Proceeding of the 18th INFATS & 27th SAE-C/STC, 2024, pp1-8
2024.403

Research on Lane-Change Policy of Autonomous Driving Based on Policy Constrained Reinforcement Learning

Rongliang ZHOU, Jiakun HUANG, Mingjun LI, Hepeng LI, Xiaolin SONG
State Key Laboratory of Advanced Design and Manufacturing Technology for Vehicle, Hunan University, Changsha , China, 410082
Email: zhourongliang@hnu.edu.cn, hjk0517@hnu.edu.cn, mingjunl@hnu.edu.cn, lhphnu@hnu.edu.cn, jqysxl@hnu.edu.cn

Abstract: A safe and efficient decision-making system is essential for autonomous vehicles. However, the complexity and variability of the driving environment limit the effectiveness of many rule-based and machine learning-based decision-making methods. The introduction of reinforcement learning (RL) into autonomous driving presents a promising solution to these challenges, but concerns regarding safety and efficiency during training have hindered its broader adoption. To address these challenges, we propose a Policy Constrained Reinforcement Learning (PCRL) approach for autonomous driving lane-change strategies, enabling the student agent to quickly assimilate the teacher model's knowledge while ensuring safe and efficient training. First, when the student agent exhibits suboptimal behavior, the teacher model intervenes to prevent dangerous situations. Then, to accelerate the student agent's learning of the teacher's policy, we constrain policy optimization using the Kullback-Leibler (KL) divergence between the teacher and student policies, transforming it into an unconstrained problem using the Lagrangian method. Finally, an annealing strategy is applied to gradually reduce the teacher's intervention, allowing the student agent to explore the environment independently in the later stages of training, thereby overcoming the limitations of the teacher model's performance. Simulation results in a highway lane-change scenario demonstrate that, compared to baseline algorithms, our approach not only improves learning efficiency and performance but also significantly enhances safety during training.

Keywords: Autonomous Vehicles; Reinforcement Learning; Policy Constraints; Lane-Change

基于策略约束强化学习的自动驾驶变道策略研究

周荣良，黄家琨，李明俊，李鹤鹏，宋晓琳
[1] 整车先进设计制造技术全国重点实验室，湖南大学，长沙，中国，410000
Email: zhourongliang@hnu.edu.cn, hjk0517@hnu.edu.cn, mingjunl@hnu.edu.cn, lhphnu@hnu.edu.cn, jqysxl@hnu.edu.cn

摘　要: 安全高效的决策系统对自动驾驶汽车至关重要。然而，驾驶环境的复杂性和多变性限制了许多基于规则和机器学习的决策方法的有效性。将强化学习(Reinforcement Learning, RL)引入自动驾驶为应对这些挑战提供了一个有前景的解决方案，但训练期间的安全性和效率问题仍然阻碍其广泛应用。为了解决这些问题，我们提出了一种基于策略约束强化学习(Policy Constrained RL, PCRL)的自动驾驶变道策略，能够使学生 agent 迅速学习教师模型的知识，实现安全高效的训练。首先，当学生 agent 表现出次优行为时，教师模型会干预其动作以避免危险情况发生。接着，为了加速学生 agent 对教师策略的学习，我们通过将教师与学生策略之间的 Kullback-Leibler (KL)散度作为策略优化的约束，并采用拉格朗日方法将其转化为无约束问题。最后，采用适当的退火策略逐步减少教师干预，确保学生代理在训练后期能够独立探索环境，克服教师模型的性能限制。在高速公路变道场景中的仿真实验结果显示，与基准算法相比，我们的算法不仅提高了学习效率和性能，还显著增强了训练期间的安全性。

关键词: 自动驾驶汽车；强化学习；策略约束；变道

1 引言

由于环境的复杂性和不可预测性，人类在驾驶汽车时容易做出次优决策，从而威胁到交通安全或降低交通效率[1]。自动驾驶系统在提高交通安全和效率方面有着巨大的潜力，其通常由感知层、决策层和规划控制层组成，而决策层通常被称为系统的“大脑”[2]。因此，确保决策算法的安全性和准确性对于自动驾驶系统的整体性能至关重要[3]。目前，自动驾驶汽车的决策方法大致分为两类：知识驱动方法和数据驱动方法。知识驱动方法，如分层状态机(Hierarchical State Machines, HSM)[4]、专家系统(Expert Systems, ES)[5]和有限状态机(Finite State Machines, FSM)[6]等，具有逻辑严谨和可解释性高的特点[7]。然而，这些方法严重依赖现有知识，难以应对知识库之外的场景[8]。随着深度学习的兴起，端到端(End-to-End, E2E)学习[9, 10]、模仿学习(Imitation Learning, IL)[11, 12]和强化学习(RL)[13, 14] 等数据驱动方法已成为自动驾驶领域的重点研究课题。数据驱动方法具有强大的自学习能力和环境适应性，使其非常容易满足各种场景的决策需求，尤其是 RL 算法。RL 算法通过 agent 与环境的交互来优化策略，从而更好地适应复杂的场景[15]，但其训练时的安全风险和效率低下阻碍了它的广泛应用[16]，特别是在自动驾驶和机器人控制等对安全要求较高的领域。

因此，在本文中我们提出了一种基于策略约束强化学习(PCRL)的自动驾驶变道策略，该算法有效地促进了从教师到学生的知识传递，提高算法性能的同时提升了训练过程中的安全性。与其他 RL 算法不同，PCRL 基于教师-学生框架，当学生策略输出的动作不同于教师策略时，训练有素的教师会对学生动作进行干预来避免危险情况发生。并且，PCRL 将教师策略和学生策略之间的 KL 散度作为策略优化的约束，使用拉格朗日方法将目标函数转化为无约束问题进行求解，使学生策略能够快速逼近教师策略。最后，通过采用适当的退火策略逐步减少教师策略的干预，确保学生在训练后期能够充分自主探索环境，克服教师模型的性能限制。本文的主要贡献总结如下：

1). 本文提出了一种名为 PCRL 的强化学习算法，它通过训练有素的教师来指导学生在复杂的驾驶环境中安全高效地学习。

2). 本文将教师和学生策略之间的 KL 散度作为策略更新的约束，并使用拉格朗日方法将其转换为无约束问题，使学生的策略快速逼近教师的策略，从而提高学习效率。

3). 本文探索了教师策略的干预和退火机制，通过逐渐降低教师的干预概率来保证学生的独立探索能力，从而突破由教师表现不佳而造成的性能限制。

2 方法

本节将全面阐述本文所涉及的理论方法。首先，我们介绍了 Proximal Policy Optimization (PPO)算法的理论基础；随后，我们对所提出的 Policy Constraint RL 决策框架进行了详细的描述。

2.1 Proximal Policy Optimization 算法

强化学习的目标是找到一个可以最大化累积折扣回报的策略$\pi(a|s)$，其中累积回报定义为$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$。在策略梯度方法中，我们希望通过优化策略$\pi_\theta(a|s)$来最大化期望的累积回报：

$$J(\theta) = E_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right] \tag{1}$$

其中$J(\theta)$是关于策略参数θ的期望累积回报; E_{π_θ}表示在策略π_θ下的回报期望; γ为折扣因子，用来权衡当前奖励与未来奖励的影响; s_t是在时间t的状态; a_t是在状态s_t下选择的动作。我们的目标是找到参数θ，使得$J(\theta)$最大化。策略梯度定理为我们提供了梯度的计算方法：

$$\nabla_\theta J(\theta) = E_{\pi_\theta}\left[\nabla_\theta log\, \pi_\theta(a_t|s_t)\widehat{A}_t\right] \tag{2}$$

其中$\nabla_\theta J(\theta)$是关于参数θ的梯度; $\widehat{A}_t$是优势函数，它衡量了当前动作相对于基准动作的改进程度，可以表示为$\widehat{A}_t = Q(s_t, a_t) - V(s_t)$，其中$Q(s_t, a_t)$是状态-动作值函数，表示在状态$s_t$采取动作$a_t$后的预期回报，$V(s_t)$是状态值函数，表示状态$s_t$下的预期回报；$log\, \pi_\theta(a_t|s_t)$是策略在状态$s_t$下选择动作$a_t$的对数概率。

在传统的策略梯度方法中，每次通过梯度：$\theta \leftarrow \theta + \alpha\nabla_\theta J(\theta)$更新策略参数，其中$\alpha$是学习率，用于控制更

新的步长。然而，直接使用梯度更新会导致策略的快速变化，可能会破坏策略的收敛性。为了解决这个问题，PPO 引入了一种策略更新方法，其基本思想是通过限制策略更新的步长，来保证更新的稳定性。PPO 的目标函数$L^{CLIP}(\theta)$可以写成：

$$L^{CLIP}(\theta) = E_t\left[\min\left(r_t(\theta)\widehat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\right)\widehat{A}_t\right] \tag{3}$$

其中$r_t(\theta) = \pi_\theta(a_t|s_t) / \pi_{\theta_{old}}(a_t|s_t)$是新旧策略的概率比；$\epsilon$是控制步长的超参数；剪切函数$clip$的作用是当$r_t(\theta)$超过 $1+\epsilon$或低于 $1-\epsilon$时，限制其值在$[1-\epsilon, 1+\epsilon]$之间，防止策略更新过大。

2.2 Policy Constraint RL 决策框架

在本节中，我们将详细介绍 PCRL 决策框架。如图 1 所示，该框架由 PCRL 高级决策层、路径规划层和控制层组成。

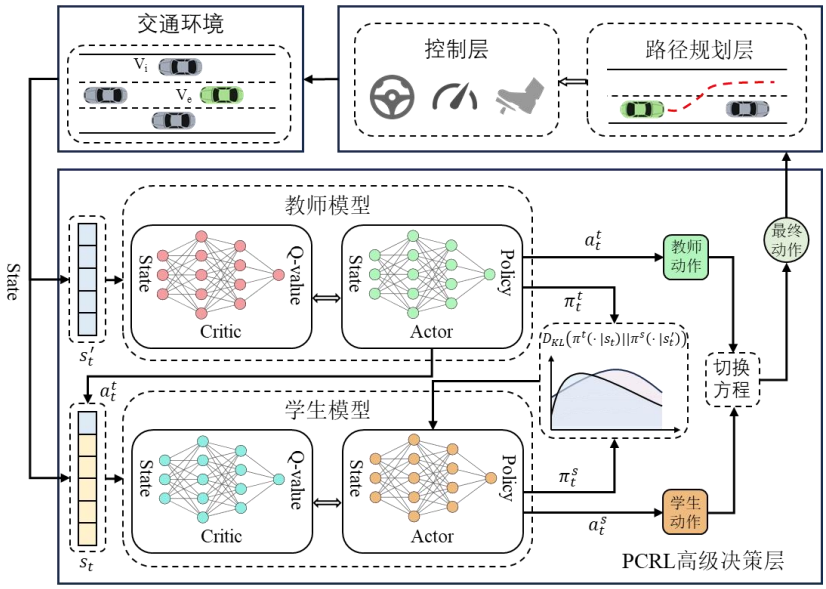


Figure 1. PCRL decision-making framework

图 1. PCRL 决策框架

2.2.1 PCRL 高级决策层

从图 1 可以看出，PCRL 决策层主要包含了教师模型和学生模型两个部分。其中“教师”是预先通过 PPO 算法在与学生相同的模拟场景中训练出来的一个性能较好的离线策略模型，而非基于预设数据集，其目的是为“学生”提供动作干预和演示。需指出，PPO 只是获取教师模型的一种方法，此外还可以通过监督学习或行为克隆（BC）算法利用人类驾驶数据训练教师模型，甚至经验丰富的人类驾驶员也可直接作为教师。而“学生”是指正在学习中的策略模型，它可以通过“教师”的指导和策略约束来提高学习能力。首先，将环境状态$s_t^{'}$输入到一个训练有素的教师策略中，策略π^t输出概率最高的动作a_t^t。a_t^t与状态$s_t^{'}$一起作为学生模型的输入：s_t。然后学生模型的策略π^s根据s_t选择概率最高的动作为a_t^s。为了保证行驶过程的安全，当教师和学生产生的动作不同时，教师应在必要时干预学生的行为。因此，我们提出了一个切换方程来确定最终动作a_t，如下所示：

$$a_t = \begin{cases} a_t^t, & if\ \ n < \tau\omega \\ a_t^s, & otherwise \end{cases} \tag{4}$$

其中n为[0,1]的随机数；ω为教师对学生进行动作干预的概率；τ为衰减系数，用来逐渐减小ω，其表达式如下：

$$\tau = \frac{1}{1 + e^{\frac{n_e}{q_1} - q_2}} \tag{5}$$

其中n_e为训练时的回合数；q_1, q_2为超参数，其值如表 1 中所示。基于该切换方程，确定最终动作a_t并传递

到下一层进行车辆控制。

当动作a_t执行时，环境会向 agent 提供反馈信息以更新神经网络。为了使学生策略能快速逼近教师策略的分布，我们采用了行为克隆的机制，在算法中以老师策略和学生策略之间的 KL 散度作为约束，利用拉格朗日方法将目标函数重构为无约束优化问题。约束函数C_t^{KL}如下所示：

$$C_t^{KL} = D_{KL}(\pi^t(\cdot|s_t^{'})||\pi^s(\cdot|s_t)) \tag{6}$$

其中D_{KL}是 Kullback-Leibler (KL)散度，用于衡量两个概率分布之间的差异；$\pi^t(\cdot|s_t^{'})$表示策略在状态$s_t^{'}$下的动作分布，$\pi^s(\cdot|s_t)$表示策略在状态s_t下的动作分布。

因此，最终 PCRL 的目标函数$L^{PCRL}(\theta)$被定义为如下所示：

$$L^{PCRL}(\theta) = E_t[\min(r_t(\theta)\widehat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon))\widehat{A}_t] - \tau\xi C_t^{KL} \tag{7}$$

其中ξ为拉格朗日乘子，它与策略同时进行更新。PCRL 的所有参数如表 1 所示：

Table 1. Parameters of PCRL decision-making framework
表 1. PCRL 决策框架参数

参数名称	值	参数名称	值
Maximal learning rate	0.0005	Discount factor	0.96
Learning rate decay	True	Lambda entropy	0.01
Total steps of one episode	5,000	Clip parameter	0.2
Total training steps	300K	Lambda advantage	0.98
Optimizer	Adam W	Hyperparameter q_1	5
Mini batch size	64	Hyperparameter q_2	10
Efficiency weight α_1	0.5	Lagrange multiplier ξ	0.01
Safety weight α_2	-1.0	Intervention probability ω	0.6

PCRL 框架的关键在于学生模型的策略学习过程。具体而言，我们使用 KL 散度作为约束，来限制学生模型的策略更新与教师策略之间的差异。为了实现这一目标，我们将该约束附加在 PPO 的目标函数（式(3)）中，并通过拉格朗日方法将带约束的优化问题转化为无约束问题进行求解。改进后的 PPO 目标函数（$L^{PCRL}(\theta)$,式(7)）即作为学生模型的优化目标。在训练过程中，学生模型更新策略时，优化过程会最小化包含 KL 散度约束项的目标函数$L^{PCRL}(\theta)$，确保学生策略能够稳定快速地向教师策略靠近，从而提高了学习效率与安全性。

2.2.2 路径规划与控制层

在 PCRL 决策框架中，我们设计了一个分层架构，以促进自动驾驶汽车在复杂和动态的交通环境中平稳行驶。首先高级决策层根据驾驶状态生成跟随或变道的高级指令，然后低级层根据指令执行路径规划并控制车辆的行驶轨迹和速度。由于本文主要关注高级决策算法，因此我们为低级的路径规划和控制层选择了经典且可靠的算法。

路径规划层：在路径规划层，我们采用三次样条函数来规划车辆的行驶路径。行驶路径的中心线点定义为$(x_0, y_0), (x_1, y_1), .., (x_n, y_n)$。每个$S_i$表示沿插值中心线的线段，并以$(x_i, y_i)$和$(x_{i+1}, y_{i+1})$为界。三次样条函数定义如下：

$$y_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{8}$$

其中$y_i(x)$是在第i段中的插值曲线，用于描述x对应的y坐标；a_i是常数项，表示样条的初始值；b_i是一次项的系数，决定了曲线在x_i处的斜率；c_i是二次项的系数，影响曲线的曲率；d_i是三次项的系数，控制曲线的变化率。参数a_i, b_i, c_i和d_i可以由文章[17]中的理论进行求解。

全局路径由 Carla 生成的导航点决定。当高级命令是换道时，我们将下一个导航点的横向位置设置在目标车道的中心线上，与当前位置保持 10 m 的纵向距离。随后，此段变道轨迹采用上述的三次样条插值进行局部规划。

控制层：在本文中，我们专注于横向决策，包括车道变换和跟车行为。为了保持简单性并防止 agent 行为过于保守，加速度和速度由 IDM 模型控制，因此本车的目标加速度v_e可使用下式进行计算：

$$\dot{v}_e = a\left[1 - \left(\frac{v_e}{v_0}\right)^{\zeta} - \left(\frac{\max\left(s_0 + v_e T + \frac{v_e(v_e - v_1)}{2\sqrt{ab}}, 0\right)}{s}\right)^2\right] \tag{9}$$

其中v_e为本车速度，v_0为期望速度，v_1为前车速度，IDM 其他参数详见表 2。通过目标加速度和当前速度，可以确定下一时刻的目标速度。

获得参考行驶路径和目标速度后，控制层利用 PID 算法计算车辆的油门、刹车、转向角，其中横向控制和纵向控制分别采用两套 PID 参数，如表 2 所示。

Table 2. Parameters of IDM model and PID controller
表 2. IDM 模型和 PID 控制器参数

IDM 模型参数		PID 控制器参数	
参数名称	值	参数名称	值
Desired speed v_0	25 m/s	Proportional Gain of lateral control K_P	0.75
Desired time gap T	0.6 s	Derivative Gain of lateral control K_D	0.01
Safety gap s_0	2.0 m	Integral Gain of lateral control K_I	0.2
Acceleration exponent ζ	4	Proportional Gain of longitudinal control K_P'	0.37
Maximal acceleration a	2 m/s	Derivative Gain of longitudinal control K_D'	0.012
Comfortable deceleration b	2 m/s	Integral Gain of longitudinal control K_I'	0.016

3 实现

本节将介绍仿真实验的实施细节。首先，我们描述了中密度高速公路变道场景的细节；随后，我们将驾驶场景建模为马尔可夫决策过程；最后，我们介绍了实验中的参数设置。

3.1 场景规范

高速公路上中密度交通场景的特点是每车道每公里约有 15 辆车，其车距范围为 50 至 90 米，这种密度通过避开拥堵交通（会阻碍车道变换）和低密度交通（无需变换车道）来达到平衡。因此，它是大多数车道变换场景的典型代表。图 2 展示了该场景在 Carla 模拟环境中的表现，其中本车在 3 车道高速公路上行驶，并在其他车辆包围的情况下执行车道变换。根据驾驶知识和交通法规，自车在做出变换车道的决策过程中必须考虑前方车辆以及左前、左后、右前和右后方车辆的位置和移动。因此，这些周围车辆的实时状态（例如速度和距离）将被作为模型的输入。

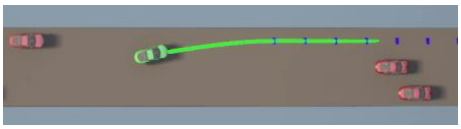


Figure 2. Lane change scenario on medium-density highway
图 2. 中密度高速公路变道场景

在实验中，本车和周围车辆的初始速度均为 0 m/s，本车的最大速度限制为 25 m/s。为了更准确地模拟真实的驾驶环境，周围车辆的目标速度在 15 到 25 m/s 的范围内随机选择。错误的变道决策可能会导致本车与其他车辆或道路边界发生碰撞。如果本车成功行驶 1 km 而没有发生任何碰撞，则认为成功完成该回合。如果行驶过程中发生碰撞，则该回合被标记为失败，环境将被重置并进行下一回合。

3.2 场景建模

接下来，我们将定义状态空间、动作空间和奖励函数，将场景建模为马尔可夫决策过程（MDP）。

3.2.1 状态空间

状态s_t包含本车环境的所有观察，包括本车和周围车辆的驾驶信息（例如速度和距离）。如第 2.2 节所述，在时间t，s_t定义如下：

$$s_t = (v_t^e, \{v_t^i, d_t^i\}) \tag{10}$$

其中v^e是本车的速度，周围车辆的速度及其相对于本车的纵向距离表示为$\{v_t^i, d_t^i\}$，其中$i = (1,2,3,4,5)$分别对应前车、左前车、左后车、右前车和右后车。

3.2.2 动作空间

在本文中，高级决策层仅提供离散的源动作，作为路径规划层和控制层的高级命令。因此，自车在时间t的动作a_t可定义如下：

$$a_t = (a_1, a_2, a_3) \tag{11}$$

其中a_1, a_2和a_3分别代表跟随，左转和右转动作。

3.2.3 奖励函数

在高速公路换道场景中，自动驾驶汽车的目标是实现高速行驶的同时避免碰撞。因此，为了鼓励本车在保持较高安全性的同时快速行驶，本文的奖励函数由以下两部分组成：

效率奖励：为了鼓励自车从 0 m/s 加速到更高的速度，我们设计了效率奖励函数如下：

$$R_e = \begin{cases} 0, & if\ \ 0 < v^e < 12.5\ m/s \\ \alpha_1\left(\dfrac{v^e}{12.5} - 1\right), & if\ 12.5\ m/s < v^e < 25 m/s \\ \alpha_1, & otherwise \end{cases} \tag{12}$$

其中$\alpha_1 > 0$ 表示用于调整效率奖励在策略中重要性的权重系数。

安全奖励：安全奖励取决于安全距离d_{safe}，它表示本车与前方或后方车辆之间的纵向距离，以及本车是否发生碰撞。安全奖励函数可以定义如下：

$$R_s = \begin{cases} \alpha_2, & if\ \ 0 < d_{safe} < 5\ m \\ \alpha_2\left(1 - \dfrac{d_{safe} - 5}{5}\right), & if\ 5\ m < d_{safe} < 10\ m \\ 1, & if\ collision \\ 0, & otherwise \end{cases} \tag{13}$$

其中$\alpha_2 < 0$ 表示用于调整安全奖励在策略中重要性的权重系数。

最终奖励函数R如式 14 所示：

$$R = R_e + R_s \tag{14}$$

3.3 实验设置

本文中所有实验均在配备 Intel i5-13600KF CPU、NVIDIA GeForce RTX 4090 GPU 和 32GB RAM 的计算机上进行。由于本文重点在于决策方法的研究，因此我们简化了感知模块，假设本车可以直接获得 50 m 半径范围内其他车辆的准确驾驶信息（例如速度、位置）。为了展示算法的泛化性，每次重置环境时都会使用不同的随机种子。除本车外所有车辆都由 Carla 的交通管理器控制，包括跟车和变道等行为。

4 结果

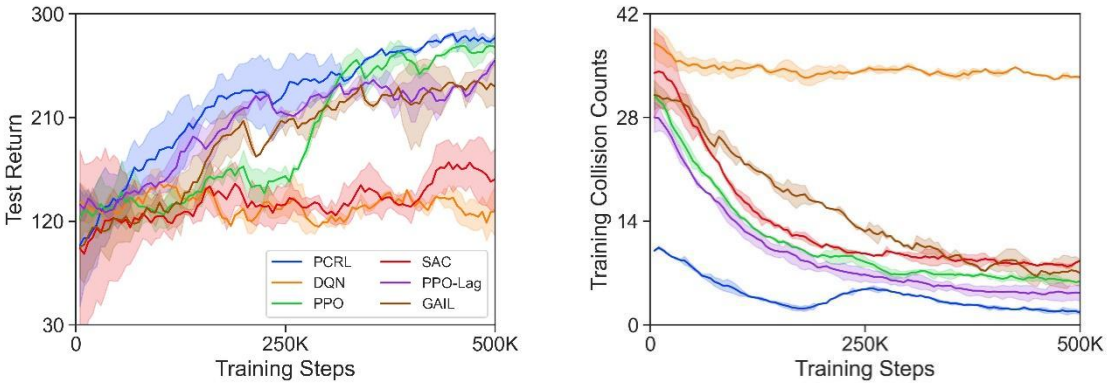
本文研究了 6 种不同算法(PCRL、DQN、PPO、SAC、PPO-Lag 和 GAIL)在 3 车道中密度高速公路交通场景中的训练和评估过程。

4.1 模型训练

对于模型训练过程，每个算法训练 2 次，使用不同的随机种子来获取它们的平均性能。在训练过程中，模型每 5,000 步进行一次测试并记录结果。

图 5 展示了各种算法在模型训练过程中的测试结果，包括测试的 Return 值和每 5,000 步 agent 与环境交互产生的碰撞次数，这两个指标主要表现了模型训练期间的综合性能和安全性。图 3(a)显示了 Return 曲线与训练步

数的关系，清晰地反映了所有算法的整体性能。如曲线所示，在教师模型的指导下，PCRL 从训练期间获得了更多准确的知识，从而从始终维持着较高的 Return 值。相比之下，其他 RL 算法则通过环境交互逐渐优化其策略。它们在测试期间经历了更频繁的碰撞，导致 Return 值几乎始终低于 PCRL。图 3(b)则表明，PCRL 在数据收集期间发生的碰撞次数不到其他算法的一半，最后每 5,000 个训练步大约只发生 2 次碰撞。因此，即使在训练的初始阶段，策略尚未完成优化时，PCRL 通过教师的指导也能使学生 agent 实现明显优于其他算法的表现和安全性，特别是对于 SAC 和 DQN 算法。



(a) Test Return **(b) Training Collision Counts**

Figure 3. Curve: result of model training
图 3. 模型训练结果曲线

4.2 模型评估

训练完成后，每个算法使用不同的随机种子进行 4 次评估，取其平均值作为结果以准确捕获模型的平均性能。测试结果主要为自动驾驶车辆行驶 1 km 的成功率、奖励值 Return、安全 Cost 以及行驶速度。其结果如表 3 所示：

Table 3. Results of model evaluation
表 3. 模型评估结果

	算法	Success Rate (%)	Return	Cost	Speed (m/s)
Value-Based RL	DQN	27.50	169.99	8.94	17.43
On-Policy RL	PPO	81.88	274.45	3.27	19.36
Off-Policy RL	SAC	20.50	147.52	6.34	18.34
Safe RL	PPO-Lag	75.25	254.78	2.36	18.83
IL	GAIL	82.63	248.48	**1.76**	18.64
Ours	PCRL	**90.75**	**283.94**	2.78	**19.41**

结果表明，PPO-based 算法(PPO, PPO-Lag, GAIL 和 PCRL)在高速公路的变道任务中表现明显优于 SAC 和 DQN 算法，这与它们在训练阶段的表现一致。值得注意的是，PCRL 在两个关键指标上表现出色：成功率和回报值 Return，成功率达到了 90.75%，回报值为 283.94，与 19.41 m/s 的行驶速度一起作为所有结果的最高值。虽然 PCRL 的安全 Cost(2.78)不是最高的，但它非常接近最佳值 1.7。需要注意的是，对于 IL 算法，必须先收集大量高质量数据来训练 GAIL 模型。因此，虽然 GAIL 性能水平接近 PCRL，并获得了最好的安全 Cost，但它的训练过程更复杂且更耗时。而 PPO 算法则不受其他策略影响，完全自主地探索环境，虽然成功率和安全性能都相对较低，但通过优先考虑更高的效率奖励，PPO 实现了 19.36 m/s 的行驶速度。而 SAC 和 DQN 算法的性能仍然较差，成功率不超过 30%，返回值低于 200。这与 PPO-based 的算法相比差距较大，也与训练阶段的结果表现一致。

5 结论

本文提出了一种基于策略约束强化学习的新方法来应对自动驾驶车辆在高速公路的变道场景。该方法通过训练有素的教师模型指导学生在复杂的驾驶环境中进行变道，提高训练安全性和效率的同时提升了算法的性能。该方法将教师策略和学生策略之间的 KL 散度作为策略更新的约束，使学生代理能够快速吸收教师的知识。并且我们设计了一个退火机制，逐渐减少教师策略的干预概率，使学生能够充分自主探索更优的策略，进一步提升性能上限。我们在 Carla 仿真环境中进行实验，其结果表明：与其他强化学习算法相比，所提出的 PCRL 显著提高了模型的学习效率和性能。更重要的是，教师模型的指导大大提高了学生训练期间的安全性，避免了大量碰撞事件的发生。

由于本研究中所有的实验均在仿真环境中进行，而真实实验对于评估所提方法的有效性和实用性是非常有必要的。因此，未来应在真实驾驶环境下进一步验证该算法的性能。并且，此研究还应扩展到涉及自动驾驶汽车和人类驾驶汽车共存的混合交通场景，增强该方法的适用性和泛化性。

参考文献 (References)

[1] Lee S E, Olsen E C B, Wierwille W W. A comprehensive examination of naturalistic lane-changes[R]. United States. Department of Transportation. National Highway Traffic Safety Administration, 2004.

[2] Chen S, Zhang S, Shang J, et al. Brain-inspired cognitive model with attention for self-driving cars[J]. IEEE Transactions on Cognitive and Developmental Systems, 2017, 11(1): 13-25.

[3] Li N, Oyler D W, Zhang M, et al. Game theoretic modeling of driver and vehicle interactions for verification and validation of autonomous vehicle control systems[J]. IEEE Transactions on control systems technology, 2017, 26(5): 1782-1797.

[4] Zheng X, Li H, Qiang Z, et al. Intelligent decision-making method for vehicles in emergency conditions based on artificial potential fields and finite state machines[J]. Journal of Intelligent and Connected Vehicles, 2024.

[5] Fu Y, Li C, Yu F R, et al. Hybrid autonomous driving guidance strategy combining deep reinforcement learning and expert system[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 11273-11286.

[6] Hwang S, Lee K, Jeon H, et al. Autonomous vehicle cut-in algorithm for lane-merging scenarios via policy-based reinforcement learning nested within finite-state machine[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 17594-17606.

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

[8] Možina M, Guid M, Krivec J, et al. Fighting knowledge acquisition bottleneck with argument based machine learning[M]//ECAI 2008. IOS Press, 2008: 234-238.

[9] Chen L, Wu P, Chitta K, et al. End-to-end autonomous driving: Challenges and frontiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[10] Teng S, Chen L, Ai Y, et al. Hierarchical interpretable imitation learning for end-to-end autonomous driving[J]. IEEE Transactions on Intelligent Vehicles, 2022, 8(1): 673-683.

[11] Ozcelik M B, Agin B, Caldiran O, et al. Decision Making for Autonomous Driving in a Virtual Highway Environment based on Generative Adversarial Imitation Learning[C]//2023 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2023: 1-6.

[12] Hu A, Corrado G, Griffiths N, et al. Model-based imitation learning for urban driving[J]. Advances in Neural Information Processing Systems, 2022, 35: 20703-20716.

[13] Yang K, Tang X, Qiu S, et al. Towards robust decision-making for autonomous driving on highway[J]. IEEE Transactions on Vehicular Technology, 2023, 72(9): 11251-11263.

[14] Jiang Y, Zhan G, Lan Z, et al. A reinforcement learning benchmark for autonomous driving in general urban scenarios[J]. IEEE Transactions on Intelligent Transportation Systems, 2023.

[15] Z. Zhu, H. Zhao, A survey of deep rl and il for autonomous driving policy learning, IEEE Transactions on Intelligent Transportation Systems 23 (9) (2021) 14043–14065.

[16] Z. Liu, Z. Cen, V. Isenbaev, W. Liu, S. Wu, B. Li, D. Zhao, Constrained variational policy optimization for safe reinforcement learning, in: International Conference on Machine Learning, PMLR, 2022, pp. 13644–13668.

[17] Y. Jiang, X. Jin, Y. Xiong, Z. Liu, A dynamic motion planning framework for autonomous driving in urban environments, in: 2020 39th Chinese Control Conference (CCC), IEEE, 2020, pp. 5429–5435.