

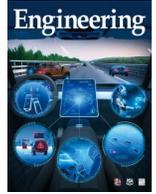


ELSEVIER

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Research
Safety for Intelligent and Connected Vehicles—Article

面向可信自动驾驶决策——一种具有安全保证的鲁棒强化学习方法

何祥坤, 黄文辉, 吕辰*

School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798, Singapore

ARTICLE INFO

Article history:

Received 15 October 2022

Revised 22 March 2023

Accepted 18 October 2023

Available online 27 November 2023

关键词

自动驾驶汽车

决策

强化学习

对抗攻击

安全保障

摘要

尽管自动驾驶汽车是智能交通系统的重要组成部分,但确保自主决策的可信性仍然是实现自动驾驶技术大规模部署的一个重大挑战。因此,我们提出了一种新颖的具有安全保证的鲁棒强化学习方法,以实现自动驾驶汽车的可信决策。该技术能够从策略鲁棒性和碰撞安全性两个方面保证自主决策的可信性。具体地说,通过逼近针对观测状态和环境动态的最优对抗摄动,可以在线学习对手模型,以模拟最坏情况下的不确定性。我们还提出了一种对抗鲁棒演员-评论家算法,使智能体能够学习针对状态观测摄动与环境动态摄动的鲁棒策略。此外,我们设计了一个基于可解释知识模型(即责任敏感安全模型)的安全掩码,保证自动驾驶智能体在训练和测试过程中的碰撞安全性。最后,通过仿真测试与实验验证对所提方法进行了评估。结果表明,基于学习到的鲁棒安全策略,自动驾驶智能体不仅能够实现可信决策,还能显著减少车辆碰撞次数。

© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. 引言

近年来,随着先进移动通信技术[1]和人工智能(AI)[2]的快速发展,自动驾驶汽车得到了迅猛发展,并有望彻底改变人类的出行和交通系统[3–5]。然而,现实世界的交通场景中存在不可预测的噪声和不确定性,这使得确保自动驾驶系统的鲁棒性和安全性变得极具挑战性。因此,自动驾驶的可信性问题引起了各机构和公众的广泛关注[6–8]。面对这些复杂的挑战,如何满足自动驾驶相关的严格要求和高期望,仍然是一个亟待解决的重要问题[9–11]。

决策系统可以被比作自动驾驶汽车的大脑,主要负责

根据感知信息确定最佳驾驶模式或策略[12–14]。许多研究已经报道了自动驾驶决策方法的进展[15–17]。有限状态机(FSM)是一种基于规则的技术,是开发决策系统中最流行的方法之一[18–19]。虽然这种方法易于实现和解释,但它在很大程度上依赖于专家的先验知识,因此在设计复杂交通情况下的驾驶规则时面临挑战。

作为现代人工智能技术的重要组成部分,强化学习(RL)通过与环境的交互,为解决复杂的序列决策任务提供了一个可行且有效的范式[20–22]。因此,一些研究尝试了各种RL方法以解决一系列的自动驾驶任务[23–25]。研究人员已经利用RL算法来学习自动驾驶的变道策略[26–27]。例如,使用风险意识优先重放深度Q-网络(RA-

* Corresponding author.

E-mail address: lyuchen@ntu.edu.sg (C. Lv).

PRDQN)方法开发了一种自动驾驶汽车的变道决策框架[28]。采用基于规则的安全验证方法,开发了一种自动驾驶安全变道决策方案[29]。此外,一些研究已经使用RL算法来学习自动驾驶车辆的最佳目标速度或速度模式(如加速、减速和维持)[30–31]。例如,利用信念状态RL方法开发了一种自动驾驶汽车的协作感知入口匝道合并决策方案[32]。采用基于状态-注意模型的分层RL方法,确定了自动驾驶汽车的基于子目标的速度模式[33]。为了保证入口匝道合并策略对环境不确定性的鲁棒性,提出了一种约束对抗RL技术的自动驾驶鲁棒决策方案[34]。许多研究人员已经利用RL算法来同时学习自动驾驶汽车的最佳变道策略和速度模式[35–37]。例如,自动驾驶的纵向和横向决策行为可以通过带有短时域安全检查器的双深度Q-网络(DDQN)学习[38],而自动驾驶车辆的目标速度和车道变化策略可以通过基于多个智能体的分层程序触发RL技术来确定[39]。在另一项研究中,开发了一种基于规则策略的可信改进RL方案,使自动驾驶智能体能够学习安全的纵向和横向驾驶速度[40]。

虽然现有的自动驾驶决策研究已经取得了许多显著成果,提高了自动驾驶汽车的性能,但在可信性方面仍有改进和完善的空间。此外,大多数研究假设交通环境中没有不确定性,或仅涉及一种特定类型的不确定性。遗憾的是,现实场景涉及大量不可避免的不确定性,这可能导致自动驾驶智能体做出不理想甚至不安全的决策。此外,现实交通环境一般同时涉及多种不确定性(如观测噪声和环境变化),可能引起复杂且具有挑战性的驾驶工况。因此,在自动驾驶领域应开展针对多重不确定性的策略鲁棒性研究。然而,很少有研究着手解决如何在充满对抗性环境不确定性的随机动态交通流中,在策略模型训练和测试两个阶段保证RL驱动的自动驾驶智能体的安全性这一挑战。

因此,以上所有见解促使我们探索一种新技术,以确保自动驾驶决策在策略鲁棒性和碰撞安全性方面的可信性。在本研究中,我们引入了一种具有安全保证的新型鲁棒RL方法(RRL-SG),旨在实现自动驾驶车辆的可信决策。本文的主要贡献总结如下:

(1) 通过逼近观测状态和环境动态的最优对抗摄动,在线训练一个对抗智能体以模拟最坏情况下的多重不确定性。提出了一种对抗鲁棒演员-评论家(ARAC)算法,使其能够学习针对观测噪声和环境变化的鲁棒策略。

(2) 使用Intel提出的可解释知识模型——责任敏感安全性(RSS)[41–42],开发了一种安全掩码,以保证自动驾驶智能体在训练和测试过程中的碰撞安全,它可以

不安全决策所对应的概率转换为零(即通过屏蔽风险动作形成安全的动作空间)。

(3) 基于城市交通模拟器SUMO的数值模拟结果[43]表明,所提出的RRL-SG方法能够保证自动驾驶汽车在对抗环境摄动影响下的随机动态交通流中的可信性。使用真实自动驾驶车辆的实验进一步证实了所提出技术的有效性。

本文的其余部分组织如下:第2节描述了所提出的RRL-SG解决方案;第3节详细介绍了技术实施的细节;第4节详细介绍了模拟和实验,并分析了由此产生的性能;最后,第5节进行了总结。

2. 方法学

2.1. 概述

在本节中,我们概述了所提出的技术。图1展示了我们的RRL-SG框架的方框图,该框图旨在实现自动驾驶汽车的可信决策。 Δ_o^* 和 Δ_d^* 分别表示对观测状态和环境动态的最优对抗摄动。 M_s 、 s 、 a 、 r 和 π 分别表示智能体的安全掩码、状态、行动、奖励和策略。 π_s 代表一个安全的策略。 t 是时间步长, T 是最后一个时间步长。 Δ 、 γ 、 β 和 Q^π 分别表示优化目标中的环境不确定性、折扣因子、权重和行为-价值函数。

对抗模型的输入是 s 智能体的状态,其输出包含对抗摄动 Δ_o^* 和 Δ_d^* 。 Δ_o^* 模拟最坏情况下的观测噪声,旨在使摄动策略上的平均变化距离最大化。此外, Δ_d^* 模拟了最坏情况下的环境动态不确定性,力求最小化智能体的预期回报。

基于RSS的安全掩码的输入是智能体的状态 s 。一个安全掩码可以通过保护其免受风险行动的影响来创造一个安全的行动空间。因此,自动驾驶智能体通过从安全策略 π_s 中采样的行动与环境进行交互。ARAC算法使智能体能够学习针对状态观测和环境动态中的摄动的鲁棒策略。

我们的自动驾驶智能体是一种金色的智能车辆,如图1所示。此外,使用基于SUMO的智能驾驶模型(IDM)对周围其他颜色的车辆进行控制。我们的自动驾驶智能体的行为空间是离散的,包括五种不同的决策行为:保持当前状态、加速、减速和向左或向右改变车道。

2.2. 对手模型

对手模型的目的是学习并逼近针对观测状态和环境动态的最优对抗摄动。

为了衡量由对抗观测摄动引起的策略变化,我们利用

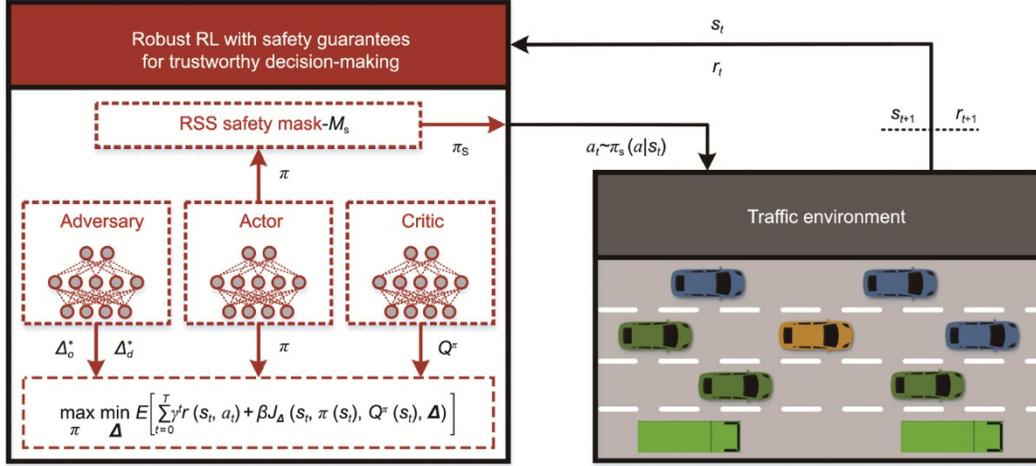


图1. 所提出的自动驾驶汽车可信决策的RRL-SG框架示意图。 Δ_o^* : 对观察状态的最优对抗摄动； Δ_d^* : 针对环境动态的最优对抗摄动； M_s : 安全掩码； s : 智能体的状态； a : 智能体的行为； r : 智能体的奖励； π : 智能体的策略； π_s : 安全策略； t : 时间步； T : 最后一步； Δ : 环境不确定性； γ : 折扣因素； β : 权重； Q^r : 行动-价值函数； E : 数学期望； J_Δ : 对手的目标函数。

了 Jensen-Shannon (JS) 散度，它可以被认为是一个对称且光滑的 Kullback-Leibler (KL) 散度[44–45]。它的一个关键特征是 JS 散度将两个概率分布之间的距离限制在 1.0 以内。因此，与观测中的摄动相关的目标函数 J_o 可定义如下：

$$\begin{aligned} J_o(s, \pi, \Delta_o) &= D_{JS}[\pi(a|s) \|\pi(\tilde{a}|\tilde{s})] \\ &= D_{JS}[\pi(a|s) \|\pi(\tilde{a}|s + \Delta_o)] \end{aligned} \quad (1)$$

$$\begin{aligned} &= \frac{1}{2} D_{KL}[\pi(a|s) \|\pi(\tilde{a}|s + \Delta_o)] + \frac{1}{2} D_{KL}[\pi(\tilde{a}|s + \Delta_o) \|\pi(a|s)] \\ &M = \frac{1}{2} [\pi(a|s) + \pi(\tilde{a}|s + \Delta_o)] \end{aligned} \quad (2)$$

式中， D_{JS} 表示基于 JS 散度的距离； D_{KL} 表示基于 KL 散度的距离； Δ_o 表示对观测值的摄动； M 是关于智能体策略和摄动策略的表达式； \tilde{s} 和 \tilde{a} 分别为被 Δ_o 摄动的状态和行为。

在本研究中，动态的对抗摄动试图最小化智能体的预期回报。当智能体遵循策略 π 时，我们利用一个行为-价值函数 $Q^\pi(s)$ 来对基于一对状态 s 和行为 a 来估计预期回报。由于我们的智能体的行为空间是离散的，行为-价值函数 $Q^\pi(s)$ 的输入不包括行为 a 。因此，与动态上的摄动相关的目标函数 J_d 可以被设计为

$$J_d(s, Q^\pi, \Delta_d) = \Delta_d Q^\pi(s) \quad (3)$$

式中， Δ_d 以概率分布的形式表示动态的摄动。此外，对手的目标函数 J_Δ 可以定义为

$$J_\Delta(s, \pi, Q^\pi, \Delta) = (\alpha - 1)J_o(s, \pi, \Delta_o) + \alpha J_d(s, Q^\pi, \Delta_d) \quad (4)$$

式中， $\alpha \in (0, 1)$ 表示一个权重； $\Delta = [\Delta_o, \Delta_d]$ 表示环境的不确定性。

关于对手模型的优化问题可以表述为

$$\begin{aligned} \Delta^* &\in \arg \min_{\Delta} E[J_\Delta(s, \pi, Q^\pi, \Delta)], \\ &\text{subject to } |\Delta_o| \leq \eta_1, |\Delta_d| \leq \eta_2 \end{aligned} \quad (5)$$

式中， Δ^* 表示最优环境不确定性； $\arg \min$ 表示最小值的参数； η_1 和 η_2 分别表示观测和动态上的摄动的界限。因此，对抗智能体的目标是最大化 J_o 和最小化 J_d 。

为了简化上述约束优化问题，我们使用双曲正切函数和 softmax 函数来约束摄动的大小。具体地说，对观测和动态的摄动可以分别用 $\Delta_o = \eta \tanh[x(s; \bar{\theta})]$ 和 $\Delta_d = \text{softmax}[x(s; \bar{\theta})]$ 表示。另外， η 表示比例因子， x 表示对手网络隐藏层的输出， $\bar{\theta}$ 表示对手模型参数。

因此，为了确定最优的对抗摄动，等式 (5) 可以转换为：

$$\bar{\theta}^* \in \arg \min_{\bar{\theta}} E[J_\Delta(s, \pi, Q^\pi; \bar{\theta})] \quad (6)$$

式中， $\bar{\theta}^*$ 表示最优对手模型的参数。显然，对观测值和动态的最优对抗摄动可以分别用 $\Delta_o^* = \eta \tanh[x(s; \bar{\theta}^*)]$ 、 $\Delta_d^* = \text{softmax}[x(s; \bar{\theta}^*)]$ 来表示。

2.3. 基于 RSS 的安全掩码

在本节中，我们使用一个可解释的 RSS 模型开发了一个安全掩码，以保证自动驾驶汽车的碰撞安全。

为了考虑驾驶舒适性，我们利用 Intel 提出的 jerk-bounded RSS 模型[42]来设计安全掩码。该模型描述了以下制动过程：车辆开始以最大抖动 j_{\max} 降低加速度，直到达到最小减速度 $a_{\min,r}$ ，然后车辆继续以减速度 $a_{\min,r}$ 制动，直到达到完全停止。根据 Jerk-bounded RSS 模型得出前后车辆最小安全距离 D_{\min}^{RSS} 的表达式：

$$D_{\min}^{\text{RSS}} = \left| v_r \bar{T} + \frac{1}{2} a_r \bar{T}^2 - \frac{1}{6} j_{\max} \bar{T}^3 + \frac{(v_r + a_r \bar{T} - \frac{1}{2} j_{\max} \bar{T}^2)^2}{2|a_{\min,r}|} - \frac{v_f^2}{2|a_{\max,f}|} \right| \quad (7)$$

式中, a_r 为后车的初始加速度; v_r 和 v_f 表示前后车辆的初始速度; $a_{\min,r}$ 表示前车的最大减速; \bar{T} 表示从开始到后车的减速首次等于 $a_{\min,r}$ 或速度降至零的时间。

我们使用图2中所示的两种情况来说明所提出的安全掩码技术。如图2(a)所示, 如果与同一车道的前方车辆的距离(表示为 D_f) 小于或等于 D_{\min}^{RSS} , 掩码将把加速度决策对应的概率(计为 a^4) 转换为零(即屏蔽风险行动 a^4 , 形成包括 a^1 、 a^2 、 a^3 和 a^5 的安全行动空间)。虽然参考文献[42]中只提供了最小纵向安全距离模型, 如果我们假

设车辆可以瞬间横向移动到目标车道, 我们仍然可以使用这个模型来评估变道风险。这样的评估是有风险的, 因为在变道过程中, 两辆车之间的距离可能会进一步缩短。在此, 我们基于 D_{\min}^{RSS} 设计了一个简单的最小横向安全距离模型, 模型如下:

$$\bar{D}_{\min}^{\text{RSS}} = \zeta D_{\min}^{\text{RSS}} \quad (8)$$

式中, ζ 表示大于1.0的比例系数; $\bar{D}_{\min}^{\text{RSS}}$ 为最小横向安全距离模型。

在图2(b)中, 如果与左车道后车的距离(表示为 D_{rl}) 小于或等于 $\bar{D}_{\min}^{\text{RSS}}$, 掩码将左车道变更决策(表示为 a^2) 对应的概率转换为零。

在图2(c)中, 当与左车道后车距离(表示为 D_{rl})、与右车道前车距离(表示为 D_{fr})、与同一车道前车距离(表示为 D_f) 小于或等于其对应的最小安全距离时, 掩码将左变道(表示为 a^2)、右变道(表示为 a^1) 和加速(表

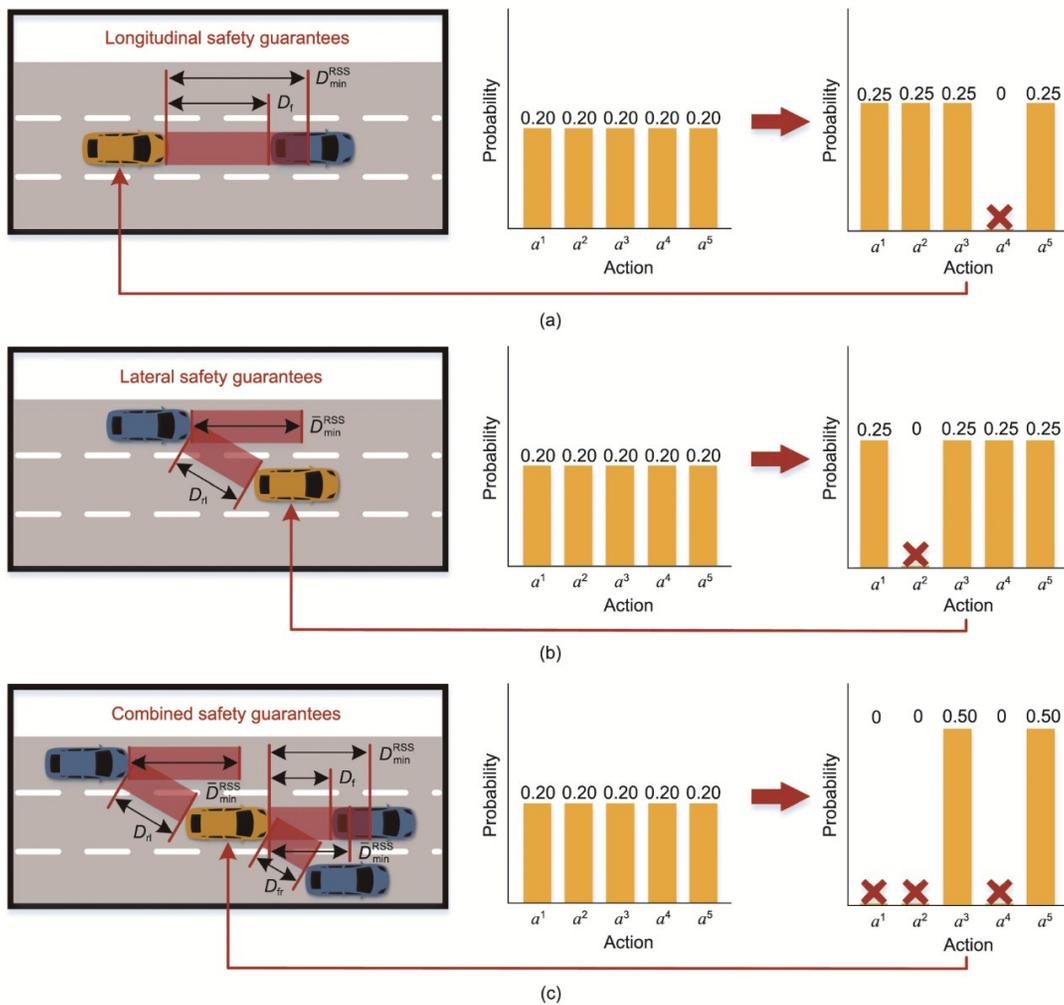


图2. 基于RSS的安全掩码的可信驾驶策略。(a)纵向安全保障示意图。(b)横向安全保障示意图。(c)组合安全保障示意图。 a_1 、 a_2 、 a_3 、 a_4 和 a_5 分别表示向右变道、向左变道、保持当前状态、加速和减速; D_f : 同一车道前方车辆的距离; $\bar{D}_{\min}^{\text{RSS}}$: 最小横向安全距离模型; D_{rl} : 与左车道后方车辆的距离; D_{fr} : 与右车道前排车辆的距离。

示为 a^4) 决策对应的概率转换为零。

算法1概述了我们基于RSS的安全掩码模块的设计, 其中, D_r 、 D_{fl} 、 D_{fr} 和 D_{rr} 分别表示与后、左前、右前、右后车辆的距离; $D_{\min, f}^{\text{RSS}}$ 、 $D_{\min, r}^{\text{RSS}}$ 、 $\bar{D}_{\min, fl}^{\text{RSS}}$ 、 $\bar{D}_{\min, rl}^{\text{RSS}}$ 、 $\bar{D}_{\min, fr}^{\text{RSS}}$ 和 $\bar{D}_{\min, rr}^{\text{RSS}}$ 分别表示与前、后、左前、左后、右前、右后车辆的最小安全距离。此外, $M_s[m]$ 表示安全掩码 M_s 中的第 m 个元素。与风险行为相关的掩码元素被分配一个负无穷值。

算法1 基于RSS的安全掩码

Input: State of the autonomous driving agent

Initialize a mask $M_s = [0, 0, 0, 0, 0]$

if $D_f \leq D_{\min, f}^{\text{RSS}}$ **then**

$M_s[4] = -\infty$ *Mask accelerating decision-making

end if

if $D_r \leq D_{\min, r}^{\text{RSS}}$ **then**

$M_s[5] = -\infty$ *Mask decelerating decision-making

end if

if $D_{fl} \leq \bar{D}_{\min, fl}^{\text{RSS}}$ **or** $D_{rl} \leq \bar{D}_{\min, rl}^{\text{RSS}}$ **then**

$M_s[2] = -\infty$ *Mask left lane-changing decision-making

end if

if $D_{fr} \leq \bar{D}_{\min, fr}^{\text{RSS}}$ **or** $D_{rr} \leq \bar{D}_{\min, rr}^{\text{RSS}}$ **then**

$M_s[1] = -\infty$ *Mask right lane-changing decision-making

end if

Output: M_s

2.4. 对抗鲁棒演员-评论家算法

2.4.1. 安全鲁棒马尔科夫决策过程

马尔科夫决策过程 (MDP) 为 RL 问题提供了一个数学范式, 旨在找到最优策略[46]。在本节中, 现有的标准 MDP 数学形式被扩展, 以明确构建自动驾驶智能体在对抗扰动和安全掩码下的决策行为。在这里, 我们介绍了一个安全鲁棒 MDP (SR-MDP), 定义如下:

SR-MDP 可以通过一个 7 元组 $[S, A, p, r, \Delta, M_s, \gamma]$ 来定义, 其中, S 为状态空间, A 为行为空间, p 为状态转移概率, r 为奖励函数, Δ 为环境不确定性, M_s 为安全掩码, $\gamma \in (0, 1)$ 为折扣因子。

在我们的研究中, SR-MDP 试图解决以下问题:

$$\max_{\pi} \min_{\Delta} E \left[\sum_{t=0}^T \gamma^t r(s_t, a_t) + \beta J_{\Delta}(s_t, \pi(s_t), Q^{\pi}(s_t), \Delta) \right] \quad (9)$$

式中, T 是最后一个时间步长; $\beta > 0$ 是一个权衡系数。

我们采用一种新的策略迭代 (PI) 算法, 称为安全鲁棒 PI (SR-PI) 来解决 SR-MDP。SR-PI 方法包括两个关键阶段: 安全鲁棒策略评估和安全鲁棒策略改进。此外, 这两个阶段都被迭代更新, 直到达到收敛。

2.4.2. 安全鲁棒策略评估

在安全鲁棒策略评估阶段, 我们的目标是估计在环境不确定性条件 Δ 下的策略 π 的预期回报。对于固定策略, 可以使用以下 Bellman 备份算子迭代逼近行为-价值函数:

$$T^{\pi, \Delta} Q^{\pi}(s_t) = r(s_t, a_t) + \gamma E[V^{\pi, \Delta}(s_{t+1})] \quad (10)$$

$$V^{\pi, \Delta}(s_{t+1}) = \pi(s_{t+1}) Q^{\pi}(s_{t+1}) + \beta J_{\Delta}(s_{t+1}, \pi(s_{t+1})), \quad (11)$$

$$Q^{\pi}(s_{t+1}, \Delta)$$

表示在对抗扰动下并基于策略 π 的价值函数。

在这里, 我们可以重写等式 (10) 如下:

$$T^{\pi, \Delta} Q^{\pi}(s_t) = r_a(s_t, a_t) + \gamma \pi(s_{t+1}) Q^{\pi}(s_{t+1}) \quad (12)$$

式中, $r_a(s_t, a_t) = r(s_t, a_t) + \gamma \beta J_{\Delta}(\cdot)$ 是增广的奖励。因此, 利用标准 RL 算法中与策略评估收敛性相关的结果, 可以保证我们的策略评估的收敛性。

为了提高模型训练的效率, 我们采用了两个参数化的行为-价值函数, 参数为 ϕ^p , $p \in \{1, 2\}$ 。这两个行为-价值函数的参数可以通过最小化关于评估网络的以下目标函数来进行优化:

$$J_Q(\phi^p) = E_{T_s \sim \mathcal{B}} [(y_t^A - Q^{\pi}(s_t; \phi^p))^2] \quad (13)$$

式中, T_s 表示从重放缓冲区 \mathcal{B} 采样的状态转换; y_t^A 表示在时间步长 t 上具有不确定性的行为-价值函数的目标值; J_Q 是优化批判网络的函数。对两个行为-价值函数都使用较小的值, 以减轻在批判网络训练时对价值函数的高估。因此, y_t^A 可以定义为:

$$y_t^A = r(s_t, a_t) + \gamma \pi(s_{t+1}) \hat{Q}_{\min}^{\pi}(s_{t+1}; \bar{\phi}^p) + \beta J_{\Delta}(s_{t+1}, \pi(s_{t+1}), \hat{Q}_{\min}^{\pi}(s_{t+1}; \bar{\phi}^p), \Delta) \quad (14)$$

式中, $\hat{Q}^{\pi}(s; \bar{\phi}^p)$ 为带有参数 $\bar{\phi}^p$ 的目标行为-价值函数; $\hat{Q}_{\min}^{\pi}(s; \bar{\phi}^p)$ 表示两个目标行为-价值函数中的较小值, 例如, $\hat{Q}_{\min}^{\pi}(s; \bar{\phi}^p) = \min_{p \in \{1, 2\}} \hat{Q}^{\pi}(s; \bar{\phi}^p)$ 。

可推导出等式 (13) 的梯度为:

$$\nabla_{\phi^p} J_Q(\phi^p) = \nabla_{\phi^p} E_{T_s \sim \mathcal{B}} [(y_t^A - Q^{\pi}(s_t; \phi^p))^2] = -2 E_{T_s \sim \mathcal{B}} [(y_t^A - Q^{\pi}(s_t; \phi^p)) \nabla_{\phi^p} Q^{\pi}(s_t; \phi^p)] \quad (15)$$

此外, 我们还可以通过 Polyak 平均的方法进行 $\bar{\phi}^p$ 更新:

$$\bar{\phi}^p \leftarrow \mu \bar{\phi}^p + (1 - \mu) \phi^p \quad (16)$$

式中, $\mu \in (0, 1)$ 为比例系数。

2.4.3. 安全鲁棒的策略改进

在安全鲁棒策略改进阶段, 我们试图对在对抗扰动下的行为-价值函数 $Q^{\pi}(\cdot)$ 的策略进行优化。由于使用行为-价值函数 $Q^{\pi}(s)$ 来估计智能体遵循策略 π 时的状态 s 和行为 a

的预期回报，因此优化问题[等式 (9)]可重写为：

$$\max_{\pi} \min_{\Delta} E[J(\pi, \Delta)] \quad (17)$$

式中， $J(\cdot)$ 表示所提出的SR-MDP的目标函数 $J(\pi, \Delta) = \pi(s)Q^{\pi}(s) + \beta J_{\Delta}(s, \pi, Q^{\pi}, \Delta)$ 。

因此，观测状态和环境动态的最优策略 π^* 和最优对抗摄动 Δ^* 可以用以下交替过程来近似：首先，确定一个策略 π ，然后通过最小化 $J(\pi, \Delta)$ 求解最优对抗摄动 Δ^* ；其次，通过最大化 $J(\pi, \Delta^*)$ 来学习最优策略 π^* 。根据等式(17)，推导出以下关系表达式：

$$\Delta^* = \arg \min_{\Delta} E[J(\pi, \Delta)] \quad (18)$$

$$\pi^* = \arg \max_{\pi} E[J(\pi, \Delta^*)] \quad (19)$$

我们观察到等式(17)表示一个零和博弈。此外，理论结果[47–49]的建立是为了保证零和博弈解的收敛性，这也保证了我们的策略改进的收敛性。

为了减少策略 π 的学习误差，我们在参考文献中使用了双重 $Q^{\pi}(\cdot)$ 技巧[50]。因此，策略模型参数 θ 可以通过最大化以下关于参与者网络的目标函数来学习：

$$J_{\pi}(\theta) = E_{\tau, \tau-B} [\pi(s; \theta) Q_{\min}^{\pi}(s; \phi^p) + \beta J_{\Delta}(s, \pi(s; \theta), Q_{\min}^{\pi}(s; \phi^p), \Delta)] \quad (20)$$

式中， $Q_{\min}^{\pi}(s; \phi^p)$ 表示两个行为-价值函数的较小值，例如， $Q_{\min}^{\pi}(s; \phi^p) = \min_{p \in \{1,2\}} Q^{\pi}(s; \phi^p)$ 是优化参与者网络的函数。

我们可以推导出等式(20)的梯度，具体内容如下：

$$\begin{aligned} \nabla_{\theta} J_{\pi}(\theta) &= \nabla_{\theta} E_{\tau, \tau-B} [\pi(s; \theta) Q_{\min}^{\pi}(s; \phi^p) + \beta J_{\Delta}(s, \pi(s; \theta), Q_{\min}^{\pi}(s; \phi^p), \Delta)] \\ &= E_{\tau, \tau-B} [\nabla_{\theta} \pi(s; \theta) Q_{\min}^{\pi}(s; \phi^p) + (\alpha - 1) \beta \nabla_{\theta} J_{\Delta}(s, \pi(s; \theta))] \\ &= E_{\tau, \tau-B} [\nabla_{\theta} \pi(s; \theta) Q_{\min}^{\pi}(s; \phi^p) + \frac{1}{2} (\alpha - 1) \beta (\nabla_{\theta} D_{\text{KL}}(\pi(a|s; \theta) \| M(s; \theta)) + \nabla_{\theta} D_{\text{KL}}(\pi(a|s + \Delta_0; \theta) \| M(s; \theta))] \end{aligned} \quad (21)$$

此外，根据方程式(4)和(5)，通过最小化以下目标函数，可以优化对手模型：

$$\begin{aligned} J_{\pi}(\bar{\theta}) &= E_{\tau, \tau-B} [J_{\Delta}(s, \pi(s; \theta), Q_{\min}^{\pi}(s; \phi^p); \bar{\theta})] \\ &= E_{\tau, \tau-B} [(\alpha - 1) J_{\Delta}(s, \pi(s; \theta); \bar{\theta}) + \alpha J_{\Delta}(s, Q_{\min}^{\pi}(s; \phi^p); \bar{\theta})] \end{aligned} \quad (22)$$

式中， $\bar{\theta}$ 表示对手模型参数； J_{π} 是优化对手网络的函数。

这里等式(22)的梯度可推导出：

$$\nabla_{\bar{\theta}} J_{\pi}(\bar{\theta}) = \nabla_{\bar{\theta}} E_{\tau, \tau-B} [J_{\Delta}(s, \pi(s; \theta), Q_{\min}^{\pi}(s; \phi^p); \bar{\theta})]$$

$$\begin{aligned} &= \nabla_{\bar{\theta}} E_{\tau, \tau-B} [(\alpha - 1) J_{\Delta}(s, \pi(s; \theta); \bar{\theta}) + \alpha J_{\Delta}(s, Q_{\min}^{\pi}(s; \phi^p); \bar{\theta})] \\ &= \nabla_{\bar{\theta}} E_{\tau, \tau-B} \left[\frac{1}{2} (\alpha - 1) (D_{\text{KL}}(\pi(a|s) \| M(s; \bar{\theta})) + D_{\text{KL}}(\pi(\tilde{a}|s + \Delta_0(s; \bar{\theta})) \| M(s; \bar{\theta}))) + \alpha J_{\Delta}(s; \bar{\theta}) Q^{\pi}(s) \right] \\ &= E_{\tau, \tau-B} \left[\frac{1}{2} (\alpha - 1) (\nabla_{\bar{\theta}} D_{\text{KL}}(\pi(a|s) \| M(s; \bar{\theta})) + \nabla_{\bar{\theta}} D_{\text{KL}}(\pi(\tilde{a}|s + \Delta_0(s; \bar{\theta})) \| M(s; \bar{\theta})) + \alpha \nabla_{\bar{\theta}} J_{\Delta}(s; \bar{\theta}) Q^{\pi}(s)) \right] \end{aligned} \quad (23)$$

3. 技术实施

3.1. 算法

在这里，我们将详细介绍所提出的技术的实现规范。算法2概述了用于可信自动驾驶决策的RRL-SG方法。演员、对手和评论家的初始模型参数是使用随机分布设置的。在与环境的交互方面，我们的智能体基于从安全策略 π_s 中取样的行为与环境进行交互。在策略学习方面，可以通过联合方程式(13)、(16)、(20)和(22)来优化智能体策略。 d_t 表示一个任务结束信号，意味着自我车辆在时间步长 t 中遇到碰撞。神经网络和超参数的细节见附录A中的表S1。

3.2. 状态空间和行为空间

设计自动驾驶智能体的状态、行为和奖励函数是实施该方案的关键。在本研究中，我们将自我车辆车道和相邻车道中最近的6个社会车辆的相关状态作为对自动驾驶主体（即自我车辆）的观察。自动驾驶智能体的状态空间有15个维度，包括周围社会车辆的相对距离和速度，以及自我车辆的速度、加速度和车道指数。车道索引是自我车辆所在的车道的索引。

我们的自动驾驶智能体的行为空间是离散的，包含5种决策行为：向右变道、向左变道、保持当前状态、加速和减速。根据参考文献[51]中的研究结果，通常，由正常驾驶员驾驶的车辆的加速度不超过 $1.47 \text{ m} \cdot \text{s}^{-2}$ ，且减速速度不小于 $-2 \text{ m} \cdot \text{s}^{-2}$ 。因此，当我们的自动驾驶智能体执行加速度决策时，自我车辆将以 $1.47 \text{ m} \cdot \text{s}^{-2}$ 的固定加速度加速。此外，如果智能体执行减速决策，则自我车辆以 $-2.00 \text{ m} \cdot \text{s}^{-2}$ 的固定减速度减速。

3.3. 奖励函数

奖励函数在RL智能体的性能中起着关键作用。我们的奖励函数是通过考虑出行效率、驾驶安全和乘客舒适度等相关因素来设计的。具体来说，我们鼓励自动驾驶智能

算法2 具有安全保证的鲁棒 RL

Initialize actor model parameters θ , adversary model parameter $\bar{\theta}$, critic model parameters φ^1 and φ^2 , target action-value function parameters $\bar{\varphi}^1 \leftarrow \varphi^1$ and $\bar{\varphi}^2 \leftarrow \varphi^2$, and an empty replay buffer B

for episode step $e = 1, 2, \dots, E$ **do**

 Reset state s_0

for time step in the environment $t = 1, 2, \dots, T$ **do**

 Determine a safe policy $\pi_s(s_t; \theta)$ via Algorithm 1:

$\pi_s(s_t; \theta) = \text{softmax}(\pi(s_t; \theta) + M_s)$

 Select an action via the safe policy $\pi_s(s_t; \theta)$:

$a_t \sim \pi_s(s_t; \theta)$

 Execute a_t in the environment and receive a transition:

$s_{t+1}, r_t, d_t \sim p(s_{t+1}|s_t, a_t)$

 Store the transition in the replay buffer B :

$B \leftarrow B \cup \{(s_t, a_t, r_t, s_{t+1}, d_t)\}$

end if

for gradient step $g = 1, 2, \dots, G$ **do**

 Sample a batch of transitions from the replay buffer B

 Update the actor model parameters via Eq. (21):

$\theta \leftarrow \nabla_{\theta} J_{\pi}(\theta)$

 Update the critic model parameters via Eq. (15):

$\varphi^1 \leftarrow \nabla_{\varphi^1} J_{\varphi}(\varphi^1), \varphi^2 \leftarrow \nabla_{\varphi^2} J_{\varphi}(\varphi^2)$

 Update the target action-value function parameters via Eq. (16):

$\bar{\varphi}^1 \leftarrow \mu \bar{\varphi}^1 + (1 - \mu) \varphi^1, \bar{\varphi}^2 \leftarrow \mu \bar{\varphi}^2 + (1 - \mu) \varphi^2$

if $g \bmod \delta$ **then**

 Update the adversary model parameters via Eq. (23):

$\bar{\theta} \leftarrow \nabla_{\bar{\theta}} J_{\bar{\pi}}(\bar{\theta})$

end if

end for

end for

体以高速行驶。此外，如果智能体的驾驶策略导致了碰撞，我们也会对其进行处罚。如果一个自动驾驶智能体执行高速变道操作，就会受到处罚。等式 (24) 是设计的奖励函数 $r(\cdot)$ ，其中， e 为自然对数， v_0 为自我车速。此外， $A = \{\text{车辆车道}\}$ 、 $B = \{v_0 > 30\}$ 和 $C = \{\text{碰撞}\}$ 是事件集。在这里，碰撞指的是自我车辆与周围的社会车辆之间的碰撞。

$$r(\cdot) = \begin{cases} e^{v_0/35-1} - v_0/350 & A \wedge B \wedge \neg C = 1 \\ e^{v_0/35-1} - 0.5 - v_0/100 & \neg(A \wedge B) \wedge C = 1 \\ e^{v_0/35-1} - v_0/350 - 0.5 - v_0/100 & A \wedge B \wedge C = 1 \\ e^{v_0/35-1} & \text{otherwise} \end{cases} \quad (24)$$

4. 仿真和实验

4.1. 基线

我们在仿真和实验中与最先进的 RL 智能体进行了比

较，以对 RRL-SG 方法进行基准测试，实现可信的自动驾驶决策。

由于对抗 DDQN (D3QN) 是一种最先进的 Q 学习算法 [52–53]，本研究采用 D3QN 作为基线之一。此外，我们还利用了近端策略优化 (PPO) [54]、软性演员-评论家 (SAC) [50,55] 和观察对抗性 RL (OARL) [56] 算法作为竞争基线，分别代表最先进的策略上、策略外和鲁棒的 RL 技术。

4.2. 指标

我们采用预期回报来评估自动驾驶智能体的综合性能。利用平均运行速度和碰撞次数对自动驾驶车辆的行驶效率和交通安全性进行了评价。此外，等式 (1) 用于衡量策略对抗抗摄动的鲁棒性，即被对手攻击的策略变化越小，策略的鲁棒性越强。

在入口匝道合并场景中，除了上述指标外，我们还使用合并成功率来评估车辆的性能。在本研究中，一个成功的入口匝道合并被定义为车辆完全从匝道进入主车道，并且在测试过程中没有发生任何碰撞。

4.3. 基于 SUMO 的仿真

为了评估所提出的自动驾驶汽车决策技术的性能，我们使用 SUMO 模拟器实现了模型训练和测试。我们利用 SUMO 在高速公路和入口匝道合并场景中创建了具有不同密度的随机动态交通流。此外，我们在正常密度的公路场景中 ($P = 0.12$)，用不同的随机种子和 400 个片段对每种方法进行了 5 次不同的训练。 P 表示以秒为单位启动车辆的概率。每个线路的最大时间步长是 200 s。所有车道的最大交通速度被设置为 $35.0 \text{ m} \cdot \text{s}^{-1}$ 。与高速公路方案不同，入口匝道合并方案仅用于模型测试。

4.3.1. 公路场景

图 3 说明了我们高速公路场景的评估方案。自我车辆是 RL 驱动的自动驾驶汽车。 P 分别设置为 0.06、0.12 和 0.24，以产生低密度、正常密度和高密度的交通流量。自动驾驶智能体只在正常密度的交通流量中进行训练。在模型测试阶段，利用低密度、正常密度和高密度的交通流量进行评估。每个被训练过的智能体 (包括不同的随机种子) 都经过 100 多次的评估。每个评估都计算了测试中 10 种情况的平均指标。当我们采用随机动态交通流时，环境动态是不断变化的。为了进一步验证策略的鲁棒性，在模型测试过程中，每个自动驾驶智能体都受到来自训练对手的最优对抗性观测摄动的攻击。换句话说，与模型训练阶段不同的是，在具有对抗性攻击的测试案例中，自动

驾驶智能体接收到受对手模型摄动的状态 δ 。

图4显示了所提出的RRL-SG方法的学习曲线和正态密度随机动态交通流量的基线。总体而言，所提出的方案在回报和安全性方面优于基线。显然，与基线相比，我们的自动驾驶智能体大大减少了碰撞次数，提高了模型训练期间的学习效率，因为所提出的基于RSS的安全掩码通过保护其免受风险行为来形成一个安全行为子空间。因此，从安全行为子空间进行的采样行为确保了决策安全，避免了冗余探索。

在模型测试过程中，对每种方法基于5个随机种子的最终策略模型进行了评估。定性地说，我们在表1中报道了模型评价结果的平均指标。粗体数字表示每个指标的最佳值。总的来说，研究表明，RRL-SG智能体在所有任务的鲁棒性和安全性方面都大大超过了基线水平。与基线相比，在三种不同密度的随机动态流量中，受到对手模型攻击的RRL-SG策略变化的JS散度约为零，这意味着RRL-SG策略几乎不受对抗性攻击的影响。此外，与D3QN、PPO、SAC和OARL自动驾驶智能体不同，RRL-SG智能体在任何测试用例中都没有发生碰撞。

更具体地说，在有或没有对抗攻击的低密度流量中，OARL自动驾驶智能体的性能与OARL智能体相当，并且在回报方面远远优于D3QN、PPO和SAC智能体。在没有对抗性攻击的正常密度流量中，与D3QN、PPO、SAC和OARL智能体相比，RRL-SG智能体分别获得了大约

22.31%、7.22%、10.34%和1.97%的改进。在没有对抗性攻击的高密度流量流中，与D3QN、PPO、SAC和OARL智能体相比，RRL-SG智能体的回报分别提高了大约78.63%、47.41%、25.45%和13.84%。此外，在具有对抗性攻击的高密度流量中，与D3QN、PPO、SAC和OARL智能体相比，RRL-SG智能体的回报率分别提高了约7669.57%、2666.25%、511.57%和8.99%。

图5显示了D3QN、PPO、SAC、OARL和RRL-SG自动驾驶智能体在不同密度和攻击情况下的随机动态交通流量中的性能。如图5所示，基于经过训练的对手模型的对抗性攻击明显影响了基线智能体驾驶的自动驾驶车辆的综合性能、行驶效率和安全性。例如，在正常密度的交通流中，与没有对抗性攻击的情况相比，受攻击的D3QN、PPO、SAC和OARL自动驾驶智能体的碰撞次数分别增加了大约358.82%、583.33%、1378.57%和5.71%。相比之下，所提出的RRL-SG自动驾驶智能体在所有测试用例中表现一致，没有记录任何碰撞事故。

在这里，我们通过计算每种方法在所有测试场景（包括不同的交通密度和攻击场景）中的回报率的均方差，实证评估了对环境动态摄动的策略鲁棒性。根据表1，在所有测试案例中D3QN、PPO、SAC、OARL和RRL-SG智能体的回报率均方差分别为31.23、16.11、39.60、12.73和7.50，表明RRL-SG智能体受环境变化的影响最小。换句话说，RRL-SG策略是鲁棒的、安全的，并表现出稳定

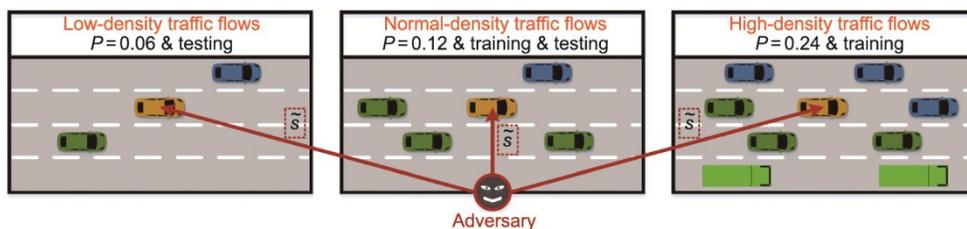


图3. 基于具有随机动态交通流和对抗性攻击的高速公路场景的评价方案。 δ : 受到对手模型干扰的状态。

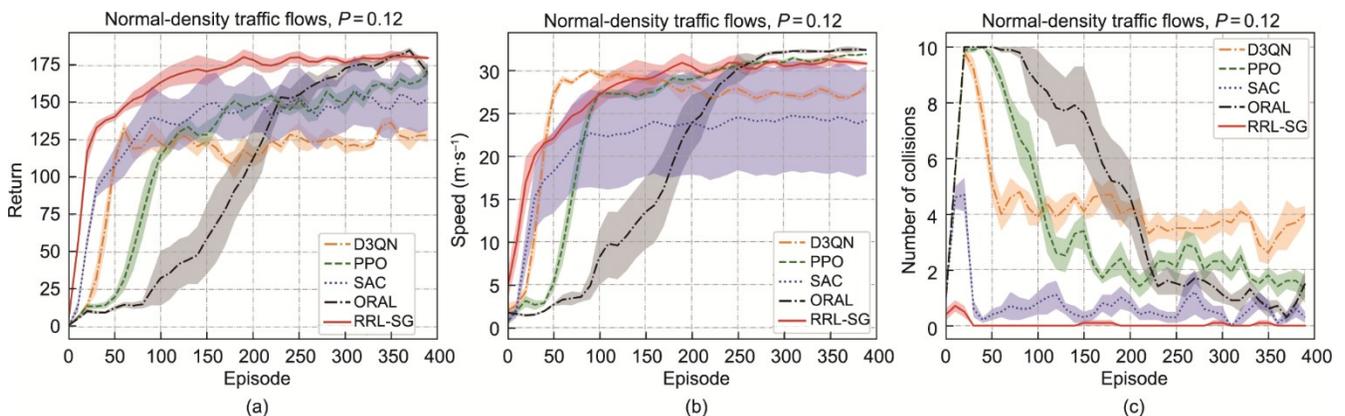


图4. 正态密度随机动态交通流中自动驾驶智能体的学习曲线。(a) 回报率；(b) 速度；(c) 碰撞次数。

表1 不同密度和攻击情况下公路随机动态交通流下自动驾驶智能体的统计结果

Method	Metric	Low-density traffic flows		Normal-density traffic flows		High-density traffic flows	
		Without attacks	With attacks	Without attacks	With attacks	Without attacks	With attacks
D3QN	Return	151.64 ± 25.67	34.89 ± 42.83	148.72 ± 32.69	17.21 ± 32.97	100.82 ± 38.53	2.30 ± 14.71
	Speed	27.78 ± 4.95	16.33 ± 9.17	27.08 ± 6.15	13.09 ± 8.04	21.82 ± 5.90	10.08 ± 6.04
	Robustness	—	0.17 ± 0.07	—	0.21 ± 0.07	—	0.22 ± 0.07
	Number of collisions	0.68 ± 1.17	3.70 ± 3.29	1.02 ± 1.45	4.68 ± 3.27	2.84 ± 2.28	6.24 ± 2.51
PPO	Return	178.99 ± 11.10	59.91 ± 21.00	169.65 ± 16.15	31.24 ± 16.70	122.17 ± 27.04	6.46 ± 4.64
	Speed	32.65 ± 0.46	23.06 ± 2.76	32.24 ± 0.56	17.45 ± 4.93	30.10 ± 1.29	8.01 ± 4.43
	Robustness	—	0.21 ± 0.03	—	0.23 ± 0.03	—	0.25 ± 0.03
	Number of collisions	0.88 ± 0.89	7.86 ± 1.22	1.32 ± 1.14	9.02 ± 1.10	4.24 ± 1.96	9.92 ± 0.27
SAC	Return	174.41 ± 23.44	128.64 ± 66.03	164.85 ± 23.63	81.20 ± 58.03	143.55 ± 22.25	29.22 ± 44.26
	Speed	30.56 ± 3.51	26.40 ± 7.61	29.09 ± 3.75	21.61 ± 9.28	25.83 ± 3.73	9.97 ± 7.87
	Robustness	—	0.21 ± 0.21	—	0.26 ± 0.22	—	0.41 ± 0.23
	Number of collisions	0.06 ± 0.24	1.30 ± 2.02	0.28 ± 0.57	4.14 ± 3.35	0.72 ± 1.11	6.42 ± 3.90
OARL	Return	190.53 ± 1.88	187.06 ± 5.44	178.38 ± 10.33	179.49 ± 14.28	158.20 ± 21.91	163.96 ± 22.56
	Speed	33.04 ± 0.26	32.83 ± 0.39	32.30 ± 0.73	32.72 ± 0.42	31.65 ± 0.95	31.98 ± 0.76
	Robustness	—	$(5.06 ± 3.16) × 10^{-4}$	—	$(7.05 ± 4.27) × 10^{-4}$	—	$(1.29 ± 0.73) × 10^{-3}$
	Number of collisions	0.02 ± 0.14	0.22 ± 0.41	0.70 ± 0.78	0.74 ± 1.00	2.18 ± 1.65	1.58 ± 1.44
RRL-SG	Return	189.91 ± 1.66	185.98 ± 8.38	181.90 ± 5.03	175.27 ± 17.30	180.09 ± 5.07	178.70 ± 7.56
	Speed	32.88 ± 0.36	32.02 ± 1.82	31.23 ± 1.07	29.87 ± 3.78	30.90 ± 1.08	30.59 ± 1.64
	Robustness	—	$(3.91 ± 6.83) × 10^{-13}$	—	$(3.94 ± 10.92) × 10^{-12}$	—	$(1.96 ± 3.69) × 10^{-12}$
	Number of collisions	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0

Bold numbers indicate the best values for each metric.

性，因此突出了本研究对实现自动驾驶汽车的可信决策的主要贡献。

4.3.2. 匝道汇入场景

为了进一步评估自动驾驶智能体决策的可信度，我们增加了入口匝道合并作为一个额外的测试场景。不适当的合并行为会导致典型的结果，包括拥塞、碰撞和行驶时间增加。

基于入口匝道合并场景的评估方案如图6(a)所示。为了测试的目的，我们直接将在高速公路场景中训练的模型部署到入口匝道合并场景中进行测试。所有自动驾驶智能体在不同攻击情况下的高密度 ($P=0.24$) 随机动态交通流量中进行了总共100种情形的评估。与高速公路场景类

似，每个模型评估计算了超过10个测试集的平均指标，每个测试集最多有200个时间步长。

如图6(b)和(c)所示，无论是否有对抗性攻击，RRL-SG自动驾驶智能体在行驶效率和合并成功率方面的性能都显著优于基线。

表2给出了在入口匝道合并场景中模型评估结果的平均指标。粗体数字表示每一列中最好的数字。例如，在没有对抗性攻击的情况下，与D3QN、PPO、SAC和OARL智能体相比，RRL-SG智能体在回报率方面分别获得了大约16.97%、21.91%、29.63%和21.18%的改进。在没有对抗性攻击的情况下，与D3QN、PPO、SAC和OARL相比，RRL-SG智能体的速度分别提高了约31.84%、

表2 不同攻击情况下随机动态流量入口匝道合并场景中自动驾驶智能体的统计结果

Method	Return		Speed		Robustness		Merging success rate	
	Without attacks	With attacks	Without attacks	With attacks	Without attacks	With attacks	Without attacks	With attacks
D3QN	128.02 ± 12.12	97.15 ± 27.49	18.28 ± 3.22	9.55 ± 7.81	—	0.39 ± 0.07	0.98 ± 0.04	0.87 ± 0.17
PPO	122.83 ± 11.73	75.42 ± 2.57	16.90 ± 3.20	3.36 ± 0.81	—	0.19 ± 0.02	0.98 ± 0.04	0.91 ± 0.10
SAC	115.51 ± 23.66	97.95 ± 17.35	14.82 ± 6.67	9.96 ± 4.89	—	0.28 ± 0.16	0.96 ± 0.08	0.94 ± 0.08
OARL	123.57 ± 6.40	123.11 ± 6.77	17.13 ± 1.81	17.02 ± 1.91	—	$(0.89 ± 1.15) × 10^{-3}$	1.00 ± 0.02	0.99 ± 0.03
RRL-SG	149.74 ± 3.93	149.61 ± 4.15	24.10 ± 1.02	24.04 ± 1.10	—	$(0.87 ± 1.11) × 10^{-8}$	1.00 ± 0.00	1.00 ± 0.00

Bold numbers represent the best in each column.

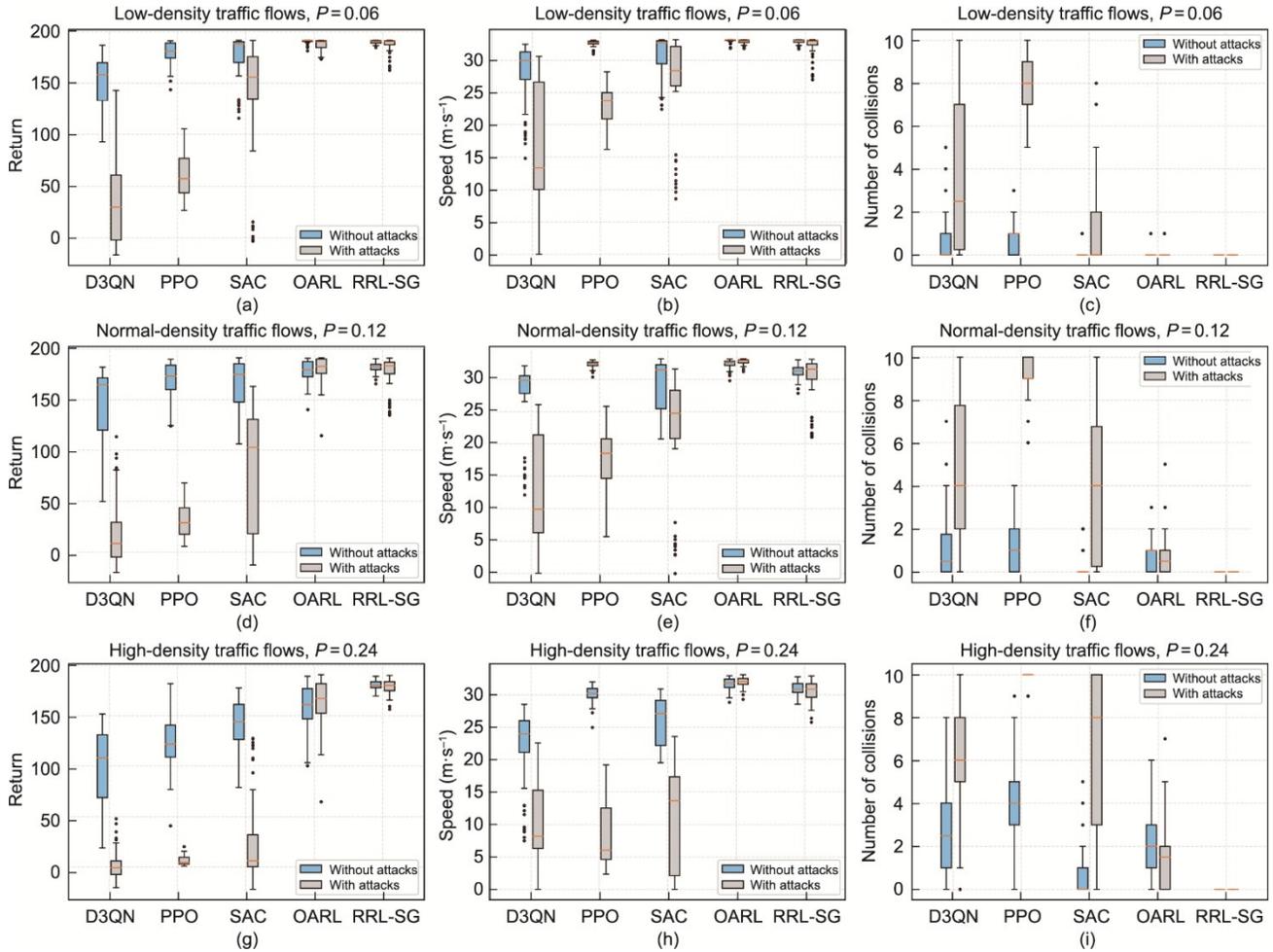


图5. 在不同交通密度和攻击情况下，公路上自动驾驶智能体的性能。(a) ~ (i) 在不同攻击情况下，自动驾驶智能体在低密度、正常密度和高密度随机动态交通流中的回报率、速度和碰撞次数。

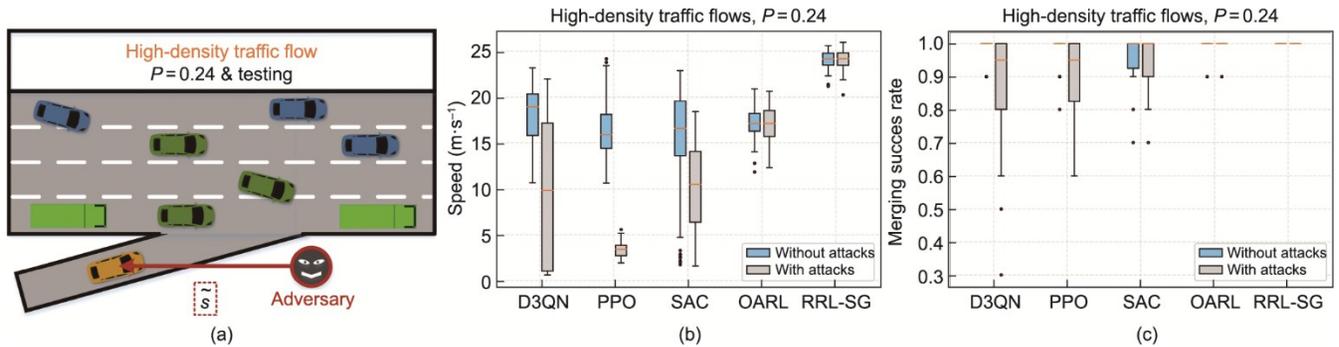


图6. 自动驾驶智能体在不同攻击情况下具有随机动态流量的入口匝道合并场景下的性能。(a) 所采用的入口匝道合并方案示意图。在不同攻击情况下的高密度随机动态交通流中，自动驾驶智能体的速度 (b) 和合并成功率 (c)。

42.60%、62.62%和40.69%。如表2所示，RRL-SG策略的鲁棒性明显优于基线策略。

如图6 (c) 和表2所示，无论是否存在对抗性攻击，我们的RRL-SG智能体都可以以100.00%的概率完成入口匝道合并任务。换句话说，对观测的对抗性攻击对RRL-SG策略几乎没有影响。此外，与D3QN、PPO、SAC和OARL智能体相比，在对抗性攻击中，RRL-SG智能体分

别获得了约13.00%，9.00%，6.00%和1.00%合并成功率的提高。

与入口匝道合并场景相关的环境动态与高速公路情景明显不同。由于我们利用了随机动态交通流，环境动态受到连续变化的影响。在这里，我们使用不同攻击情况下，每个智能体的回报率均方差来评估策略对环境动态摄动的鲁棒性。根据表2，D3QN、PPO、SAC、OARL和RRL-

SG 智能体在不同攻击条件下的回报率均方差分别为 19.81、7.15、20.51、6.59 和 4.04，说明 RRL-SG 智能体对环境变化表现出优越的策略鲁棒性，使其与基线相比最不容易受到环境变化的影响。这些结果突出了我们对自动驾驶汽车可信决策的关键贡献。

4.4. 用真实的自动驾驶汽车进行实验

我们使用一种真实的低速自动驾驶车辆 Hunter (AgileX Robotics, 中国) 进行了物理平台实验，以进一步验证该方法的可信度。如图 7 (a) 所示，Hunter 配备了 16 线光探测和测距 (LiDAR)、2 台立体摄像机、8 个超声波传感器和 1 个 “Jetson Xavier NX 16GB” 边缘计算系统 (NVIDIA, 美国)。因此，RL 策略模型可以根据车载传感器的感知状态实时生成决策命令，所有计算都在 NVIDIA Jetson 平台上执行。所有在 SUMO 模拟器中训练的模型都直接部署在 Hunter 中，并在自由空间为 $8\text{ m} \times 8\text{ m}$ 的实验室环境中进行测试。这里只对训练过的策略模型进行了测试，而没有进行进一步的训练 (即模型参数是固定的)。策略模型需要大约 0.002 s 来执行一个推断。Hunter 的采样频率为 30 Hz。由于评估的策略模型在接收到一组采样状态后执行决策，因此 Hunter 的决策频率为 30 Hz。

图 7 (b) 和 (c) 举例说明了实验方案。与模拟器中的模型测试类似，我们使用 5 个不同的随机种子实例化由每个算法训练的 5 个最终策略模型，并在不同的条件下 (有和没有对抗性攻击) 对每个模型进行评估。在图 7 (b) 所示的实验案例中，Hunter 的感知信息仅包含原始的环境观测结果，没有任何对抗性的观测扰动。相比之下，如图 7 (c) 所示，Hunter 感知的驾驶环境信息包含原始环境观测和训练后的对抗模型产生的对抗扰动。此外，从模拟环境到现实世界的物理平台，环境动态发生了显著的

变化。

实验空间没有静态或动态的障碍，这意味着 Hunter 应该能够在不受对手模型攻击的情况下保持直线运行。我们在 Hunter 从一边开车到另一边的时期 (150 个时间步长) 中评估了每个策略模型。在具有对抗性攻击的测试用例中，攻击开始于第 75 个时间步。Hunter 可以执行五种决策行为：右转、左转、保持当前状态、加速和减速。

图 8 显示了不同智能体在不同攻击情况下驾驶的全局运动轨迹，其中所有策略模型都使 Hunter 能够在没有对抗性攻击的情况下继续直线运行。然而，在具有对抗性攻击的测试用例中，基线模型的性能受到了不同程度的影响。具体来说，所有的 D3QN 自动驾驶智能体、五分之四的 PPO 智能体、所有的 SAC 智能体和五分之一的 OARL 智能体在对抗性攻击下做出转向决策。相比之下，提出的 5 种 RRL-SG 策略模型在所有情况下都表现一致，例如，RRL-SG 驱动的 Hunter 即使受到对手模型的攻击，也可以保持直线运行。更多有关视觉结果可参考附录 A 中的视频 S1。

为了说明对抗性攻击对策略模型的影响，图 9 显示了基于 D3QN 和 RRL-SG 策略在遇到对抗扰动前后的行为概率分布。我们利用 softmax 函数将 D3QN 策略模型的输出转换为每个行为的 Q 值，即每个行为的概率分布。基于 RRL-SG 策略的行为分布与基于 D3QN 策略的行为分布相比几乎没有任何变化。具体来说，在没有对抗性攻击的情况下，基于 D3QN 策略的 5 个决策行为的概率分别约为 12.77%、19.55%、20.61%、34.45% 和 12.63%。在对抗性攻击下，基于 D3QN 策略的 5 个决策行为的概率分别约为 38.14%、15.30%、17.95%、15.40% 和 13.21%，从而解释了为什么对抗性攻击会导致 D3QN 驱动的 Hunter 突然直线转弯继续运行。此外，在没有对抗扰动的情况下，关于 RRL-SG 策略的 5 个决策行为的概率约为 $3.97 \times 10^{-13}\%$ 、

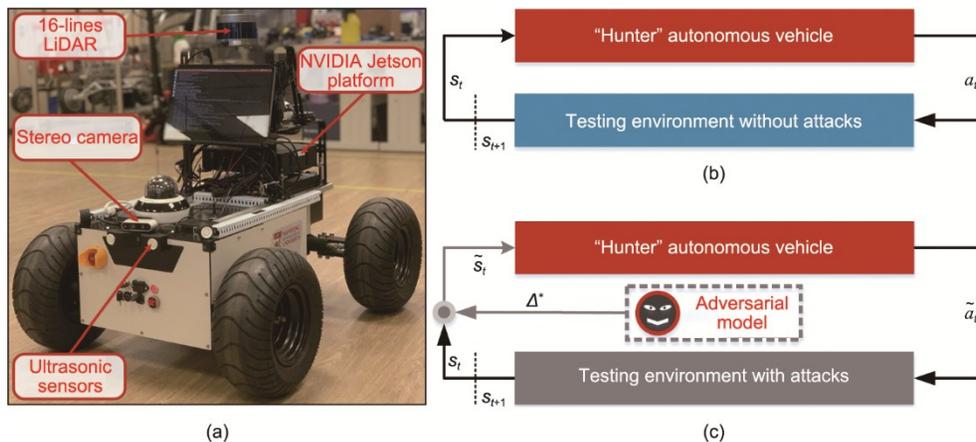


图 7. 真实物理系统的实验设置。(a) “Hunter” 自动驾驶汽车用于实验验证；(b) 无对抗性攻击的实验方案说明；(c) 具有对抗性攻击的实验方案的说明。LiDAR：光探测和测距。

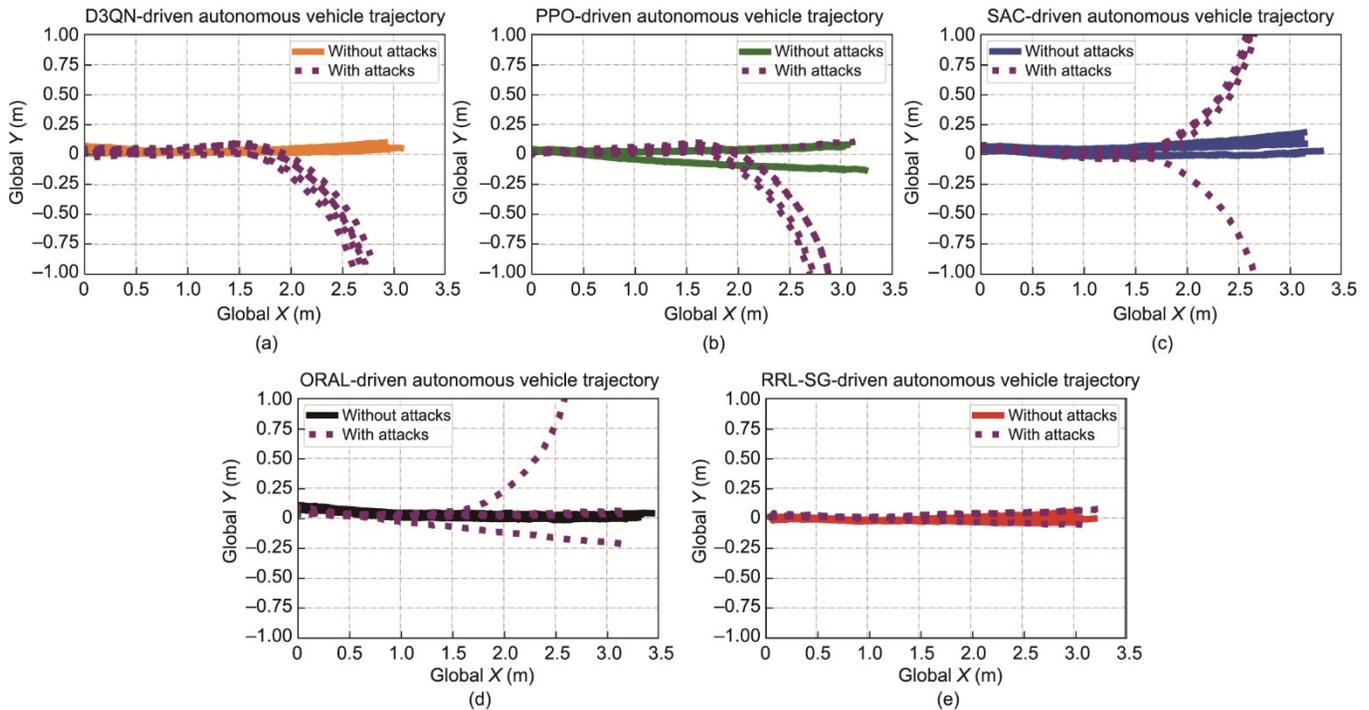


图8. 不同智能体驾驶的自动驾驶车辆在不同攻击情况下的全局运动轨迹。不同攻击情况下基于D3QN (a)、PPO (b)、SAC (c)、OARL (d)和RRL-SG (e) 智能体的自动驾驶车辆的全局运动轨迹。

$1.74 \times 10^{-13}\%$ 、 $2.48 \times 10^{-7}\%$ 、 100% 和 $5.57 \times 10^{-13}\%$ 。在对抗摄动下，关于RRL-SG策略的5个决策行为的概率分别近似为 $6.83 \times 10^{-8}\%$ 、 $3.74 \times 10^{-8}\%$ 、 $2.09 \times 10^{-7}\%$ 、 100% 和 $6.45 \times 10^{-8}\%$ 。因此，基于RRL-SG策略的Hunter可以保持其运行状态，而不受对抗摄动的影响。

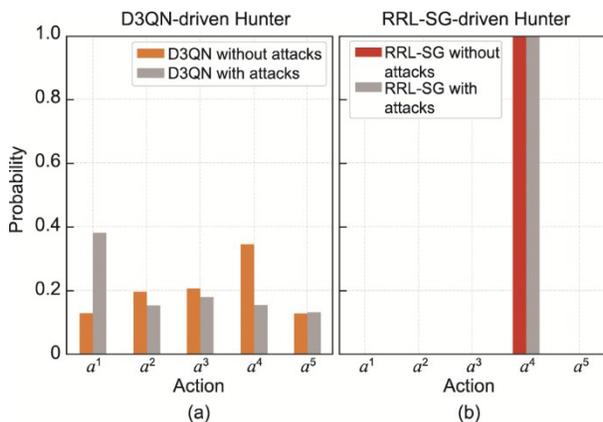


图9. 不同攻击情况下基于不同自动驾驶策略的行为概率分布。在有和没有对抗性攻击的情况下，D3QN (a)和RRL-SG (b)策略下的行动概率分布。 a_1 、 a_2 、 a_3 、 a_4 和 a_5 分别表示向右转、向左转、保持当前状态、加速和减速。

5. 结论

在本研究中，我们提出了RRL-SG技术以保证自动驾驶汽车做出可信的决策。所提出的方法旨在确保策略鲁棒

性和碰撞安全性两方面的可信性。具体地说，通过逼近针对观测状态和环境动态的最优对抗摄动，可以在线训练对手模型，以模拟最坏情况下的不确定性。同时，我们提出了一种对抗鲁棒演员-评论家算法，以促使智能体学习应对对手模拟的多重不确定性鲁棒策略。此外，一个基于可解释知识模型RSS的安全掩码被设计以保证自动驾驶智能体在训练和测试过程中的碰撞安全性。

基于随机动态交通流的仿真和实际自动驾驶车辆的实验结果表明，所提出的RRL-SG方案使自动驾驶智能体能够学习可信策略来应对环境不确定性。此外，与四个基线相比，RRL-SG自动驾驶智能体在鲁棒性和安全性方面表现优越。值得注意的是，在仿真和实验中，我们的RRL-SG方法始终提供了比基线更稳定的性能。

尽管我们展示了所提出方法的潜力，但仍存在一个局限性。虽然RRL-SG解决方案利用了最坏情况下的设置和可解释的知识模型，但为自动驾驶模型的鲁棒性和安全性提供理论保证仍是未来研究的一个关键课题。因此，未来我们将研究可验证和可解释的决策技术，以进一步提高自动驾驶系统的可信性。

Acknowledgements

This work was supported in part by the Start-Up Grant-

Nanyang Assistant Professorship Grant of Nanyang Technological University, the Agency for Science, Technology and Research(A*STAR) under Advanced Manufacturing and Engineering (AME)Young Individual Research under Grant (A2084c0156), the MTC Individual Research Grant (M22K2c0079), the ANR-NRF Joint Grant (NRF2021-NRF-ANR003 HM Science), and the Ministry of Education (MOE) under the Tier 2 Grant (MOE-T2EP50222-0002).

Compliance with ethics guidelines

Xiangkun He, Wenhui Huang, and Chen Lv declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary material

Supplementary material to this article can be found online at <https://doi.org/10.1016/j.eng.2023.10.005>.

References

- [1] Yang B, Cao X, Xiong K, Yuen C, Guan YL, Leng S, et al. Edge intelligence for autonomous driving in 6G wireless system: design challenges and solutions. *IEEE Wireless Commun* 2021;28(2):40–7.
- [2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017 Dec 4–9; Long Beach, CA, USA. New York City: Curran Associates Inc.; 2017. p. 6000–10.
- [3] Wang J, Huang H, Li K, Li J. Towards the unified principles for level 5 autonomous vehicles. *Engineering* 2021;7(9):1313–25.
- [4] Mollah MB, Zhao J, Niyato D, Guan YL, Yuen C, Sun S, et al. Blockchain for the internet of vehicles towards intelligent transportation systems: a survey. *IEEE Internet Things J* 2021;8(6):4157–85.
- [5] Li J, Shao W, Wang H. Key challenges and Chinese solutions for SOTIF in intelligent connected vehicles. *Engineering* [in press].
- [6] Feng S, Yan X, Sun H, Feng Y, Liu HX. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nat Commun* 2021;12(1):e748.
- [7] Liu J, Luo Y, Zhong Z, Li K, Huang H, Xiong H. A probabilistic architecture of long-term vehicle trajectory prediction for autonomous driving. *Engineering* 2022;19(12):228–39.
- [8] He X, Wu J, Huang Z, Hu Z, Wang J, Sangiovanni-Vincentelli A, et al. Fear-neuroinspired reinforcement learning for safe autonomous driving. *IEEE Trans Pattern Anal Mach Intell* 2023 Oct;:1–13.
- [9] Yuan K, Huang Y, Yang S, Zhou Z, Wang Y, Cao D, et al. Evolutionary decisionmaking and planning for autonomous driving based on safe and rational exploration and exploitation. *Engineering* 2023;30(11):1313–25.
- [10] Huang W, Zhou Y, He X, Lv C. Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation. *IEEE Trans Intell Transp Syst* 2023 Sep;:1–14.
- [11] Zhang Y, Li C, Luan TH, Yuen C, Fu Y. Collaborative driving: learning-aided joint topology formulation and beamforming. *IEEE Veh Technol Mag* 2022; 17(2):103–11.
- [12] Wu J, Huang Z, Hu Z, Lv C. Toward human-in-the-loop AI: enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering* 2023;21(2):75–91.
- [13] Wang H, Khajepour A, Cao D, Liu T. Ethical decision making in autonomous vehicles: challenges and research progress. *IEEE Intell Transp Syst Mag* 2022; 14(1):6–17.
- [14] He X, Lv C. Toward personalized decision making for autonomous vehicles: a constrained multi-objective reinforcement learning technique. *Transp Res Part C Emerging Technol* 2023;156:104352.
- [15] Tang X, Yang K, Wang H, Wu J, Qin Y, Yu W, et al. Prediction-uncertainty-aware decision-making for autonomous vehicles. *IEEE Trans Intell Veh* 2022; 7(4):849–62.
- [16] Liu J, Wang H, Peng L, Cao Z, Yang D, Li J. PNUAD: perception neural networks uncertainty aware decision-making for autonomous vehicle. *IEEE Trans Intell Transp Syst* 2022;23(12):24355–68.
- [17] Li G, Qiu Y, Yang Y, Li Z, Li S, Chu W, et al. Lane change strategies for autonomous vehicles: a deep reinforcement learning approach based on transformer. *IEEE Trans Intell Veh* 2023;8(3):2197–211.
- [18] Urmson C, Anhalt J, Bagnell D, Baker C, Bittner R, Clark MN, et al. Autonomous driving in urban environments: boss and the urban challenge. *J Field Rob* 2008;25(8):425–66.
- [19] Montemerlo M, Becker J, Bhat S, Dahlkamp H, Dolgov D, Ettinger S, et al. Junior: the Stanford entry in the urban challenge. *J Field Rob* 2008; 25(9): 569–97.
- [20] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature* 2015; 518(7540):529–33.
- [21] Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 2019;575(7782):350–4.
- [22] He X, Chen H, Lv C. Robust multiagent reinforcement learning toward coordinated decision-making of automated vehicles. *SAE Int J Veh Dyn Stab NVH* 2023;7(4):2023.
- [23] Hieu NQ, Hoang DT, Niyato D, Wang P, Kim DI, Yuen C. Transferable deep reinforcement learning framework for autonomous vehicles with joint radardata communications. *IEEE Trans Commun* 2022;70(8):5164–80.
- [24] Duan J, Li SE, Guan Y, Sun Q, Cheng B. Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data. *IET Intell Transp Syst* 2020;14(5):297–305.
- [25] Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, et al. Deep reinforcement learning for autonomous driving: a survey. *IEEE Trans Intell Transp Syst* 2022;23(6):4909–26.
- [26] Ye F, Wang P, Chan CY, Zhang J. Meta reinforcement learning-based lane change strategy for autonomous vehicles. In: *Proceedings of 2021 IEEE Intelligent Vehicles Symposium IV*; 2021 Jul 11–17; Nagoya, Japan. Piscataway: IEEE; 2021. p. 223–30.
- [27] Wang G, Hu J, Li Z, Li L. Harmonious lane changing via deep reinforcement learning. *IEEE Trans Intell Transp Syst* 2022;23(5):4642–50.
- [28] Li G, Yang Y, Li S, Qu X, Lyu N, Li SE. Decision making of autonomous vehicles in lane change scenarios: deep reinforcement learning approaches with risk awareness. *Transp Res Part C* 2022;134:e103452.
- [29] Mirchevska B, Pek C, Werling M, Althoff M, Boedecker J. High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning. In: *Proceedings of 2018 21st International Conference on Intelligent Transportation Systems*; 2018 Nov 4–7; Maui, HI, USA. Piscataway: IEEE; 2018. p. 2156–62.
- [30] Lubars J, Gupta H, Chinchali S, Li L, Raja A, Srikant R, et al. Combining reinforcement learning with model predictive control for on-ramp merging. In: *Proceedings of 2021 IEEE International Intelligent Transportation Systems Conference*; 2021 Sep 19–22; Indianapolis, IN, USA. Piscataway: IEEE; 2021. p. 942–7.
- [31] Wang H, Gao H, Yuan S, Zhao H, Wang K, Wang X, et al. Interpretable decision-making for autonomous vehicles at highway on-ramps with latent space reinforcement learning. *IEEE Trans Veh Technol* 2021;70(9):8707–19.
- [32] Bouton M, Nakhaei A, Fujimura K, Kochenderfer MJ. Cooperation-aware reinforcement learning for merging in dense traffic. In: *Proceedings of 2019 IEEE Intelligent Transportation Systems Conference*; 2019 Oct 27 – 30; Auckland, New Zealand. Piscataway: IEEE; 2019. p. 3441–7.
- [33] Qiao Z, Tyree Z, Mudalige P, Schneider J, Dolan JM. Hierarchical reinforcement learning method for autonomous vehicle behavior planning. In: *Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*; 2020 Oct 24–2021 Jan 24; Las Vegas, NV, USA. Piscataway: IEEE; 2021. p. 6084–9.

- [34] He X, Lou B, Yang H, Lv C. Robust decision making for autonomous vehicles at highway on-ramps: a constrained adversarial reinforcement learning approach. *IEEE Trans Intell Transp Syst* 2023;24(4):4103–13.
- [35] Hoel CJ, Driggs-Campbell K, Wolff K, Laine L, Kochenderfer MJ. Combining planning and deep reinforcement learning in tactical decision making for autonomous driving. *IEEE Trans Intell Veh* 2020;5(2):294–305.
- [36] Zhang Y, Gao B, Guo L, Guo H, Chen H. Adaptive decision-making for automated vehicles under roundabout scenarios using optimization embedded reinforcement learning. *IEEE Trans Neural Networks Learn Syst* 2021;32(12):5526–38.
- [37] He X, Lv C. Toward intelligent connected e-mobility: energy-aware cooperative driving with deep multiagent reinforcement learning. *IEEE Veh Technol Mag* 2023;18(3):101–9.
- [38] Nagesh Rao S, Tseng HE, Filev D. Autonomous highway driving using deep reinforcement learning. In: *Proceedings of 2019 IEEE International Conference on Systems, Man and Cybernetics*; 2019 Oct 6–9; Bari, Italy. Piscataway: IEEE; 2019. p. 2326–31.
- [39] Gangopadhyay B, Soora H, Dasgupta P. Hierarchical program-triggered reinforcement learning agents for automated driving. *IEEE Trans Intell Transp Syst* 2022;23(8):10902–11.
- [40] Cao Z, Xu S, Jiao X, Peng H, Yang D. Trustworthy safety improvement for autonomous driving using reinforcement learning. *Transp Res Part C* 2022;138:103656.
- [41] Shalev-Shwartz S, Shammah S, Shashua A. On a formal model of safe and scalable self-driving cars. 2017. arXiv:1708.06374.
- [42] Shalev-Shwartz S, Shammah S, Shashua A. Vision zero: can roadway accidents be eliminated without compromising traffic throughput? 2018. arXiv:1901.05022.
- [43] Lopez PA, Behrisch M, Bieker-Walz L, Erdmann J, Flötteröd YP, Hilbrich R, et al. Microscopic traffic simulation using SUMO. In: *Proceedings of 2018 21st International Conference on Intelligent Transportation Systems*; 2018 Nov 4–7; Maui, HI, USA. Piscataway: IEEE; 2018. p. 2575–82.
- [44] Lin J. Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 1991;37(1):145–51.
- [45] Huszár F. How (not) to train your generative model: scheduled sampling, likelihood, adversary? 2015. arXiv:1511.05101.
- [46] Huang W, Zhang C, Wu J, He X, Zhang J, Lv C. Sampling efficient deep reinforcement learning through preference-guided stochastic exploration. *IEEE Trans Neural Networks Learn Syst* 2023 Oct:1–12.
- [47] Hoffman AJ, Karp RM. On nonterminating stochastic games. *Manage Sci* 1966; 12(5):359–70.
- [48] Hansen TD, Miltersen PB, Zwick U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J ACM* 2013;60(1):1–16.
- [49] Mazalov V. *Mathematical game theory and applications*. Chichester: John Wiley & Sons Ltd; 2014.
- [50] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: *Proceedings of the 35th International Conference on Machine Learning*; 2018. p. 1861–70.
- [51] Bae I, Moon J, Jung J, Suk H, Kim T, Park H, et al. Self-driving like a human driver instead of a robocar: personalized comfortable driving experience for autonomous vehicles. 2020. arXiv:2001.03908.
- [52] Wang Z, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N. Dueling network architectures for deep reinforcement learning. In: Balcan MF, Weinberger KQ, editors. *ICML '16: Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*; 2016 Jun 19–24; YorkNew, NY, USA. JMLR.org; 2016. p. 1995–2003.
- [53] Hessel M, Modayil J, van Hasselt H, Schaul T, Ostrovski G, Dabney W, et al. Rainbow: combining improvements in deep reinforcement learning. In: McIlraith SA, Weinberger KQ, editors. *AAAI' 18/IAAI' 18/EAAI' 18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*; 2018 Feb 2–7; New Orleans, LA, USA. Palo Alto: AAAI Press; 2018. p. 3215–22.
- [54] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017. arXiv:1707.06347.
- [55] Christodoulou P. Soft actor-critic for discrete action settings. 2019. arXiv:1910.07207.
- [56] He X, Yang H, Hu Z, Lv C. Robust lane change decision making for autonomous vehicles: an observation adversarial reinforcement learning approach. *IEEE Trans Intell Veh* 2023;8(1):184–93.