驾驶系统的开发依赖于车载传感器、路侧感知单元、航拍数据等多种数据来源，这些数据源共同构成了对环境的全方位、多级别的理解，为自动驾驶系统的决策和控制提供了强大的支持。针对不同的自动驾驶任务，研究人员通常会结合实际的需求选取合适的数据。已有的研究中涉及了多种自然驾驶数据集，为强化学习模型的训练和测试提供了丰富的、真实的驾驶情境数据，其中部分代表性数据集如表1所示。

在已有的自然驾驶数据集中，以NGSIM和HighD为代表的一些数据集采用鸟瞰视角的航拍视频数据，通过计算机视觉技术提取车辆位置。这些数据集通常涵盖了多车道的交通流信息，包含大量车辆轨迹和车辆之间的交互行为，非常适合研究自动驾驶车辆在交通流中的驾驶策略；另一方面，KITTI和ApolloScape等数据集主要借助车载定位设备获取车辆位置坐标，侧重于感知数据的采集，包含了图像、3D点云等感知信息，可以为强化学习模

**表1 自然驾驶数据集**

**Tab.1 Datasets of natural driving**

| 数据集 | 国家 | 场景类型 | 数据类型 | 规模 | 年份 |
|---|---|---|---|---|---|
| NGSIM[22] | 美国 | 高速/城市 | 轨迹数据 | 4个路段，11 285条轨迹 | 2007 |
| KITTI[23] | 德国 | 城市/郊区/乡村 | 轨迹、图像和点云 | 采集里程39 km，22条轨迹 | 2013 |
| Cityscapes[24] | 德国 | 城市 | 图像、轨迹数据 | 采集于50个城市 | 2016 |
| HighD[25] | 德国 | 高速 | 轨迹数据 | 110 000条轨迹 | 2018 |
| ApolloScape[26] | 中国 | 高速/城市 | 轨迹、图像和点云数据 | 采集里程1 000 km | 2018 |
| BDD100K[27] | 美国 | 高速/城市 | 图像、轨迹数据 | 100 000条轨迹 | 2018 |
| INTERACTION[28] | 多国 | 交叉口/环岛 | 轨迹数据 | 40 054条轨迹 | 2019 |
| TJRD TS[29] | 中国 | 高速 | 轨迹数据 | 3个路段，全长18.39 km | 2019 |
| InD[30] | 德国 | 交叉口 | 轨迹数据 | 11 500条轨迹 | 2020 |
| RoundD[31] | 德国 | 环岛 | 图像、轨迹数据 | 13 700条轨迹 | 2020 |
| nuScenes[32] | 美国，新加坡 | 城市 | 轨迹、图像、点云和雷达数据 | 1 000条轨迹 | 2020 |
| Waymo Open[33] | 美国 | 高速/城市 | 轨迹、图像和点云数据 | 103 354条轨迹 | 2020 |
| Lyft Level 5[34] | 美国 | 郊区 | 轨迹、图像、点云和雷达数据 | 170 000条轨迹 | 2020 |
| Oxford Radar[35] | 英国 | 城市 | 轨迹、图像、点云和雷达数据 | 采集里程280 km，32条轨迹 | 2020 |
| DAIR－V2X[36] | 中国 | 高速/城市 | 轨迹、图像和点云数据 | 大于200 000条轨迹 | 2021 |
| HighSim[37] | 美国 | 高速 | 轨迹数据 | 采集路段长2.44 km，采集时间2 h | 2021 |
| PandaSet[38] | 中国 | 城市 | 轨迹、图像和点云数据 | 大于100条轨迹 | 2021 |
| ExiD[39] | 德国 | 匝道 | 轨迹数据 | 69 172条轨迹 | 2022 |
| Boreas[40] | 加拿大 | 城市 | 轨迹、图像和点云数据 | 采集里程350 km，44条轨迹 | 2023 |

型利用车辆感知信息学习驾驶策略提供支持。这些数据集的车辆轨迹数量相对较少，多为单车轨迹序列，更适合研究自动驾驶车辆在不同道路环境下的驾驶策略。在场景复杂度方面，HighD等数据集主要收集了高速路段上的车辆轨迹，这些数据集的车辆行为以跟驰为主，车辆速度变化较小，场景相对较简单且仅包含车辆行驶数据，没有其他交通参与者信息。因此，这类数据集适用于特定场景下，如高速道路，研究自动驾驶车辆轨迹控制。然而，在涉及车辆之间高度交互和复杂交通环境的自动驾驶策略开发方面，这些数据集存在一定的限制。INTERACTION、InD、RoundD等数据集涵盖了环岛和交叉口等复杂场景的车辆轨迹数据，还包括行人、非机动车等其他交通参与者。这些数据集的场景更为复杂，适用于研究自动驾驶车辆在复杂场景下的控制策略。其中，NGSIM和KITTI等数据集由于发布年份较早，数据量较小且数据质量相对较低，数据可靠性也存在一定问题。随着自动驾驶技术的进步，对大规模高精度自然驾驶数据集的需求逐渐增加，HighD、TJRD TS、HighSim、nuScenes等数据集相继提出，在扩大数据规模、提高数据质量的同时，引入了更多样化的驾驶场景，如不同地点、不同天气条件、多种车辆类型以及各种道路标志和驾驶规则等，对于推动自动驾驶技术的发展发挥了重要作用。

由于强化学习依赖于大量的试错和探索，直接在实际道路环境中训练是不切实际的，在自动驾驶系统的开发中通常需要借助仿真环境进行训练。目前用于自动驾驶系统开发的仿真环境可以分为驾驶模拟器与交通流仿真软件两大类，驾驶模拟器用于模拟自动驾驶车辆在各种环境和情境下的驾驶行为，为自动驾驶算法的开发和验证提供真实且安全的实验环境。交通流仿真软件则用于模拟整个交通系统的运行，主要用于研究和分析自动驾驶汽车与其他车辆的交互行为以及对交通流的影响，从宏观角度评估和优化自动驾驶系统的性能。部分代表性仿真环境及其简介如表2所示。

### 2.3 应用案例

#### 2.3.1 规划决策与控制

在自动驾驶领域，规划决策与控制模块是实现安全、高效驾驶的核心环节，相比于传统方法，强化学习方法在应对充满不确定性的道路交通环境时具有天然的优势。根据驾驶环境的特性，已有关于自动驾驶规划决策控制的研究通常聚焦于两种路况：高速路况下的自动驾驶任务和城市路况下的自动驾驶任务，如图4所示。

**表2　自动驾驶仿真环境**

**Tab.2　Simulation environments of autonomous driving**

| 仿真环境 | 简介 |
|---|---|
| SUMO[41] | 开源、高度可定制的城市交通流仿真框架,支持道路网络建模,提供丰富的车辆控制模型和可调用接口供仿真使用 |
| CARLA[42] | 开源自动驾驶模拟器,具有丰富的天气和环境模拟功能以及详细的传感器模型 |
| Apollo[43] | 百度开源的自动驾驶平台,提供了完整的硬件、软件以及云服务解决方案,包括高精地图、感知、规划、控制等模块,适用于各种自动驾驶场景的研究 |
| Flow[44] | 开源的微观交通仿真框架,可以与深度强化学习库进行集成,用于研究混合交通和智能交通系统 |
| TORCS[45] | 开源赛车模拟器,常用于自动驾驶相关人工智能算法的开发与研究 |
| OnSite[46] | 为高等级自动驾驶汽车的感知、决策、规划、控制等模块提供测评服务的公共平台,用于自动驾驶汽车规划控制算法测试、驾驶能力评价等相关研究 |
| AirSim[47] | 微软开发的一个开源、跨平台的人工智能研究平台,可以为自动驾驶汽车和无人机提供贴近实际的仿真环境,用于试验深度学习、计算机视觉和强化学习算法 |
| Highway-env[48] | 为自动驾驶研究和开发设计的开源仿真环境,提供一系列交通场景供强化学习算法训练使用 |
| Unity[49] | 一款实时3D游戏引擎,已被广泛用作自动驾驶仿真工具 |
| MetaDrive[50] | 一款轻量且易于安装的模拟器,支持路网和交通流建模,并且能够导入真实数据进行强化学习算法开发 |

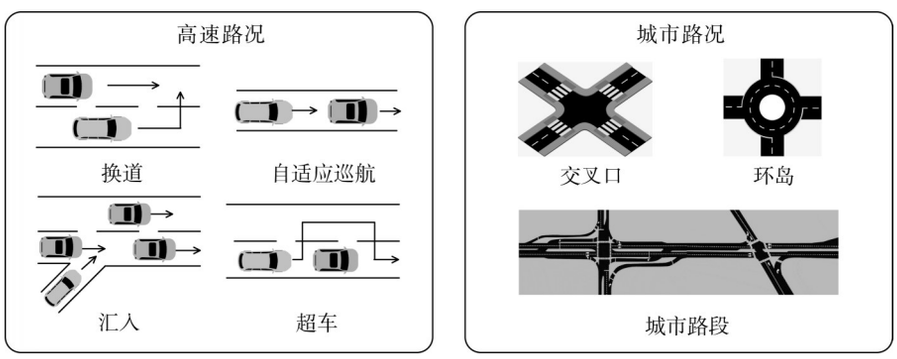



**图4　自动驾驶规划决策控制中的主要任务图示**

**Fig.4　Illustration of main tasks in autonomous driving planning, decision-making, and control**

(1)高速路况下的自动驾驶任务。由于高速公路的交通环境较为单一,研究人员通常将高速公路下的自动驾驶任务分解为换道、自适应巡航、汇入等场景规划决策控制子任务,有针对性地提取对应的场景数据集用于训练强化学习模型学习不同任务下的控制策略。在自动驾驶汽车换道策略的相关研究中,冯耀等[51]通过设计分层式智能网联车辆换道轨迹规划架构,基于完全信息纯策略博弈的换道行为决策模型,以及基于TD3算法的横纵向换道轨迹规划方法,实现了安全、节能、舒适的换道轨迹优化;Alizadeh等[52]针对自动驾驶系统中的自主换道场景开发了用于训练强化学习模型的仿真环境,所训练的换道模型在复杂不确定性环境中的表现优于传统启发式方法;李文礼等[53]为解决自动驾驶汽车在高速公路安全换道问题,通过融合车辆换道路径模型和深度强化学习框架,引入DDPG算法来优化车辆在高速环境下的换道跟踪控制模型,在横向位置误差和角误差等方面均优于传统模型预测控制(model predictive control, MPC)方法。自适应巡航策略是自动驾驶技术研究的热门方向,朱冰等[54]通过结合深度强化学习模型与前车的运动随机过程模型,进行迭代式的交互学习,来形成适应于运动不确定性跟驰环境下的主车纵向控制策略,这种策略可以实现稳定的跟随,而无需依赖对前车运动的额外预测;He等[55]设计了一个综合考虑安全、效率以及舒适性的奖励函数,基于DDPG算法建立自适应巡航控制模型,采用NGSIM数据集进行训练,使自动驾驶车辆能够在安全行驶的同时优化多项行驶指标;Zhao等[56]通过从状态转换数据中学习最优策略,利用强化学习解决在模型未知、运行环境不确定的自动驾驶系统中实现自适应巡航控制最优控制的难题。匝道路段作为高速公路的事故多发路段,解决自动驾驶系统汇入策略的设计问题尤为重要。乔良等[57]基于DQN算法开发的自动驾驶匝道汇入模型可以实现根据车速调整控制策略,提高了自动驾驶系统在汇入场景下的智能化决策水平;Lubars等[58]结合强化学习方法与MPC方法的优势,设计了一种在多方面指标中达到折中的匝道汇入算法;提出了一种基于强化学习的自适应车辆汇入策略网络,嵌套在有限状态机中,在保证安全的前提下实现了高效的车辆自动汇入。除此之外,强化学习还被应用于自动驾驶车辆车道保持、超车等任务,并取得了良

好的表现[59-60]。

(2)城市路况下的自动驾驶任务。由于城市路况复杂多变,交通参与者众多,驾驶任务主要由不同的道路场景或驾驶任务进行驱动,例如在交叉口、环岛等复杂场景进行综合决策控制,或者在综合城市道路中行驶。Deshpande等[61]提出了一种基于DQN算法的方法,使用基于网格的状态表示方法和特定的奖励函数,使得自动驾驶车辆能够在结构化的城市环境中安全地通过交叉口并遵循交通规则;欧阳卓等[62]基于分布式深度强化学习的方法,解决无信号灯交叉路口自动驾驶车辆控制的问题,通过高效奖励函数、迁移学习和适应所有自动驾驶车辆比例的策略,在仿真平台上提升了交叉口的通行效率;Qiao等[63]将自动驾驶技术在现实环境的应用问题建模为马尔可夫决策过程,并使用分层选项的强化学习方法来学习策略,解决了部分可观测马尔可夫决策过程对大量数据的依赖和对于连续动作计算效率低下的问题,有效提高了自动驾驶车辆在带有双向停车标志的四路交叉口中的性能;Liang等[64]提出了一种通用且有原则的可控模仿强化学习方法,通过模仿人类经验在合理约束的动作空间中进行探索,从而缓解大型连续动作空间中低探索效率的问题,在CARLA模拟器中,仅基于视觉输入使自动驾驶车辆在交叉口、环岛等驾驶任务中取得优于有监督学习方法的表现。Guo等[65]为了提升自动驾驶车辆在复杂场景下的行驶表现,采用DDPG算法和DQN算法分别控制车辆纵向加减速和横向换道决策,提出一种混合强化学习方法使得自动驾驶车辆在连续通过多个信号交叉口的同时保证通行效率并显著减少能耗。

尽管强化学习在自动驾驶的规划、决策与控制各个子任务上有所突破,如高速公路和城市环境下的轨迹规划与决策控制,但目前在复杂交通环境中的自主驾驶能力依然有待提升。为了使自动驾驶车辆能够应对不同风险等级的场景,已有研究尝试建立高覆盖度、风险等级完善的自动驾驶场景库用于训练强化学习模型[55]。同时,目前自动驾驶车辆决策、规划与控制模块多数通过接收经过感知端处理得到的低维状态信息来与感知端进行信息交互,并且在车辆的驾驶控制中通常依赖传统方法或其他技术手段进行分层控制,为了提升自动驾驶系统的表现,需要越来越庞大的算力来支撑系统的正常运行。在此背景下,越来越多的研究正在尝试通过端到端的控制方法,直接将传感器感知信号转换为车辆控制输出,以简化系统并提高其运行效率。但与此同时,超高维度的输入状态空间也给强化学习算法中奖励函数的设计及其训练和收敛带来了前所未有的挑战。

### 2.3.2 自动驾驶测试

强化学习在自动驾驶测试中的应用具有重要的意义和作用。自动驾驶技术的发展和实际应用离不开大规模的测试和验证。传统的测试方法只能覆盖有限的场景和情况,难以涵盖复杂多变的真实交通环境,为了验证自动驾驶汽车的安全性能,往往需要在自然驾驶环境中测试数亿甚至数千亿英里[10]。而强化学习作为一种自主学习的方法,可以用于生成多样化的交通场景,模拟各种复杂交通情况和现实测试中难以遇到的边缘场景,从而帮助自动驾驶系统适应各种挑战性的驾驶情况[66-68]。此外,强化学习还可以用于加速自动驾驶测试过程,通过智能体在仿真环境中自主学习和优化驾驶策略,减少在真实道路上进行试验的次数和成本[10]。

在自动驾驶测试场景生成相关研究中,李江坤等[66]为解决高风险边缘测试场景的稀疏问题,利用场景动力学和强化学习,模拟了车辆之间的对抗和博弈行为,从而自动生成边缘场景,并在场景交互博弈、覆盖率和可重复性测试等方面表现出色;Feng等[67]使用DQN算法生成离散对抗性交通场景,并采用时间差分强化学习对该团队搭建的自动驾驶系统测试场景库生成框架进行增强,能够有效且高效地生成测试场景库并显著加速评估过程[68];Chen等[69]使用DDPG算法生成对抗性策略来控制周围车辆以生成自动驾驶车辆与周围车辆频繁交互的典型变道场景。Feng等[10]为了解决自动驾驶汽车安全测试问题中的"维度灾难"和"稀疏度灾难"复合挑战,通过识别和删除非安全关键状态,对马尔可夫过程进行编辑,提出了密集强化学习方法,将自然驾驶环境中的自动驾驶汽车安全性测试评估速度提高了$10^3$~$10^5$倍。

值得注意的是,在生成稀疏场景方面,其真实性和可靠性尤为关键。为评价这些场景,需要引入多维度的评价指标。表3介绍了一些典型场景评价指标,为生成场景的评价提供了指导,但是如何有效地评价自动生成的测试场景仍是一个需要持续探索的课题。同时,在运用强化学习生成自动驾驶测试场景的过程中还存在许多挑战。例如,可能生成风险过高的无效场景,导致碰撞不可避免;生成场景重复度过高,从而影响自动驾驶测试效率等。

表3　自动驾驶测试生成场景评价指标

Tab.3　Evaluation metrics for generated autonomous driving test scenarios

| 评价指标 | 描述 |
| --- | --- |
| JS散度、KL散度[70] | 评价生成场景与真实场景相似度，衡量真实性 |
| 碰撞时间TTC、车辆间距等碰撞替代指标[71] | 可用于评价生成场景危险程度，也可用于衡量生成场景与真实场景相似程度 |
| 测试任务（如换道）成功率[72] | 通过被测车辆能否完成测试任务来衡量场景可靠性 |
| 换道切入、冲突、碰撞、紧急制动等事件发生的次数和时间[73] | 通过对危险事件的发生进行统计从而量化场景风险，同时可对场景可靠性进行分析 |

### 2.3.3　自动驾驶车辆性能提升

强化学习在自动驾驶领域的应用不止于规划决策控制和自动驾驶测试，在自动驾驶车辆性能提升中也有多方面的应用，如车辆能量管理系统、车辆动力系统和车身姿态控制等。在车辆能量管理系统方面，强化学习可用于优化混合动力车辆、增程式电动汽车和电动汽车的能量消耗，以最大程度地提高燃油经济性或电池续航里程，帮助自动驾驶汽车实现更智能化和高效的动力分配，以适应不同的行驶场景和驾驶习惯。陈泽宇等[74]建立了基于DDPG算法的车载能源功率分配策略，减少了车辆行驶状态突变所造成的车辆能耗；Lian等[75]将专家知识与DDPG算法巧妙结合，并将最佳制动比燃油消耗曲线和混合动力汽车的电池特性纳入综合考虑，从而进行多目标能量管理优化，实现了更好的燃油经济性，提高了能量管理系统的相对稳定性；Xiong等[76]基于强化学习对插电式混合动力电动汽车进行实时功率管理，并在不同条件下验证了该策略不仅可以限制最大放电电流并减少电池组的充电频率，还可以降低能量损失并优化系统效率。在车身姿态控制方面，强化学习算法可以用于自动驾驶车辆的稳定性控制和操纵，以确保在各种驾驶条件下车辆的安全性和舒适性。殷国栋等[77]基于多智能体强化学习来解决新型分布式轮毂电机驱动纯电动汽车的底盘控制系统问题，有效改善了车辆的横向操纵稳定性能；江洪等[78]采用汤普森抽样算法构建强化学习模型，建立空气悬架车身高度智能控制系统，在单一工况和混合工况的实验测试中均取得了出色的表现；Li等[79]将考虑乘坐舒适性、道路操纵性和执行限制的安全强化学习框架应用于主动悬挂系统，并获得了优于常规控制器的性能。

在最新的研究进展中，强化学习在自动驾驶车辆性能提升方面已经做出了一系列突破。不仅关注单一性能指标的优化，而且能够综合处理更复杂、更全面的问题，例如将车载传感器收集的路面感知信息作为强化学习算法的输入来实时调整车辆底盘的各项参数，并与速度规划算法相结合，多维度地提升乘客的舒适性[80]。在此背景下，强化学习应用于自动驾驶车辆性能提升方面的一大挑战在于如何设计端到端、轻量级的深度强化学习框架，最大限度地利用现有车载传感器的感知信息，并将这些信息与车辆的转向舒适性与稳定性、能耗优化等更多维度的车辆性能指标相关联，从而简化现有的复杂分层控制体系。通过这种方式，强化学习不仅能够对车辆性能进行综合优化，还能与车辆的规划决策和控制算法综合作用，从而实现自动驾驶车辆性能的全方位提升。

## 2.4　挑战与展望

尽管强化学习在自动驾驶领域吸引了众多研究者的关注，但在实际应用中仍然面临许多诸如安全性挑战和实际落地应用等方面的挑战。

安全性对于强化学习算法在自动驾驶系统的应用至关重要。相较于传统的基于规则的方法，机器学习方法的可解释性较差，一直受到广泛的质疑。其“黑箱”特性使得强化学习中智能体可能在难以预料的情况下采取不安全的策略，这严重违背自动驾驶系统中的安全性原则。为了确保强化学习方法的安全性，研究人员通常将基于强化学习的方法与基于规则的方法相结合来获取安全的驾驶策略。例如，代珊珊等[81]通过将使无人车偏离轨道或者发生碰撞的动作标记为约束动作来提高自动驾驶过程中的安全性；Shalev-Shwartz等[82]将汇入问题分解为可学习的深度强化学习策略和不可学习的硬约束，提供了比纯粹的端到端框架更好的解释性。尽管已有许多研究通过约束动作避免强化学习执行不安全的策略，但是要做到在学习到的最优策略与选择安全约束之间寻求平衡仍然是一项巨大的挑战，在未来研究中还需要进一步的探索。

目前，自动驾驶领域的大部分强化学习实验都在低成本的仿真环境中进行。由于机器学习通常默认训练数据和测试数据遵循独立同分布假设，而直接将在仿真环境中学习到的策略部署到实际的自动驾驶车辆上可能会违背独立同分布假设，这使得模型的实际应用效果存在不确定性。为解决此问题，

已有部分研究尝试使用迁移学习的方法，例如Bewley等[83]利用图像转换领域的先进技术实现了将基于视觉的驾驶策略从仿真环境迁移至现实中的乡村道路，但是此类方法的实际效果和稳定性还有待验证。此外，在实际应用中自动驾驶车辆可能会面临各种瞬息万变的场景，需要处理许多全新的任务，然而大部分强化学习模型难以根据有限的已有经验应对全新的场景。连续学习是解决该问题的一种方案，在机器人控制领域已经积累了一定的经验，未来有希望应用于自动驾驶的实际落地。另一方面，自动驾驶系统的测试评价是一个极其重要的挑战。虽然高保真模拟器例如CARLA可以为强化学习模型提供虚拟平台进行部署和评估，但其在实际部署前的详尽验证仍是一项难题。同时，实际车辆的测试需要高昂的时间、金钱和人力成本，也存在一定的危险性。因此，未来的研究需要找到实用、有效、低风险和经济的自动驾驶车辆测试评价方法，以促进强化学习在自动驾驶领域的实际落地应用。

## 3 结论

强化学习作为人工智能领域的重要分支对自动驾驶领域的发展起到了推动作用。基于马尔可夫决策过程的理论，强化学习模型能够在自动驾驶环境中学习最优策略，实现自主决策与控制。其中，基于价值的强化学习方法和基于策略的强化学习方法分别在规划决策控制和性能提升方面发挥重要作用。深度强化学习的出现进一步拓展了强化学习的应用领域，使其能够处理更复杂、高维度的驾驶任务。在自动驾驶领域，高质量、多样化的数据集和贴近真实的仿真环境推动强化学习在规划决策与控制、自动驾驶测试和车辆性能提升等方面都取得了令人瞩目的成果。例如，强化学习模型通过优化换道、自适应巡航等场景规划决策控制子任务，实现了安全高效的驾驶。在自动驾驶测试方面，强化学习生成多样化的交通场景，从而加速测试过程，降低测试成本。在车辆性能提升方面，强化学习通过优化能量消耗、车身姿态控制等，提高驾驶稳定性和经济性。然而，强化学习在自动驾驶领域存在安全性以及实际落地应用难等尚未完全解决的挑战，仍需进一步的研究和探索。

综上所述，强化学习为自动驾驶技术赋能，为其实现智能化和高效化提供了重要手段。在未来的研究中，应进一步探索强化学习在自动驾驶中的可解释性和安全性，完善实际落地应用的方案以及的测试评价方法，并寻求与其他领域的融合，共同推动自动驾驶技术的发展。

**作者贡献声明：**

何逸煦：框架设计，写作和修改。

林泓熠：查阅资料，论文写作。

刘　洋：学术指导，论文修改。

杨　澜：论文审阅，论文修改。

曲小波：学术指导，论文审阅。

**参考文献：**

[1] 张雷，沈国琛，秦晓洁，等.智能网联交通系统中的信息物理映射与系统构建[J].同济大学学报(自然科学版)，2022，50(1)：79.
ZHANG Lei，SHEN Guochen，QIN Xiaojie，*et al*. Information physical mapping and system construction of intelligent network transportation [J]. Journal of Tongji University(Natural Science)，2022，50(1)：79.

[2] 林泓熠，刘洋，李深，等.车路协同系统关键技术研究进展[J].华南理工大学学报(自然科学版)，2023，51(10)：46.
LIN Hongyi，LIU Yang，LI Shen，*et al*. Research progress on key technologies in the cooperative vehicle infrastructure system [J]. Journal of South China University of Technology (Natural Science)，2023，51(10)：46.

[3] LIU Y，WU F，LIU Z，*et al*. Can language models be used for real-world urban-delivery route optimization? [J]. The Innovation，2023，4(6)：1.

[4] LIU Y，LYU C，ZHANG Y，*et al*. DeepTSP：deep traffic state prediction model based on large-scale empirical data [J]. Communications in Transportation Research，2021，1：100012.

[5] 刘兵，王锦锐，谢济铭，等.微观轨迹数据驱动的交织区换道概率分布模型[J].汽车安全与节能学报，2022，13(2)：333.
LIU Bing，WANG Jinrui，XIE Jiming，*et al*. Microscopic trajectory data-driven probability distribution model for weaving area of channel change [J]. Journal of Automotive Safety and Energy，2022，13(2)：333.

[6] YANG Z，CHAI Y，ANGUELOV D，*et al*. Surfelgan：synthesizing realistic sensor data for autonomous driving [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2020：11118-11127.

[7] HU Y，YANG J，CHEN L，*et al*. Planning-oriented autonomous driving [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2023：17853-17862.

[8] MINSKY，LEE M. Theory of neural-analog reinforcement systems and its application to the brain-model problem [R]. Princeton：PrincetonUniversity，1954.

[9] BROWN T，MANN B，RYDER N，*et al*. Language models

are few-shot learners [J]. Advances in Neural Information Processing Systems, 2020, 33: 18771.
[10] FENG S, SUN H, YAN X, *et al*. Dense reinforcement learning for safety validation of autonomous vehicles [J]. Nature, 2023, 615(7953): 620.
[11] BELLMAN R. A Markovian decision process [J]. Indiana University Mathematics Journal, 1957, 6(4):15.
[12] HOWARD R A. Dynamic programming and Markov process [M]. Cambridge: MIT Press, 1960.
[13] SUTTON R S. Learning to predict by the methods of temporal differences[J]. Machine Learning, 1988, 3: 9.
[14] WATKINS C J C H. Learning from delayed rewards [D]. Cambridge:University of Cambridge, 1989.
[15] RUMMERY G A, NIRANJAN M. On-line Q-learning using connectionist systems [R]. Cambridge: University of Cambridge, 1994.
[16] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529.
[17] SCHULMAN J, LEVINE S, ABBEEL P, *et al*. Trust region policy optimization [C]//International Conference on Machine Learning. Cambridge : JMLR, 2015: 1889-1897.
[18] SCHULMAN J, WOLSKI F, DHARIWAL P, *et al*. Proximal policy optimization algorithms [J]. arXiv preprint arXiv:1707.06347, 2017.
[19] HAARNOJA T, ZHOU A, ABBEEL P, *et al*. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor [C]//International Conference on Machine Learning. Cambridge: JMLR, 2018: 1861-1870.
[20] LILLICRAP T P, HUNT J J, PRITZEL A, *et al*. Continuous control with deep reinforcement learning [J]. arXiv preprint arXiv:1509.02971, 2015.
[21] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods [C]//International Conference on Machine Learning. Cambridge: JMLR, 2018: 1587-1596.
[22] ALEXIADIS V, COLYAR J, HALKIAS J, *et al*. The next generation simulation program [J]. Institute of Transportation Engineers, 2004, 74(8): 22.
[23] GEIGER A, LENZ P, STILLER C, *et al*. Vision meets robotics: the kitti dataset [J]. The International Journal of Robotics Research, 2013, 32(11): 1231.
[24] CORDTS M, OMRAN M, RAMOS S, *et al*. The cityscapes dataset for semantic urban scene understanding [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 3213-3223.
[25] KRAJEWSKI R, BOCK J, KLOEKER L, *et al*. The highd dataset: a drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE, 2018: 2118-2125.
[26] HUANG X, CHENG X, GENG Q, *et al*. The apolloscape dataset for autonomous driving [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2018: 954-960.
[27] YU F, CHEN H, WANG X, *et al*. Bdd100k: a diverse driving dataset for heterogeneous multitask learning [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2636-2645.
[28] ZHAN W, SUN L, WANG D, *et al*. Interaction dataset: an international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps [J]. arXiv preprint arXiv:1910.03088, 2019.
[29] WANG J H, FU T, XUE J T, *et al*. Realtime wide-area vehicle trajectory tracking using millimeter-wave radar sensors and the open TJRD TS dataset [J]. International Journal of Transportation Science and Technology, 2023, 12(1): 273.
[30] BOCK J, KRAJEWSKI R, MOERS T, *et al*. The ind dataset: a drone dataset of naturalistic road user trajectories at german intersections [C]//2020 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE, 2020: 1929-1934.
[31] KRAJEWSKI R, MOERS T, BOCK J, *et al*. The round dataset: a drone dataset of road user trajectories at roundabouts in germany [C]//2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE, 2020: 1-6.
[32] CAESAR H, BANKITI V, LANG A H, *et al*. nuscenes: a multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11621-11631.
[33] SUN P, KRETZSCHMAR H, DOTIWALLA X, *et al*. Scalability in perception for autonomous driving: Waymo open dataset [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2446-2454.
[34] HOUSTON J, ZUIDHOF G, BERGAMINI L, *et al*. One thousand and one hours: self-driving motion prediction dataset [C]//Conference on Robot Learning. Cambridge: JMLR, 2021: 409-418.
[35] BARNES D, GADD M, MURCUTT P, *et al*. The oxford radar robotcar dataset: a radar extension to the oxford robotcar dataset [C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2020: 6433-6438.
[36] YU H, LUO Y, SHU M, *et al*. Dair-v2x: a large-scale dataset for vehicle-infrastructure cooperative 3d object detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 21361-21370.
[37] SHI X, ZHAO D, YAO H, *et al*. Video-based trajectory extraction with deep learning for High-Granularity Highway Simulation (HIGH-SIM) [J]. Communications in Transportation Research, 2021, 1: 100014.

[38] XIAO P, SHAO Z, HAO S, *et al*. Pandaset: advanced sensor suite dataset for autonomous driving [C]//2021 IEEE International Intelligent Transportation Systems Conference (ITSC). Piscataway: IEEE, 2021: 3095-3101.

[39] MOERS T, VATER L, KRAJEWSKI R, *et al*. The exiD dataset: a real-world trajectory dataset of highly interactive highway scenarios in Germany [C]//2022 IEEE Intelligent Vehicles Symposium (Ⅳ). Piscataway: IEEE, 2022: 958-964.

[40] BURNETT K, YOON D J, WU Y, *et al*. Boreas: a multi-season autonomous driving dataset [J]. The International Journal of Robotics Research, 2023, 42(1/2): 33.

[41] LOPEZ P A, BEHRISCH M, BIEKER-WALZ L, *et al*. Microscopic traffic simulation using sumo [C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE, 2018: 2575-2582.

[42] DOSOVITSKIY A, ROS G, CODEVILLA F, *et al*. CARLA: an open urban driving simulator[C]//Conference on Robot Learning. Cambridge: JMLR, 2017: 1-16.

[43] FAN H Y ,ZHU F ,LIU C C , *et al*. Baidu apollo em motion planner[J]. arXiv preprint arXiv:1807.08048, 2018.

[44] WU C, KREIDIEH A, PARVATE K, *et al*. Flow: architecture and benchmarking for reinforcement learning in traffic control[J]. arXiv preprint arXiv:1710.05465, 2017, 10.

[45] WYMANN B, ESPIÉ E, GUIONNEAU C, *et al*. Torcs, the open racing car simulator [EB/OL]. [2020-02-06]. https://onsite.run.

[46] SUN J, TIAN Y. OnSite[EB/OL]. [2022-08-30]. https://onsite.run.

[47] SHAH S, DEY D, LOVETT C, *et al*. Airsim: high-fidelity visual and physical simulation for autonomous vehicles [C]// Field and Service Robotics: Results of the 11th International Conference. Cham: Springer International Publishing, 2018: 621-635.

[48] EDOUARD L. An environment for autonomous driving decision-making [EB/OL]. [2022-08-30]. https://github.com/eleurent/ highway- env.

[49] LEURENT E. A collection of environments for autonomous driving and tactical decision-making tasks [EB/OL]. [2022-08-30]. https://github.com/eleurent/highway-env.

[50] LI Q, PENG Z, FENG L, *et al*. Metadrive: composing diverse driving scenarios for generalizable reinforcement learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3461.

[51] 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.
FENG Yao, JING Shoucai, HUI Fei, *et al*. Deep reinforcement learning-based lane-changing trajectory planning method of intelligent and connected vehicles [J]. Journal of Automotive Safety and Energy,2022,13(4):705.

[52] ALIZADEH A, MOGHADAM M, BICER Y, *et al*. Automated lane change decision making using deep reinforcement learning in dynamic and uncertain highway environment [C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). Piscataway: IEEE, 2019: 1399-1404.

[53] 李文礼,邱凡珂,廖达明,等.基于深度强化学习的高速公路换道跟踪控制模型[J].汽车安全与节能学报,2022,13(4):750.
LI Wenli, QIU Fanke, LIAO Daming, *et al*. Highway lane change decision control model based on deep reinforcement learning[J]. Journal of Automotive Safety and Energy,2022,13(4):750.

[54] 朱冰,蒋渊德,赵健,等.基于深度强化学习的车辆跟驰控制[J].中国公路学报,2019,32(6):53.DOI:10.19721/j.cnki.1001-7372.2019.06.005.
ZHU Bing, JIANG Yuande, ZHAO Jian, *et al*. A car-following control algorithm based on dep reinforcement learning [J]. China Journal of Highway and Transport,2019,32(6):53.DOI:10.19721/j.cnki.1001-7372.2019.06.005.

[55] HE Y, LIU Y, YANG L, *et al*. Deep adaptive control: deep reinforcement learning-based adaptive vehicle trajectory control algorithms for different risk levels [J]. IEEE Transactions on Intelligent Vehicles, 2023, 9(1):1654.

[56] ZHAO D, XIA Z, ZHANG Q. Model-free optimal control based intelligent cruise control with hardware-in-the-loop demonstration research frontier [J]. IEEE Computational Intelligence Magazine, 2017, 12(2): 56.

[57] 乔良,鲍泓,玄祖兴,等.基于强化学习的无人驾驶匝道汇入模型[J].计算机工程,2018,44(7):20.DOI:10.19678/j.issn.1000-3428.0050990.
QIAO Liang, BAO Hong, XUAN Zuxing, *et al*. Autonomous driving ramp merging model based on reinforcement learning [J]. Computer Engineering,2018,44(7):20.DOI:10.19678/j.issn.1000-3428.0050990.

[58] LUBARS J, GUPTA H, CHINCHALI S, *et al*. Combining reinforcement learning with model predictive control for on-ramp merging [C]//2021 IEEE International Intelligent Transportation Systems Conference (ITSC). Piscataway: IEEE, 2021: 942-947.

[59] KIM M, SEO J, LEE M, *et al*. Vision-based uncertainty-aware lane keeping strategy using deep reinforcement learning [J]. Journal of Dynamic Systems, Measurement, and Control, 2021, 143(8): 084503.

[60] LU H, LU C, YU Y, *et al*. Autonomous overtaking for intelligent vehicles considering social preference based on hierarchical reinforcement learning[J]. Automotive Innovation, 2022, 5(2): 195.

[61] DESHPANDE N, SPALANZANI A. Deep reinforcement learning based vehicle navigation amongst pedestrians using a grid-based state representation [C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). Piscataway: IEEE, 2019: 2081-2086.

[62] 欧阳卓,周思源,吕勇,等.基于深度强化学习的无信号灯交叉路口车辆控制[J].计算机科学,2022,49(3):46.
OUYANG Zhuo, ZHOU Siyuan, LÜ Yong , *et al*. DRL-

based vehicle control strategy for signal-free intersections[J]. Computer Science,2022,49(3):46.

[63] QIAO Z, MUELLING K, DOLAN J, *et al*. Pomdp and hierarchical options mdp with continuous actions for autonomous driving at intersections [C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE, 2018: 2377-2382.

[64] LIANG X, WANG T, YANG L, *et al*. Cirl: controllable imitative reinforcement learning for vision-based self-driving [C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer International Publishing, 2018: 584-599.

[65] GUO Q, ANGAH O, LIU Z, *et al*. Hybrid deep reinforcement learning based eco-driving for low-level connected and automated vehicles along signalized corridors[J]. Transportation Research Part C: Emerging Technologies, 2021, 124: 102980.

[66] 李江坤,邓伟文,任秉韬,等.基于场景动力学和强化学习的自动驾驶边缘测试场景生成方法[J].汽车工程,2022,44(7):976.DOI:10.19562/j.chinasae.qcgc.2022.07.004.
LI Jiangkun, DENG Weiwen, REN Bingtao, *et al*. Automatic driving edge scene generation method based on scene dynamics and reinforcement learning[J]. Automotive Engineering, 2022, 44(7):976.DOI:10.19562/j.chinasae.qcgc.2022.07.004.

[67] FENG S, YAN X, SUN H, *et al*. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment[J]. Nature Communications, 2021, 12(1): 748.

[68] FENG S, FENG Y, SUN H, *et al*. Testing scenario library generation for connected and automated vehicles, part II: case studies [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(9): 5635.

[69] CHEN B, CHEN X, WU Q, *et al*. Adversarial evaluation of autonomous vehicles in lane-change scenarios [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 10333.

[70] BARZ B, RODNER E, GARCIA Y G, *et al*. Detecting regions of maximal divergence for spatio-temporal anomaly detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(5): 1088.

[71] SUN H, FENG S, YAN X, *et al*. Corner case generation and analysis for safety assessment of autonomous vehicles [J]. Transportation Research Record, 2021, 2675(11): 587.

[72] CHEN B, CHEN X, WU Q, *et al*. Adversarial evaluation of autonomous vehicles in lane-change scenarios [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 10333.

[73] FENG S, YAN X, SUN H, *et al*. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment[J]. Nature Communications, 2021, 12(1): 748.

[74] 陈泽宇,方志远,杨瑞鑫,等.基于深度强化学习的混合动力汽车能量管理策略[J].电工技术学报,2022,37(23):6157.DOI:10.19595/j.cnki.1000-6753.tces.211342.
CHEN Zeyu, FANG Zhiyuan, YANG Ruixin, *et al*. Energy management strategy for hybrid electric vehicle based on the deep reinforcement learning method[J]. Transactions of China Electrotechnical Society, 2022, 37(23): 6157.DOI: 10.19595/j.cnki.1000-6753.tces.211342.

[75] LIAN R, PENG J, WU Y, *et al*. Rule-interposing deep reinforcement learning based energy management strategy for power-split hybrid electric vehicle [J]. Energy, 2020, 197: 117297.

[76] XIONG R, CAO J, YU Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle [J]. Applied Energy, 2018, 211: 538.

[77] 殷国栋,朱侗,任祖平,等.基于多Agent的电动汽车底盘智能控制系统框架[J].中国机械工程,2018,29(15):1796.
YIN Guodong, ZHU Tong, REN Zuping, *et al*. Intelligent control system framework for multi-agent based electric vehicle chassises [J]. China Mechanical Engineering, 2018, 29(15): 1796.

[78] 江洪,王鹏程,李仲兴.基于智能体理论的空气悬架车身高度智能控制系统研究[J].重庆理工大学学报(自然科学),2019,33(4):17.
JIANG Hong, WANG Pengcheng, LI Zhongxing. Research on air suspension vehicle height intelligent control system based on agent theory [J]. Journal of Chongqing University of Technology(Natural Science),2019,33(4): 17.

[79] LI Z, CHU T, KALABIĆ U. Dynamics-enabled safe deep reinforcement learning: Case study on active suspension control [C]//2019 IEEE Conference on Control Technology and Applications (CCTA). Piscataway: IEEE, 2019: 585-591.

[80] DU Y, CHEN J, ZHAO C, *et al*. A hierarchical framework for improving ride comfort of autonomous vehicles via deep reinforcement learning with external knowledge[J]. Computer-Aided Civil and Infrastructure Engineering, 2023, 38(8): 1059.

[81] 代珊珊,刘全.基于动作约束深度强化学习的安全自动驾驶方法[J].计算机科学,2021,48(9):235.
DAI Shanshan, LIU Quan. Action constrained deep reinforcement learning based safe automatic driving method[J]. Computer Science,2021,48(9):235.

[82] SHALEV-SHWARTZ S, SHAMMAH S, SHASHUA A. Safe , multi-agent, reinforcement learning for autonomous driving[J]. arXiv preprint arXiv:1610.03295, 2016.

[83] BEWLEY A, RIGLEY J, LIU Y, *et al*. Learning to drive from simulation without real world labels [C]//2019 International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2019: 4818-4824.