https://doi.org/10.26599/JICV.2023.9210021

Multi-level objective control of AVs at a saturated signalized intersection with multi-agent deep reinforcement learning approach

Wenfeng Lin[1], Xiaowei Hu[1,✉], Jian Wang[2]

[1]*School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 15000, China*
[2]*School of Management, Harbin Institute of Technology, Harbin 15000, China*

Received: June 15, 2023; Revised: September 18, 2023; Accepted: October 9, 2023

ABSTRACT: Reinforcement learning (RL) can free automated vehicles (AVs) from the car-following constraints and provide more possible explorations for mixed behavior. This study uses deep RL as AVs' longitudinal control and designs a multi-level objectives framework for AVs' trajectory decision-making based on multi-agent DRL. The saturated signalized intersection is taken as the research object to seek the upper limit of traffic efficiency and realize the specific target control. The simulation results demonstrate the convergence of the proposed framework in complex scenarios. When prioritizing throughputs as the primary objective and emissions as the secondary objective, both indicators exhibit a linear growth pattern with increasing market penetration rate (MPR). Compared with MPR is 0%, the throughputs can be increased by 69.2% when MPR is 100%. Compared with linear adaptive cruise control (LACC) under the same MPR, the emissions can also be reduced by up to 78.8%. Under the control of the fixed throughputs, compared with LACC, the emission benefits grow nearly linearly as MPR increases, it can reach 79.4% at 80% MPR. This study employs experimental results to analyze the behavioral changes of mixed flow and the mechanism of mixed autonomy to improve traffic efficiency. The proposed method is flexible and serves as a valuable tool for exploring and studying the behavior of mixed flow behavior and the patterns of mixed autonomy.

KEYWORDS: autonomous vehicles (AVs), mixed autonomy, traffic efficiency, deep reinforcement learning (RL), saturated signalized intersection

1 Introduction

Intersections are the bottleneck that affects the traffic efficiency of urban road networks (Wu and Qu, 2022; Yu et al., 2021). Under saturated flow conditions, congestion and queue overflow are more likely to occur, resulting in greater loss of traffic efficiency (Vahidi and Sciarretta, 2018). The autonomous driving technology has been considered to have great potential in increasing road network capacity, reducing emissions, and stabilizing traffic flow over the past 10 years (Phan et al., 2020). Various control policies represented by Automatic Cruise Control (ACC) have been widely used in newly-produced high-level vehicles (Xiao and Gao, 2011), which means that it is technically feasible to directly control the actions of Automated Vehicles (AVs). Therefore, the design of control policies for AVs in mixed autonomy is particularly important and meaningful, from the perspective of system-level optimization.

Currently, some studies have focused on control methods for improving traffic efficiency at intersections with mixed autonomy. These studies can be roughly divided into two types: modeling optimization methods based on traditional mathematics and model-free methods based on Artificial Intelligence (AI).

The control idea of the modeling optimization methods can mainly include three categories: single-agent trajectory optimization (Dai et al., 2016; Han et al., 2020; Jiang et al., 2017; Zhang and Yang, 2021), multi-agent distributed cooperation (Stryszowski et al., 2021; Wang et al., 2021)), and multi-agent global planning (Fayazi and Vahidi, 2018; Feng et al., 2018; Morales Medina et al., 2018). The single-agent methods will introduce more randomness with the market penetration rates (MPR) increasing, and the multi-agent methods are usually more difficult to solve. All these methods inevitably design subjective rules such as dynamics assumptions (Meng and Cassandras, 2022) and communication structure (Jing et al., 2019; Liu et al., 2018) in the modeling process. Theoretically, high-level AVs can surpass human-driven vehicles (HDVs) at the reaction time and the decision-making ability (Chen et al., 2021). Using the modeling optimization methods may ignore some special traffic patterns due to design rules, so the results obtained by modeling optimization methods are not guaranteed to be optimal at the system level.

AI methods have an advantage over modeling optimization methods in that they can understand real-world constraints under model-free (Liu et al., 2023). Reinforcement learning (RL) can decouple the dynamic characteristics of AVs (Zhou et al., 2020), and its Markov characteristics also have good adaptability to the optimization problem of sequential decision-making (Aragon-Gómez and Clempner, 2020). Deep Reinforcement Learning (DRL) methods developed from RL have also been used in some

✉ Corresponding author.
E-mail: xiaowei_hu@hit.edu.cn

related works (He et al. 2023; Kiran et al., 2022). Most of them are the optimization of traffic signal (Aslani and Mesgari, 2017; Mushtaq et al., 2021; Zhang et al., 2021), and the optimization of one vehicle's intelligent trajectory on the road (Ding et al., 2022; Hoel et al., 2019; Makantasis et al., 2020; Ye et al., 2019). Some of works applied the DRL at AVs are oriented to the fully AVs flow (Li et al., 2020), and other works are oriented to the mixed traffic autonomy, which focuses on the conflict resolution at non-signalized intersection (Isele et al., 2017; Guan et al., 2020).

There are two challenges in applying multi-agent DRL to saturated signal intersections. Firstly, although the multi-agent architecture can improve the learning ability of RL, it poses new convergence challenges in complex environments. Secondly, mixed vehicles' behaviors are different from traditional driving behaviors. Constructing saturated traffic conditions at a single intersection is difficult without constraints, due to queued overflow upstream and saturated input downstream.

This study designs a multi-level objectives framework for AVs' trajectory decision-making based on multi-agent DRL. The proposed method can realize the flexible multi-level objective control, and we used it to explore the upper limit of the throughputs during the green phase of saturated signalized intersections. It can make decisions based on predetermined objectives.

The rest of this paper is organized as follows: Section 2 introduces the decision-making framework. Section 3 introduces the environment selection and the setting parameters of the simulation experiments. In Section 4, we show the results and discussion of the experiments results, including the discussion of the algorithm convergence performance, the superiority, and the flexibility of the proposed RL method. According to the results, some phenomena worth discussing are also be found. In Section 5, we conclude the paper and propose future research.

2 Method

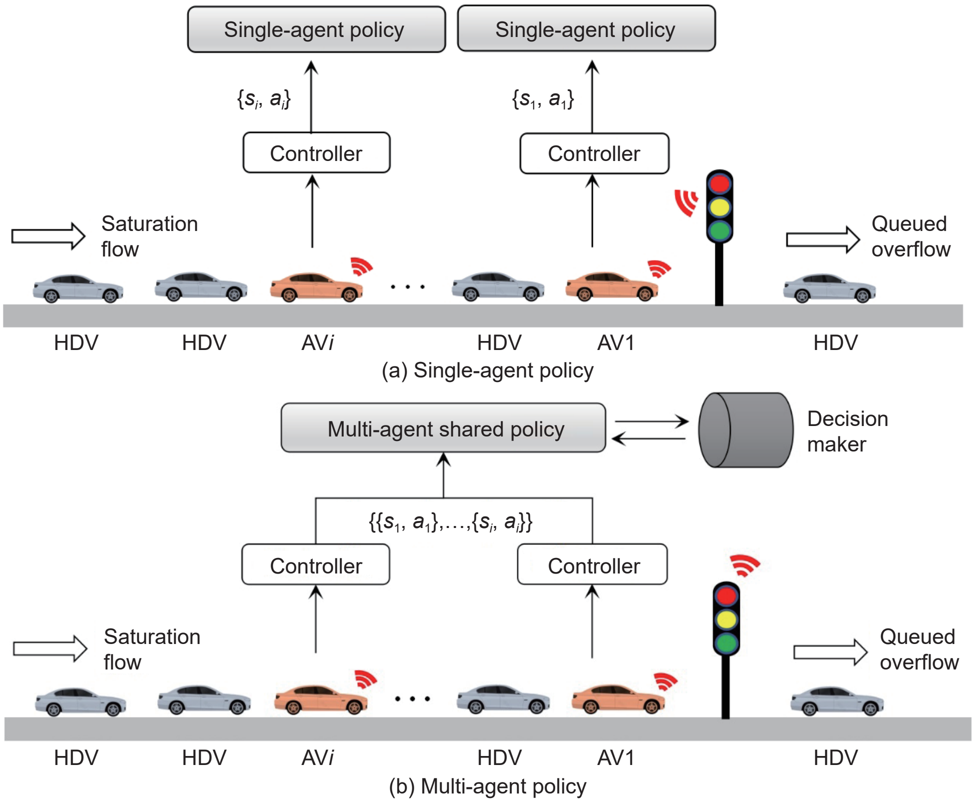
2.1 Markov decision processe (MDP) problem statement

This work considers a simplest urban four-way saturated signalized intersection, which only allows straight travel. When there is only one lane in the same direction, vehicles cannot make lateral lane changes. The acceleration action along the longitudinal direction of the lane can be regarded as the dynamic behavior of the vehicle. We used a DRL algorithm to control the acceleration of AVs in mixed autonomy and a classical micro car following model to simulate HDVs. We used the intelligent driver model (IDM) for HDVs:

$$a_{\text{IDM}} = a\left[1-(v/v_{\text{o}})^{\delta}-\left(s^{*}(v,\Delta v)/s\right)^{2}\right] \tag{1}$$

where a_{IDM} is the acceleration of a HDV; a is the maximum acceleration of the HDV; v is the speed; v_{o} is the expected speed; δ is the acceleration index; s is the distance from the front vehicle; $s^{*}(v,\Delta v)$ is the expected following distance.

Fig. 1 shows the difference between the decision-making mechanisms of single-agent and multi-agent at a saturated signalized intersection with mixed autonomy. Multi-agent decision-making is more flexible and comprehensive for complex tasks aimed at improving traffic efficiency through system-level optimization compared to single-agent decision-making. In the multi-agent architecture, system-level objectives like total throughputs can be added to the multi-level objectives, besides the individual-level objectives of the vehicle's own trajectory. The AVs have specific perception capabilities while in a queue, including speed and distance. The road infrastructure is equipped to receive this information, as well as the current traffic phase. In addition, this work does not consider the cooperation of AVs, which means any other AVs will regard other AV as HDV units.



(a) Single-agent policy

(b) Multi-agent policy

Fig. 1 Difference decision-making mechanisms between the single-agent and multi-agent. (a) Single-agent policy, (b) multi-agent policy.

https://doi.org/10.26599/JICV.2023.9210021

The total environment is modeled as MDP, which is defined as $\{S, A, P, R, \gamma\}$. At each time step t, the state of the environment is $s_t \in S$. The agent obtains the state s_t by observing the environment, and then executes an action $a_t \in A$ according to its own policies. Performing this action will enable the agent to get the next state according to the state transition probability $s_{t+1} \sim P(s_{t+1}|s_t, a_t) \in [0, 1]$, and the environment will feed back the agent's immediate rewards through the defined reward function $r(s_t, a_t, s_{t+1})$.

The aim of the agent is to solve the MDPs and obtain the optimal policies. By sampling, it can maximize its cumulative discount expected reward. The agent's policies are defined asπ_θ: $S \to A$, which can be parameterized by θ. The cumulative reward function Q^π is defined as shown in Eq. (2), and the optimal policy π_θ^* can be calculated from the expectation of the reward as Eq. (3):

$$Q^\pi(s_t, a_t) = r_t + \gamma^t \max_{a_{t+1}} Q^\pi(s_{t+1}, a_{t+1}) \tag{2}$$

$$\pi_\theta^* = \underset{\pi}{\mathrm{argmax}} \mathbb{E}_{s_t \sim \rho^\pi, a \sim \pi_\theta} \left[Q^\pi(s_t, a_t) \,|_{a_t \sim \pi_\theta(a_t|s_t), s_0} \right] \tag{3}$$

where the discount factor γ is introduced to express that the longer the time, the less impact the reward; ρ_π is the stable distribution of the MDPs under the policy π.

2.2 Proximal policy optimization algorithm

The goal of the standard RL is to maximize the cumulative expected reward of the agent (Eq. (2)). The gradient descent algorithm is used to calculate the estimator of the parameterized of the policy. Compared with the value-based RL, the advantage of policy-based RL is that the algorithms can learn random policies. which are constructive for non-communicating agent environments. In addition, the policy-based RL can also avoid the convergence problem of some estimation function due to nonlinear approximation and partial observation.

According to the Deterministic Policy Gradient (Silver et al., 2014), when the agent's policy is deterministic and the action space is continuous, there is a deterministic policy. The gradient of policy parameters $J(\cdot)$ is shown as Eq. (4) and Eq. (5) is the update rule of $J(\cdot)$:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s_t \sim \rho^\pi, a \sim \pi_\theta} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \, Q^\pi(s_t, a_t) \,|_{a = \pi_{\theta(s)}} \right] \tag{4}$$

$$\theta_{\mathrm{new}} \leftarrow \theta_{\mathrm{now}} + \beta \cdot \nabla J(\theta_{\mathrm{now}}) \tag{5}$$

where β is the learning rate of the policy gradient $J(\cdot)$; θ is the parameter of policy π_θ.

In training process, the Actor-Critic (AC) framework (Bhatnagar et al., 2009) can be used to reduce the variance caused by insufficient interaction. AC framework has two types of deep neural networks. The actor-network is used to approximate the policy function $\mu_\theta \to \pi_\theta$ and perform interaction with the environment and produce actions. The critic-network is used to approximate evaluate the value function $Q_w \to Q^\pi$ and guide the policy promotion. For the actor-network, the policy gradient is still used to update the parameters. For the critic-network, the loss function $L(\omega)$ is as Eqs. (6) and (7):

$$L(\omega) = \left[Q_\omega(s_t, a_t) - Q_\omega(s_{t+1}, a_{t+1}) \right]^2 \tag{6}$$

$$Q_\omega(s_{t+1}, a_{t+1}) = r_t + \gamma^t Q_\omega\left(s_{t+1}, \mu_\theta\left(s_{t+1}|\theta^\mu\right) |\omega^Q\right) \tag{7}$$

where the approximate value function Q_ω is parameterized by the parameter ω; the policy functionπ_θ is parameterized by μ_θ.

Policy gradient is on policy, and its data utilization is inefficient. It can add a replay buffer to AC, and use sampling to convert on-policy to off-policy to improve the sample utilization (Wang et al., 2016). The AC framework with a replay buffer is shown in Fig. 2.

For Eq. (6), the advantage function A^π can be used to approximate the cumulative reward function Q^π. A^π is designed as

$$A^\pi(s_t, a_t) = Q_\omega(s_t, a_t) - V_\omega(s_t) \tag{8}$$

where $V_\omega(s_t)$ is the expected reward function according s_t.

To reduce the variance of Q^π, the generalized advantage estimator (GAE) (Schulman et al., 2015) can be used to estimate the advantage function A^π:

$$\begin{aligned} \hat{A}_t^{\mathrm{GAE}(\chi,\lambda)} &= (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \cdots\right) \\ &= \sum_{l=0}^{\infty} (\chi\lambda)^l \left(r_{t+1} + \chi V(s_{t+l+1}) - V(s_{t+l})\right) \end{aligned} \tag{9}$$

where χ is discount rate; $\lambda \in [0, 1]$ is the step size of the updates.

Proximal Policy Optimization (PPO) (Schulman et al., 2017) is a widely used RL algorithm. Compared with the classic policy gradient RL algorithms with AC architecture such as TRPO, PPO is a first-order optimization algorithm that utilizes Generalized Advantage Estimation (GAE) to achieve faster convergence than classic policy gradient RL algorithms with AC architecture such as Trust Region Policy Optimization (TRPO). Furthermore, PPO is straightforward to implement. PPO is effective in high-dimensional state space environments and has been proven to perform well in specific multi-agent engineering tasks. PPO enhances the architecture displayed in Fig. 2 by adding an actor network and Algorithm 1 shows the pseudocode for PPO.

Algorithm 1 Proximal policy optimization

```
For $i \in \{1, \cdots, N\}$ do
    Run policy $\pi_{\theta_{\mathrm{old}}}$ for $T$ timesteps, collecting$\{s_t, a_t, r_t\}$
    Estimate advantage $\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_T$
    $\theta_{\mathrm{old}} \leftarrow \theta$
    For $j \in \{1, \ldots, M\}$ do
        $J_{\mathrm{PPO}}(\theta) = \sum_{t=1}^{T} \left(\pi_\theta(a_t|s_t) / \pi_{\theta_{\mathrm{old}}}(a_t|s_t)\right) \hat{A}_t - \lambda KL[\pi_{\theta_{\mathrm{old}}}|\pi_\theta]$
        Update $\theta$ by a gradient method w.r.t $J_{\mathrm{PPO}}(\theta)$
    end for
    For $j \in \{1, \ldots, B\}$ do
        $L_{BL}(\phi) = -\sum_{t=1}^{T} \left( \sum_{t'>t} \gamma^{t'-t} r_{t'} - V_\varphi(s_t) \right)^2$
        Update $\theta$ by a gradient method w.r.t $L_{\mathrm{BL}}(\phi)$
    end for
    if $KL\left[\pi_{\mathrm{old}}|\pi_\theta\right] > \beta_{\mathrm{high}} KL_{\mathrm{target}}$ then
        $\lambda \leftarrow \alpha\lambda$
    else if $KL\left[\pi_{\mathrm{old}}|\pi_\theta\right] < \beta_{\mathrm{low}} KL_{\mathrm{target}}$ then
        $\lambda \leftarrow \lambda / \alpha$
    end if
end for
```

2.3 Design of multi-level objective decision-making process for multi-agent

This work integrates the decision-making process into the flexible multi-level objectives framework, based on the PPO algorithm. The framework is as shown in Fig. 3.

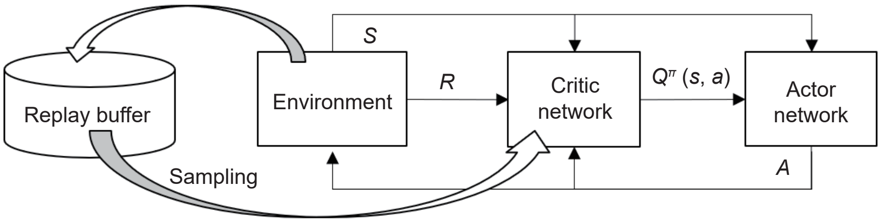


Fig. 2 AC framework with a replay buffer.

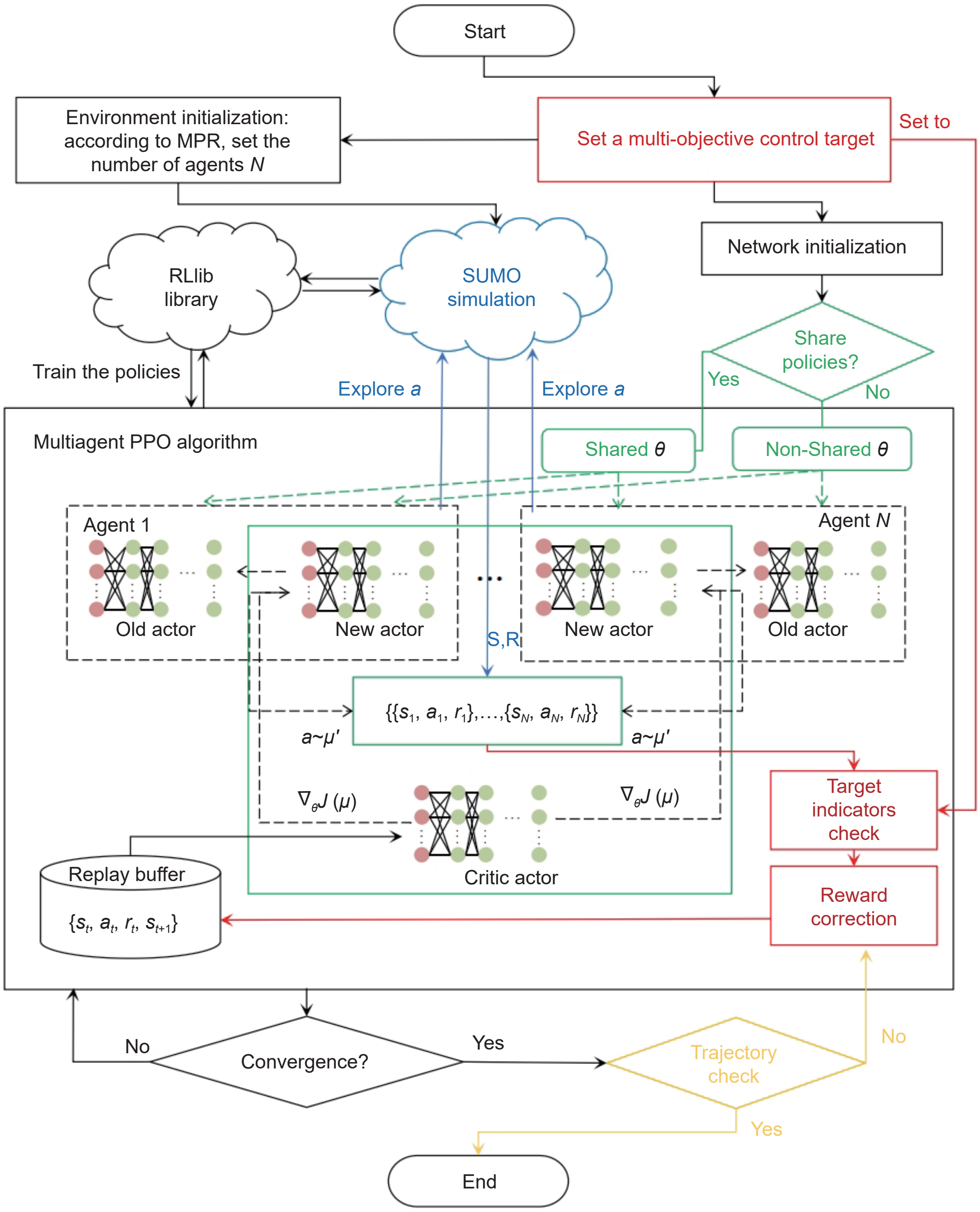


Fig. 3 Framework of the proposed RL method.

Firstly, we set a multi-level objective (such as throughputs, fuel consumptions, emissions, and average speed) for the multi-agents and assign these objectives to the reward correction unit. Secondly, the number of agents N is set according to MPR, and the training networks are initialized according to whether the policies are shared between agents. Then, the policies are learned through the interaction between the simulation environment and the multi-agents. Finally, the trajectories are checked before convergence.

More specifically, this work introduces the following units to the standard PPO algorithm for solving the MDPs.

2.3.1 Multi-agent framework unit

Multi-agent framework is based on the idea of centralized training and decentralized exploration. Centralized training is carried out

https://doi.org/10.26599/JICV.2023.9210021

with global information, while each agent interacts with the environment in independently manner. We assumed there were no communication rules between agents, and we did not consider the scenario where an agent malfunctions.

As shown in the green part of Fig. 3, each agent is equipped with two actor-networks, and shares the same critic-network. When the agent shares policies, the actor-networks of all agents share their parameters. Each agent only uses the local information it can observes to choose actions according to its own actor-networks. The critic-network uses the information of all agents that can be observed to update.

For N agents, the gradient of policy parameters θ of the agent is

$$\begin{aligned}\nabla_{\theta_i} J\left(\mu_{\theta_i}\right) = {} & \mathbb{E}_{s\sim\rho^{\pi},a_i\sim\mu_{\theta_i}} \times \\ & \left[\nabla_{\theta_i}\log\mu_{\theta_i}\left(a_i|s_i\right) Q_{\omega_i}\left(\{s_i\}, a_1, \cdots, a_N\right)|_{a_i=\mu_{\theta_i}(s_i)}\right]\end{aligned} \tag{10}$$

The replay buffer records the experience of all agents, and the critic-network is updated as Eqs. (11) and (12):

$$L\left(\omega_i\right) = \mathbb{E}_{\{s_i\}\sim\rho^{\mu_{\theta_i}},a_i\sim\mu_{\theta_i}}\left[\left(Q^{t}_{\omega_i}\left(\left\{s^{t}_{i}\right\}, a^{t}_{1}, \cdots, a^{t}_{N}\right) - Q^{t+1}_{\omega_i}\right)^2\right] \tag{11}$$

$$Q^{t+1}_{\omega_i} = r^{t}_{i} + \gamma^{i} Q^{t+1}_{\omega_i}\left(\left\{s^{t+1}_{i}\right\}, a^{t+1}_{1}, \ldots, a^{t+1}_{N}\right) \tag{12}$$

2.3.2 Reward correction unit

The method of single-agent DRL multi-objective control usually designs a weighted reward function. The reward is

$$r_t = \sum\nolimits_{n} \eta_n r^{n}_{t} \tag{13}$$

where η_n is the weighted parameters of the n-th control objective; and r^{n}_{t} is the instantaneous reward related to the n-th control objective.

There are two issues with using the weighted reward.

Firstly, the linear weighted reward function has the risk of local convergence or non-convergence at the multi-agent environment. To illustrate the problem, we analyze a control scenario with multi-level objectives, where throughput is considered at the first level and acceleration change at the second level. If the weight of the instantaneous reward about the throughputs is set to the maximum, the agent will give priority to explore the action that increases the throughputs, while ignoring other actions like the speed. Although the higher the speed usually means the higher the throughputs, the agent cannot link the two variables together. During the initial stages of training, the effects are quite prominent. Even the slightest alteration in acceleration may not be adequate to increase throughputs. In fact, it may result in a penalty due to the acceleration change, which will accumulate until the end of the simulation. Consequently, the agent may choose to halt as it cannot discover efficient actions to improve throughputs.

Secondly, the throughputs increase monotonically with respect to the traffic time, while the change of acceleration is fluctuating in a relatively stable interval. A practical method for achieving a balance of various objectives involves developing a segmented function for the throughputs.

To solve the problems, we add a reward correction unit (the red part in Fig. 3) in the design-making progress. After finishing an episode, the agent checks the realization the multi-level objectives, and corrects the accumulated rewards according to

$$R_i = \prod_{l} \eta'_{l}\left(\varphi^{l}_{\text{record}} - \varphi^{l}_{i}\right) \sum_{t\geq 0} r^{i}_{t} \tag{14}$$

where R_i is the cumulative reward of the i-th episode; r^{i}_{t} is the instantaneous reward of the i-th episode at t step; η'_{l} is the penalty coefficient of the l-level objective; $\varphi^{l}_{\text{record}}$ is the expected value of the l-level objective.

Through the reasonable design of $\varphi^{l}_{\text{record}}$ and δ_l, flexible multi-level objectives control can be realized. If the first-level objective is to maximum one index, $\varphi^{l}_{\text{record}}$ can record the maximum value and keep update during the training process; if there has a specific control index, $\varphi^{l}_{\text{record}}$ can be designed to the index. η'_{l} can be designed as positively correlated with l.

In addition, when setting multi-level control objectives, it is important to consider the potential relationship between indicators to avoid contradictory settings that could affect algorithm efficiency.

2.3.3 Trajectory check unit

The agent should prioritize safety as a hard constraint for all vehicle actions. In the modeling optimization methods, a minimum safety distance is usually designed as a constraint, but the distance is also not easy to design properly. A small distance cannot guarantee the safety risks caused by the reaction time of HDVs. A large distance can ensure safety, but the vehicles cannot achieve more active car-following behaviors. Especially for AVs, an improperly safety distance may cause the algorithm to ignore the optimal solution located at the boundary of the constrained space.

Considering the aforementioned differences, we designed a minimum safety distance of 2.5 m for HDVs to comply with its dynamic characteristics. Simultaneously, we set the minimum safety distance of AVs to 0.1 m to encourage the behaviors of accelerating and exploring. The value of AVs' minimum safety distance is small compared to that of HDVs. The purpose of the small distance is to help the RL agent make the most efficient following actions during trajectory planning, which increases traffic efficiency. Under a reasonable control model, the smaller the distance, the higher the upper limit of performance improvement. Based on a reasonable control model, AV can be flexibly controlled to achieve its minimum safety distance.

To reconfirm the rationality of the planning scheme in terms of safety, we designed that RL can use penalty to assurance the constraint of safety. When a collision occurs, a large penalty will be immediately obtained by agents and the current simulation episode will be ended.

We also added a trajectory check unit to guarantee the safety, as shown in the yellow part of Fig. 3. For the episodes where the simulation is close to convergence, safety checks can be performed at the end of each episode based on the spatial–temporal trajectories. If the trajectories are not meeting the requirements of the safety index, we use the reward correction unit to correct the cumulative rewards. Then we put the experiences into the replay buffer for retraining.

In addition, based on the same considerations, it is not advisable to incorporate comfort objectives into the multi-level objectives.

3 Experimental

We use an open-source framework called "Flow" (Wu et al., 2017) to link the transport simulator SUMO (Lopez et al., 2018) and python library Rllib (Liang et al., 2017). We visualize the AV policies in SUMO, and interact the simulation data with the agent within each step. After a certain period of learning iteration, the control objectives can be achieved. In addition, this work chooses the discount factor γas large as possible (close to 1) to approximate the non-discount problem.

3.1 Simulation environment

This work takes saturated signalized intersections as the research object, focusing on the through process of the vehicles in an entrance lane during a green phase. In order to improve the efficiency of the algorithm and reduce hardware consumption, we need to use a simpler simulation environment, including shorter simulation time and fewer agents.

The simulation of signalized intersections generally requires input control at the signal entrance. At a single intersection, no input distribution can accurately reproduce the saturated traffic behavior. The queued overflow phenomenon at the exit cannot be achieved by imposing constraints.

To accurately simulate saturated signalized intersections, we need to set up an environment with at least 3 consecutive signalized intersections. The exit of the first intersection is also the entrance of the second intersection, and the entrance of the third intersection is also the exit of the second intersection. Then, the control of the entrance and exit sections in the second intersection is equivalent to the saturation constraint control. On these settings, we can only focus on the control policies of the vehicles on the middle section of the road through the second signal intersection.

As shown in Fig. 4, we only focus on the second intersection along the advancing direction of the flow, whose trajectories are highlighted by a red box. The blank at the bottom left of the box is due to vehicle delays caused by the signal, and the blank at the bottom right is because the number of vehicles entered is insufficient. We can fill in the bottom right blank by increasing the number of continuous signals, so we conducted a simulation with 5 consecutive intersections. As shown by the blue box in Fig. 4, it is the spatial-temporal trajectories of the fifth intersection along the direction of the flow.

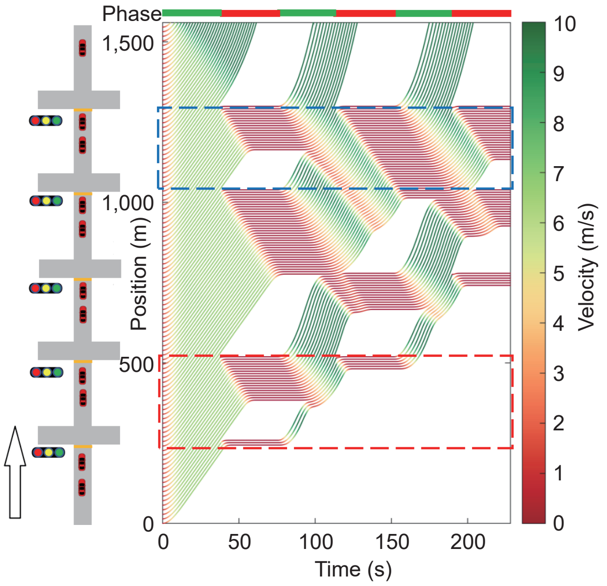


Fig. 4 Spatial–temporal trajectories of 5 continuous saturated signalized intersections. Vehicles are equally distributed on all road sections at the saturation density and start synchronously, at the beginning of the simulation.

Through Fig. 4, it can be seen that if we want to fully simulate the vehicles' behavior during a duration of green phase under saturation constraints, at least 5 consecutive saturated signalized intersections need to be designed. The second signal cycle of the fifth intersection can be used as the observation object of the study. If the number of vehicles on each road segment is N, the number of agents at 100% MPR is $5N$. The computational burden is even more of an exponential order of $5N$.

All vehicles are HDVs, which using the IDM model as the car-following model. All intersections have the same fixed signal cycle. The yellow time is included in the duration of green phase.

To simplify the problems, we consider finding an alternative environment. Inspired by the work in Kreidieh et al. (2018), their work confirmed that the stop-and-go waves in a ring road are similar to that of the road section. We designed a ring road with a signal and conducted a simulation of the same traffic conditions as Fig. 4 again. The spatial–temporal trajectories of the ring road are shown in Fig. 5.

Comparing Figs. 4 and 5, it can be seen that the spatial-temporal characteristics of the vehicles in a signalized ring road are completely similar to that of a saturated signalized intersection. In addition, except the first signal cycle is slightly different, the second and third signal cycles are also completely similar, both of which are the same as the saturated signalized intersections. The differences in the first cycle may be caused by the different initial acceleration conditions of the two environments, which can be ignored.

Opting for a signalized ring road can result in a substantial reduction in the number of agents, up to five times less compared to saturated signalized intersections. Moreover, the spatial-temporal aspect of a signalized ring road is akin to that of a saturated signalized intersection. Hence, in this work, a signalized ring road is employed as a simulation environment instead of saturated signalized intersections.

3.2 Experiment settings

To perform experiments in a signalized ring road, we designed MDPs elements, simulation settings, and hyper-parameters of the training networks.

3.2.1 MDP elements

State. The State Space $Space = \{x_i, v_i, d_i, L\}_{i \in AV}$ includes the collection of the spatial position, speed, distance from the preceding vehicle, and signal control state (period and phase) of all AVs in the environment. Because in this work we do not consider

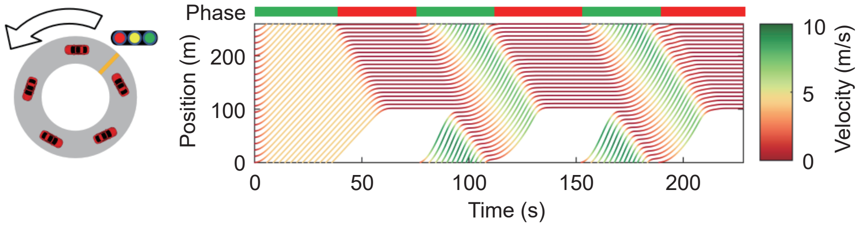


Fig. 5 Spatial–temporal trajectories of a signalized ring road. The signal cycle is set to the same as the experiment in Fig. 4.

https://doi.org/10.26599/JICV.2023.9210021

cooperative driving, no relevant states are designed for AVs.

Action. The Action Space Action $= \{a_i\}_{i \in AV}$ consists of the collection of the acceleration vectors of all AVs. At each time step, RL-agent can modify the real-time acceleration of the AVs in its control range. If the acceleration interval of the vehicle is limited to $[a_{\min}, a_{\max}]$, and its action is $a_i \in [a_{\min}, a_{\max}]$.

Reward. To encourage high throughputs and punish high acceleration, we design the reward function as Eqs. (15) and (16):

$$Reward = (\|v_{\text{des}}\| - \|v_{\text{des}} - v_t\|)\sum_{T_{\text{green}}} n_{\text{through}} + \alpha P \tag{15}$$

$$P = v\left(Mva + Mg\left(f_r + i\right) + 0.5\rho_a C_D A_f (v + v_{\text{wind}})^2\right) \tag{16}$$

where v_{des} is the desired speed; v_t is the speed of the AVs in the environment at step; n_{through} is the number of vehicles passing the green phase T_{green}; $\alpha$is the weight factor; and P is the instantaneous emission, which is defined by Eq. (16) (Abousleiman and Rawashdeh, 2015).

The meanings and values of symbols in Eq. (16) are shown in Table 1.

Table 1 Coefficients of Eq. (8)

Description	Symbol	Value	Unit
Aerodynamic drag coefficient	C_D	0.3	—
Frontal area	A_f	2.6	m^2
Rolling resistant coefficient	f_r	0.005	—
Vehicle mass	M	1,200	kg
Gravitational acceleration	g	9.8	m/s^2
Air mass density	ρ_a	1.225	kg/m^3
Rotational inertia dactor (mass factor)	v_{wind}	0	—
Regenerative braking factor	δ	1	—

3.2.2 *Simulation settings*

Detailed settings of the simulation settings are shown in Table 2 and the parameters of the IDM model used throughout this paper are shown in Table 3. When selecting vehicle parameters, we tended to set thresholds as high as possible to increase randomness and flexibility. Expanding the action spaces can enhance traffic efficiency to a greater possibility. Our determination of parameter values was primarily based on the work of Treiber et al. (2000).

Table 2 Simulation settings

Parameter	Value	Unit
Time step	0.1	s
Road length	260	m
Vehicle length	5	m
Minimal gap for HDVs	2.5	m
Minimal gap for AVs	0.1	m
Max vehicle speed	40	km/h
Acceleration ability of vehicles	3.5	m/s^2
Deceleration ability of vehicles	−3.5	m/s^2
Duration of green phase (including yellow time)	35	s
Duration of red phase	35	s

3.2.3 *Hyper-parameters of the networks*

Table 4 shows the detailed settings of hyper-parameters for the networks at the experiments.

Table 3 Parameters of the IDM model used throughout this paper

Parameter	Value	Unit
Desired velocity	40	km/h
Maximum acceleration	3.5	m/s^2
Acceleration index	4	—
Expected following distance	2.5	m

Table 4 Hyper-parameters settings

Hyper-parameter	Value
Actor- networks and critic-networks	3 layers with 32 units each and ReLU non-linearities
Batch size	20 × 1,520
Policy initialization	Standard Gaussian
Discount factorγ	0.999
Learning rate	0.01

4 Results and discussion

This work evaluates the proposed RL method from three aspects: (1) algorithm convergence, (2) superiority: a case to find max throughputs under the saturation flow and a case to find max throughputs under saturation flow to prove the superiority of the proposed RL method, and (3) flexibility: a case to achieve the specific throughputs under the saturation flow to prove the flexibility of the proposed RL method.

4.1 Algorithm convergence

Fig. 6 shows reward-curve of the standard PPO and the proposed RL method. Generally speaking, the proposed RL method is not prone to local optima or non-convergence. The proposed RL method is more stable, and its curves will not appear a sudden drop like the curve of MPR = 40% using standard PPO. While the standard PPO has a risk of local convergence and non-convergence in the complex tasks (MPR = 60% and 80%).

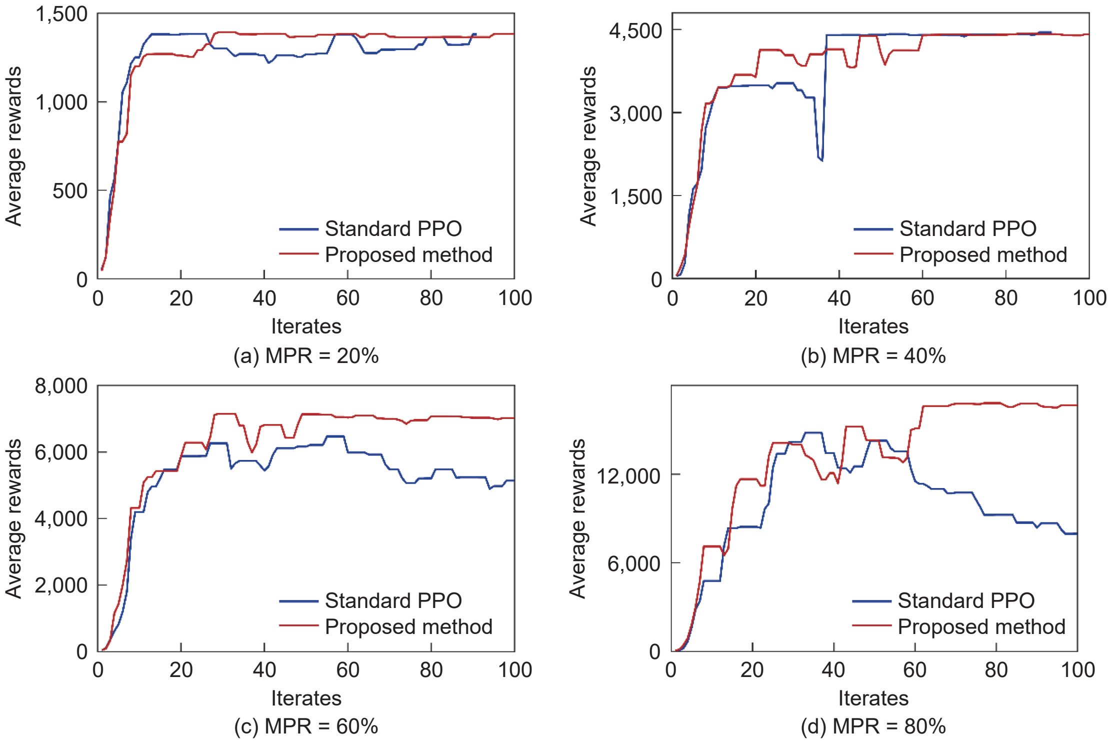
In the simple tasks (MPR = 20% and 40%), the efficiency of the two algorithms is similar, which can converge within 100 iterations. The jagged fluctuation of the proposed RL method's reward curve in the early stage of training is caused by the secondary punishment of the accumulated reward. The curves of the proposed RL method become stable when the training process enters the convergence stage.

The proposed algorithm is more computationally efficient than standard PPO. This is due to the secondary dynamic reward correction (Eq. (14)), which adjusts based on real-time accumulated training experience. It's similar to incorporating the knowledge of experts to expedite the process of exploration and training.

4.2 Superiority: A case to find max throughputs under the saturation flow

Adaptive cruise control (ACC) can actively avoid congestion and is often used as AVs' following model to study the mixed flow behavior of mixed autonomy (Kesting et al., 2008). Linear adaptive cruise control (LACC) is used as benchmarks for performance comparison with our RL approach, since our research does not consider the cooperation between vehicles. We set a two-level control objective for RL agent to explore the upper limit of the throughputs: The first level is maximum the throughputs, and the second level is the minimum emissions.

Fig. 7a shows the throughputs using RL-Control and LACC-



(a) MPR = 20%

(b) MPR = 40%

(c) MPR = 60%

(d) MPR = 80%

Fig. 6 Reward-curve of the standard PPO and the proposed RL method. (a) MPR = 20%, (b) MPR = 40%, (c) MPR = 60%, (d) MPR = 80%.

Control at various MPR. Fig. 7b shows the throughput benefits of using RL-Control at various MPR. Fig. 8 shows the emissions and benefits of the emissions using RL-Control and LACC-Control at various MPR. To eliminate randomness, the data in Figs. 7a and 7b are the mean of 10 simulation results.

As shown in Fig. 7a, RL-Control has a obvious improvement on the throughputs compared with LACC-Control. When MPR is 100%, 22 vehicles both can pass through the signal at a cycle. LACC-Control will fall into a “control bottleneck” in the range of MPR = 40%–70%, which will not significantly increase the throughputs.

As shown in Fig. 7b, with the increase of MPR, the throughput benefits have linearly improvement compared with the fully HDVs (MPR = 0%). When MPR is 100%, the throughput benefits can reach 69.2%.

The throughput benefits also have improvement compared with the LACC-Control with the same MPR. The curves of the benefits show a segmented fluctuation rising pattern, and there both exit a jump in the range of MPR = 30%–40%, 70%–80%.

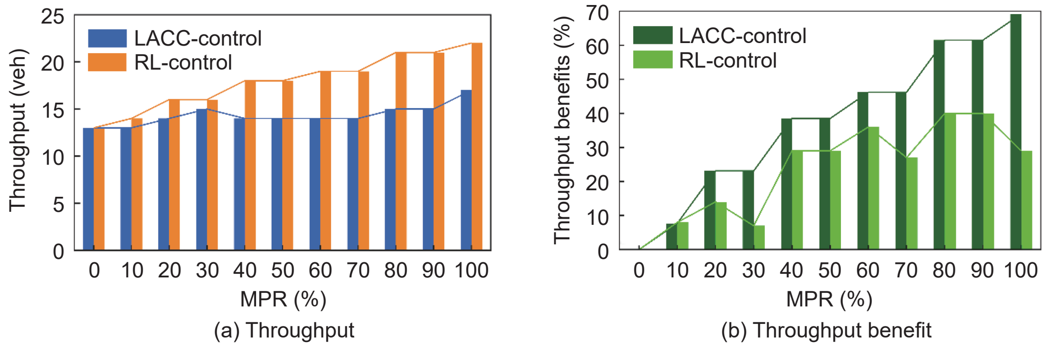
As shown in Figs. 8a and 8b, with the increase of MPR, the emission benefits of RL-Control have near linearly improvement compared with LACC-Control at the same throughputs, even if the emissions are set as the second-level multi-target setting.

Compared Figs. 7 and 8, LACC-Control can achieve a certain degree of throughput increase when MPR is greater than 70%, but it is at the expense of high emissions. The poor performance of LACC-control at high MPR is also consistent with our inference in Fig. 1: the method based on single-agent control is not suitable for complex conditions.

In addition, the proposed RL method is based on multi-agent, and we do not set cooperation rules for agents. The benefits under the high MPR show that the agent seems to have explored a certain undesigned cooperation rule, which also shows the great potential of the proposed RL method to adapting to the complex conditions.

To analyze the spatial-temporal characteristics at mixed autonomy in more detail, we show the spatial-temporal trajectories of LACC-Control and RL-Control at various MPR in Figs. 9–11. According to the conclusions in in Section 3.1, we only focus on the spatial-temporal characteristics of the mixed flow in the second signal cycle.

Generally speaking, whether using RL-Control or LACC-Control, the AVs tend to maintain a smaller headway to improve the traffic efficiency. While the throughputs are increasing, the



(a) Throughput

(b) Throughput benefit

Fig. 7 Improvements of throughput at various MPR. (a) Throughput, (b) throughput benefit.

https://doi.org/10.26599/JICV.2023.9210021

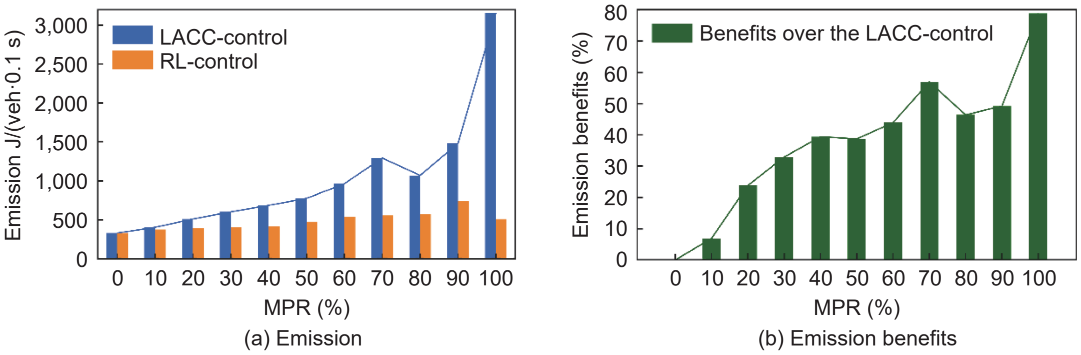


Fig. 8 Improvements of emissions at various MPR. (a) Emission, (b) emission benefits.

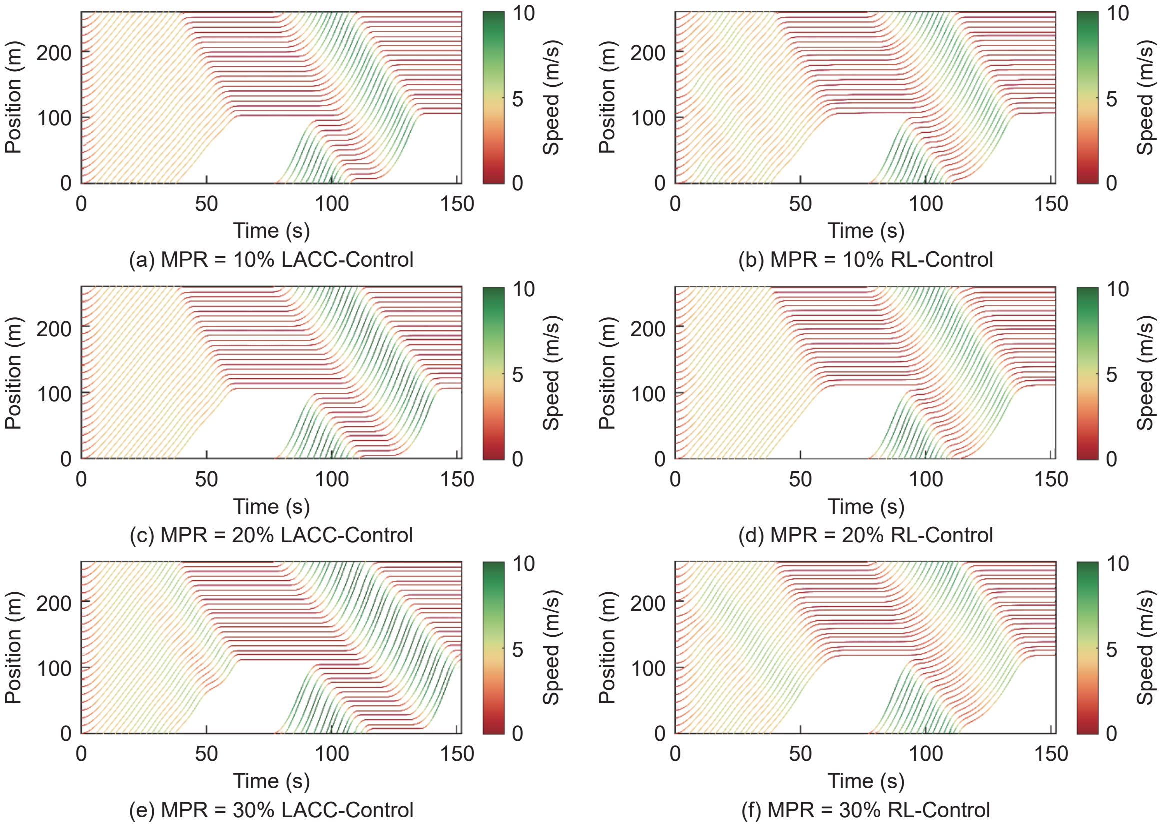


Fig. 9 Spatial–temporal trajectories of experiment at various MPR (10%–30%). (a) MPR = 10% LACC-Control, (b) MPR = 10% RL-Control, (c) MPR = 20% LACC-Control, (d) MPR = 20% RL-Control, (e) MPR = 30% LACC-Control, (f) MPR = 30% RL-Control.

speed of passing the intersection is also improving.

RL-Control can eliminate the stop–and–go fluctuations on the road to a certain extent and it has the higher comfort of the trajectories, compared with LACC-Control.

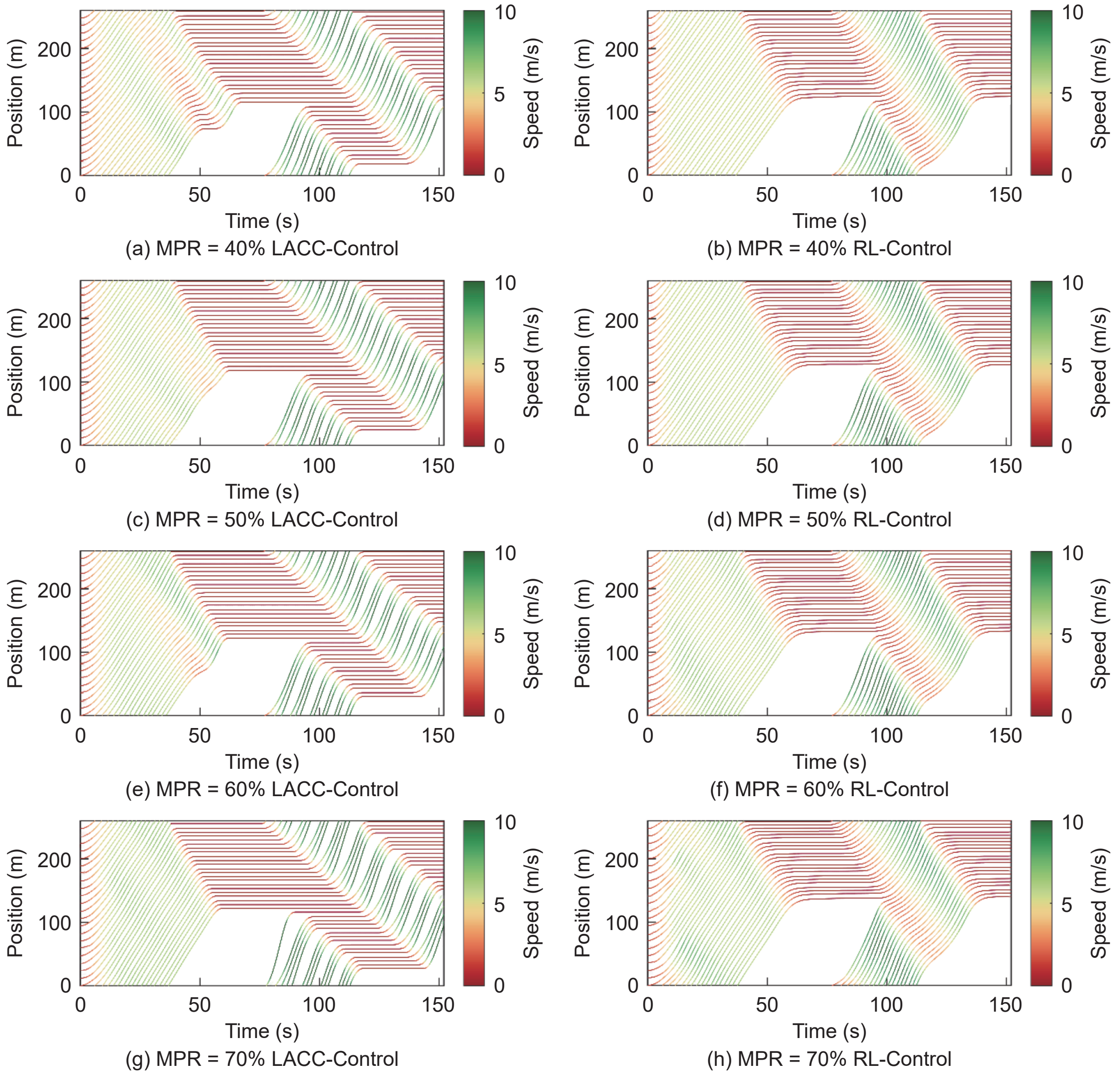
As shown in Fig. 9, when MPR is less than 30%, Compared with LACC control, there is little difference in the distribution of overall spatiotemporal trajectories. This shows that it is with the similar mechanisms to improve traffic efficiency when use RL-Control or LACC-control as an AVs car-following model. RL-Control adopts a smoother acceleration change when passing the green phase. The optimization is without sacrificing throughput, so it can be used as an eco-driving policy.

As shown in Fig. 10, when MPR is in the 40%–70% range, the spatial-temporal trajectories of LACC-Control and RL-Control begin to show obvious differences. LACC-Control will cause large waves, which is caused by the acceleration principle of LACC. The actions of the AVs in pursuit of a smaller distance will increase the waves. This means that when the MPR is greater than 40%, it is necessary to consider the system-level optimizing, such as RL-Control, in order to reduce waves.

As shown in Fig. 11, when the MPR is higher than 80%, the spatial-temporal trajectories of LACC-Control and RL-Control gradually become analogously again. It means that the choice of control policies has less influence on mixed traffic behavior. There are easier for AVs to achieve the effect of cooperation. From the spatial-temporal trajectories, the cooperation mode may be the synchronized acceleration and deceleration. Similar to MPR less than 30%， it is with the similar mechanisms to improve traffic efficiency when use RL-Control or LACC-control as an AVs car-following model. However using RL-Control is much better than using LACC-control.

In addition, as shown in Fig. 8, when MPR reaches 100%, two interesting phenomena are worth discussing:

1) Although both 22 vehicles can pass through the signal in one cycle, the stop-and-go waves still cannot be completely resolved. The phenomenon is caused by the insufficient shifting distance.



(a) MPR = 40% LACC-Control

(b) MPR = 40% RL-Control

(c) MPR = 50% LACC-Control

(d) MPR = 50% RL-Control

(e) MPR = 60% LACC-Control

(f) MPR = 60% RL-Control

(g) MPR = 70% LACC-Control

(h) MPR = 70% RL-Control

Fig. 10 Spatial–temporal trajectories of experiment at various MPR (40%–70%).(a) MPR = 40% LACC-Control, (b) MPR = 40% RL-Control, (c) MPR = 50% LACC-Control, (d) MPR = 50% RL-Control, (e) MPR = 60% LACC-Control, (f) MPR = 60% RL-Control, (e) MPR = 70% LACC-Control, (f) MPR = 70% RL-Control.

This may mean that the value of the saturated flow with mixed autonomy needs to be re-corrected at the high MPR conditions.

2) The comparison of passing speeds between LACC-Control and RL-Control during the green phase reveals that achieving higher throughputs does not depend on a higher passing speed or the system limit speed. The related studies on speed control usually set the maximum speed limit as the exit speed of the intersection. From our results, the setting seems inappropriate under saturated flow.

4.3 Flexibility: A case to achieve the specific throughputs under the saturation flow

To prove the flexibility of the proposed RL method, we set two-level objectives for RL-Control: the first-level objective is the same throughputs as the LACC-Control experiment in Section 4.2, and the second-level objective is the minimum emissions.

According to the results in Section 4.2, the spatial-temporal trajectories of LACC-Control and RL-Control are similar at low MPR, so we carried out three experiments with MPR of 40%, 60%, and 80%.

As shown in Fig. 12, RL-control can reduce emissions by 64.5%, 72.0%, and 79.4%, with MPR of 40%, 60% and 80%, respectively, compared with LACC-Control. The emissions benefits grow nearly linearly with MPR increasing.

Figs. 13–15 are spatial–temporal trajectories of LACC-Control and RL-Control at MPR = 40%, MPR = 60%, and MPR = 80%. It can be seen that RL-Control can greatly alleviate the stop-and-go waves in the mixed flow, when the throughputs are the same.

It should be noted that the green circle in Figs. 13–15 shows that the AVs controlled by RL actively adjust the speed to reduce emissions. When the green phase starts, an AV can be the leader of the platoon passing the green phase. The higher the MPR in the platoon, the smoother the spatial–temporal trajectories of the platoon passing the green phase.

There is also a phenomenon can be found in Figs. 13 and 14. In LACC-Control, there is a high probability that the first one to pass the green phase is an HDV, since MPR is not very high. The phenomenon has a negative impact on the overall emissions. In RL-Control, the agents will control the last AV from the upstream intersection to actively decelerate and make the AV be the leader of the platoon at current green phase. Meanwhile, RL-Control does not affect the overall throughputs compared with LACC-Control.

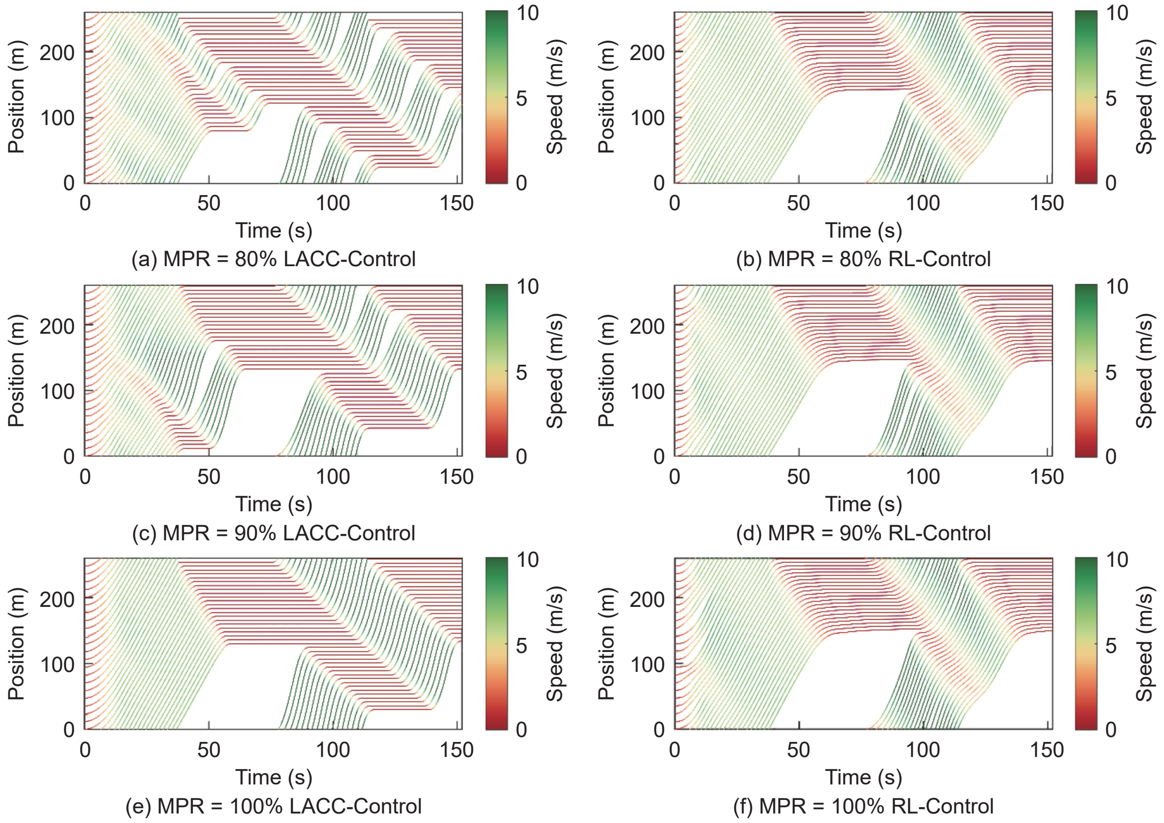


Fig. 11 Spatial–temporal trajectories of experiment at various MPR (80%–100%). (a) MPR = 80% LACC-Control, (b) MPR = 80% RL-Control, (c) MPR = 90% LACC-Control, (d) MPR = 90% RL-Control, (e) MPR = 100% LACC-Control, (f) MPR = 100% RL-Control.

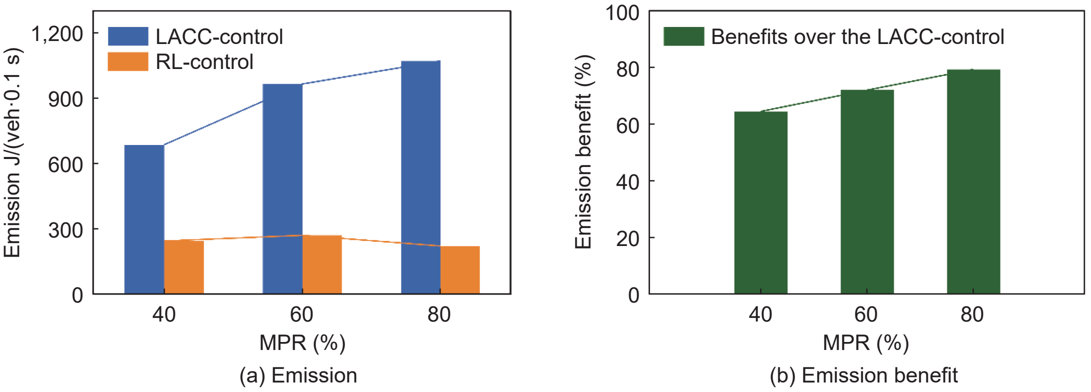


Fig. 12 Improvements of emissions at MPR = 40%, 60%, and 80%. (a) Emission, (b) emission benefit.

5 Conclusions and future work

This work proposed a multi-level objectives framework for AVs' trajectories decision-making based on multi-agent DRL. The proposed method has stable convergence performance in complex conditions. The proposed method is superior and flexible, and it can be used to analyze the behavior changes of mixed flow and the mechanism of mixed autonomy to improve traffic efficiency.

We took the saturated signalized intersection with mixed autonomy as the research object. At the experiment of superiority, we come to the following conclusions:

1) Using RL-Control based on the proposed method, the throughputs increase linearly as MPR increases, compared with fully HDVs. When MPR is 100%, the throughputs benefits can reach 69.2%.

2) It is with the similar behavioral characteristics of the reliability to use RL-Control and LACC-control as an AVs car-following model when MPR is below 30% and more than 80%. The former (below 30%) has similar traffic efficiency between RL-Control and LACC-control, while the latter (more than 80%) has different traffic efficiency.

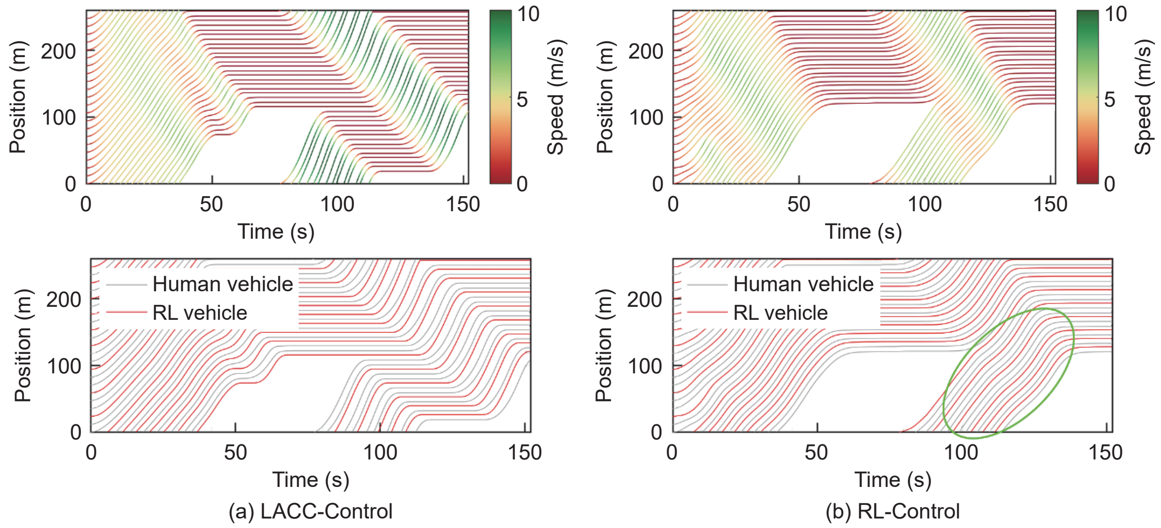
3) When MPR is above 40%, it is necessary to study the cooperation of AVs under the system-level optimizations.

4) If the cooperation is not considered, when MPR is higher than 80%, AVs' behaviors are less affected by the choice of control policies.

At the experiment of flexibility, we come to the following conclusions:

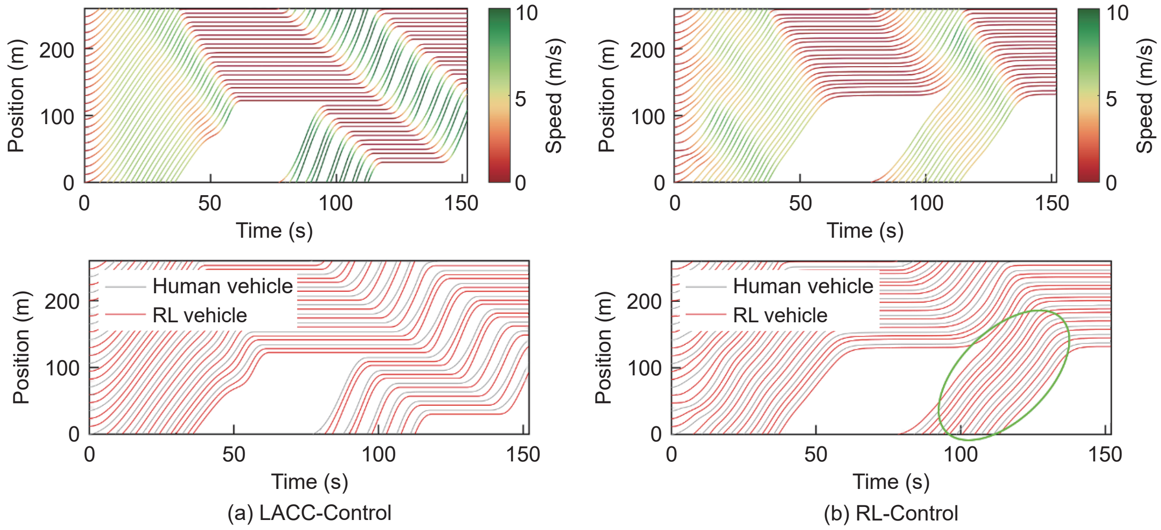
1) Using RL-Control based on the proposed method, the emission benefits increase nearly linearly as MPR increases at 40%, 60%, and 80%. RL-control can reduce emissions by 64.5%, 72.0% and 79.4%, respectively, compared with the LACC-Control with the same MPR.

2) The traffic efficiency can be improved by flexibly distributing



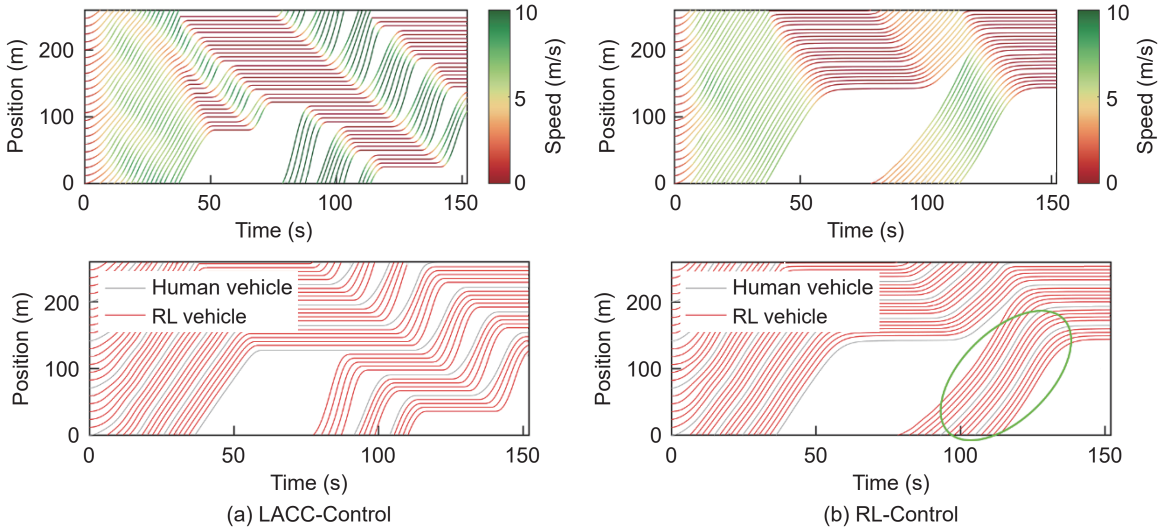
(a) LACC-Control (b) RL-Control

Fig. 13 Spatial–temporal trajectories of the same throughputs experiment at MPR = 40%. (a) LACC-Control, (b) RL-Control.



(a) LACC-Control (b) RL-Control

Fig. 14 Spatial–temporal trajectories of the same throughputs experiment at MPR = 60%. (a) LACC-Control, (b) RL-Control.



(a) LACC-Control (b) RL-Control

Fig. 15 Spatial–temporal trajectories of the same throughputs experiment at MPR = 80%. (a) LACC-Control, (b) RL-Control.

the vehicles' distribution between the upstream and downstream intersections.

We also found some interesting phenomena of mixed autonomy worthy of discussion, which can provide some references and directions for the studies of mixed autonomy at saturated signalized intersections.

1) The benefits under the high MPR show that the agent seems to have explored a certain undesigned cooperation rule.

2) The value of the saturated flow with mixed autonomy at a high MPR needs to be re-corrected.

3) Setting the exit speed of the intersection as the maximum speed limit, this constraint setting seems inappropriate under the saturated flow.

The focus of our follow-up research will be on modeling and adjusting engineering parameters based on the theoretical threshold, specifically in the study of robustness and cooperative adaptability.

Replication and data sharing

Code and experimental data are available at https://doi.org/10.26599/ETSD.2023.9190025.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 52272332 and 51578199), Heilongjiang Provincial Natural Science Foundation (Grant No. YQ2021E031), and the Fundamental Research Funds for the Central Universities (Grant No. HIT.OCEF.2022026).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Abousleiman, R., Rawashdeh, O., 2015. Energy consumption model of an electric vehicle. In: 2015 IEEE Transportation Electrification Conference and Expo (ITEC), 1–5.

Aragon-Gómez, R., Clempner, J. B., 2020. Traffic-signal control reinforcement learning approach for continuous-time Markov games. Eng Appl Artif Intell, 89, 103415.

Aslani, M., Mesgari, M. S., Wiering, M., 2017. Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events. Transp Res Part C Emerg Technol, 85, 732–752.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., Lee, M., 2009. Natural actor-critic algorithms. Automatica, 45, 2471–2482.

Chen, C., Wang, J., Xu, Q., Wang, J., Li, K., 2021. Mixed platoon control of automated and human-driven vehicles at a signalized intersection: Dynamical analysis and optimal control. Transp Res Part C Emerg Technol, 127, 103138.

Dai, P., Liu, K., Zhuge, Q., Sha, E. H. M., Lee, V. C. S., Son, S. H., 2016. Quality-of-experience-oriented autonomous intersection control in vehicular networks. IEEE Trans Intell Transport Syst, 17, 1956–1967.

Ding, H., Li, W., Xu, N., Zhang, J., 2022. An enhanced eco-driving strategy based on reinforcement learning for connected electric vehicles: Cooperative velocity and lane-changing control. J Intell Connect Veh, 5, 316–332.

Fayazi, S. A., Vahidi, A., 2018. Mixed-integer linear programming for optimal scheduling of autonomous vehicle intersection crossing. IEEE Trans Intell Veh, 3, 287–299.

Feng, Y., Yu, C., Liu, H. X., 2018. Spatiotemporal intersection control in a connected and automated vehicle environment. Transp Res Part C Emerg Technol, 89, 364–383.

Guan, Y., Ren, Y., Li, S. E., Sun, Q., Luo, L., Li, K., 2020. Centralized cooperation for connected and automated vehicles at intersections by proximal policy optimization. IEEE Trans Veh Technol, 69, 12597–12608.

Han, X., Ma, R., Zhang, H. M., 2020. Energy-aware trajectory optimization of CAV platoons through a signalized intersection. Transp Res Part C Emerg Technol, 118, 102652.

He, Y., Liu, Y., Yang, L., Qu, X., 2023. Deep adaptive control: Deep reinforcement learning-based adaptive vehicle trajectory control algorithms for different risk levels. IEEE Trans Intell Veh. Http://doi.org/10.1109/TIV.2023.3303408.

Hoel, C. J., Driggs-Campbell, K., Wolff, K., Laine, L., Kochenderfer, M. J., 2019. Combining planning and deep reinforcement learning in tactical decision making for autonomous driving. https://arxiv.org/abs/1905.02680.pdf

Isele, D., Rahimi, R., Cosgun, A., Subramanian, K., Fujimura, K., 2017. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning. https://arxiv.org/abs/1705.01196.pdf

Jiang, H., Hu, J., An, S., Wang, M., Park, B. B., 2017. Eco approaching at an isolated signalized intersection under partially connected and automated vehicles environment. Transp Res Part C Emerg Technol, 79, 290–307.

Jing, S., Hui, F., Zhao, X., Rios-Torres, J., Khattak, A. J., 2019. Cooperative game approach to optimal merging sequence and on-ramp merging control of connected and automated vehicles. IEEE Trans Intell Transport Syst, 20, 4234–4244.

Kesting, A., Treiber, M., Schönhof, M., Helbing, D., 2008. Adaptive cruise control design for active congestion avoidance. Transp Res Part C Emerg Technol, 16, 668–683.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., *et al.*, 2022. Deep reinforcement learning for autonomous driving: A survey. IEEE Trans Intell Transport Syst, 23, 4909–4926.

Kreidieh, A. R., Wu, C., Bayen, A. M., 2018. Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 1475–1480.

Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Gonzalez, J., *et al.*, 2017. Ray rLLib: A composable and scalable reinforcement learning library. http://arxiv.org/abs/1712.09381

Li, G., Li, S., Li, S., Qin, Y., Cao, D., Qu, X., *et al.*, 2020. Deep reinforcement learning enabled decision-making for autonomous driving at intersections. Automot Innov, 3, 374–385.

Liu, C., Lin, C. W., Shiraishi, S., Tomizuka, M., 2018. Distributed conflict resolution for connected autonomous vehicles. IEEE Trans Intell Veh, 3, 18–29.

Liu, Y., Wu, F., Liu, Z., Wang, K., Wang, F., Qu, X., 2023. Can language models be used for real-world urban-delivery route optimization? Innovation, 4, 100520.

Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., *et al.*, 2018. Microscopic traffic simulation using SUMO. In: 2018 21st international conference on intelligent transportation systems (ITSC), 2575–2582.

Makantasis, K., Kontorinaki, M., Nikolos, I., 2020. Deep reinforcement-learning-based driving policy for autonomous road vehicles. IET Intell Transp Syst, 14, 13–24.

Morales Medina, A. I., van de Wouw, N., Nijmeijer, H., 2018. Cooperative intersection control based on virtual platooning. IEEE Trans Intell Transport Syst, 19, 1727–1740.

Meng, X., Cassandras, C. G., 2022. Eco-driving of autonomous vehicles for nonstop crossing of signalized intersections. IEEE Trans Automat Sci Eng, 19, 320–331.

Mushtaq, A., Haq, I. U., Imtiaz, M. U., Khan, A., Shafiq, O., 2021. Traffic flow management of autonomous vehicles using deep reinforcement learning and smart rerouting. IEEE Access, 9, 51005–51019.

Phan, T. T., Ngoduy, D., Le, L. B., 2020. Space distribution method for autonomous vehicles at a signalized multi-lane intersection. IEEE Trans Intell Transport Syst, 21, 5283–5294.

Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P., 2015. High-dimensional continuous control using generalized advantage estimation. https://arxiv.org/abs/1506.02438.pdf

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. https://arxiv.org/abs/1707.06347.pdf

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.,

2014. Deterministic policy gradient algorithms. 31st Int Conf Mach Learn ICML 2014, 1, 605–619.

Stryszowski, M., Longo, S., Velenis, E., Forostovsky, G., 2021. A framework for self-enforced interaction between connected vehicles: Intersection negotiation. IEEE Trans Intell Transport Syst, 22, 6716–6725.

Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 62, 1805–1824.

Vahidi, A., Sciarretta, A., 2018. Energy saving potentials of connected and automated vehicles. Transp Res Part C Emerg Technol, 95, 822–843.

Wang, K., Wang, Y., Du, H., Nam, K., 2021. Game-theory-inspired hierarchical distributed control strategy for cooperative intersection considering priority negotiation. IEEE Trans Veh Technol, 70, 6438–6449.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., *et al.*, 2016. Sample efficient actor-critic with experience replay. http://arxiv.org/abs/1611.01224

Wu, C., Kreidieh, A., Parvate, K., Vinitsky, E. and Bayen, A. M., 2017. Flow: Architecture and benchmarking for reinforcement learning in traffic control. https://arxiv.org/abs/1710.05465

Wu, J., Qu, X., 2022. Intersection control with connected and automated vehicles: A review. J Intell Connect Veh, 5, 260–269.

Xiao, L., Gao, F., 2011. Practical string stability of platoon of adaptive cruise control vehicles. IEEE Trans Intell Transport Syst, 12, 1184–1194.

Ye, Y., Zhang, X., Sun, J., 2019. Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. Transp Res Part C Emerg Technol, 107, 155–170.

Yu, H., Jiang, R., He, Z., Zheng, Z., Li, L., Liu, R., *et al.*, 2021. Automated vehicle-involved traffic flow studies: A survey of assumptions, models, speculations, and perspectives. Transp Res Part C Emerg Technol, 127, 103101.

Zhang, Y., Zhou, Y., Lu, H., Fujita, H., 2021. Cooperative multi-agent actor-critic control of traffic network flow based on edge computing. Future Gener Comput Syst, 123, 128–141.

Zhang, Z., Yang, X. T., 2021. Analysis of highway performance under mixed connected and regular vehicle environment. J Intell Connect Veh, 4, 68–79.

Zhou, M., Yu, Y., Qu, X., 2020. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach. IEEE Trans Intell Transport Syst, 21, 433–443.

Xiaowei Hu received the B.E. degree in transportation engineering from Harbin Institute of Technology, Harbin, China, in 2006, and the M.S. and Ph.D. degrees in transportation planning and management from Harbin Institute of Technology, Harbin, China, in 2008 and 2013, respectively. He is currently an associate professor with the School of Transportation Science and Engineering, Harbin Institute of Technology, China. His research interests include travel behavior analysis, transportation big data analysis, passenger transportation policy, etc. He has authored/coauthored six books and more than 40 academic papers in many mainstream transportation journals, including *Transportation Research Part A* and *Journal of Transportation Geography.*