Computer Science and Application 计算机科学与应用, 2025, 15(9), 63-72
Published Online September 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.159224

Hans汉斯

基于启发式强化学习自动驾驶的多目标协同速度控制研究

李　琦

上海电科智能系统股份有限公司，上海

收稿日期：2025年7月25日；录用日期：2025年8月23日；发布日期：2025年9月1日

摘　要

近年来，自动驾驶技术快速发展，但速度控制的安全性、效率与舒适性仍是关键挑战。本文提出一种基于深度强化学习(Deep Deterministic Policy Gradient, DDPG)的车辆跟驰控制模型。通过IDM模型约束强化学习输出，融合安全、效率和舒适性的多目标奖励函数，利用下一代仿真(NGSIM)数据集中提取的1341个跟驰事件对模型进行了训练和测试。将所提出的模型与没加约束的DDPG算法进行比较，以评估所提出的模型的性能。结果表明，该方法有助于开发更好的自动驾驶系统，具有一定的实用价值，能够为自主驾驶系统的开发提供参考。

关键词

车辆跟驰，速度控制，强化学习，启发式

Heuristic Reinforcement Learning-Based Multi-Objective Speed Coordination for Autonomous Vehicles

Qi Li

Shanghai SEARI Intelligent System Co., Ltd., Shanghai

Received: Jul. 25th, 2025; accepted: Aug. 23rd, 2025; published: Sep. 1st, 2025

Abstract

In recent years, the rapid development of autonomous driving technology has continued to present key challenges in achieving safe, efficient, and comfortable speed control. This paper proposes a vehicle car-following control model based on the Deep Deterministic Policy Gradient (DDPG) algorithm.

文章引用: 李琦. 基于启发式强化学习自动驾驶的多目标协同速度控制研究[J]. 计算机科学与应用, 2025, 15(9): 63-72.
DOI: 10.12677/csa.2025.159224

To enhance performance, the model's output is constrained by the Intelligent Driver Model (IDM), and it utilizes a multi-objective reward function that integrates safety, efficiency, and comfort. The model was trained and tested on 1341 car-following events extracted from the Next Generation Simulation (NGSIM) dataset. A comparative analysis was conducted against an unconstrained DDPG algorithm to evaluate the proposed model's performance. The results demonstrate that this method contributes to the development of more effective autonomous driving systems, holds significant practical value, and can serve as a reference for future autonomous system design.

Keywords

Car Following, Velocity Control, Reinforcement Learning, Heuristic

1. 引言

跟车是日常驾驶中最常见的场景之一，其主要任务是控制车速，以确保与前车保持安全且舒适的车距。自动跟车速度控制有望减轻驾驶员的负担，提升交通安全性，同时也能增加道路的通行效率[1]。驾驶员模型是速度控制系统的关键要素[2]。近年来，研究人员提出了各种车辆跟驰模型，一般可分为两类。第一类是传统的基于规则的模型，如 Gipps 模型[3]、智能驾驶员模型(IDM) [4]。[4]中模型能够模拟各种拥堵交通状态，并通过调整相关参数来呈现不同状态之间的转换过程，进一步解释这些状态在不同交通流量和道路条件下的表现。它可以预测道路瓶颈(如匝道、坡道、车道减少、交通事故等)对交通流的影响。研究表明，瓶颈的存在通常会导致上游发生拥堵，而拥堵的类型和严重程度则取决于瓶颈的特性和交通流量的大小。通过分析不同交通状态的形成条件及其演变规律，智能驾驶员模型(IDM)可以为交通控制策略的制定提供理论支持。该模型考虑了车辆之间的相对速度，因此能够模拟避免碰撞的驾驶行为，其加速和减速规律更符合实际驾驶情况，避免了其他模型中可能出现的极端加减速现象。此外，IDM 模型的参数具有明确的物理意义，并可通过实地数据进行校准，计算效率较高，适合大规模交通仿真应用。

基于规则的模型通常依赖于手工设定的规则和标准来模拟人类驾驶员的驾驶行为。这些规则通常是根据驾驶员的经验和行为模式进行设计的，因此具有较高的可解释性。然而，由于这些模型依赖于手工制定的标准和规则，基于规则的模型往往缺乏足够的灵活性和泛化能力[5]。在现实交通环境中，复杂和不规则的场景是不可避免的，因此，单纯依靠这些模型无法满足自动驾驶汽车在多变道路条件下的需求。

为了解决基于规则的模型存在的局限性，研究人员将重点转向了第二类模型，即基于学习的模型。这些模型通过借助机器学习的方法，尤其是深度神经网络，来提升泛化能力，以应对更复杂的驾驶场景。在这种情况下，我们可以根据是否利用人类驾驶员的真实数据，将这些模型进一步划分为基于监督学习(SL)和基于深度强化学习(DRL)两类模型。基于监督学习(SL)的模型通常采用深度神经网络对人类驾驶员的车辆轨迹数据进行回归，旨在学习与人类驾驶员相似的跟车模式[6] [7]。然而，训练基于监督学习(SL)的模型需要大量的人类驾驶轨迹数据，而收集和标注这些数据既昂贵又费时。此外，由于人类驾驶车辆很难收集事故或接近事故情境的真实数据，这对提高基于 SL 模型的性能至关重要[8]。另一方面，模仿人类驾驶员的跟车决策并不是自动驾驶汽车的最优选择，用户可能不希望自动驾驶汽车以他们这样的方式行驶[9]。首先，人类驾驶员在跟车决策中存在许多缺点，例如驾驶员性格差异可能导致过于保守或过于激进的跟车行为[10]。其次，相比于人类驾驶员，自动驾驶汽车配备了更先进的传感器，能够更精准、

更全面地感知周围环境，从而有潜力做出更加高效和合理的跟车决策。

为此，研究人员致力于使用 DRL 方法开发跟驰决策模型[11] [12]，这可以有效地减少基于 SL 模型对人类驾驶员真实数据的依赖。基于 DRL 的模型旨在通过与模拟交通环境的试错交互来学习车辆跟驰决策。借助奖励函数的反馈指导，模型期望能够学习到表现优异的跟驰决策策略。同时，由于 DRL 模型能够通过在模拟环境中进行大量可承受的试错学习来不断优化，因此智能体在学习过程中会遭遇各种潜在风险场景。这促使训练好的模型能够在面对罕见的风险情况下，依然作出更安全的决策[13]。

然而，现有的基于 DRL 的跟驰模型存在以下关键缺陷。深度强化学习模型(DRL)应用于加速度搜索时，由于其探索空间过于广泛，可能导致计算资源的浪费以及优化效率的低下。传统的 DRL 方法通过不断试探和调整参数来寻找最优解，但这一过程往往需要大量的训练和反复试验，尤其在加速度调节这种具有复杂约束和多重目标的场景中，探索空间的广泛性会使得学习过程难以迅速收敛。

因此，我们提出了一种基于 DRL 智能体的启发式跟驰模型 DDPG with IDM。通过在 DRL 算法的搜索范围内引入启发式搜索策略，我们能够在合理的搜索空间内进行快速且高效的优化。启发式搜索利用经验法则和简单的规则，引导搜索过程避免无谓的盲目探索，从而在较小的范围内快速找到最优解[14]。相比传统的 DRL 方法，这种启发式搜索不仅减少了计算复杂度，还能有效缩小搜索空间，提高优化精度和收敛速度。总的来说，结合启发式搜索与深度强化学习的方法，不仅保持了 DRL 在处理复杂动态环境中的强大适应性，还通过引导搜索过程实现了更加高效的优化。

2. 深度强化学习方法

深度强化学习是指使用神经网络来近似值函数 $V(s;\theta)$，策略 $\pi(a|s;\theta)$，或系统模型的强化学习算法。

2.1. 强化学习

强化学习(RL)通过让 RL 智能体与环境交互来优化顺序决策问题。在时间步 t，智能体观察状态，并基于从状态 s_t 映射到动作 a_t 的策略 $\pi(a_t|s_t)$ 从某个动作空间 A 中选择动作 a_t。同时，系统给智能体一个奖励 r_t，并转移到下一个状态 s_{t+1}。此过程将继续，直到达到终端状态，然后智能体将重新启动。智能体打算获得最大折扣，累积奖励 $R_t=\sum_{k=0}^{\infty}\gamma^k r_{t+k}$，折扣因子为 $\gamma\in(0,1]$。一般来说，有两种类型的强化学习方法：基于值的和基于策略的。

2.2. 深度 Q 网络

深度 Q 学习使用神经网络作为函数近似器来估计动作－值函数，而不是为每个状态－动作对计算 $Q(s,a)$。选择具有最大 $Q(s,a)$ 值的操作。深度 Q 网络(DQN)在离散的动作空间中工作得很好，但在连续的动作空间中失败了，就像我们的例子一样。为了解决这个问题，Lillicrap 等人[15]开发了一种称为深度确定性策略梯度(DDPG)的算法。DDPG 在 DQN 中引入了一种行动者－批评者机制，可用于连续控制问题。

2.3. 深度确定性政策梯度

DDPG 使用两个独立的网络来分别近似演员和评论家。权值为 θ^Q 的评价网络负责估计行动价值函数 $Q(s,a|\theta^Q)$。权重为 θ^μ 的行动者网络负责显式地表示智能体的策略 $\mu(s|\theta^\mu)$。该算法采用了 DQN 中的经验回放和目标网络技术，提高了学习的稳定性和鲁棒性。

• 经验回放

应用重放缓冲器以避免从顺序生成的相关经验样本中学习。重放缓冲器是存储从环境采样的转换 (s_t,a_t,r_t,s_{t+1}) 的有限大小的高速缓存 D。重放缓冲区通过用新样本替换旧样本而不断更新。在每个时间

步，演员和评论家网络都是在来自重放缓冲区的随机小批量转换上训练的。

• 目标网络

目标网络用于表示主网络的目标值，以避免算法的发散。两个目标网络 $Q'\left(s,a\left|\theta^{Q'}\right.\right)$ 和 $\mu'\left(s\left|\theta^{\mu'}\right.\right)$ 被分别建立为主要的评论者和演员网络。它们具有与主网络相同的体系结构，但具有不同的网络参数。目标网络的参数通过让它们缓慢地跟踪主网络来更新：$:\theta'=\tau\theta+\left(1-\tau\right)\theta'$ with $\tau\ll 1$。这样，目标值被约束为缓慢更新，极大地增强了学习的稳定性。

DDPG 算法首先初始化重放缓冲区和演员，评论家和相应的目标网络。在每个时间步，根据探索性策略采取动作 a。然后，观察奖励 r_t 和新状态 s_{t+1} 并将其存储在重放存储器 D 中。评论家是用从重放存储器中采样的小批量来训练的。之后，通过对采样的策略梯度执行梯度上升步骤来更新动作器。最后，更新具有权重 $\theta^{Q'}$ 和 $\theta^{\mu'}$ 的目标网络，以缓慢跟踪演员和评论家网络。

3. 数据准备

下一代仿真(NGSIM)项目。如图 1 所示，轨迹数据是 2005 年 4 月 13 日在加利福尼亚州埃莫里维尔的弗朗西斯科湾区从 I-80 东行获取的。调查区域长约 500 米(1640 英尺)，由六条高速公路车道组成，其中包括一条高载客率车辆(HOV)车道。在整个数据集中，可访问 45 分钟的数据集合，分为三个 15 分钟的时间段：下午 4:00 至下午 4:15；下午 5 时至 5 时 15 分；以及下午 5:15 到 5:30。这些时段包含拥堵累积，或非拥堵和拥堵交通状态之间的州际，以及高峰时段期间的完全拥堵。该数据提供了每辆车的精确位置信息，采样率为 10 Hz。为了提高数据质量，使用了重建的 NGSIM I-80 数据[16]。

通过应用 Wang 等人[17]所述的车辆跟踪过滤器提取车辆跟踪事件。跟车事件定义为：

• 引导车辆和跟随车辆保持在同一车道上；

• 事件持续时间 ＞15 秒：确保车辆跟随持续足够长的时间以进行分析。

本研究共提取并使用了 1341 个跟车事件。

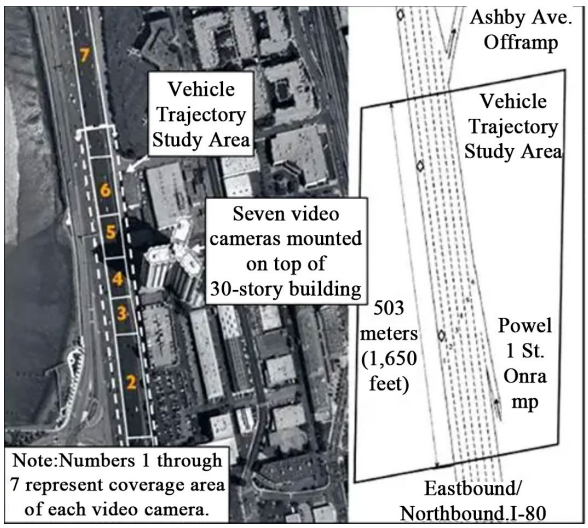

Figure 1. I-80 Aerial photos and schematics of the research area
图 1. I-80 研究区域的航拍照片和示意图

4. 奖励函数特点

在这一节中，提出了捕捉车辆跟驰速度控制的相关目标的特征，最终目的是构建适当的奖励函数。

4.1. 安全

安全性应该是自动跟车的最重要的因素。碰撞时间(TTC)用于表示安全性。TTC 作为一种广泛使用的安全指示器，代表了两辆车相撞前所剩的时间。其计算公式如下：

$$\mathrm{TTC}(t)=-\frac{S_{n-1,n}(t)}{\Delta V_{n-1,n}(t)} \tag{1}$$

其中 t 表示时间；$n-1$ 和 n 分别表示前车和后车；$n-1$，n 组合表示与前车和后车相关的变量：$S_{n-1,n}$ 是间隙距离，$\Delta V_{n-1,n}$ 是相对速度(前车速度 - 后车速度)。

TTC 与碰撞风险成反比。为了将 TTC 应用为反映安全性的特征，应确定安全限值(TTC 的下限)。然而，文献中报告了不同的阈值(从 1.5 s 到 5 s) [18]。在本研究中，我们尝试了从 1 s 到 9 s 的安全限制，发现它们对最终的跟车性能没有太大影响。使用 4 秒的最终限制是因为它导致最佳的整体性能。TTC 功能构建为：

$$F_{\mathrm{TTC}}=\begin{cases}\log\left(\dfrac{\mathrm{TTC}}{4}\right) & 0\le \mathrm{TTC}\le 4\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

这样，如果 TTC 小于 4 s，TTC 特性将为负。当 TTC 接近零时，TTC 特性将接近负无穷大，这代表了对接近撞车情况的严重惩罚。

4.2. 效率

在本研究中，高效驾驶是指保持安全和短时间的车头时距。车头时距定义为前车(LV)和后车(FV)到达指定点之间经过的时间。在安全范围内保持短的车头时距可以提高交通流效率，因为短车头时距对应于大的道路通行能力。

本研究基于经验 NGSIM 数据确定了适当的车头时距。对所有提取的 1341 个跟车事件中的车头时距数据对数正态分布拟合。对数正态分布是对数具有正态分布的概率分布。对数正态分布的概率密度函数为：

$$f_{\mathrm{lognorm}}(x\mid\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\mathrm{e}^{\frac{-(\ln x-\mu)^2}{2\sigma^2}};x>0 \tag{3}$$

其中 x 是分布变量，即本研究中的车头时距，μ 分别是变量 x 的平均值和对数标准差。基于经验数据，估计的 μ 和分别为 0.4226 和 0.4365。

车头时距特征被构建为估计的车头时距对数正态分布的概率密度值：

$$F_{headway}=f_{\mathrm{lognorm}}(headway\mid\mu=0.4226,\sigma=0.4365) \tag{4}$$

根据该车头时距特征，车头时距为 1.26 s 对应于最大车头时距特征值(约 0.65)；而车头时距过长或过短对应于低特征值。通过这种方式，RL 代理被鼓励保持大约 1.26 s 的恒定时间间隔。请注意，为了鼓励恒定的时间间隔，也可以使用正态分布的密度函数，但我们发现拟合的对数正态密度函数恰好优于正态密度函数，这导致模型性能不稳定。

4.3. 舒适性

被定义为加速度的变化率的加加速度被用于测量驾驶舒适度，因为它对乘客的舒适度有很大的影响。加加速度特征构造为：

$$F_{jerk} = -\frac{jerk^2}{3600}, \tag{5}$$

加加速度特征的较小值对应于较不舒适的驾驶。将加加速度的平方除以基值(3600)以将特征缩放到[0, 1]的范围内。基础值由以下直觉确定：

1) 数据的采样间隔为 0.1 s；

2) 基于所有跟车事件的观测 FV 加速度，加速度被限制在−3 到 3 m/s^2 之间；

3) 因此，最大加加速度值为$\frac{3-(-3)}{0.1}$，如果平方，我们得到 3600。

5. 启发式 DDPG 算法

在本节中，解释了使用 DDPG 学习速度控制策略的方法。

5.1. 状态和动作

在一定的时间步长 t 处，车辆跟随过程的状态由 FV 速度 $V_n(t)$，间隙距离 $S_{n-1,n}(t)$，相对速度 $\Delta V_{n-1,n}(t)$。动作是 FV 的纵向加速度 $a_n(t)$。给定时间步 t 的状态和动作，下一步状态由运动学点质量模型更新：

$$V_n(t+1) = V_n(t) + a_n(t) * \Delta T \tag{6}$$

$$\Delta V_{n-1,n}(t+1) = V_{n-1}(t+1) - V_n(t+1) \tag{7}$$

$$S_{n-1,n}(t+1) = S_{n-1,n}(t) + \frac{\Delta V_{n-1,n}(t) + \Delta V_{n-1,n}(t+1)}{2} * \Delta T \tag{8}$$

其中 T 是模拟时间间隔，在本研究中设置为 0.1 s，V_{n-1} 是外部输入的领头车辆(LV)的速度。

5.2. 仿真设置

为了使 RL 智能体能够从试错中学习，实现了一个简单的数值跟驰仿真环境。仿真仅涉及 LV 和 FV 两个智能体，LV 遵循经验数据，FV 由 RL 算法控制。用经验给定的下列车辆速度、间隙距离和速度差进行初始化，$V_n(t=0) = V_n^{data}(t=0)$，$S_{n-1,n}(t=0) = S_{n-1,n}^{data}(t=0)$，和 $\Delta V_{n-1,n}(t=0) = \Delta V_{n-1,n}^{data}(t=0)$，RL 智能体用于计算 FV 的加速度 $a_n(t)$。给定加速度，未来 FV 速度，相对速度和间隙距离，然后根据公式(6) (7) (8)代生成。在每个时间步，仿真环境向 RL 代理提供奖励值(基于车间时距、TTC 和加加速度计算)作为反馈。一旦跟车事件到达其终点，则利用下一事件的经验数据重新初始化状态。事件被随机打乱，以避免顺序的影响。

5.3. 奖励函数

奖励函数 $r(s,a)$ 用作训练信号，以在期望任务的上下文中鼓励或阻止行为。对于自动跟车任务，基于第 4 节中构建的特征的线性组合建立了奖励函数：

$$r = w_1 F_{\mathrm{TTC}} + w_2 F_{headway} + w_3 F_{jerk} \tag{9}$$

其中 w_1， w_2 和 w_3 是特征的系数，在当前研究中均设为 1。

5.4. 动作探测噪声

通过在原 Actor 策略中加入从噪声过程中采样的噪声，构造了一种探测策略。如[15]所建议的，使用 $\theta = 0.15$ 且 $\sigma = 0.2$ 的 Ornstein-Uhlenbeck 过程[19]。Ornstein-Uhlenbeck 过程模拟具有摩擦的布朗粒子的速

度，产生以零为中心的时间相关值。时间相关的噪声使智能体能够在具有动量的物理环境中很好地探索。

5.5. IDM 模型启发式约束

智能驾驶员模型(IDM)考虑了车辆之间的相对速度，因此能够模拟避免碰撞的驾驶行为，其加速和减速规律更符合实际驾驶情况，避免了其他模型中可能出现的极端加减速现象。此外，IDM 模型的参数具有明确的物理意义，并可通过实地数据进行校准，计算效率较高，适合大规模交通仿真应用。我们设计两种 IDM 风格来限定智能体的输出加速度，使智能体输出的加速度限定于这个区间中。

根据数据测试，选出两组风格参数，激进型 IDM 风格：期望速度 = 25 m/s，安全时间间隔 = 1 s，最大加速度 = 3 m/s^2，舒适减速度 = 4.5 m/s^2，加速度指数 = 4，最小间距 = 2 m，最大减速限制 = −9 m/s^2；保守型 IDM 风格：期望速度 = 25 m/s，安全时间间隔 = 3 s，最大加速度 = 1.2 m/s^2，舒适减速度 = 2 m/s^2，加速度指数 = 4，最小间距 = 2 m，最大减速限制 = −9 m/s^2；IDM 模型具体计算公式如下：

$$d^* = d_0 + \tau v + \frac{v\Delta v}{2\sqrt{ab}} \tag{10}$$

$$a_{\text{IDM}} = a\left[1 - \left(\frac{v}{v_0}\right)^{\delta} - \left(\frac{d^*}{d}\right)^2\right] \tag{11}$$

5.6. 训练 IDM 约束的 DDPG 速度控制模型

对于提取的 1341 个跟车事件，70% (938)用于训练，30%用于测试。在训练阶段，RL 代理顺序地模拟训练数据中随机混洗的跟车事件。也就是说，当跟车事件终止时，从 938 个训练事件中随机选择一个新事件，并且用新事件的经验数据初始化智能体的状态。重复训练 3000 次。本研究中的事件是指跟车事件。图 2 展示了是否加了 IDM 约束的 DDPG 算法的碰撞对比，可以看出蓝色实线加了 IDM 约束的模型比没加约束的模型大大避免了碰撞。图 3 示出了滚动平均事件奖励相对于训练事件的变化。平均事件奖励是在跟车事件的所有时间步长(采样间隔 = 0.1 s)上聚合的平均奖励，滚动平均事件奖励是大小为 100 的滚动窗口上的平均事件奖励的平均值。进行多次训练，并汇总结果：蓝色实线表示没加 IDM 约束模型的多次训练的平均值与橙色虚线是加了 IDM 约束的模型进行对比，可以看出，DDPG with IDM 模型在训练集达到约 550 时开始收敛，当模型收敛时，智能体收到的奖励值约为 0.18，对比没加 IDM 约束的模型有更高的奖励值。这是通过以使 TTC 和加加速度特征值接近 0 并获得最大车头时距特征的方式选择动作来实现的。

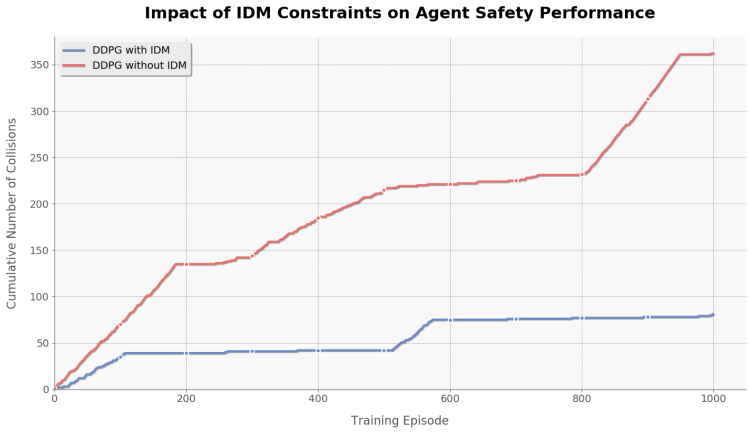

Figure 2. Comparison chart of collisions in different constrained DDPG algorithms
图 2. 不同约束 DDPG 算法的碰撞对比图

Figure 3. Comparison of average rewards of different constrained DDPG algorithms
图 3. 不同约束 DDPG 算法的平均奖励对比图

6. 验证

在本节中，将加入 IDM 模型进行约束的 DDPG 模型与无约束的模型进行比较，以证明该模型安全、高效和舒适地跟随领先车辆的能力。所有的分析都是基于测试数据。DDPG 模型通过将领先车辆轨迹作为输入来产生跟随车辆轨迹。

6.1. 安全驾驶

在跟车事件中，基于 TTC 来评价驾驶安全性。图 4 显示了有无 IDM 模型约束 DDPG 算法模拟的 TTC 累积分布。为了更好地解释，仅列出了 0 至 50 s 范围内的 TTC 值。可以看出，0~3 s 的高危险区域内，IDM 约束的 DDPG 模型比无 IDM 约束的 DDPG 算法具有更低的累积概率。这意味着由 IDM 约束的 DDPG 模型生成的跟车行为比没约束的 DDPG 算法中观察到的驾驶员行为更安全。

6.2. 高效驾驶

在跟车过程中，基于车头时距评价行车效率。在跟车事件的每个时间步长计算车头时距，这些车头时距的累积分布如图 5 所示。DDPG with IDM 模型、DDPG without IDM 模型的平均车头时距分别为 1.24 s、和 1.61 s。可以看出，无约束的 DDPG 模型具有更宽的时间间隔分布范围(0 s 至 8 s)。其中包括一些小于 1 秒的危险车头时距，也包括一些大于 3 秒的低效车头时距。因此，可以得出结论，IDM 约束的 DDPG 模型可以跟随领先车辆与一个有效的和安全的时间车头时距。

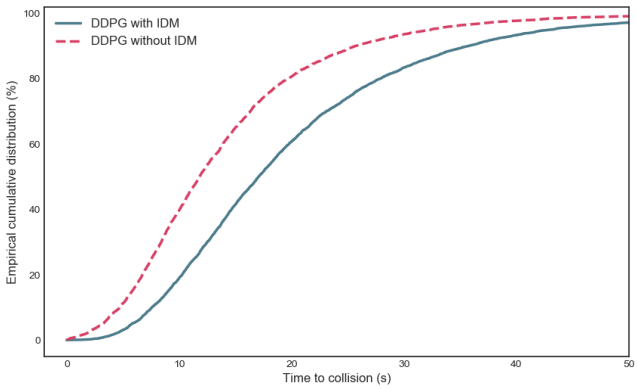

Figure 4. Empirical cumulative distribution of TTC during car following
图 4. 跟车过程中 TTC 的经验累积分布

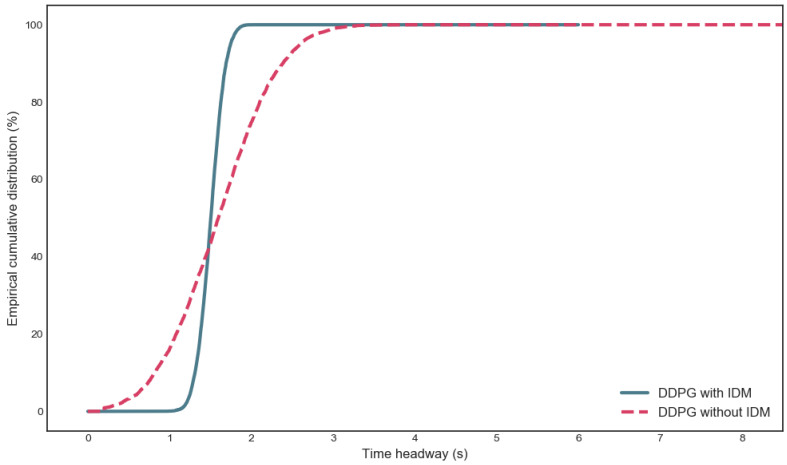

Figure 5. Empirical cumulative distribution of time headway during car following
图 5. 跟驰过程中车头时距的经验累积分布

6.3. 舒适驾驶

根据跟车过程中的加加速度值评价驾驶舒适性。与车头时距类似，它是针对跟车事件的每个时间步长计算的。图 6 示出了在车辆跟随事件期间加加速度值的累积分布。DDPG with IDM 模型、DDPG without IDM 模型的加加速度的平均值分别为 0.67 m/s^3 和 1.68 m/s^3。由于加加速度的绝对值越小，驾驶越舒适，因此可以得出结论，在 NGSIM 数据中，加了 IDM 约束的 DDPG 模型可比没加约束的模型以更舒适的方式控制车辆速度。

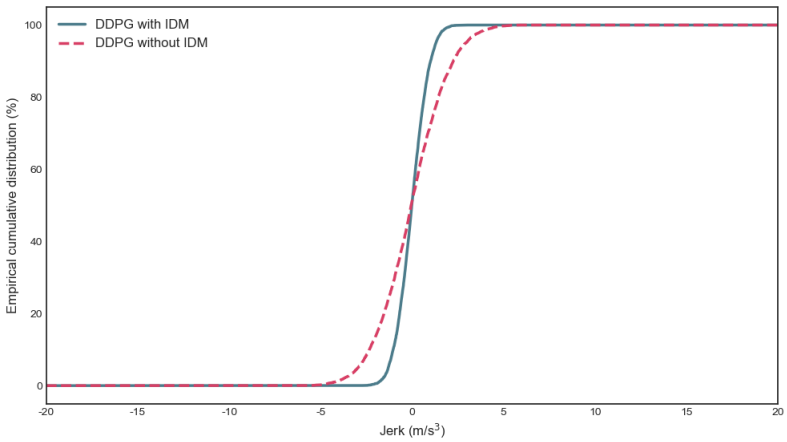

Figure 6. Empirical cumulative distribution of jerk during car following
图 6. 跟驰过程中加加速度的经验累积分布

7. 结论

综上所述，本研究使用启发式 RL 来学习如何以安全、有效和舒适的方式控制车辆在跟驰过程中的速度。使用 NGSIM 研究中的真实的世界的人类驾驶数据来训练模型。将该模型与没加 IDM 约束的模型进行比较，以评估模型的性能。仿真结果表明，该模型具有安全、高效、舒适的驾驶性能。与 DDPG without IDM 算法相比，所提模型在安全性、舒适性、尤其是运行速度等方面明显优于没加约束的算法。研究结果表明，启发式 RL 方法有助于自动驾驶系统的发展。

参考文献

[1] Zhu, M., Wang, X., Tarko, A. and Fang, S. (2018) Modeling Car-Following Behavior on Urban Expressways in Shanghai: A Naturalistic Driving Study. *Transportation Research Part C*: *Emerging Technologies*, **93**, 425-445. https://doi.org/10.1016/j.trc.2018.06.009

[2] Wang, X., Chen, M., Zhu, M. and Tremont, P. (2016) Development of a Kinematic-Based Forward Collision Warning Algorithm Using an Advanced Driving Simulator. *IEEE Transactions on Intelligent Transportation Systems*, **17**, 2583-2591. https://doi.org/10.1109/tits.2016.2528508

[3] Gipps, P.G. (1981) A Behavioural Car-Following Model for Computer Simulation. *Transportation Research Part B*: *Methodological*, **15**, 105-111. https://doi.org/10.1016/0191-2615(81)90037-0

[4] Treiber, M., Hennecke, A. and Helbing, D. (2000) Congested Traffic States in Empirical Observations and Microscopic Simulations. *Physical Review E*, **62**, 1805-1824. https://doi.org/10.1103/physreve.62.1805

[5] Zhang, Y., Xu, Q., Wang, J., Wu, K., Zheng, Z. and Lu, K. (2023) A Learning-Based Discretionary Lane-Change Decision-Making Model with Driving Style Awareness. *IEEE Transactions on Intelligent Transportation Systems*, **24**, 68-78. https://doi.org/10.1109/tits.2022.3217673

[6] Jia, H.F., *et al*. (2003) Develop a Car-Following Model Using Data Collected by "Five-Wheel System". *Proceedings of the* 2003 *IEEE International Conference on Intelligent Transportation Systems*, Vol. 1, 346-351.

[7] Wei, D. and Liu, H. (2013) Analysis of Asymmetric Driving Behavior Using a Self-Learning Approach. *Transportation Research Part B*: *Methodological*, **47**, 1-14. https://doi.org/10.1016/j.trb.2012.09.003

[8] Li, G., Yang, Y., Li, S., Qu, X., Lyu, N. and Li, S.E. (2022) Decision Making of Autonomous Vehicles in Lane Change Scenarios: Deep Reinforcement Learning Approaches with Risk Awareness. *Transportation Research Part C*: *Emerging Technologies*, **134**, Article ID: 103452. https://doi.org/10.1016/j.trc.2021.103452

[9] Basu, C., Yang, Q., Hungerman, D., Singhal, M. and Dragan, A.D. (2017) Do You Want Your Autonomous Car to Drive like You? *Proceedings of the* 2017 *ACM/IEEE International Conference on Human-Robot Interaction*, Vienna, 6-9 March 2017, 417-425. https://doi.org/10.1145/2909824.3020250

[10] Ossen, S. and Hoogendoorn, S.P. (2011) Heterogeneity in Car-Following Behavior: Theory and Empirics. *Transportation Research Part C*: *Emerging Technologies*, **19**, 182-195. https://doi.org/10.1016/j.trc.2010.05.006

[11] Yang, X., Zou, Y., Zhang, H., Qu, X. and Chen, L. (2023) Improved Deep Reinforcement Learning for Car-Following Decision-Making. *Physica A*: *Statistical Mechanics and Its Applications*, **624**, Article ID: 128912. https://doi.org/10.1016/j.physa.2023.128912

[12] Zhu, M., Wang, X. and Wang, Y. (2018) Human-Like Autonomous Car-Following Model with Deep Reinforcement Learning. *Transportation Research Part C*: *Emerging Technologies*, **97**, 348-368. https://doi.org/10.1016/j.trc.2018.10.024

[13] Hart, F., Okhrin, O. and Treiber, M. (2024) Towards Robust Car-Following Based on Deep Reinforcement Learning. *Transportation Research Part C*: *Emerging Technologies*, **159**, Article ID: 104486. https://doi.org/10.1016/j.trc.2024.104486

[14] Hart, P., Nilsson, N. and Raphael, B. (1968) A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, **4**, 100-107. https://doi.org/10.1109/tssc.1968.300136

[15] Lillicrap, T.P., Hunt, J.J., Pritzel, A., *et al*. (2015) Continuous Control with Deep Reinforcement Learning.

[16] Montanino, M. and Punzo, V. (2015) Trajectory Data Reconstruction and Simulation-Based Validation against Macroscopic Traffic Patterns. *Transportation Research Part B*: *Methodological*, **80**, 82-106. https://doi.org/10.1016/j.trb.2015.06.010

[17] Wang, X., Jiang, R., Li, L., Lin, Y., Zheng, X. and Wang, F. (2018) Capturing Car-Following Behaviors by Deep Learning. *IEEE Transactions on Intelligent Transportation Systems*, **19**, 910-920. https://doi.org/10.1109/tits.2017.2706963

[18] Vogel, K. (2003) A Comparison of Headway and Time to Collision as Safety Indicators. *Accident Analysis & Prevention*, **35**, 427-433. https://doi.org/10.1016/s0001-4575(02)00022-2

[19] Uhlenbeck, G.E. and Ornstein, L.S. (1930) On the Theory of the Brownian Motion. *Physical Review*, **36**, 823-841. https://doi.org/10.1103/physrev.3f6.823