Transportation Research Part C 158 (2024) 104445

Contents lists available at ScienceDirect

Transportation Research Part C

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

A multi-agent reinforcement learning-based longitudinal and lateral control of CAVs to improve traffic efficiency in a mandatory lane change scenario

Shupei Wang [a], Ziyang Wang [a,*], Rui Jiang [b,*], Feng Zhu [c], Ruidong Yan [b], Ying Shang [b]

[a] School of Traffic and Transportation, Beijing Jiaotong University, 100044 Beijing, PR China
[b] School of Systems Science, Beijing Jiaotong University, 100044 Beijing, PR China
[c] School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Mandatory lane change
Connected autonomous vehicles
Reinforcement learning
Traffic flow

ABSTRACT

Bottleneck areas are prone to severe traffic congestion due to the sudden drop in capacity. To improve traffic efficiency in the bottleneck area, this paper proposes a multi-agent deep reinforcement learning framework integrating collision avoidance strategies to improve traffic efficiency in a mandatory lane change scenario. The proposed method considers distance-keeping and lane-changing coordination in a connected autonomous vehicle (CAV) environment, by controlling vehicles' longitudinal and lateral movement to effectively reduce traffic congestion in a mandatory lane change scenario. This framework was trained and tested in a simulation environment that is the same as the natural driving environment. Compared with real-world data and the benchmark model (a Dueling Double Deep Q-Network-based model), the proposed model shows better performance in terms of average speed, travel time, throughput, and safety in the bottleneck area. The results show that the proposed model can effectively reduce traffic congestion and improve traffic efficiency in a mandatory lane change scenario.

1. Introduction

Traffic congestion often occurs in bottleneck areas due to reduced throughput. (Daganzo et al., 1997; Chowdhury et al., 2000; Helbing et al., 2001; Treiber et al., 2011; Chen et al., 2012). There are various typical nonrecurrent bottleneck scenarios, such as lane closures in work zones or traffic accidents. The frequent mandatory lane changes caused by lane closures lead to substantial disturbances in traffic flow, which would further lead to severe traffic congestion or secondary crashes (Zheng et al., 2010; Memarian et al., 2019; Han et al., 2020; Zhu et al., 2021). With the development of Connected Autonomous Vehicles (CAVs) technology, vehicles can communicate and collaborate through vehicle-to-vehicle or vehicle-to-infrastructure technologies. CAVs have better sensing capabilities than conventional vehicles while being able to follow precise instructions to control the vehicle's trajectory, which opens unprecedented potential for traffic management. Many studies have been conducted on how to smoothly merge vehicles in bottleneck areas to reduce traffic congestion. These studies can be divided methodologically into three main categories: rule-based methods, optimization-based methods, and learning-based methods.

* Corresponding authors.
E-mail addresses: wangzy@bjtu.edu.cn (Z. Wang), jiangrui@bjtu.edu.cn (R. Jiang).

https://doi.org/10.1016/j.trc.2023.104445
Received 5 May 2023; Received in revised form 21 October 2023; Accepted 27 November 2023
Available online 2 December 2023

Table 1
Summary of learning-based methods.

Papers	Simulation Scenario		RL algorithm		collision avoidance strategy		Action			
	SS	RS	SA	MA	1D	2D	LO	LA	CO	DI
Bouton et al. (2019)	√	–	√	–	√	–	√	–	–	√
Nishi et al. (2019)	–	√	√	–	√	–	√	–	–	√
Wu et al. (2020)	–	√	√	–	√	–	√	–	–	√
Ren et al. (2020)	√	–	√	–	√	–	√	–	√	–
Chen et al. (2021)	√	–	–	√	√	–	–	√	–	√
Guo et al. (2021)	–	√	√	–	√	–	√	√	√	√
Wang et al. (2021)	√	–	–	√	√	–	–	√	–	√
Han et al., (2022a)	–	√	√	–	√	–	√	–	–	√
Han et al., (2022b)	√	–	√	–	√	–	√	–	–	√
Jiang et al. (2022)	–	√	√	–	√	–	√	–	√	–
Wang et al. (2022)	–	√	√	–	√	–	–	√	–	√
Li et al. (2022)	√	–	√	–	–	√	–	√	–	√
Proposed	–	√	–	√	–	√	√	√	√	–

SS - Simulations based on simulated scenarios; RS - Simulations based on real-world scenarios; SA - Single-agent; MA - Multi-agent;
1D - One-dimensional; 2D - Two-dimensional; LO - longitudinal; LA - lateral; CO - continuous; DI - discrete;

Rule-based approaches usually manage traffic by establishing reasonable and simple rules, which have the advantages of clear instructions, easy implementation, and high interpretability. For example, Zhang et al. (2013) developed control strategies of Variable Speed Limits (VSL) for improving traffic efficiency at freeway merge bottleneck areas by preventing capacity drop. Rios-Torres and Malikopoulos (2016) presented the virtual mapping technique by comparing each vehicle's path length with a fictitious fixed merging point, and vehicles close to the merging point are given an earlier sequence. Ding et al. (2020) proposed a rule-based cooperative merging strategy to coordinate CAVs going through the merging zone and the proposed method can achieve a near-optimal merging sequence. However, rule-based approaches are usually suitable for specific scenarios and can be applied to generate simple trajectories, but would face challenges in optimization.

A considerable amount of literature has been published on optimization-based methods (Hu et al., 2019; Fukuyama et al., 2020; Tajalli et al., 2022; Markakis et al., 2022;). Usually, the optimization model will optimize the merging sequence at the upper level and plan the motion trajectory at the lower level, while the objective is mostly to maximize the traffic efficiency of the system, such as travel time, fuel consumption, and traffic volume. Zhang et al. (2019) used hyper-heuristic optimization to obtain the lane-changing advisory proportion for each segment upstream of the bottleneck area. Karimi et al. (2020) established a set of control algorithms for cooperative CAV trajectory optimization in different merging scenarios in mixed traffic. Sun et al. (2020) developed a deterministic cooperative ramp merging mechanism to mitigate traffic conflicts and improve merging efficiency for mixed traffic. Cao et al. (2021) presented a cooperative traffic control strategy to increase the throughput of nonrecurrent bottlenecks such as work zones by making full use of the spatial resources upstream of work zones. Xiong et al. (2022) proposed a control strategy for a freeway merging bottleneck consisting of a CAV-exclusive lane and a human-driven vehicle lane. Tang et al. (2022) proposed a novel hierarchical system optimal cooperative merging control model considering flexible merging positions (CMC-FMP) to realize safe and efficient merging processes. Zhu et al. (2022) presented a flow-level CAV coordination strategy to facilitate merging operations in multilane freeways. Xue et al. (2023) proposed a platoon-based cooperative optimal control algorithm for CAVs at highway on-ramps under heavy traffic, which extends individual CAV merging control to platoon cooperative control. However, the optimization-based methods have computational time challenges in solving nonlinear optimization problems and may not be suitable for complex scenarios with high real-time requirements.

In recent years, learning-based methods have received increasing attention (Wu et al., 2020; Chen et al., 2021; Guo et al., 2021), as summarized in Table 1. Learning-based models have advantages in computational speed while being able to handle rich structured and unstructured features, bringing good environment awareness to CAVs. Bouton et al. (2019) presented a reinforcement learning (RL) approach to solve the problem of autonomously merging in dense traffic. Their study confirms that an autonomous agent can benefit from reasoning about the interaction with other drivers. Nishi et al. (2019) presented a freeway merging approach based on multi-policy decision-making coupled with an RL technique called passive actor-critic, which learns with less knowledge of the system and without active exploration. Ren et al. (2020) proposed a cooperative highway work zone merge control strategy based on RL. Each vehicle in the closed lane learns how to optimize its longitudinal position to find a safe gap in the open lane using an off-policy soft actor-critic RL algorithm. Wang et al. (2021) proposed a harmonious lane-changing strategy based on the deep RL method. Instead of focusing only on the efficiency of an individual vehicle, the proposed method balances it with the overall traffic efficiency by introducing a global traffic indicator (flow rate) into the reward function. Han et al., (2022a) proposed a physics-informed RL-based ramp metering strategy, which trains the RL model using a combination of historic data and synthetic data generated from a traffic flow model. In another scenario, Han et al., (2022b) proposed a new RL-based VSL control approach to improve traffic efficiency by using VSL against freeway jam waves. Jiang et al. (2022) proposed a cooperative longitudinal control based on Soft Actor-Critic (SAC) RL to dampen the stop-and-go waves by generating smaller oscillation growth. Wang et al. (2022) implemented an intensive microscopic simulation based on RL to determine the desired ego-efficient lane-changing strategy. Li et al. (2022) proposed a lane change decision-making framework based on deep RL to find a risk-aware driving decision strategy with the minimum expected risk for autonomous driving. However, most studies lack a comparison of simulation results with naturalistic driving data, which may result in simulation

Fig. 1. Overhead view of the test site.

scenarios that are far from the actual scenario. At the same time, the process of lane change is ignored or simplified in the study of vehicle merging, usually involving only a lane change decision and longitudinal action. Both longitudinal and lateral effects are critical in the process of lane change and need to be considered jointly.

In summary, there are still two gaps in the existing literature. Firstly, due to the limited availability of natural driving data in the mandatory lane change scenario, most studies are unable to compare the simulation results with the natural driving results; secondly, most studies ignore or simplify the lane change process, resulting in simplifying the interaction of vehicles between different lanes. Typically, fixed lateral speeds or fixed lane-changing durations are used for implementation, which significantly differs from real-world scenarios, especially in congested bottleneck areas. To fill these gaps, this paper conducts a two-lane circular experiment with 30 human-driven vehicles to collect real-world data under a mandatory lane change scenario. Subsequently, we have developed a novel control strategy that integrates a multi-agent reinforcement learning framework (MARL) to mitigate congestion in a bottleneck area. On one hand, unlike previous single-policy RL research that mainly focuses on vehicle interaction factors, the proposed framework not only sets different driving intentions for vehicles in different lanes to facilitate coordination but also incorporates a buffer zone strategy. This allows CAVs to maintain a certain distance before entering the bottleneck area, reducing the vehicle density and increasing the lane-changing gaps. On the other hand, considering the interactions during the lane-changing process, we have proposed a 2-dimensional inverse time-to-collision (2D-iTTC) metric to quantify safety during lane-changing. Based on the safety threshold from real data, we extend the traditional Gipps model to a 2D version of the collision avoidance strategy. Compared to the conservative strategy in Simulation of Urban Mobility (SUMO), the proposed strategy is more proactive and equally safe, effectively aiding the RL model in continuous control both in the longitudinal and lateral directions. In the simulation experiments, the proposed framework demonstrated a significant improvement in traffic efficiency in the bottleneck area compared to SUMO's rule-based model, real-world data, and the baseline discrete decision model based on Dueling Double Deep Q-Network (DDDQN).

The main contributions of this paper are summarized as follows: (1) a mandatory lane change scenario was created on a two-lane circular road, and real driving data were collected from 30 human-driven vehicles as a benchmark; (2) we propose a multi-agent reinforcement learning framework that integrates the buffer zone strategy and a 2D collision avoidance strategy to eliminate congestion in the bottleneck area, as observed in natural driving data; (3) simulation results indicate that the proposed framework shows significant improvements in safety, average speed, and traffic throughput compared to SUMO's rule-based model, real-world data, and the baseline discrete decision model based on DDDQN.

The rest of the paper is organized as follows: Section 2 describes real experiments and then presents two strategies to mitigate the congestion; Section 3 introduces the specifics of the proposed framework and the baseline model for comparison; Section 4 provides a qualitative and quantitative analysis of the experimental results; and Section 5 concludes the paper and presents future directions.

2. Problem statement

In this section, we first introduce the specifics of the real experiments of the mandatory lane change scenario, then present the traffic congestion problem in the bottleneck area due to capacity drop followed by two strategies to mitigate traffic congestion.

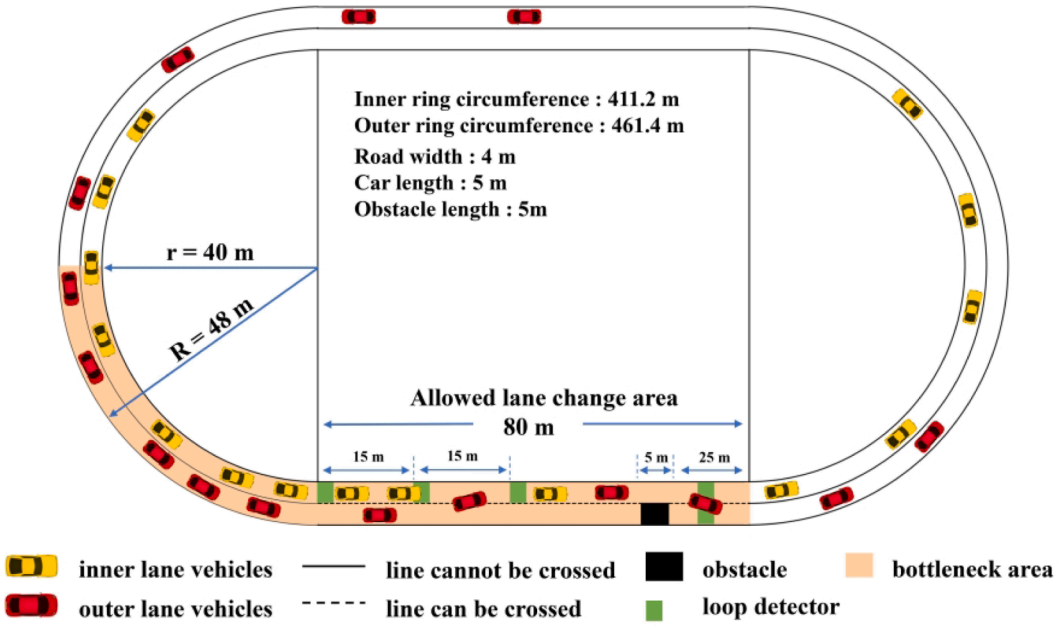


Fig. 2. Schematic diagram of the experimental site.

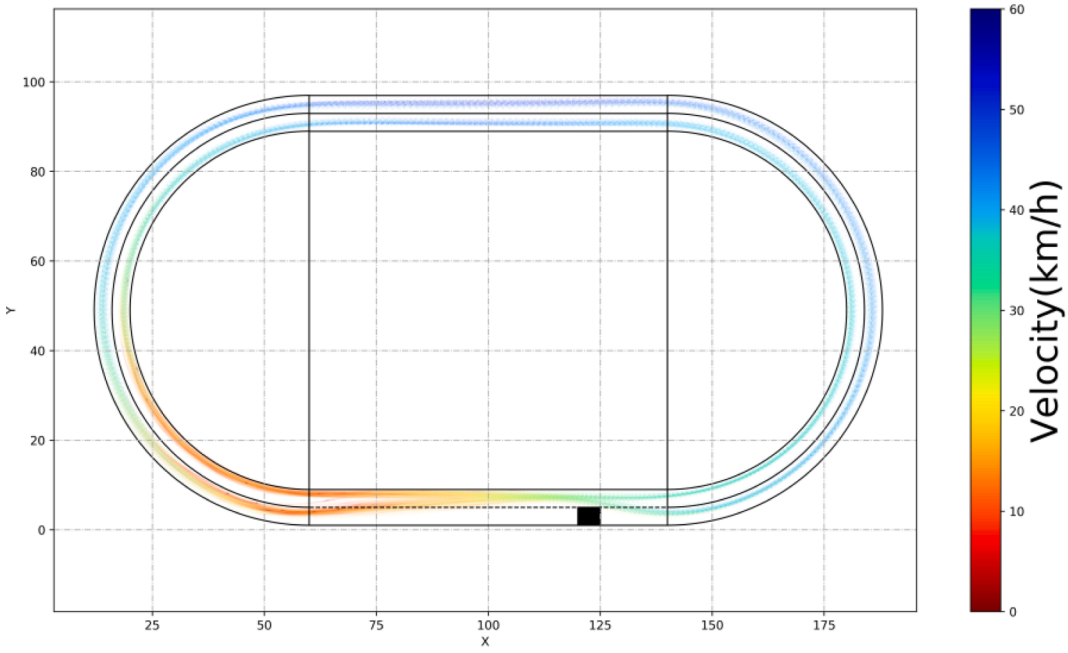


Fig. 3. Velocity-trajectory diagram of real data.

2.1. Experiment description

This study aims to reduce the traffic congestion generated by lane changes in a mandatory lane change scenario to improve traffic efficiency. The majority of lane change studies in the literature are based on NGSIM (Next Generation Simulation) data. This dataset contains vehicle trajectories on US-101, I-80, and other roadways, providing a large number of examples of car-following and lane-changing. However, this dataset does not cover mandatory lane change scenarios in bottleneck areas.

To compensate for the data, this study conducted a two-lane experiment on a ring road at the road traffic test site belonging to the Ministry of Transport of China. The experimental runway design is shown in Fig. 1 and Fig. 2. The circular runway is with an outer ring

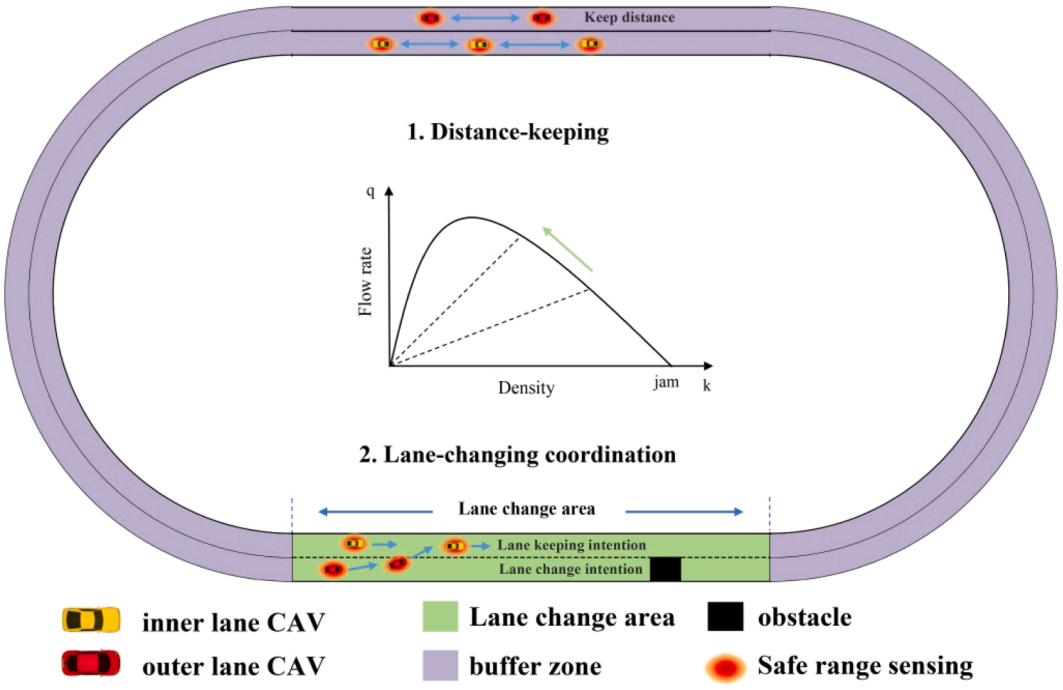


Fig. 4. Basic idea schematic diagram.

of 461.4 m, an inner ring of 411.2 m, an inner ring radius of 40 m, and an outer ring radius of 48 m. Further, the lane width is 4 m, the vehicle length is 5 m, the obstacle length is 5 m, and the allowable lane change area is 80 m. Specifically, the 5-meter-wide obstacle area on the outside lane simulates the bottleneck scenario in reality (due to work zone or vehicle accidents). A total of 30 vehicles were used in the experiment, all driven by human drivers. Microscopic traffic flow data were collected through GPS devices in the vehicles. The average location and velocity errors of the GPS device are within 1 m and 1 km/h, respectively. During the experiment, inner-ring vehicles were kept on the inner ring, and outer-ring vehicles needed to merge into the inner ring after entering the allowed lane change area and change lanes again to return to the outer ring after passing through the forbidden area. A total of 4 sets of experiments were conducted, each lasting 25 min. Different sets were achieved by changing the number of vehicles in the inner and outer rings. In the four sets, there are 24 vehicles, 26 vehicles, 28 vehicles, and 30 vehicles, respectively. We selected the group of 30 vehicles as the object of analysis in the follow-up experiment. Because the proposed model was applied to the other three groups with similar effects but the congestion was not as severe in the other three groups, the 30-vehicle group was able to illustrate the effectiveness of the method. The collected data includes the vehicle ID, the time stamp, latitude, and longitude, and the speed of the vehicle. The experiments were conducted from 9:30–12:00 including a total of 2.5 h of data, with 0.1 s as a time step.

2.2. Traffic congestion problem and two strategies

After collecting real data, we found that the problem of traffic congestion in real scenarios emerged. As shown in Fig. 3, a low-velocity area indicated in red appears in the lane change area. The traffic oscillation propagates further upstream, and a clear deceleration area appears. We refer to this 150-meter-long low-velocity area as the bottleneck area. The reason for the formation of the bottleneck area, we believe, is mainly the high density of vehicles in the area accompanied by a large number of lane-changing behaviors, which generate much disturbance and cause the stop-and-go phenomenon. And the oscillation is not dampened but propagated upstream, which further increases the vehicle density in the bottleneck area. This eventually led to the traffic congestion problem in the bottleneck area.

To address the traffic congestion problem in the bottleneck area, we have two strategies: distance-keeping and lane-changing coordination. As shown in Fig. 4, on the one hand, we use distance keeping strategy to reduce the density of the bottleneck area, while reserving some space for the lane-changing vehicles. According to the fundamental diagram of traffic, the density of the area is reduced, increasing the traffic flow rate and average speed. Therefore, the simple idea is to create buffer zones so that vehicles are kept at a certain distance before entering the lane change area.

On the one hand, we use a lane-changing coordination strategy to reduce conflicts between vehicles. Considering different driving intentions of inner and outer ring vehicles. Inner ring vehicles desire to keep their lanes to pass the bottleneck area as soon as possible, while outer ring vehicles desire to change lanes to the target lane to pass the bottleneck area quickly. We use a multi-agent deep reinforcement learning method coupled with collision avoidance strategies to coordinate lane-changing behavior. The multi-agent deep reinforcement learning method can take into account the driving intentions of different vehicles, encourage coordinated lane

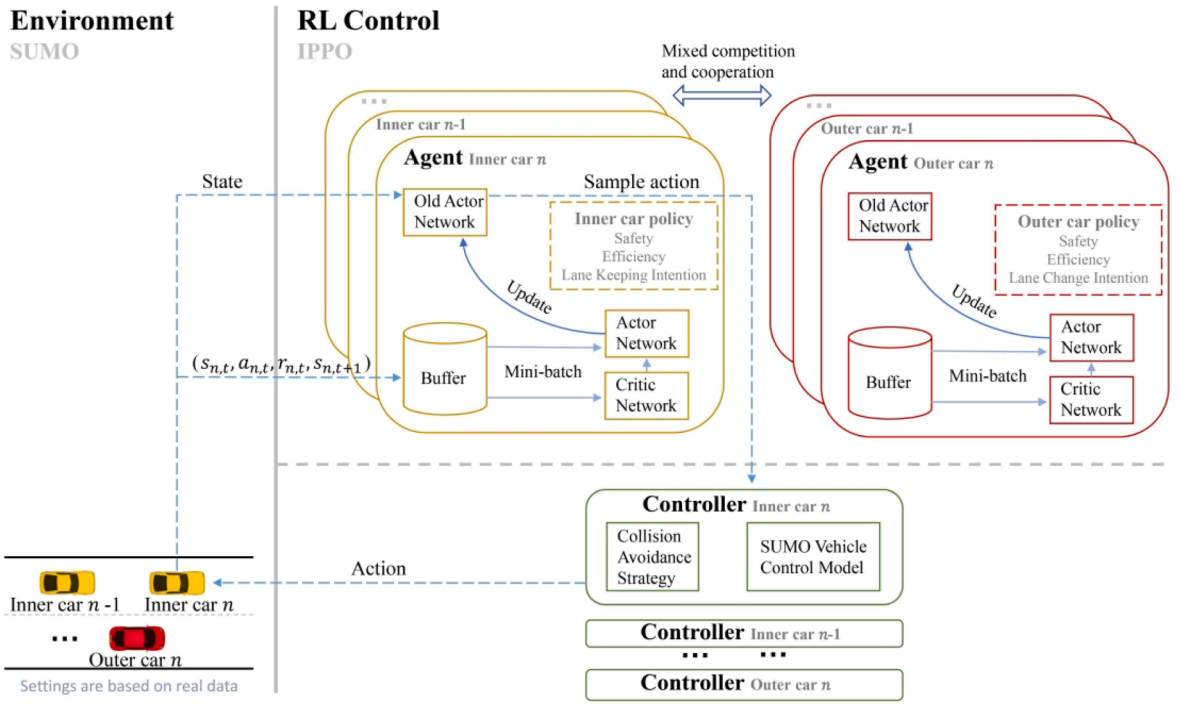


Fig. 5. Training framework diagram.

changes, and penalize unreasonable behavior. The added collision avoidance strategy determines the priority of passing. In addition, action masking will also be applied to speed up the convergence of the model.

To implement these two strategies, we have adopted the Independent Proximal Policy Optimization (IPPO) framework (Witt et al., 2020). IPPO is a decentralized MARL framework, a multi-agent variant of proximal policy optimization (Schulman et al., 2017). It decomposes the n-agent MARL problem into n single-agent problems, treating all other agents as part of the environment. The agents learn the policy through local observations. While centralized joint learning reduces or eliminates issues of partial observability and environment non-stationarity, it must cope with the joint action space that grows exponentially with the number of agents (Wei and Luke, 2016). A decentralized structure is easier to construct and expand. Considering the scalability and the size of the vehicles, we chose the decentralized IPPO framework. We use proximal policy optimization to learn decentralized policies $\pi_\theta^n(a_t^n|s_t^n)$ parameterized by θ for n agents with individual policy clipping, where s_t^n is the state of agent n at time step t and a_t^n is the action of agent n at time step t. Each agent n learns a local observation based on critic $V_\varphi^n(s_t^n)$ parameterized by φ using Generalized Advantage Estimation (Schulman et al., 2016) with discount factor γ and λ. In each iteration, T steps of empirical data are collected in parallel, and then the advantage function A_t^n and the loss function $L^n(\theta)$ are calculated for each step to form the minibatch. The advantage function is calculated by:

$$A_t^n = \sum_{i=0}^{T} (\gamma\lambda)^i \delta_{t+i}^n \tag{1}$$

where $\delta_t^n = r_t + \gamma V_\varphi^n(s_{t+1}^n) - V_\varphi^n(s_t^n)$is the temporal difference error at time step t. r_t is the team reward at time step t. It can be interpreted that the advantage estimation function A_t^n represents whether the action is good or bad in the future. The update of parameter θ for each agent n is implemented by the following loss function:

$$L^n(\theta) = \mathbb{E}_{s_t^n, a_t^n}\left[min\left(\frac{\pi_\theta^n(a_t^n|s_t^n)}{\pi_{\theta_{old}}^n(a_t^n|s_t^n)} A_t^n, clip\left(\frac{\pi_\theta^n(a_t^n|s_t^n)}{\pi_{\theta_{old}}^n(a_t^n|s_t^n)}, 1-\epsilon, 1+\epsilon \right) A_t^n \right) \right] \tag{2}$$

where ϵ is a hyperparameter which roughly indicates how far away the new policy π_θ^n is allowed to go from the old $\pi_{\theta_{old}}^n$. By clipping in the range of $[1-\epsilon, 1+\epsilon]$, it prevents the overestimation or underestimation of the current policy and maintains the stability of the policy update. After updating the parameter θ with minibatch, replace the old parameter θ_{old} will be replaced with the new one, and the next iteration continues.

Table 2
Input features of a vehicle.

Symbol	Description	Unit
$v_x(t)$	longitudinal speed of the current vehicle at time t	m/s
$v_y(t)$	lateral speed of the current vehicle at time t	m/s
$\theta(t)$	steering angle of the current vehicle	°
$x(t)$	horizontal axis coordinates of the current vehicle at time t	–
$y(t)$	vertical axis coordinates of the current vehicle at time t	–
$d(t)$	distance to obstacle	m
$\Delta v_{lead}^{inner}(t)$	relative speed between the current vehicle and the leading vehicle in the inner ring	m/s
$\Delta v_{follow}^{inner}(t)$	relative speed between the current vehicle and the following vehicle in the inner ring	m/s
$\Delta v_{lead}^{outer}(t)$	relative speed between the current vehicle and the leading vehicle in the outer ring	m/s
$\Delta v_{follow}^{outer}(t)$	relative speed between the current vehicle and the following vehicle in the outer ring	m/s
$\Delta x_{lead}^{inner}(t)$	clearance distance between the current vehicle and the leading vehicle in the inner ring	m
$\Delta x_{follow}^{inner}(t)$	clearance distance between the current vehicle and the following vehicle in the inner ring	m
$\Delta x_{lead}^{outer}(t)$	clearance distance between the current vehicle and the leading vehicle in the outer ring	m
$\Delta x_{follow}^{outer}(t)$	clearance distance between the current vehicle and the following vehicle in the outer ring	m
$\tilde{t}_{lead}^{inner}(t)$	2-dimensional inverse time-to-collision between the current vehicle and the leading vehicle in the inner ring	s^{-1}
$\tilde{t}_{follow}^{inner}(t)$	2-dimensional inverse time-to-collision between the current vehicle and the following vehicle in the inner ring	s^{-1}
$\tilde{t}_{lead}^{outer}(t)$	2-dimensional inverse time-to-collision between the current vehicle and the leading vehicle in the outer ring	s^{-1}
$\tilde{t}_{follow}^{outer}(t)$	2-dimensional inverse time-to-collision between the current vehicle and the following vehicle in the outer ring	s^{-1}

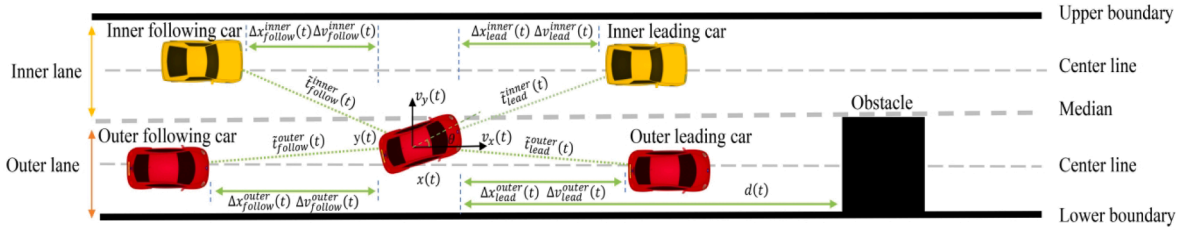


Fig. 6. Schematic diagram of input features.

3. Proposed model

3.1. Training setup and hyperparameters

This study further built the two-lane simulation environment (the same as in Section 2.1) on the open-source simulation platform SUMO and trained the RL agent using the IPPO model. This multi-agent training framework is developed based on the open-source DRL framework FLOW. The policy parameters of each agent are obtained, and then the actions are output. There are 30 CAVs in total in the simulation environment. There are 15 vehicles in the inner ring and 15 vehicles in the outer ring. The simulation sampling time interval is 0.1 s. The maximum time step for one episode is 3000 steps. This study conducted a total of 300 episodes of training on a computer with 16 GB of RAM and an NVIDIA 3080Ti GPU. The average computation time for the trained model to generate action commands for all CAVs per step is 0.026 ~ 0.031 s. The simulation environment is the same as the natural driving environment. The maximum speed is set to 60 km/h and the acceleration to [-4 m/s^2, 4 m/s^2].

IPPO is an on-policy DRL algorithm and is suitable for continuous action space environments, such as the acceleration control of vehicles in this study. It is built on the architectures of the Actor-Critic networks. Both networks have three layers: input, hidden, and output. The hidden layer we built has two layers, each of which is 128 units. The training framework is shown in Fig. 5. In this architecture, each vehicle has its policy network to control and make decisions, and the outer and inner ring vehicles have different rewards and goals. Inner ring vehicles want to pass the bottleneck area quickly, while outer ring vehicles want to change lanes to the inner ring and avoid long waiting times. Some of the key hyperparameters are set as follows: train batch size = 6000, minibatch size = 128, discount rate = 0.99, learning rate = 0.0003, and lambda = 0.95.

In one time step, the state of each vehicle in the environment is input separately to the respective policy network in the RL control, and then the sampled actions are obtained for the next time step. The framework then feeds the sampled actions to the lower-level safety controller, which first judges whether the constraints of the collision avoidance strategy are satisfied. If the constraints are satisfied, the actions are handed over to the vehicle control module for execution; if not, the actions will be replaced with safe actions. Finally, the environment states will be updated. After several time steps, the buffer collects the specified length of data. The framework divides the data in the buffer into several batches that are used to update the actor-network and critic network. After the actor-network is updated, its parameters are copied to the old actor network, and then data collection continues for the next episode.

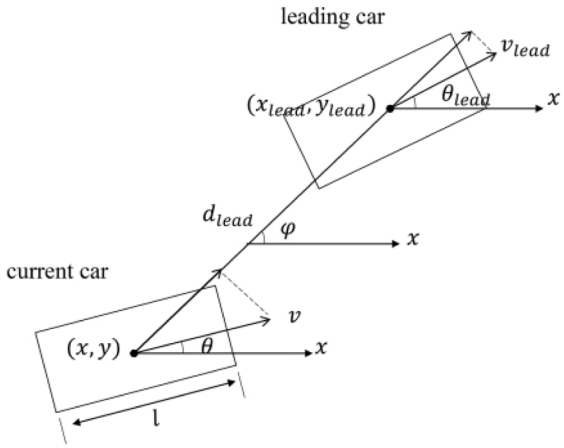


Fig. 7. Schematic diagram of 2-dimensional inverse time-to-collision(2D-iTTC).

3.2. State and action

Well-defined state features help agents better understand their surroundings and make decisions. Therefore, we use interpretable structured features as input. These basic features are also consistent with the key variables in traditional car-following or lane-changing models, such as speed, relative speed, relative distance, time-to-collision (TTC), etc. The overall state is defined as a matrix of size $N \times W$, where N is the number of vehicles and W is the number of features used to represent the vehicle state. The state of the n th vehicle is denoted as s_n, and the overall state space of the system is the joint state of all CAVs, i.e., $s = s_1 \times s_2 \times \cdots \times s_N$. The selected features are detailed in Table 2 and Fig. 6. State features include the state of the vehicle itself and the state of the surrounding environment. To begin with, the state of the vehicle itself includes longitudinal speed $v_x(t)$, lateral speed $v_y(t)$, steering angle $\theta(t)$, X-axis coordinates $x(t)$, and Y-axis coordinates $y(t)$. These features help vehicles understand their intent and location while driving. In addition, the surrounding environment features include the distance to the obstacle $d(t)$, relative velocity Δv, relative distance Δx, and 2D-iTTC $\widetilde{t}$. The relative relationship is between the current vehicle and the four nearest surrounding vehicles. Taking $\Delta v_{lead}^{inner}(t)$ as an example, it indicates the relative speed between the current car and the car in front of it in the inner ring. The superscript indicates which lane it is in, and the subscript indicates the car-following relationship with the current car. As for the 2-dimensional inverse time-to-collision $\widetilde{t}$, it is calculated as follows:

$$\begin{cases} d_{lead} = \sqrt{(x_{lead} - x)^2 + (y_{lead} - y)^2} \\ \varphi = \arccos[(x_{lead} - x)/d_{lead}] \\ \widetilde{t} = max\left(0, \dfrac{vcos(\varphi - \theta) - v_{lead}cos(\varphi - \theta_{lead})}{d_{lead} - l}\right) \end{cases} \tag{3}$$

where x and y represent the horizontal and vertical coordinates of the current vehicle, respectively. The similar x_{lead} and y_{lead} represent the horizontal and vertical coordinates of the leading vehicle. d_{lead} is the distance between the centers of the current and leading vehicle. l represents the length of the vehicle. v and v_{lead} are the speed of the current vehicle and the leading vehicle, respectively. θ and θ_{lead} are the steering angle of the current vehicle and the leading vehicle, respectively. φ is the angle between the current vehicle and the leading vehicle. The demonstration of the symbols is shown in Fig. 7. More details can be referred to Fu et al. (2017). $\widetilde{t}$ helps vehicles be well aware of longitudinal and lateral collision risks. The smaller the value of $\widetilde{t}$, the smaller the risk.

In this study, the action space A_n of agent n is defined as a continuous action space to more realistically reflect the movement of the vehicle in reality, which is calculated as follows:

$$a_n = \{a_X, a_Y\} \tag{4}$$

where a_X and a_Y are the longitudinal acceleration and lateral acceleration of the current vehicle, respectively. The range of a_X is [-4 m/s^2, 4 m/s^2], and the range of a_Y is [-1 m/s^2, 1 m/s^2]. The overall action space of the system is the joint action of all CAVs, which is $a = a_1 \times a_2 \times \cdots \times a_N$. All vehicles in our simulation follow a kinematic bicycle model (Kong et al., 2015).

3.3. Reward function

A well-designed reward function can help RL agents better learn the value of actions and obtain a good policy. The reward function designed in this paper consists of three parts, including safety, efficiency, and the behavioral intent of vehicles on different lanes.

First, for safety, the reward function concerns 2D-iTTC and collision. For 2D-iTTC, the maximum threshold $\widetilde{t}_{max}$ is set at 0.5 s^{-1}.

Values below this threshold will be penalized. This reward takes into account safety in both longitudinal and lateral directions. The specific reward is calculated by the following equation:

$$r_{2DiTTC} = max\left(min\left(-\frac{\widetilde{t}}{\widetilde{t}_{max}}, 0\right), -1\right) \tag{5}$$

where $\widetilde{t}$ denotes the 2D-iTTC between the current vehicle and the surrounding vehicles. Additionally, a large penalty of -100 is given if the vehicle collides with another vehicle or hits the obstacle or boundary:

$$r_{collision} = -100, ifcollision \tag{6}$$

Second, for efficiency, vehicles are encouraged to travel at a faster speed in the bottleneck area. We set the maximum speed v_{max} to 60 km/h (to be consistent with the naturalistic driving data in the experiment). The specific speed reward is given in the following equation:

$$r_{speed} = \frac{\sqrt{v_x(t)^2 + v_y(t)^2}}{v_{max}} \tag{7}$$

where $v_x(t)$ and $v_y(t)$ are the longitudinal and lateral speeds of the current vehicle at time t. To reduce the density of the bottleneck area, vehicles are encouraged to keep a certain distance from other vehicles in the longitudinal direction before entering the lane change area. This will allow more space to be reserved during lane changes, reducing the stop-and-go phenomenon. We analyzed the cases where the following car did not have to stop during the actual lane change and configured 7 m being an appropriate distance. The specific distance reward is as follows:

$$r_{distance} = \begin{cases} 1, ifinbufferzoneand\Delta x > 7 \\ 0, otherwise \end{cases} \tag{8}$$

where Δx indicates the relative distance between the current vehicle and the surrounding vehicles.

Third, for the different behavioral intentions of vehicles in different lanes, it is stipulated that vehicles on the outer ring would try to change lane to the inner ring because of the obstacle, while vehicles on the inner ring do not need to change lane but would try to get through the bottleneck area as quickly as possible. Therefore, for the outer ring, the closer the vehicles get to the center line of the inner ring, the higher the reward, while being close to the road boundary line would be penalized. The specific reward of the outer ring vehicle is:

$$r_{outer} = \begin{cases} 1 - \dfrac{y - y_{center}}{y_{upper} - y_{center}}, ifinlanechangeareaandy_{center} < y < y_{upper} \\ 1 - \dfrac{y_{center} - y}{y_{center} - y_{lower}}, ifinlanechangeareaandy_{lower} < y < y_{center} \end{cases} \tag{9}$$

where y_{upper} denotes the upper boundary of the inner ring, y_{center} is the center line of the inner ring, y_{lower} is the lower boundary of the outer ring, and y is the vertical axis coordinate of the outer ring vehicle.

On the other hand, for the inner ring, vehicles are encouraged to drive in the center line of the inner ring, but no penalty is imposed on them for temporarily approaching the border to avoid collision with lane-change vehicles. The specific reward of the inner ring vehicle is:

$$r_{inner} = \begin{cases} 1, ifinlanechangeareaandy = y_{center} \\ 0, otherwise \end{cases} \tag{10}$$

where y_{center} is the center line of the inner ring, and y is the vertical axis coordinate of the current inner-ring vehicle. Combining the above considerations, we obtain the rewards for inner- and outer-ring vehicles separately, as follows:

$$r_n = \begin{cases} \alpha^* r_{2DiTTC} + \beta^* r_{collision} + \gamma^* r_{speed} + \delta^* r_{distance} + \mu^* r_{outer}, ifinnerringvehicle \\ \alpha^* r_{2DiTTC} + \beta^* r_{collision} + \gamma^* r_{speed} + \delta^* r_{distance} + \mu^* r_{inner}, ifouterringvehicle \end{cases} \tag{11}$$

where α, β, γ, δ, and μ are weights for different components of the reward function. After fine tuning of the parameters, their values are as follows: $\alpha = 1$, $\beta = 1$, $\gamma = 1$, $\delta = 2$, and $\mu = 2$.

3.4. Collision avoidance strategy

In this paper, an additional collision avoidance strategy is added for faster convergence of the model. The technique we use is called action masking, which masks out bad actions to accelerate learning and improve policy. Kometani and Sasaki (1959) first proposed a safety distance model, which assumes that the current vehicle takes action to avoid a collision by sensing the distance between it and the preceding vehicle. However, unlike Kometani and Sasaki's idea of maintaining a minimum safe distance from the preceding

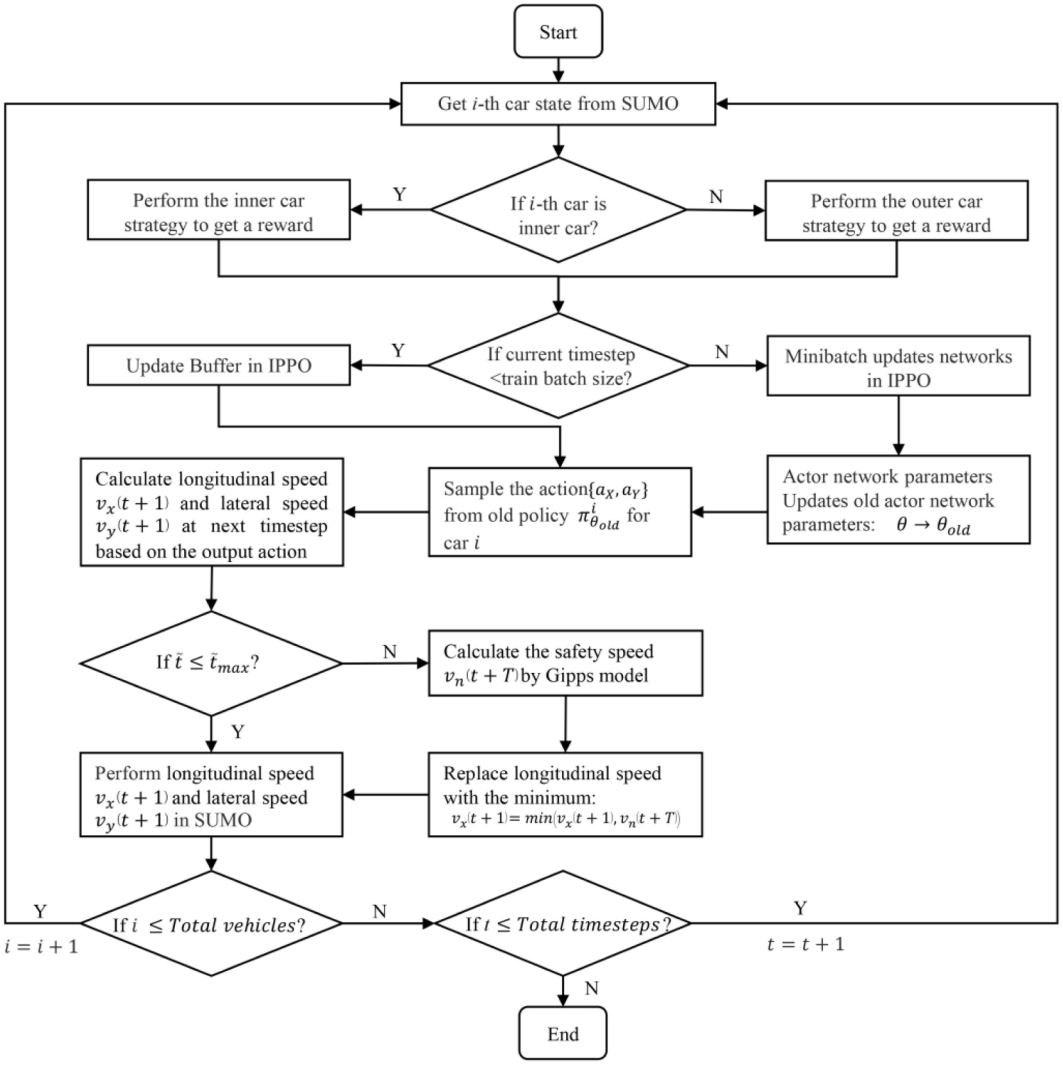


Fig. 8. Flowchart of the proposed framework.

vehicle, Gipps (1981) believes that the following vehicle should travel at a speed at which it can safely stop even when the lead vehicle brakes suddenly. Gipps' model is that the driver travels at a smaller speed in the two states of free flow and congested flow. The model is shown below:

$$v_n(t+T) = min\left\{ v_n(t) + 2.5\widetilde{a}_n T\left(1 - \frac{v_n(t)}{\widetilde{v}_n}\right)\sqrt{0.025 + \frac{v_n(t)}{\widetilde{v}_n}\widetilde{b}_n T + \sqrt{\left(\widetilde{b}_n T\right)^2 - \widetilde{b}_n\left\{2[x_{n-1}(t) - x_n(t) - S_{n-1}] - v_n(t)T - \frac{v_{n-1}^2(t)}{\widehat{b}}\right\}}}\right. \tag{12}$$

where $v_n(t+T)$ is the longitudinal speed of vehicle n at time $t+T$. T is the reaction time set to 0.5 s. $v_n(t)$ is the longitudinal speed of vehicle n at time t. $v_{n-1}(t)$ is the longitudinal speed of preceding vehicle $n-1$ at time t. $\widetilde{v}_n$ is the desired speed of vehicle n set to 60 km/h. $\widetilde{a}_n$ is the desired acceleration of the vehicle n set to 4 m/s^2. $\widetilde{b}_n$ is the desired deceleration of vehicle n set to -4 m/s^2. $\widehat{b}$ is the estimated deceleration of the preceding vehicle set to -4 m/s^2. $x_n(t)$ is the longitudinal position of vehicle n at time t. $x_{n-1}(t)$ is longitudinal position of preceding vehicle $n-1$ at time t. S_{n-1} is the sum of vehicle length and safety distance set to 6.5 m.

The safety speed from the Gipps model is conservative. Because vehicles may change lanes, whether cutting in or out, following them too cautiously can exacerbate the stop-and-go phenomenon. Meanwhile, the right of way in the lane change area is based on a

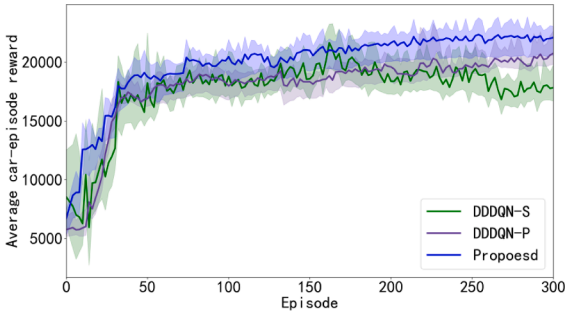


Fig. 9. Convergence of training.

first-come, first-served concept to fully improve space utilization. The collision avoidance strategy only works when the 2D-iTTC is greater than the safety threshold, so that safety and efficiency are well balanced. The current vehicle will identify the closest vehicle to its front bumper as the preceding vehicle to calculate Gipps' safe speed. As a result, the following collision avoidance strategies are proposed:

$$v_x(t+1) = min(v_x(t+1), v_n(t+T))when\widetilde{t} > \widetilde{t}_{max} \tag{13}$$

where $v_x(t+1)$ is the longitudinal speed of the next timestep calculated from the longitudinal acceleration output from the RL model at time t. $v_n(t+T)$ is the safety speed calculated from the Gipps model at time t. $\widetilde{t}$ is the 2D-iTTC calculated from Eq. (3). $\widetilde{t}_{max}$ is the maximum threshold of the 2D-iTTC set to 0.5 s^{-1}. With this collision avoidance strategy as shown in Fig. 8, unreasonable actions are masked and faster model convergence can be achieved while ensuring the safety of vehicles.

3.5. Baseline model

In this paper, a commonly used decision model is implemented as a baseline to compare the difference between the lane-change decision model and the integrated model. The lane-change decision model has been successfully applied in several studies in lane change scenarios, such as Qi et al. (2019), Li et al. (2021), and Wang et al. (2022). In this study, the state-of-the-art DDDQN model is chosen as the baseline model. DDDQN has promising performance in stability and accuracy compared to Deep Q-Network. During the training process of the baseline model, the environmental parameters are consistent with the proposed model, including the number of vehicles, speed limit, etc. Different from our proposed model, the DDDQN-based model's action is a discrete lane-change decision, and the action execution is handed over to the SUMO rule-based model. The hidden layer we built has two layers, 128 units each. Some of the key hyperparameters are set as follows: train batch size = 128, learning rate = 0.0001, and buffer size = 50000. As for the state and reward functions, no changes were made to facilitate the comparison with the proposed model. The action of the baseline model is distinguished from the proposed model (for more details please refer to Chen et al. (2021)) and is set to the following equation:

$$action = \{keeplane, changetoleft, changetoright\} \tag{14}$$

The execution of the action is left to SUMO. The Intelligent Diver Model (IDM) proposed by Treiber et al. is used for the car-following model, and the key parameters are as follows: maximum acceleration = 2.5 m/s^2, maximum speed = 60 km/h, desired deceleration = 1.5 m/s^2, safe time headway = 0.8 s, and standstill safety distance = 0.5 m. The randomized slowdown probability is set to 0.2. The lane change model used is SL2015, and the specific parameters are in SUMO's documentation (Lopez et al., 2018). With this baseline model, we can compare the performance difference between the state-of-the-art lane-change decision model and the proposed model. The main difference between the DDDQN-based model and the proposed model is the different execution framework. The DDDQN-based model is only a lane change decision model. While the DDDQN-based model makes discrete decisions which are then handed over to the SUMO rule-based model for execution, the proposed model coordinates the longitudinal and lateral actions and then executes them with collision avoidance strategies.

4. Analysis of results

In this section, we will analyze and discuss the results of the SUMO-based simulation, the baseline model, and the proposed model in terms of training results, safety performance, congestion reduction effect, and traffic efficiency.

4.1. Training results

To compare the differences in training between the DDDQN-based model and the proposed model, we performed a total of 300 training episodes (900,000 steps) under the same experimental conditions. We will refer to the control group trained using the default collision avoidance model in SUMO as "DDDQN-S" and the control group trained using the proposed collision avoidance strategy as "DDDQN-P". Fig. 9 shows the variation of the average reward per vehicle with the number of episodes during the training. In the

Table 3
Traffic efficiency under different reward designs.

Reward design	Average speed (**km/h**)	Difference (%)	Average travel time (*s*)	Difference (%)
real data	18.4	–	29.3	–
$r_{2DiTTC} + r_{collision} + r_{speed}$	19.8	+7.6 %	27.3	−6.8 %
$r_{2DiTTC} + r_{collision} + r_{speed} + r_{distance}$	22.4	+21.7 %	24.1	−17.7 %
proposed	31.9	+73.4 %	16.9	−42.3 %

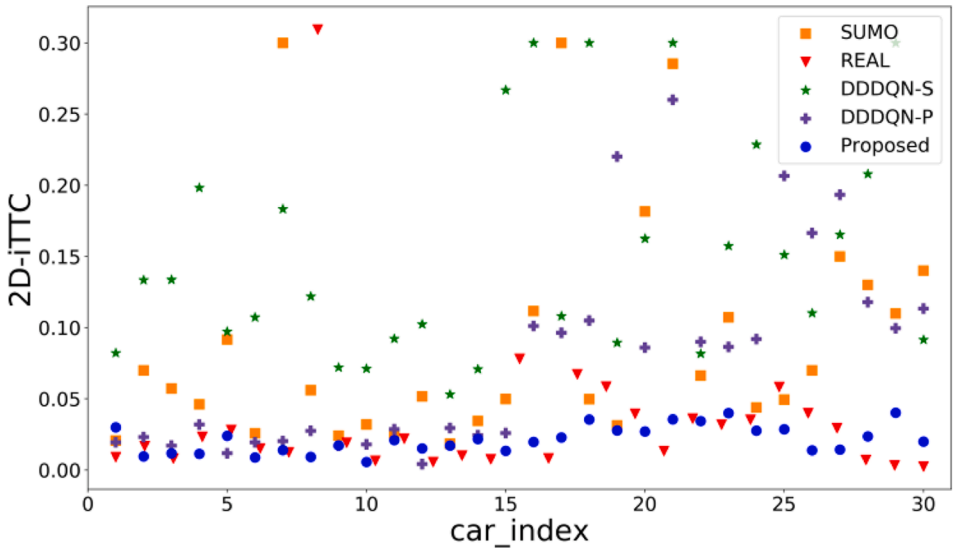


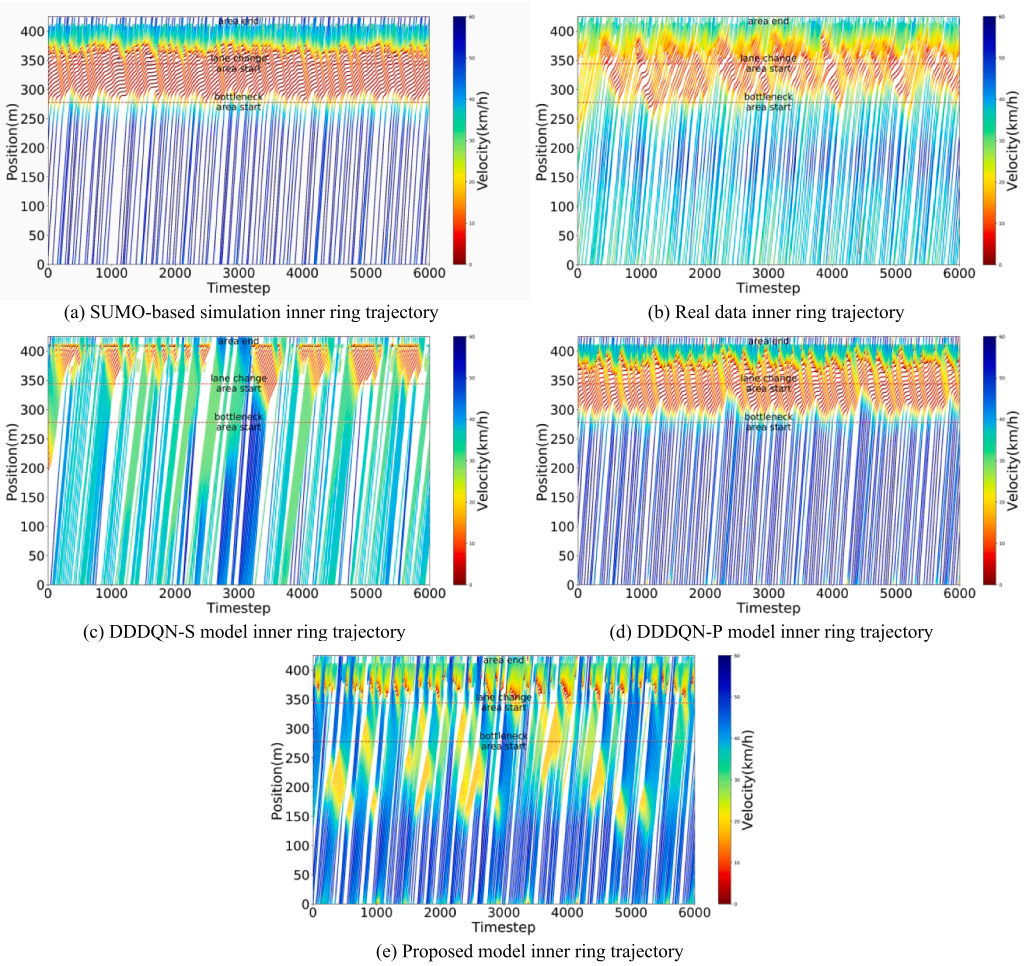
Fig. 10. Comparison of average 2D-iTTC of SUMO-based simulation, real data, DDDQN-S model, DDDQN-P model, and proposed model.

training process, the proposed model starts to converge at around 50 episodes, while the DDDQN-S and DDDQN-P models start to converge at around 70 and 55 episodes, respectively. This shows that the proposed model starts to converge faster with the help of a collision avoidance strategy compared to the DDDQN-based model. Meanwhile, it is further observed that the proposed model is smoother in terms of reward enhancement compared to the DDDQN-based model and has a higher final average reward value. It indicates that the model proposed in this paper performs better during the training period.

To further evaluate the effect of reward design on traffic efficiency, we add two kinds of rewards to compare with real data, including $r_{2DiTTC} + r_{collision} + r_{speed}$ and $r_{2DiTTC} + r_{collision} + r_{speed} + r_{distance}$. The average speed and travel time used to compare these three rewards were averaged over a 150-meter-long bottleneck area. A total of 10 experiments were conducted, and their average values were finally reported. The real data we selected is the average speed and average travel time of 30 vehicles in the bottleneck area. From Table 3, it can be seen that compared with the real data, the proposed model improves the average speed in the bottleneck area by 73.4 %, which is significantly better than the models of the other two reward designs. In addition, it is worth noting that the second model, which takes into account the larger distance before the lane change area, improves the average speed by 21.7 % compared to 7.6 % in the first model, which only takes into account the speed reward in terms of efficiency. This shows that reserving the distance before the mandatory lane change area helps to improve traffic efficiency. As for the average travel time, the proposed model has a 42.3 % reduction compared to the real data, while the other two models have reductions of 6.8 % and 17.7 %, respectively. Combining the above results, the proposed model significantly improves traffic efficiency compared to real data and outperforms the other two models that only consider safety and efficiency without considering vehicle intent.

4.2. Comparison of safety performance

In this section, we compare the 2D-iTTC between the SUMO-based simulation, real data, the DDDQN-S model, the DDDQN-P model, and the proposed model to evaluate safety performance. 2D-iTTC is a two-dimensional extension of TTC, considering vehicle direction, spacing, and speed. Moreover, compared with TTC, 2D-iTTC overcomes the limitation that the denominator is 0 when the speed of the lead and following vehicles is equal. A larger 2D-iTTC value implies higher risk. The average 2D-iTTC comparison of the 30 vehicles in the experiment is shown in Fig. 10. Yellow squares represent the SUMO-based simulation, red triangles indicate the real data, green stars indicate the DDDQN-S model, purple pluses indicate the DDDQN-P model, and blue circles indicate the proposed model. From Fig. 10, it can be seen that the performance of the proposed model is similar to the real data, and no collision occurred, indicating that the proposed model performs well in terms of safety. In addition, the 2D-iTTC value from the DDDQN-S model is larger than the other models, indicating that the risk of driving is higher (but does not exceed the safety threshold



(a) SUMO-based simulation inner ring trajectory

(b) Real data inner ring trajectory

(c) DDDQN-S model inner ring trajectory

(d) DDDQN-P model inner ring trajectory

(e) Proposed model inner ring trajectory

Fig. 11. Comparison of inner ring trajectories of SUMO-based simulation, real data, DDDQN-S model, DDDQN-P model, and proposed model.

and the traffic is still safe). The DDDQN-P model, which incorporates our proposed collision avoidance strategy, performs close to SUMO, lying between DDDQN-S and the proposed model. This suggests that the proposed collision avoidance strategy provides some improvement in safety. The reason for the higher risk of the DDDQN-S model is that it is only a lane-changing decision model, and the rule-based lane-changing model does not take into account the lane-changing process. In order to investigate the activation frequency of the collision avoidance strategy, we set up triggers during the testing process. Based on the results from the triggers, we calculated the average activation frequency per vehicle during the simulation time. The proposed model showed an average of 0.04 activations per vehicle, indicating a low frequency of activation. In general, the performance of the proposed model in terms of safety is basically aligned with real data and better, outperforming the DDDQN-based model and SUMO-based simulation, while the 2D-iTTC of all models is smaller than the safety threshold.

4.3. Comparison of traffic congestion reduction

In this section, we will compare the SUMO-based simulation, the real data, the DDDQN-S model, the DDDQN-P model, and the proposed model in terms of congestion reduction. To avoid the bias of the data during the warm-up and end phases of the experiment, we intercepted a 600-second segment of the trajectory for analysis, using 120 s after the start of the experiment as the starting point. The analysis focuses on the difference in trajectory between the 150-meter bottleneck area and the 80-meter lane change area.

Fig. 11(a) and Fig. 12(a) represent the inner and outer ring trajectories based on SUMO simulation, respectively. From Fig. 11(b) and 12(b), it can be seen that the real data show severe traffic congestion in the bottleneck area both in the inner and outer rings. This is due to the outer ring vehicles changing lanes to the inner ring in the lane change area, causing disturbance to the rear of the inner ring and thus producing the stop-and-go wave that propagates further upstream. As shown in Fig. 11(c) and 12(c), the DDDQN-S model still has significant traffic congestion in the lane change area of the inner ring, but the affected area is reduced compared to the real data, almost exclusively within the lane change area. This is because the design of the reward function takes into account the buffer distance before the lane change area, leaving space for the lane change. The effect can be seen in the reduction of congestion in the outer ring. In

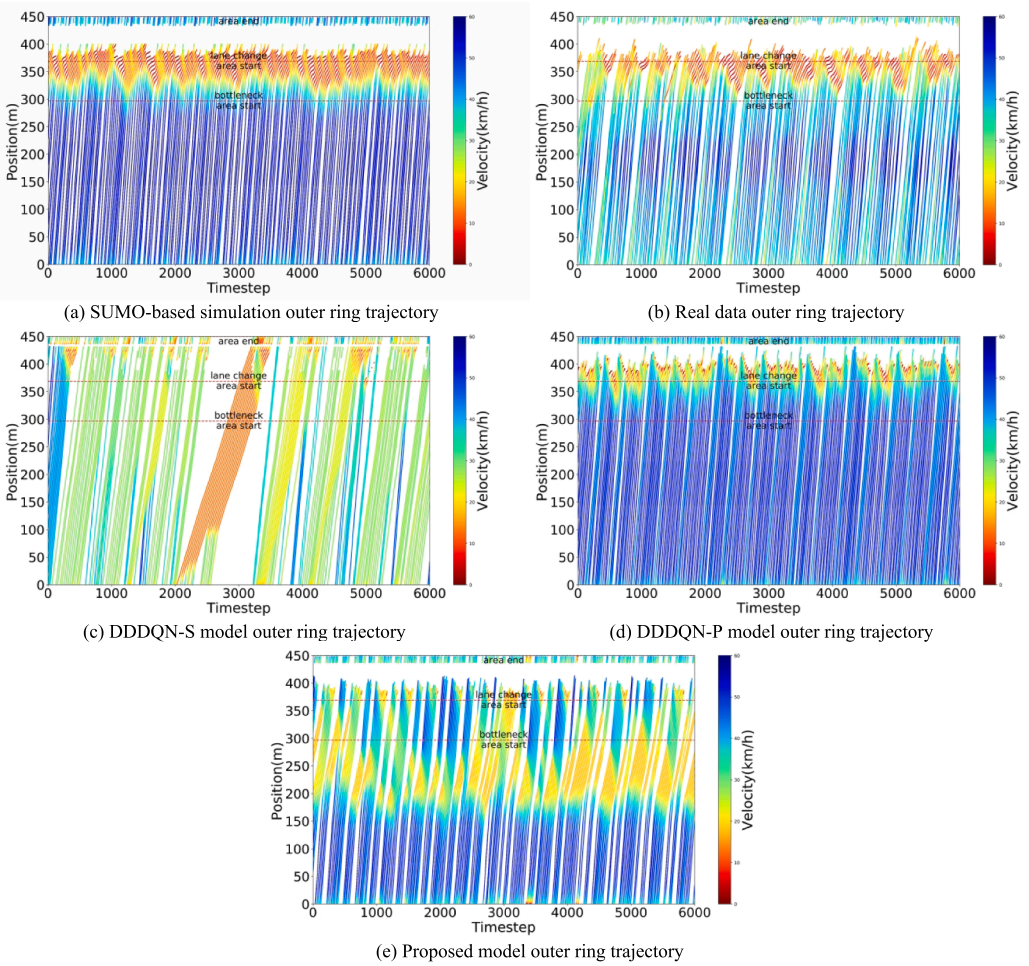


Fig. 12. Comparison of outer ring trajectories of SUMO-based simulation, real data, DDDQN-S model, DDDQN-P model, and proposed model.

addition, we also notice the frequent lane changes in the lane change area and the wide range of deceleration between 2000 and 3000 steps for the vehicles in Fig. 12(c). This is caused by the discrete decisions from the DDDQN-S model. The discrete decisions consist of only lane keeping, left turn, and right turn actions. The commands frequently change at each time step: it may be a left turn at one step, and the next step it might be the opposite command. Moreover, the reward function does not include a penalty for frequent lane changes (an additional lane change penalty term is included in Chen et al., 2021). The reward function is the same as the proposed model for a fair comparison.

As for the wide range of deceleration, it is due to the speed and interval of inner ring vehicles not meeting the lane change rules, causing the outer ring vehicles unable to successfully change lanes. This also indicates that simple discrete decisions combined with rule-based models have greater challenges in practical applications. From Fig. 11(d) and Fig. 12(d), it can be observed that the DDDQN decision model trained with our proposed collision avoidance strategy appears to be more proactive compared to the conservative SUMO rule-based model. The DDDQN-P model leads to faster speeds for both inner and outer ring vehicles outside the bottleneck area. Inside the bottleneck area, the inner ring experiences more severe congestion, while the outer ring remains smoother. This difference is attributed to DDDQN-P being a discrete lane-changing decision model that cannot directly control lateral acceleration. Therefore, we set a fixed lateral speed when executing lane-changing decisions. From the experimental results, it can be seen that this setting leads to significant interference from outer ring vehicles on inner ring vehicles. This also indicates that there are limitations combining the discrete decision model with the proposed 2D collision avoidance model.

Compared with the SUMO-based simulation, the real data, the DDDQN-S model, and the DDDQN-P model, the proposed model performs much better in the lane change area. From Fig. 11(e), only a small amount of deceleration occurs in the inner ring trajectory. But unlike the other models with a wide range of congestion, the congestion area is quickly reduced in the proposed model. From Fig. 12(e), there is almost no serious deceleration in the outer ring trajectory. In addition, the proposed model does not produce frequent lane changes compared to the DDDQN-S model. At the same time, the proposed model shows superior lane-changing coordination in congested areas compared to the DDDQN-P model. It benefits from the continuous action space and execution being integrated, with good feedback from the interaction with the environment. Before entering the bottleneck area, agents start slowing

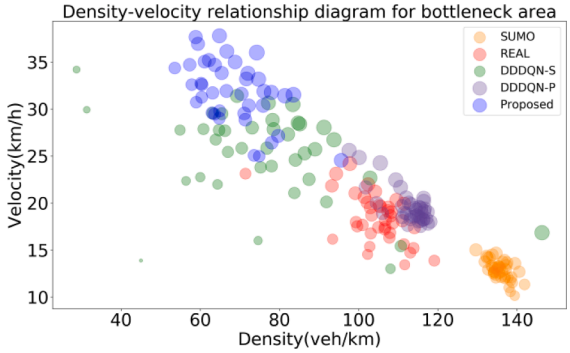


Fig. 13. Density-velocity relationship diagram for bottleneck area.

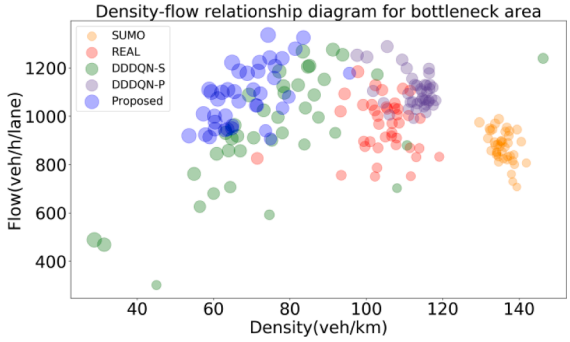


Fig. 14. Density-flow relationship diagram for bottleneck area.

down early to maintain proper spacing, allowing some distance for outer ring vehicles to change lanes.

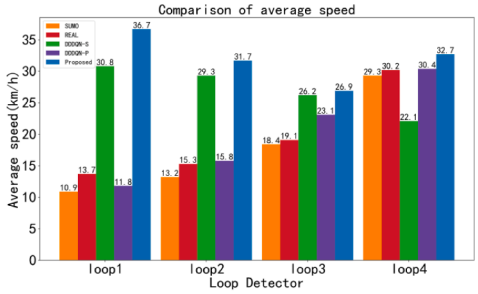
4.4. Comparison of traffic efficiency

In this section, we compare both qualitatively and quantitatively the SUMO-based simulation, the real data, the DDDQN-S model, the DDDQN-P model, and the proposed model in terms of traffic efficiency improvement. The experimental setup is the same as in the previous section, and the same 600-second segment is used for comparison.

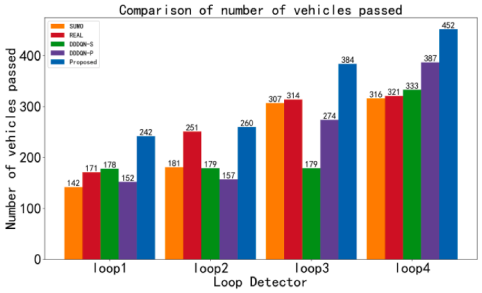
From the perspective of qualitative analysis, we will discuss the relationship between density, velocity, and flow in the bottleneck area. During the experiment, the average density, average velocity, and average flow rate in the 150-meter bottleneck area were counted at 15-second intervals. The size of the circles indicates the magnitude of the flow in Fig. 13 and the magnitude of the velocity in Fig. 14. From Fig. 13, it can be seen that the proposed model has lower density and faster speed in the bottleneck area compared to the other models. This result is also consistent with the observations of the trajectories in the previous section. This is because the strategy of reserving lane change distance reduces the density of the bottleneck area, which further lessens the impact of lane change on the following vehicles, reduces the stop-and-go traffic, and thus increases the average speed. From Fig. 14, the flow rate from the proposed model is higher than the DDDQN-S model at the same density, while at the same flow rate, the density is lower than the DDDQN-P model. In addition, the density of the proposed model is lower than that of the other models for the same flow rate. This also shows that the road traffic conditions are smoother in the proposed model compared to the other models.

The qualitative analysis alone is insufficient to illustrate the proposed model's performance in terms of traffic efficiency. Hence, we further evaluate the impact on traffic efficiency with quantitative analysis. Four 2-meter-long virtual loop detectors were set up to detect the average speed, number of vehicles passing, and average travel time. The position of loop detectors is shown in Fig. 2. Loop 1 is set at the beginning of the lane change area, and loops 2, 3, and 4 are set 15 m, 30 m, and 60 m from the beginning, respectively. Loops 1, 2, and 3 are set in the inner lane, and loop 4 is set in both lanes. From the real data, it is observed that lane change behavior is concentrated in the inner lane, so more loops (1, 2, and 3) are set in the inner lane. For loop 4, it detects the traffic condition of both lanes after vehicles pass the bottleneck.

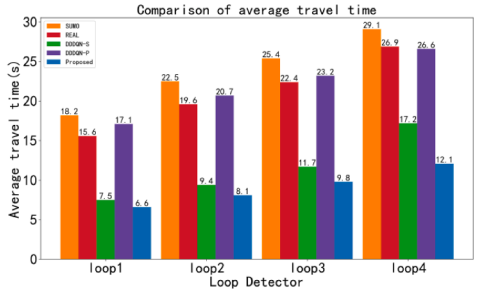
Fig. 15(a) shows the comparison of the average speed of vehicles passing through the different loop detectors in the five scenarios. We can observe that the average speed of the proposed model at loop 1 is the highest, at 36.7 km/h, which is 168 % higher than the real data (13.7 km/h), 19 % higher than the DDDQN-S model (30.8 km/h), and 211 % higher than the DDDQN-P model (11.8 km/h). Meanwhile, the average speed of the proposed model at each loop is greater than the other models. This indicates that the proposed model performs better in terms of average speed. From Fig. 15(b), the total number of vehicles passing loop 4 for the proposed model reaches a maximum of 452, which is 41 % higher than 321 for real data, 36 % higher than 333 for the DDDQN-S model, and 17 %



(a) Comparison of average speed under different loop detectors



(b) Comparison of the number of vehicles passed under different loop detectors



(c) Comparison of average travel time under different loop detectors

Fig. 15. Comparison of average speed, number of vehicles passed, and average travel time under different loop detectors for SUMO-based simulation, real data, the DDDQN-S model, the DDDQN-P model, and the proposed model.

higher than 387 for the DDDQN-P model. We note that the number of vehicles in the real data does not change much from loop 3 to loop 4, while there is an increase in the other scenarios. This implies that the lane change is mostly completed in the real data before loop 3, while the lane change from the DDDQN-S model is concentrated between loop 3 and loop 4, and the lane change from the DDDQN-P model and the proposed model is concentrated between loop 2 and loop 4. Combining Fig. 15(a) and 15(b), we can see that the frequent lane changes affect the average speed in the DDDQN-S model. The speed of the real data, the DDDQN-P model, and the proposed model increase after passing the obstacle, while the speed of the DDDQN-S model decreases. This shows that the proposed model is more reasonable in terms of the distribution of lane changes. In addition, we observe that the proposed model has a higher number of vehicles passing at each loop than the other cases. This indicates that the proposed model performs better than the other models in terms of improving road throughput. The comparison of the average travel time is also presented, where the travel time refers to the travel time from the beginning of the bottleneck area to the loop. From Fig. 15(c), the average travel time of the proposed model is smaller than the other two at each loop. Taking loop 4 near the end of the bottleneck area as an example, the travel time of the proposed model is 12.1 s, which is 55 %, 30 %, and 55 % less than 26.9 s of the real data, 17.2 s of the DDDQN-S model, and 26.6 s of the DDDQN-P model, respectively.

The analysis of the combined qualitative and quantitative results shows that the proposed model has a qualitatively smoother traffic condition compared to the other scenarios and quantitatively faster average speeds, more passing vehicles, and shorter travel times through different loops. This is due to the well-coordinated framework that integrates the buffer zone strategy and the 2D collision avoidance strategy, which enhances both longitudinal and lateral coordination. Overall, the proposed model significantly improves traffic efficiency compared to the SUMO-based simulation, the real data, the DDDQN-S model, and the DDDQN-P model.

5. Conclusions

In this study, a multi-agent reinforcement learning method coupled with the buffer zone strategy and collision avoidance strategies

is proposed to address the challenge of traffic congestion in a mandatory lane change scenario. This method uses the two strategies of distance-keeping and lane-changing coordination to improve traffic efficiency. On the one hand, the distance-keeping strategy creates buffer zones and lane change areas separately. A suitable spacing between vehicles in the buffer zone is maintained to reduce the density of the bottleneck area. It allows for more space for lane change maneuvers in the bottleneck area, reducing vehicle conflicts and increasing average speed. On the other hand, lane-changing coordination is achieved by considering driving intentions and collision avoidance strategies of vehicles on different lanes. The driving intentions of vehicles are different for the inner ring and outer ring vehicles, while the proposed 2D collision avoidance strategy determines the priority between vehicles and coordinates the passing order to make the lane change smoother. The proposed model is trained and tested in a scenario based on a real two-lane circular bottleneck experimental setup. Results from time-space and fundamental diagrams showed that the proposed model significantly reduced traffic congestion, increased the average speed and flow rate, and decreased the density compared to both the real data and the state-of-the-art baseline model (DDDQN-based model). The proposed model also improved the number of passed vehicles by 41 %, 36 %, and 17 %, and reduced the average travel time by 55 %, 30 %, and 55 % compared to the real data, the DDDQN-S model, and the DDDQN-P model, respectively. Both qualitative and quantitative results show that the proposed model effectively reduced congestion and improved traffic efficiency in the bottleneck area.

The proposed model in the study is trained and tested in a pure CAV environment. It is not readily applicable to be extended to a mixed traffic scenario with human-driven vehicles. Especially, the randomness of human drivers in a mixed traffic flow scenario will pose challenges to the robustness of the model. There are extensive studies in the literature on modeling the lane change behavior of human drivers. Our future work intends to first model the lane change behavior of human drivers in the bottleneck area based on real data, accounting for the stochasticity and heterogeneity of human behavior, and then extend the proposed model to incorporate the interaction between human drivers and CAVs in a mixed traffic flow.

CRediT authorship contribution statement

Shupei Wang: Conceptualization, Methodology, Formal analysis, Software, Writing – original draft. **Ziyang Wang:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Rui Jiang:** Conceptualization, Methodology, Data curation, Writing – review & editing. **Feng Zhu:** Conceptualization, Methodology, Data curation, Writing – review & editing. **Ruidong Yan:** Conceptualization, Writing – review & editing. **Ying Shang:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 72288101, 71931002).

References

Bouton, M., Nakhaei, A., Fujimura, K., Kochenderfer, M.J., 2019. Cooperation-aware reinforcement learning for merging in dense traffic. 2019 IEEE Intell Transp. Syst. Conf. ITSC 2019, 3441–3447. https://doi.org/10.1109/ITSC.2019.8916924.

Cao, D., Wu, J., Wu, J., Kulcsár, B., Qu, X., 2021. A platoon regulation algorithm to improve the traffic performance of highway work zones. Comput. Civ. Infrastruct. Eng. 36, 941–956. https://doi.org/10.1111/mice.12691.

Chen, S., Dong, J., Ha, P., Li, Y., Labi, S., 2021. Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles. Comput. Civ. Infrastruct. Eng. 36, 838–857. https://doi.org/10.1111/mice.12702.

Chen, D., Laval, J.A., Ahn, S., Zheng, Z., 2012. Microscopic traffic hysteresis in traffic oscillations: a behavioral perspective. Transp. Res. Part b: Methodol. 46, 1440–1453.

Chowdhury, D., Santen, L., Schadschneider, A., 2000. Statistical physics of vehicular traffic and some related systems. Phys. Rep. 329 (4–6), 199–329.

Ding, J., Li, L., Peng, H., Zhang, Y., 2020. A rule-based cooperative merging strategy for connected and automated vehicles. IEEE Trans. Intell. Transp. Syst. 21, 3436–3446. https://doi.org/10.1109/TITS.2019.2928969.

Fu, X., Jiang, Y., Huang, D., Huang, K., Wang, J., 2017. Trajectory planning for automated driving based on ordinal optimization. Tsinghua Sci. Technol. 22, 62–72. https://doi.org/10.1109/TST.2017.7830896.

Gipps, P.G., 1981. A behavioural car-following model for computer simulation. Transp. Res. Part B 15, 105–111. https://doi.org/10.1016/0191-2615(81)90037-0.

Guo, Q., Angah, O., Liu, Z., Ban, X., (Jeff),, 2021. Hybrid deep reinforcement learning based eco-driving for low-level connected and automated vehicles along signalized corridors. Transp. Res. Part C Emerg. Technol. 124, 102980 https://doi.org/10.1016/j.trc.2021.102980.

Han, Y., Ramezani, M., Hegyi, A., Yuan, Y., Hoogendoorn, S., 2020. Hierarchical ramp metering in freeways: an aggregated modeling and control approach. Transp. Res. Part C Emerg. Technol. 110, 1–19. https://doi.org/10.1016/j.trc.2019.09.023.

Han, Y., Wang, M., Li, L., Roncoli, C., Gao, J., Liu, P., 2022a. A physics-informed reinforcement learning-based strategy for local and coordinated ramp metering. Transp. Res. Part C Emerg. Technol. 137, 103584 https://doi.org/10.1016/j.trc.2022.103584.

Han, Y., Hegyi, A., Zhang, L., He, Z., Chung, E., Liu, P., 2022b. A new reinforcement learning-based variable speed limit control approach to improve traffic efficiency against freeway jam waves. Transp. Res. Part C Emerg. Technol. 144, 103900 https://doi.org/10.1016/j.trc.2022.103900.

Jiang, L., Xie, Y., Evans, N.G., Wen, X., Li, T., Chen, D., 2022. Reinforcement Learning based cooperative longitudinal control for reducing traffic oscillations and improving platoon stability. Transp. Res. Part C Emerg. Technol. 141, 103744 https://doi.org/10.1016/j.trc.2022.103744.

Karimi, M., Roncoli, C., Alecsandru, C., Papageorgiou, M., 2020. Cooperative merging control via trajectory optimization in mixed vehicular traffic. Transp. Res. Part C Emerg. Technol. 116, 102663 https://doi.org/10.1016/j.trc.2020.102663.

Kong, J., Pfeiffer, M., Schildbach, G., Borrelli, F., 2015. Kinematic and dynamic vehicle models for autonomous driving control design. IEEE Intell. Veh. Symp. Proc. 2015-August, 1094–1099. Doi: 10.1109/IVS.2015.7225830.

Li, G., Yang, Y., Li, S., Qu, X., Lyu, N., Li, S.E., 2022. Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness. Transp. Res. Part C Emerg. Technol. 134, 103452 https://doi.org/10.1016/j.trc.2021.103452.
Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wiessner, E., 2018. Microscopic traffic simulation using SUMO. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 2575–2582. Doi: 10.1109/ITSC. 2018.8569938.
Markakis, M.G., Talluri, K., Tikhonenko, D., 2022. Managing lane-changing of algorithm-assisted drivers. Transp. Res. Part C Emerg. Technol. 138, 103586 https://doi.org/10.1016/j.trc.2022.103586.
Memarian, A., Rosenberger, J.M., Mattingly, S.P., Williams, J.C., Hashemi, H., 2019. An optimization-based traffic diversion model during construction closures. Comput. Civ. Infrastruct. Eng. 34, 1087–1099. https://doi.org/10.1111/mice.12491.
Nishi, T., Doshi, P., 2019. Merging in congested freeway traffic using multipolicy decision making and passive actor-critic learning. IEEE Trans. Intell. Veh. 3, 453–462. https://doi.org/10.1109/TIV.2018.2873899.
Ren, T., Xie, Y., Jiang, L., 2020. Cooperative highway work zone merge control based on reinforcement learning in a connected and automated environment. Transp. Res. Rec. 2674, 363–374. https://doi.org/10.1177/0361198120935873.
Schulman, J., Moritz, P., Levine, S., Jordan, M.I., Abbeel, P., 2016. High-dimensional continuous control using generalized advantage estimation. 4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc. 1–14.
Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal Policy Optimization Algorithms 1–12.
Sun, Z., Huang, T., Zhang, P., 2020. Cooperative decision-making for mixed traffic: A ramp merging example. Transp. Res. Part C Emerg. Technol. 120, 102764 https://doi.org/10.1016/j.trc.2020.102764.
Tajalli, M., Niroumand, R., Hajbabaie, A., 2022. Distributed cooperative trajectory and lane changing optimization of connected automated vehicles: Freeway segments with lane drop. Transp. Res. Part C Emerg. Technol. 143, 103761 https://doi.org/10.1016/j.trc.2022.103761.
Tang, Z., Zhu, H., Zhang, X., Iryo-Asano, M., Nakamura, H., 2022. A novel hierarchical cooperative merging control model of connected and automated vehicles featuring flexible merging positions in system optimization. Transp. Res. Part C Emerg. Technol. 138, 103650 https://doi.org/10.1016/j.trc.2022.103650.
Wang, G., Hu, J., Li, Z., Li, L., 2021. Harmonious lane changing via deep reinforcement learning. IEEE Trans. Intell. Transp. Syst. 1–9 https://doi.org/10.1109/TITS.2020.3047129.
Wang, Y., Wang, L., Guo, J., Papamichail, I., Papageorgiou, M., Wang, F.Y., Bertini, R., Hua, W., Yang, Q., 2022. Ego-efficient lane changes of connected and automated vehicles with impacts on traffic flow. Transp. Res. Part C Emerg. Technol. 138, 103478 https://doi.org/10.1016/j.trc.2021.103478.
Wei, E., Luke, S., 2016. Lenient learning in independent-learner stochastic cooperative games. J. Mach. Learn. Res. 17, 1–42.
Wu, Y., Tan, H., Qin, L., Ran, B., 2020. Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm. Transp. Res. Part C Emerg. Technol. 117, 102649 https://doi.org/10.1016/j.trc.2020.102649.
Xiong, B.K., Jiang, R., Li, X., 2022. Managing merging from a CAV lane to a human-driven vehicle lane considering the uncertainty of human driving. Transp. Res. Part C Emerg. Technol. 142, 103775 https://doi.org/10.1016/j.trc.2022.103775.
Xue, Y., Zhang, X., Cui, Z., Yu, B., Gao, K., 2023. A platoon-based cooperative optimal control for connected autonomous vehicles at highway on-ramps under heavy traffic. Transp. Res. Part C Emerg. Technol. 150, 104083 https://doi.org/10.1016/j.trc.2023.104083.
Zhang, H., Li, Z., Liu, P., Xu, C., Yu, H., 2013. Control strategy of variable speed limits for improving traffic efficiency at merge bottleneck on freeway. Procedia - Soc. Behav. Sci. 96, 2011–2023. https://doi.org/10.1016/j.sbspro.2013.08.227.
Zhang, C., Sabar, N.R., Chung, E., Bhaskar, A., Guo, X., 2019. Optimisation of lane-changing advisory at the motorway lane drop bottleneck. Transp. Res. Part C Emerg. Technol. 106, 303–316. https://doi.org/10.1016/j.trc.2019.07.016.
Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. Accid. Anal. Prev. 42, 626–636. https://doi.org/10.1016/j.aap.2009.10.009.
Zhu, J., Tasic, I., Qu, X., 2022. Flow-level coordination of connected and autonomous vehicles in multilane freeway ramp merging areas. Multimodal Transport. 1, 100005 https://doi.org/10.1016/j.multra.2022.100005.