计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210700010

基于深度强化学习的无信号灯交叉路口车辆控制

欧阳卓[1]　周思源[1,2]　吕　勇[1]　谭国平[1,2]　张　悦[1]　项亮亮[1]
1 河海大学计算机与信息学院　南京 211100
2 江苏智能交通及智能驾驶研究院　南京 210019
(191307020022@hhu.edu.cn)

摘　要　利用深度强化学习技术实现无信号灯交叉路口车辆控制是智能交通领域的研究热点。现有研究存在无法适应自动驾驶车辆数量动态变化、训练收敛慢、训练结果只能达到局部最优等问题。文中研究在无信号灯交叉路口,自动驾驶车辆如何利用分布式深度强化方法来提升路口的通行效率。首先,提出了一种高效的奖励函数,将分布式强化学习算法应用到无信号灯交叉路口场景中,使得车辆即使无法获取整个交叉路口的状态信息,只依赖局部信息也能有效提升交叉路口的通行效率。然后,针对开放交叉路口场景中强化学习方法训练效率低的问题,使用了迁移学习的方法,将封闭的8字型场景中训练好的策略作为暖启动,在无信号灯交叉路口场景继续训练,提升了训练效率。最后,提出了一种可以适应所有自动驾驶车辆比例的策略,此策略在任意比例自动驾驶车辆的场景中均可提升交叉路口的通行效率。在仿真平台Flow上对TD3强化学习算法进行了验证,实验结果表明,改进后的算法训练收敛快,能适应自动驾驶车辆比例的动态变化,能有效提升路口的通行效率。

关键词:自动驾驶;无信号灯交叉路口;深度强化学习;V2X
中图法分类号　TP391

DRL-based Vehicle Control Strategy for Signal-free Intersections

OUYANG Zhuo[1],ZHOU Si-yuan[1,2],LYU Yong[1],TAN Guo-ping[1,2],ZHANG Yue[1] and XIANG Liang-liang[1]
1 School of Computer and Information,Hohai University,Nanjing 211100,China
2 Jiangsu Intelligent Transportation and Intelligent Driving Research Institute,Nanjing 210019,China

Abstract　Using deep learning technology to control vehicles at intersections is a research hotspot in the field of intelligent transportation. Previous studies suffer from the inability to adapt to dynamic changes in the number of self-driving vehicles,slow convergence of training,and locally optimal training results. This work focuses on how autonomous vehicles can use distributed deep reinforcement methods to improve the efficiency of intersections at unsignalized intersections. First,an efficient reward function is proposed to apply the distributed reinforcement learning algorithm to the unsignalized intersection scenario,which can effectively improve the efficiency of intersection passage by relying on only local information even if the vehicle cannot obtain the whole intersection state information. Then,to address the problem of inefficient training of reinforcement learning methods in open intersection scenarios,a transfer learning approach is used to improve the training efficiency by using the trained strategy in the closed figure-of-eight scenario as a warm start and continuing the training in the unsignalized intersection scenario. Finally,this paper proposes a strategy that can be adapted to all proportions of autonomous vehicles,and this strategy can improve intersection access efficiency in scenarios with any proportion of autonomous vehicles. The algorithm is validated on the simulation platform Flow,and the experimental results show that the proposed smart body model converges quickly in training,can adapt to dynamic changes in the proportion of self-driving vehicles,and can effectively improve the efficiency of intersections.

Keywords　Autonomous vehicles,Signal-free intersections,Deep reinforcement learning,V2X

1　引言

无信号灯交叉路口场景下车辆的调度是自动驾驶领域最困难的任务之一,也是本文研究的主题。

早期关于无信号交叉路口自动驾驶车辆的通行策略方面的研究使用的是基于规则的方法。基于规则的决策方法中基于用户体验、交通规则等建立规则库,自动驾驶车辆根据所处环境选择最优行为策略[1]。Lv等[2]提出了一种基于顺序

到稿日期:2021-07-01　返修日期:2021-08-28
基金项目:国家自然科学基金(61701168,61832005);中国博士后科研基金(2019M651546);江苏省交通运输厅重大科技项目(2019Z07)
This work was supported by the National Natural Science Foundation of China(61701168,61832005),China Postdoctoral Science Funded Project(2019M651546) and Major Technological Projects of Jiangsu Province Transportations Department(2019Z07).
通信作者:谭国平(gptan@hhu.edu.cn)

选择的调度机制，为每辆即将到达十字路口的车辆规划路口通行方案，提升了路口的通行效率。Rios-Torres 等[3]将车辆在路口的通行问题建模为无约束最优控制问题，并利用哈密尔顿分析推导出解决方案，该方案降低了油耗，减少了车辆平均行驶时间。Wang 等[4]提出了基于分布式共识的协议来解决不同方向车辆汇流的问题，减少了车辆平均行驶时间、能耗和污染物排放，并保证了行车安全。然而，基于规则的方法高度依赖于对车辆和地图、障碍物等环境的精确建模，无法应用到高度复杂和动态的驾驶环境中[5-6]。

深度强化学习利用深度神经网络作为策略函数逼近器，在部署过程中需要的计算时间很少，在解决决策问题时具有天然优势[7]。强化学习已被用于处理交叉路口自动驾驶车辆的通行问题。文献[8-9]使用了分布式自动驾驶车辆在十字路口基于深度 Q 网络（Deep Q Network，DQN）的导航策略，可以避免碰撞，提升了通行成功率。然而，DQN 只能处理离散低维动作的控制，而自动驾驶车辆的控制是连续高维动作空间问题，需要输出连续的方向盘转角值和加速度值，DQN 无法解决[10]。文献[11] 建立了基于 DDPG（Deep Deterministic Policy Gradient）深度强化学习算法的端到端无人驾驶决策控制模型，在 TORCS 平台上通过输入连续的车辆行驶环境感知信息，决策控制模型直接输出加速、刹车、转向等驾驶动作，实现了端到端无人驾驶决策。文献[12]针对 DRL 算法的无约束探索导致的不安全问题，提出了动作约束安全强化学习方法，此方法在智能体执行决策时针对不安全问题进行了有效的约束，提升了智能体完成任务的成功率。多自动驾驶车辆环境的强化学习方法研究中，现有的强化学习方法可以分为集中式强化学习[13]和分布式强化学习[14]。集中式强化学习场景下，所有智能体采用联合状态和联合奖励，集中控制设备对所有智能体进行联合控制，文献[15]在有少量自动驾驶车辆和人类驾驶车辆的混合交通流场景中，使用了集中式强化学习的训练方法，以提升所有车辆在环形道路上行驶的平稳性。然而，由于联合状态和联合奖励的维度会随着智能体数量的增加而增加，因此集中式强化学习控制方法存在训练难度大、对实时通信要求高等缺点。分布式强化学习采用了集中式训练，分布式执行的方法独立采集数据，所有智能体将采集到的数据收集到经验缓冲区统一进行训练，得到最终策略。由于每个智能体都使用局部状态进行独立决策，因此分布式强化学习方法具有可拓展性强、通信成本低的优点。文献[16]在混合自主环境下使用了一种完全分布式的强化学习方法来控制自主驾驶车辆，提升了瓶颈路口的吞吐量。文献[17]将混合交通高速公路入口匝道合并问题描述为一个多智能体强化学习（Multi-Agent Reinforcement Learning，MARL）问题，汇流和直行车道上的自动驾驶车辆共同学习一个策略，使交通吞吐量最大化。文献[18]利用近端策略优化（Proximal Policy Optimization，PPO）算法来解决自动驾驶车辆在无信号灯交叉路口的通行问题，提升了路口车辆的平均速度。现有的无信号灯交叉路口强化学习算法存在以下缺陷：1）对无信号灯交叉路口建模理想化，例如文献[18]中车辆做决策时只考虑前后车和自身的状态，使所有车辆只能直行；2）奖励函数设计困难，单一的奖励函数经常会使训练陷入局部最优，训练后达不到预期的效果。

本文的主要工作如下：

（1）将混合交通流无信号灯交叉路口自动驾驶车辆的通行问题转化为一个分布式强化学习问题，设计了一种混合奖励函数，有效提升了路口吞吐量。

（2）为了解决自动驾驶车辆数量动态变化的问题，本文将自动驾驶车辆比例参数化，训练得到的车辆控制策略能适应不同比例自动驾驶车辆的混合交通流场景。

（3）使用了迁移学习的方法，将封闭网络中训练的分布式强化学习策略迁移到开放交叉口网络中，提升了训练效率。

2 系统模型

本研究考虑无信号灯交叉路口混合交通流场景，首先构建了一个 4 向交叉路口，然后给出了自动驾驶车辆及人类驾驶车辆的动力学模型，自动驾驶车辆使用强化学习训练得到策略控制车辆加速度。

2.1 交叉路口建模

如图 1 所示，十字路口由 4 个方向组成，每个方向各有 2 个车道。每条车道的长度为 100m。车辆有左转、右转和直行 3 种可能的驾驶行为。

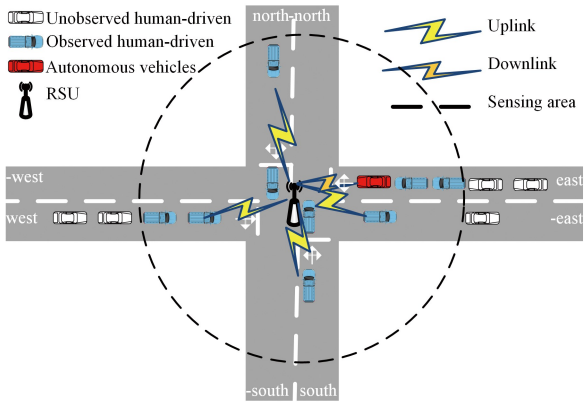

图 1 系统模型图（电子版为彩色）

Fig. 1 System model

在本场景中，每辆自动驾驶车辆都配备有传感设备，并具有短距离通信能力。假设自动驾驶车辆可以通过无线通信、激光雷达等方式获取感知区域内最近的 9 辆车的状态信息。状态信息包括车道、绝对位置、转向信息和行驶速度。红色车辆代表自动驾驶车辆，蓝色车辆代表自动驾驶车辆感知范围内的车辆，白色车辆代表人类驾驶的车辆。

2.2 人类驾驶车辆控制

本文使用智能驾驶员模型（Intelligent Driver Model，IDM）[19]模拟人类驾驶车辆的纵向动力学，该模型假设车辆在限速条件下以最大速度驾驶，同时与前车保持至少 1s 的制动时间车头间距。

2.3 性能指标

本文研究混合自主交通：同时存在自主驾驶车辆和人类驾驶车辆时，自动驾驶车辆如何优化交通系统。在开放交叉路口，一个好的控制策略应该是在固定的时间间隔内最大化通过交叉口的车辆总数。因此将文献[20]中的 Outflow 作为

性能指标，Outflow 表示每小时离开路口的车辆数量。

Outflow 性能指标：

$$Outflow=\frac{\sum_{t=1}^{H} o_t}{H} \tag{1}$$

其中，H 是回合长度，O_t 表示在 t 时刻离开网络的车辆数量。

3 基于 TD3 的强化学习控制策略

本节主要介绍深度强化学习方法的实现。首先将无信号灯交叉路口的通行问题公式化为部分可观测的马尔可夫决策过程(Partially Observable Markov Decision Process，POMDP)，然后介绍 TD3 强化学习算法，最后构建状态空间。动作空间和奖励函数求解公式化后的马尔可夫决策过程。

3.1 问题定义

本小节主要介绍使用的强化学习算法中需要用到的数学表征。将混合交通流问题建模为时间的马尔可夫决策过程(Markov Decision Process，MDP)。MDP 可以用元组〈S，A，T，R〉表示。其中，S 表示状态空间，A 表示动作空间，T 为状态转移概率。$T(s_t, a_t, a_{t+1})$表示在状态 s_t 下采取动作 a_t，转移到状态 s_{t+1} 的概率；$R(s_t, a_t)$表示在 s_t 状态下采取动作 a_t 所获得的奖励值。强化学习的最终目标是找到一个策略，使累计奖励的期望值最大。

自动驾驶车辆利用深度强化学习进行加减速控制。假设动作是连续的，正值代表加速，负值代表减速。完整的工作流程如图 2 所示。假设所有自动驾驶车辆都是同质的，具有相同的车长、制动时间等。

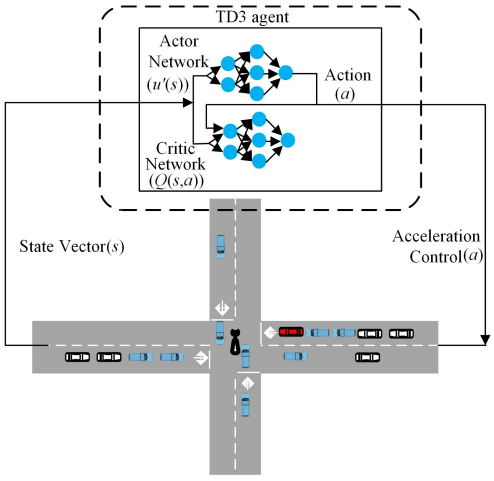

图 2 TD3 架构

Fig. 2 Architecture of TD3

3.2 TD3 算法

TD3 是一种基于 Actor-Critic(AC)框架的强化学习算法。相比同属 AC 架构的 DDPG 算法，TD3 算法作出了以下几个方面的改进：

(1)TD3 使用两个用于评估动作值函数的神经网络以及相对应的两个目标神经网络，两个深度神经网络同时评估此时刻下的状态-动作对，并取两个估计值之间的较小值计算 TD 误差。

(2)相比 Critic 动作值函数单步更新策略，Actor 网络参数更新频率更低，通过此方法可以减小近似动作值函数的方差。

(3)通过在目标动作值函数的输出动作上添加服从正态分布的噪声，来提高算法的鲁棒性。

3.3 分布式车辆控制策略

分布式强化学习训练时，每辆自动驾驶车辆作为一个智能体独立采集数据，所有智能体将采集到的数据收集到经验缓冲区统一进行训练，得到最终策略。由于智能体执行决策时只需要获取局部状态信息，所有智能体并行决策，因此称此策略为分布式驾驶策略。

本节主要阐述了如何将分布式深度强化学习算法应用到路口车辆控制问题中，并且给出了算法的具体定义及网络结构。针对强化学习算法的 3 要素，即状态空间、动作空间和奖励函数进行了定义。

图 2 描述了在仿真环境中智能体(自动驾驶车辆)与环境进行交互的过程。如图所示，在每个时隙 t，智能体通过观测来获取环境的状态 s_t，并以此状态作为其策略网络的输入，随后智能体会基于当前自身的策略执行一个动作 a_t，随之环境会跳转至下一状态 s_{t+1}，并反馈给智能体一个奖励 r_t。〈s_t，a_t，r_t，s_{t+1}〉元组被存储在经验缓冲区，作为训练的样本。在积累了足够的样本后，传递给优化器计算新策略。

3.3.1 状态空间的定义

对于无信号灯交叉路口的自动驾驶车辆，在 t 时刻状态输入如下：

$$S=\left[\frac{v}{V_{\max}}, e, s, \frac{x}{X_{\max}}, \frac{y}{Y_{\max}}, \frac{v_1}{V_{\max}}, e_1, s_1, \frac{x_1}{X_{\max}}, \frac{y_1}{Y_{\max}}, \cdots, \frac{v_9}{V_{\max}}, e_9, s_9, \frac{x_9}{X_{\max}}, \frac{y_9}{Y_{\max}}\right] \tag{2}$$

其中，v_i, e_i, s_i, x_i, y_i 分别表示车辆 i 的速度(m/s)、车道编号、转向编号、横坐标(m)、纵坐标(m)。$V_{\max}, X_{\max}, Y_{\max}$ 分别表示最大可能速度、最大横向坐标、最大纵向坐标。v_j, e_j, s_j, x_j, y_j 分别表示车辆 i 附近第 j 辆车的速度(m/s)、车道编号、转向编号、横坐标(m)，纵坐标(m)。

3.3.2 控制动作的定义

强化学习中的控制动作为自动驾驶车辆的加速度。不考虑车辆动力学延迟，每个时隙，利用加速度对车辆速度进行更新。从车辆安全性和舒适性角度考虑，对车辆的加速度做如下限制 $a\in[-4.5, 2.6]$(m/s^2)。

3.3.3 奖励函数

奖励的定义是深度强化学习是能否收敛并取得良好效果的关键。适当的奖励有助于自动驾驶车辆采取最佳的行动策略。在无信号灯交叉路口，本文关注的是路口吞吐量，每个时隙的奖励被定义为以下因素的加权平均。

(1)第 t 时隙，离开网络的车辆数量的标准值 n_t，n_t 的定义如下：

$$n_t=0.1*\left(\frac{N_t}{N_{\max}}-1\right) \tag{3}$$

其中，N_t 为第 t 时隙离开路口的车辆数量，$N_{\max}$ 为路口能容纳的车辆数。这一部分奖励用来指导自动驾驶车辆并提升路口的整体车辆通行效率。

(2)第 t 时隙，车辆到终点的距离的标准值 d_t 的定义如下：

$$d_t = -0.1 * \frac{D_t}{D_{\max}} \tag{4}$$

其中，D_t 为第 t 时隙到终点的距离，$D_{\max}$ 表示从起点到终点的距离。这一部分奖励用来指导自动驾驶车辆向终点行驶。

综合考虑以上要素和权重，如果车辆还没到达终点，奖励如下：

$$r_t = \lambda n_t + (1-\lambda) d_t \tag{5}$$

其中，λ 表示权重因子，取值范围为 $0<\lambda<1$，用来表征奖励中用于提升路口吞吐量所占的比例，λ 越大，智能体就更倾向于学习提升路口吞吐量的策略，即本文中的全局目标；λ 越小，智能体就更倾向于学习如何尽快到达终点，即本文中的个体目标。

(3)如果第 t 时隙自动驾驶车辆到达终点，则此时的时隙奖励为 20。

$$r_t = 20 \tag{6}$$

3.4 通用控制器

在 3.3 节，我们针对每一种自动驾驶汽车比例训练得到一种控制策略，称为独立控制策略。但这种策略具有两种缺陷：1)计算开销大；2)自动驾驶车辆需要实时得知自动驾驶车辆的比例，以切换到不同的控制器，这在高度动态的混合交通流场景中并不现实。因此，需要一个能够适应渗透率动态变化的混合交通流环境的通用控制策略。

3.5 迁移学习方法

迁移学习技术提供了利用从训练一项任务中获得的经验来改进另一项任务的训练方法。强化学习中，常见的迁移学习方式包括共享策略参数 θ 和状态动作对$\langle s, a, r, s' \rangle$。开放道路车辆数量高度动态变化，训练效率低。封闭道路中车辆环境相对稳定，训练效率更高。此外，封闭道路中车辆数量固定，从而避免了自动驾驶车辆通过控制道路上的车辆数量来增加奖励。本文研究了在封闭道路上训练得到的控制策略能否通过迁移和微调的方式应用到开放网络环境中。

本节详细介绍了本文研究的学习场景。为此，我们设计了类似于第 2.1 节中描述的交叉路口网络的 8 字形道路设置(见图 3)。

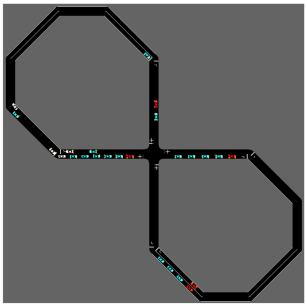

图 3 8 字型封闭网络

Fig. 3 Figure 8 closed network

封闭网络与开放网络不同，封闭网络中车辆数量固定，这避免了自动驾驶车辆通过控制网络中的车辆数量来增加奖励，因此在封闭网络中，我们采用以下奖励函数来鼓励车辆提高整体速度。

封闭网络奖励函数：

$$r_{i,t} = \| v_{\text{des}} \| - \| v_{\text{des}} - v(t) \| - \alpha \sum_{i \in AV} \max[h_{\max} - h_i(t), 0] \tag{7}$$

前两项鼓励系统级速度接近期望速度，同时保证奖励值为正。第三项用来惩罚小的车头间距。当车头间距大于阈值 $h_{\max}$ 时，这一项惩罚为 0，α 用来控制惩罚的幅度。这里，我们选择了以下常量：

$v_{\text{des}} = 30\,\text{m/s}$，$h_{\max} = 1\,\text{s}$，$\alpha = 0.1$

开放网络中仍然使用 3.3 节中提出的混合奖励。本文研究通过深度强化学习在封闭网络开发的控制策略能否被转移和微调，来改善现实开放网络环境中的交通状况。

训练过程中，自动驾驶车辆最初在环形道路上接受训练，其动作、观察和奖励与上述相同。然后，在预定次数的迭代之后，网络被替换为十字交叉路口，并且继续训练。

4 实验结果与分析

本节首先介绍实验的仿真环境和相关参数设置，然后在 Flow 框架[21]中对每个配置执行 3 次训练策略，选择训练得到奖励最高的策略，并使用 100 个随机种子在 100 次仿真中评估其性能，记录相关指标的平均值。

4.1 仿真环境与参数设置

实验在开源框架 Flow 进行，Flow 提供了分布式强化学习库(RLlib)[22]和微观交通仿真器(SUMO)[23]之间的接口。在 Flow 中可以通过 SUMO 创建各种交通场景，然后利用 RLlib 提供的强化学习算法为自动驾驶车辆、智能交通灯等训练生成控制策略。

本文以 Intel Core i7-8700 CPU 作为硬件环境，所有实验都使用 TD3 算法用相同的参数集进行训练。各项任务都以回合制方式进行训练，一个回合 2 000 个时隙，每个时隙长为 0.5 s。

强化学习训练过程中主要超参数如表 1 所列。

表 1 深度强化学习主要超参数

Table 1 Hyper-parameters for training

参数	值
Algorithm	TD3
Horizon	200
Stepsize/s	0.5
Optimizer	SGD
Actor Learning Rate	0.001
Critic Learning Rate	0.001
Discount Factor	0.99
Actor Hiddens	[400,300]
Critic Hiddens	[400,300]
Buffer Size	1 000 000
Policy Frequency	2
Policy Noise	0.1
Max Acceleration/(m/s^2)	2.6
Max Deceleration/(m/s^2)	−4.5
Training Iterations	200

图 4 给出了无信号灯交叉口的进入路口车流量与驶出路口车流量的关系。以每个方向 200，400，…，1 600 vehs/h 的速度生成车辆，统计车辆驶出流量。用 Flowrate 表示每个方向向路口发车的速度(vehs/h)，Outflow 表示单位时间内驶出路口的车辆数量(vehs/h)。图 4 显示，当 Flowrate 低于

400 vehs/h 时，不发生拥堵，Outflow 与 Flowrate 基本成正比；当高于 800 vehs/h 时，会发生拥堵。后续的实验配置设置为 800 vehs/h。

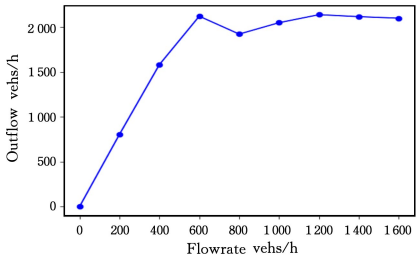

图 4 进入路口车流量与驶出路口车流量的关系

Fig. 4 Outflow vs Flowrate

4.2 算法性能分析

为了验证本文所使用的深度强化学习策略在车辆控制上应用的有效性，本文分别将混合奖励函数与目前常用的奖励函数、迁移学习与未使用迁移学习的训练方式、通用策略与定制策略进行了性能对比。

4.2.1 奖励函数的比较

这里将本文提出的混合奖励函数与目前常用的两种奖励函数在不同的自动驾驶车辆比例下进行了性能比较。

在分布式策略中，主要考虑了自动驾驶车辆对路口吞吐量性能的影响，可以看到，渗透率分别为 10%，20%，30%，…，100%，与纯人类驾驶车辆相比，路口的通行效率都有显著提升，这证明了训练得到的控制策略在提高交通效率方面的有效性。

如图 5 所示，随着强化学习车辆在道路上的比例增加，单位时间内通过路口的车辆数量都有增加。reward-mix 表示本文构造的混合奖励函数，reward-outflow 表示只使用式(3)中的奖励函数，reward-speed 表示使用式(7)中的奖励函数。对比发现，在所有渗透率下，混合奖励函数训练得到的策略提升路口吞吐量的能力优于式(3)和式(7)两种奖励函数。

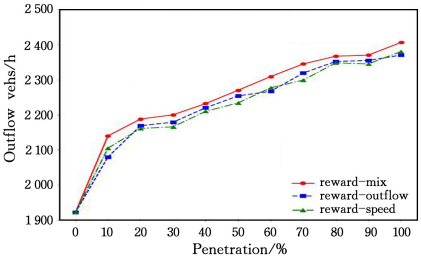

图 5 不同奖励函数的比较

Fig. 5 Comparision between different reward functions

4.2.2 通用策略与定制策略对比

分别以 $p=10\%,30\%,50\%,70\%,90\%$ 渗透率训练了 5 种不同的定制策略，以及一个通用策略。将这些策略移植到 10%～100%渗透率的环境中，比较其驶出流量性能。

如图 6 所示，所有控制器的性能在任何渗透率下都优于纯人类驾驶车辆。但是定制控制器在不同渗透率混合交通流下效果不稳定，例如，将 10%渗透率下训练得到的控制器移植到 60%渗透率的场景中，Outflow 小于 10%。通用控制器虽然不是在每一种渗透率下都能做到性能最优，但是随着自动驾驶车辆渗透率的提高，Outflow 都在增加，这说明通用控制器对渗透率具有鲁棒性。

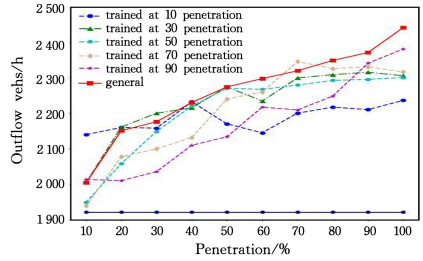

图 6 通用控制器 vs. 定制控制器

Fig. 6 General controller vs. customized controller

4.2.3 使用迁移学习与未使用迁移学习对比

图 7 给出了在自动驾驶车辆渗透率为 10%的情况下，智能体训练的收敛趋势。

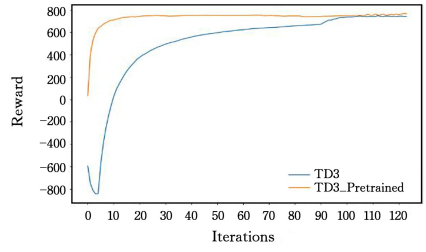

图 7 迁移学习性能

Fig. 7 Transfer learning performance

将 3.5 节中提到的迁移学习方法与在开放交叉网络中直接训练强化学习智能体进行比较，结果显示，采用在 8 字型封闭网络中迁移学习得到的策略收敛时的奖励值与从头开始训练收敛时的奖励值近似，这表明封闭道路训练得到的策略可以应用到开放道路。然而，使用了迁移学习方法之后，训练收敛的迭代次数仅仅约为不使用迁移学习的 20%，这说明迁移学习可以极大地提升无信号灯交叉路口自动驾驶车辆的训练效率。

结束语 本文研究了无信号灯交叉路口自动驾驶车辆强化学习控制策略，并提出了一个多目标奖励函数，它不仅鼓励自主车辆尽快到达交叉路口的终点，而且鼓励车辆提高交叉口的整体通行效率。实验结果表明，使用强化学习方法训练得到的驾驶策略，在平均速度、吞吐量方面相比纯人类驾驶车辆都有明显提升。

针对单一渗透率环境下训练出的策略无法适用于渗透率动态变化的环境的问题，本文研究提出了一种能适应不同自动驾驶车辆渗透率的通用控制策略，节省了训练时间和计算开销。

本文进一步研究了密度相近时封闭场景中生成的控制器可以转移到开放网络中，并且发现，即使在两个网络中使用不同的奖励函数，迁移学习也可以节省大量的训练时间。本文通过仿真实验结果证明，当自主车辆的普及率相对较低时，交叉路口的交通拥堵也可以得到缓解。

然而，仍然有一些问题没有解决。自主车辆的协同问题是一个典型的多智能体任务。在本文中，使用的是单智能体算法，它不能连接多个代理的动作。下一步的工作是将多智能体强化学习算法应用到车辆控制器中。本文研究只考虑了

单车道场景，决策中没有加入换道动作，在下一步的工作中，将考虑在多车道交叉口增加车辆换道动作，并联合控制横向和纵向动作，以提高交叉口吞吐量。

致谢 感谢这篇论文所涉及的各位学者，特别感谢本论文中文献[20]的作者 Jiaxun Cui 和她所在的团队，Jiaxun Cui 在我刚接触这一领域时，很用心的给我提供了实验指导，对于本论文的发表有比较大的贡献。

参 考 文 献

[1] MA M，LI Z. A time-independent trajectory optimization approach for connected and auto-nomous vehicles under reservation-based inte-rsection control[J]. Transportation Research Interdisciplinary Perspectives，2021，9(5)：100312.

[2] LV P，HE Y B，XU J. An Improved Trust Evaluation Model Based on Bayesian for WSNs[J]. Acta Electronica Sinica，2021，49(5)：912-919.

[3] RIOS-TORRES J，MALIKOPOULOS A A. Automated and Cooperative Vehicle Merging at Highway On-Ramps[J]. IEEE Transactions on Intelligent Transportation Systems，2016，18(4)：1-10.

[4] WANG Z，KIM B G，KOBAYASHI H，et al. Agent-Based Modeling and Simulation of Connected and Automated Vehicles Using Game Engine：A Cooperative On-Ramp Merging Study[J]. arXiv：1810.09952，2018.

[5] MAITLAND A，MCPHEE J. Quasi-translations for fast hybrid nonlinear model predictive control[J]. Control Engineering Practice，2020，97(4)：104352.1-104352.9.

[6] DING J，LI L，PENG H，et al. A Rule-Based Cooperative Merging Strategy for Connected and Automated Vehicles[J]. IEEE Transactions on Intelligent Transportation Systems，2020，21(8)：3436-3446.

[7] XIONG L，KANG Y C，ZHANG P Z，et al. Research on behavior decision-making system for unmanned vehicle[J]. Automobile Technology，2018，515(8)：1-9.

[8] KAMRAN D，LOPEZ C，LAUER M，et al. Risk-aware high-level decisions for automated driving at occluded intersections with reinfor-cement learning[J]. arXiv：2004.04450，2020.

[9] ISELE D，RAHIMI R，COSGUN A，et al. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning[C]// 2018 IEEE ICRA. Brisbane：IEEE，2018：2034-2039.

[10] XU G Y，ZONG X P，YU G Z，et al. A research on intelligent obstacle avoidance of unmanned vehicle based on DDPG algorithm[J]. Automotive Engineering，2019，41(2)：206-212.

[11] ZHANG B，HE M，CHEN X L，et al. Self-driving via improved DDPG algorithm[J]. Computer Engineering and Applications，2019，55(10)：264-270.

[12] DAI S S，LIU Q. Action Constrained Deep Reinforcement Learning Based Safe Automatic Driving Method[J]. Computer Science，2021，48(9)：235-243.

[13] SUN C Y，MU C X. Important scientific probems of multi-agent deep reinforcement learning[J]. Acta Automatica Sinica，2020，46(7)：1301-1312.

[14] SUN H，CHEN C L，LIU Q，et al. Constrained Deep Reinforcement Learning Based Safe A-utomatic Driving Method[J]. Computer Science，2020，47(2)：169-174.

[15] WEI H，LIU X，MASHAYEKHY L，et al. Mixed-Autonomy Traffic Control with Proximal Policy Optimization[C]// 2019 IEEE Vehicular Networking Conference (VNC). IEEE，2019.

[16] VINITSKY E，LICHTLE N，PARVATE K，et al. Optimizing Mixed Autonomy Traffic Flow With Decentralized Autonomous Vehicles and Multi-Agent RL[J]. arXiv：2011.00120，2020.

[17] CHEN D，LI Z J，WANG Y Q，et al. Deep Multi-agent Reinforcement Learning for High-way On-Ramp Merging in Mixed Traffic[J]. arXiv：2105.05701v1，2021.

[18] TRAN D Q，BAE S H. Proximal Policy Optimization Through a Deep Reinforcement Learning Framework for Multiple Autonomous Vehicles at a Non-Signalized Intersection[J]. Applied Sciences，2020，10(16)：5722.

[19] TREIBER M，HENNECKE A，HELBING D. Congested traffic states in empirical observations and microscopic simulations[J]. Physical Review E，2000，62(2)：1805.

[20] CUI J，MACKE W，YEDIDSION H，et al. Scalable Multiagent Driving Policies For Reducing Traffic Congestion[J]. arXiv：2103.00058，2021.

[21] WU C，KREIDIEH A，PARVATE K，et al. Flow：A Modular Learning Framework for Autonomy in Traffic[J]. arXiv：1710.05465v2，2007.

[22] LIANG E，LIAW R，NISHIHARA R，et al. Ray RLLib：A Composable and Scalable Reinforcement Learning Library[J]. arXiv：1712.09381，2017.

[23] KRAJZEWICZ D，ERDMANN J，BEHRISCH M，et al. Recent Development and Applications of SUMO Simulation of Urban MObility[J]. International Journal on Advances in Systems and Measurements，2012，12(3/4/5)：128-138.

OUYANG Zhuo，born in 1995，postgraduate. His main research interests include wireless communication theory and cooperative communications.

TAN Guo-ping，born in 1975，Ph.D，professor，Ph.D supervisor. His main research interests include Internet of vehicles，mobile edge computing，and wireless distributed machine learning.

（责任编辑：李亚辉）