Multi-agent Reinforcement Learning for Cooperative On-Ramp Merging of Connected Automated Vehicles

Boyuan Zhao[1] and Jianxun Cui[1,2](✉)

[1] School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150000, China
cuijianxun@hit.edu.cn

[2] Chongqing Research Institute of HIT, 618 Liangjiang Avenue, Longxing Town, Yubei District, Chongqing, China

Abstract. Ramp merging areas on highways often serve as bottleneck areas, leading to frequent interactions and accidents between vehicles on the ramp and the arterial road. This results in severe congestion and reduced traffic performance. The emergence of Connected Autonomous Vehicles (CAVs) offers advanced solutions to address these issues and improve traffic operations at ramp merging areas. While previous studies have explored CAV decision-making approaches such as optimization control, model predictive control, and reinforcement learning, they face difficulties in accurately modeling the complex and dynamic scenarios of ramp merging. To overcome these challenges, this paper proposes a collaborative decision-making and control model based on Multi-agent Reinforcement Learning (MARL) for mixed vehicles (CAV-HDV) in multi-lane ramp merging scenarios on arterial roads. The paper introduces three novel MARL algorithms and conducts simulations in six different scenarios to evaluate traffic performance under various lane numbers and traffic densities. The results demonstrate the effectiveness of the proposed collaborative model for ramp merging vehicles. The proposed algorithms significantly reduce collision rates and improve traffic efficiency.

Keywords: Highway Ramp Merging · Multi Agent Reinforcement Learning · Connected Automatic Vehicle · Decision-making and Control

1 Introduction

The merging of ramp vehicles can cause chaos in the arterial road traffic flow, leading to a decrease in driving speed and an increase in delay, even more accidents. It can be seen that the center of merging on highway ramps involves that vehicles cooperate with each other to ensure traffic safety and efficiency, which is a practical research issue and a "strong interaction" problem.

For highway ramp merging, there have been traditional traffic flow level control methods in the past, such as ZONE [1] and ALINEA [2], which are both macro level

X. Chen et al. (Eds.): IoTaaS 2023, LNICST 585, pp. 200–221, 2025.
https://doi.org/10.1007/978-3-031-70507-6_16

controls. Although they have improved the traffic efficiency of ramp merging vehicles, insufficient considerations and limitations of technology at the time still cause shortcomings, for example, less of accuracy and randomness. On the basis of the above, Connected Autonomous Vehicle (CAV) has become the best choice to solve the problem of highway ramp merging. Unlike traditional cars, CAVs are equipped with advanced sensors, controllers, integrating environmental awareness, intelligent decision-making, and collaborative control. CAV provides two characteristics: automatic driving controllability and Vehicle-to-everything connectivity. Compared to traffic flow level control, CAV achieves more collaborative control, transitioning from traffic flow level control to individual level control, providing the possibility of further improving the safety and efficiency of ramp merging.

In recent years, optimization control [3], MPC [4] and Single-agent Reinforcement learning methods [5] are widely used in the research of CAV decision-making and control problems. By optimizing the decision-making and control performance of individual CAV, the overall traffic flow performance is improved to a certain extent. Despite the achievement, for freeway merging scenarios, there are complex, dynamic and nonlinear interactive decision-making among CAVs, resulting in difficulties for efficient decision-making and cooperation control.

The existing highway ramp merging control methods considering CAV are divided into two categories based on traffic flow composition conditions: complete CAV scenario and mixed scenario of CAV and Human Driving Vehicle (HDV). For the research in the complete CAV scenario, the high-level decisions of CAV mainly include the optimization of confluence gap [6, 7], the arrangement of merging sequence [8], and Game theory coordination decisions [9, 10]. However, there will be a long period of coexistence between CAVs and HDVs before all vehicles on the road become CAVs, where HDV is uncontrollable. Therefore, in the actual traffic scenario, the control of CAV needs to resist the interference generated by HDV. At present, the mainstream research methods in mixed scenarios mainly include trajectory optimization method [11–15], Game theory method [16, 17] and Reinforcement learning method [18–20]. Among many research methods, Reinforcement learning has great potential in the high-level decision-making learning of CAVs, which can significantly improve the traffic efficiency and safety of ramp merging, and has received extensive attentions from researchers in recent years. Despite the advantages, few researchers have combined low-level control of CAV with high-level decision-making to design a collaborative comprehensive system. At the same time, there is also insufficient consideration for the lane changing behavior of mixed vehicles in the scene of multi-lane ramp merging on the arterial road.

In view of this, this paper aims to conduct cooperative merging decision making modeling of CAVs based on MARL and constructs a comprehensive framework for vehicle merge decision and control, which realizes the coordination of low-level control and high-level decision-making. We take MAPPO as a baseline MARL algorithm and also improve algorithm performance and scene adaptability respectively from the perspective of group collaborative optimization, proposing 2 specified versions of MAPPO. Finally, the two mechanisms were combined to obtain the DDNRCC-MAPPO algorithm, which further improves the overall traffic flow operation efficiency and safety characteristics. At the same time, through extensive experiments on the Highway simulation platform

in six scenarios with different number of lanes and vehicle density on the arterial road, 3 novel algorithms and 2 classic algorithms, were used for training and comparison. We systematically verified the efficiency and safety of the three novel algorithms from quantitative analysis of algorithm performance, traffic performance and training mechanism effectiveness.

2 Methods

The main content of this section is the collaborative decision-making modeling of highway ramp merging vehicles based on MARL. Firstly, we give the detail description of the problem we are addressing in this research, and a systematic framework for decision-making and control of merging vehicles on highway ramps is designed. Then, the control of HDV and the low-level control and high-level collaborative decision-making parts of CAV are elaborated separately. Next, we construct a Markov model for the collaborative decision-making problem of merging vehicles, in which the state space, action space, and reward function are presented. Finally, based on MARL, the algorithm and training mechanism of highway ramp merging vehicle collaborative decision-making are carried out.

2.1 Problem Statement

This research delves into the issue of ramp merging in a situation involving both CAV and HDV. The scene of ramp merging is presented in Fig. 1, showcasing a mix of CAV (depicted in red) and HDV (depicted in grey) in a composite driving environment. Cars traversing the secondary roadway amalgamate onto the arterial road via the ramp.

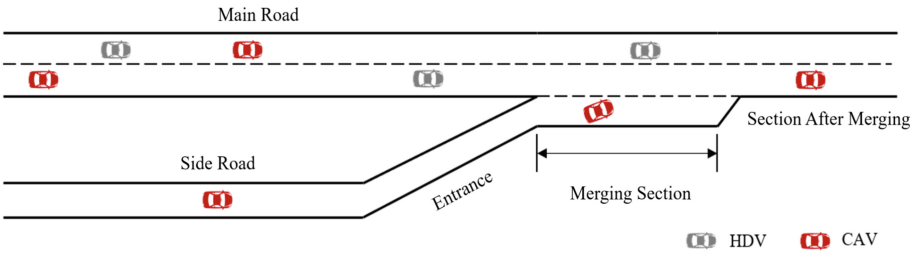

Fig. 1. A figure of highway ramp merging scenario.

In this context, CAVs are able to establish network communication. HDV is driven by the car following model IDM [21] and lane changing model MOBIL [22] for control. While MOBIL handles the lateral control, IDM manages the longitudinal control - with additional information on these coming up in the subsequent sections. A two-layer collaborative framework (see Fig. 2) consisting of low-level control and high-level decision-making is proposed for the collaborative decision-making control problem of CAVs. Therefore, the collaborative decision-making of CAV in the process of highway ramp merging is the primary research focus of this paper.

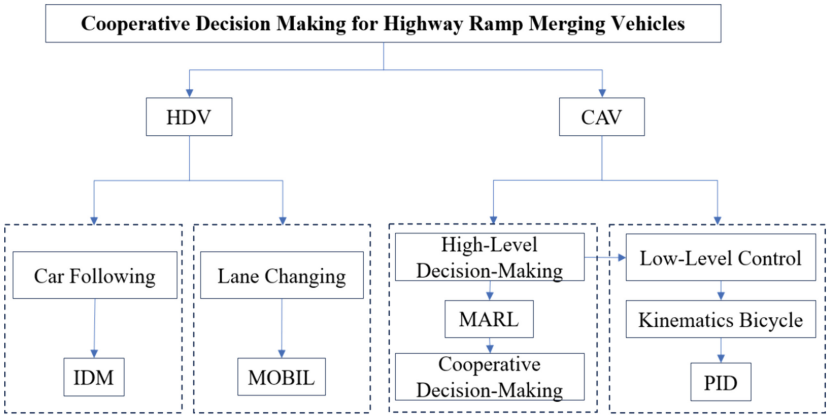

Fig. 2. Framework of vehicle decision-making and control method.

2.2 Decision Making and Control of HDV

Longitudinal Control (IDM). IDM Which is a rule-based car following model is employed to model longitudinal control of HDV. IDM was originally proposed in the field of adaptive cruise control (ACC) to generate appropriate acceleration for the ego vehicle based on its relative driving state with the leading on a single lane. the longitudinal control formulas described by IDM are shown in Eq. 1–2.

$$a = a_{max}(1 - \left(\frac{v}{v_d}\right)^{\delta} - (\frac{d^*(v, \Delta v)}{d})^{2}) \tag{1}$$

$$d^*(v, \Delta v) = d_{min} + vT + \frac{v\Delta v}{2\sqrt{ba_{max}}} \tag{2}$$

where, a is the instant accerlation of ego vehicle, which is needed to be determined in each decision step; a_{max} is the maximum acceleration of ego vehicle; v, v_d is the current and desired speed of ego vehicle; Δv is the speed difference between ego vehicle and its leading vehicle; d is the gap between ego vehicle and its leading vehicle; d_{min} is the minimum safety gap between ego vehicle and its leading vehicle; T is safe time headway; b is the desired acceleration of ego vehicle.

As is seen in the Eq. 1–2, original IDM model only restricted the acceleration of ego vehicle by maximum acceleration a_{max}, however the minimum deceleration is not indicated. So, a condition depicted by Eq. 3 is added by us to limit minimum deceleration of ego vehicle.

$$a = \left\{ \begin{array}{c} a, a \geq a_{min} \\ a_{min}, otherwise \end{array} \right\} \tag{3}$$

where, a_{min} is the minimum deceleration allowed.

In practice, the HDVs on each single lane execute the IDM longitudinal decision-making model respectively, and then generate their own acceleration decisions in each

time interval. If there is no leading vehicle in front of an HDV, its Δv and d is set to 0 and d_{max} (maximum gap for empty lane).

Lateral control (MOBIL). MOBIL is a rule-based lane change model and is adopted to make lateral decision of HDV. MOBIL determines whether lane change is safe and accessible according to the relative acceleration between the ego vehicle and the vehicles on the adjacent lanes. MOBIL's control process is divided into two steps: First, according to the limit of safety standards, the deceleration of new following vehicles should not be too low when lane changing occurs, which is described in Eq. 4.

$$\hat{a}_{new-follower} > b_{safe} \tag{4}$$

where, $\hat{a}_{new-follower}$ is the acceleration of new following vehicles after lane change of ego vehicle, which can be calculated by IDM; b_{safe} is the maximum safe deceleration. Second, if the first condition defined in Eq. 4 is met, MOBIL will check the second condition defined in Eq. 5 to make final decision about whether trigger a lane change of ego vehicle.

$$\hat{a}_{ego} - a_{ego} + p\left(\hat{a}_{new-follower} - a_{new-follower}\right) + q\left(\hat{a}_{old-follower} - a_{old-follower}\right) > a_{th} \tag{5}$$

where, $\hat{a}_{ego}, a_{ego}$ are the new acceleration of ego vehicle calculated by IDM after lane change and the old acceleration before lane change; $\hat{a}_{new-follower}, a_{new-follower}$ are the new and old accelerations respectively of the new follower vehicle when lane change of ego vehicle occurs. $\hat{a}_{old-follower}, a_{old-follower}$ Are the new and old accelerations respectively of the old follower vehicle when lane change of ego vehicle occurs; p and q are politeness factors respectively of the new and old following vehicles; a_{th} is a predefined threshold value.

2.3 Decision Making and Control of HDV

High-level Decision Making of CAV. The collaborative decision-making problem of merging vehicles can be defined as a Decentralized Partial Observable Markov Decision Process (Dec-POMDP). Each agent can only observe nearby agents, meaning that CAVs can only perceive and communicate with nearby vehicles. In addition, this problem can be described using $\{\{S_i, A_i, R_i\}, T\}$, in which T represents the state transition function and S_i, A_i, R_i respectively represent the state, action, and reward of the i-th vehicle. In this section, we will give a detailed introduction to the design of these parts.

(1) **State space.** Effective state representation directly affects the performance of deep reinforcement learning algorithm. The state space is designed as a matrix $N \times F$, where N represents the number of vehicles observed by the ego vehicle and F represents the number of features of each vehicle's state. We assume that the ego vehicle can only observe adjacent vehicles, which are defined as the nearest N vehicles within the range of 75 m forward to 75 m backward in the longitudinal direction of it. Because this paper focus on the ramp merging scenario, we set $N \leq 5$. In addition, when $F = 5$, the 5 features of each vehicle state are designed as follows:

"Whether there are surrounding vehicles", the longitudinal and lateral relative distance between the observed vehicle and the ego vehicle, and the longitudinal and lateral relative speed of the observed vehicle and the ego vehicle. If the state of the i-th agent is defined as S_i, then the joint state of all agents is $S = S_1 \times S_2 \times \cdots \times S_N$.

(2) **Action space**. The action space of this article includes five discrete actions for vehicle speeds and lane changes in ramp merging scenarios: "Acceleration", "Deceleration", "Turn Left", "Keep Current Speed and Lane", and "Turn Right". Define the action of the i-th agent as Ai, and the joint action of all agents is $A = A_1 \times A_2 \times \cdots \times A_N$.

(3) **Reward space**. The design of rewards is crucial to the effectiveness of a reinforcement learning algorithm. In order to encourage safer and more efficient merging process, a multi-objective reward function is proposed from 4 perspectives: safety, stability, efficiency, and constraints, which are defined separately in Eq. 6–10.

$$r_1 = \begin{cases} 0 & no\ collision \\ -1 & otherwise \end{cases} \tag{6}$$

Stability-related reward:

$$r_2 = log\left(\frac{d}{t_h v_t}\right) \tag{7}$$

where, d is the headway of the vehicle, t_h is a predefined time threshold. If the time interval is less than t_h, the agent will be punished.

Efficiency-related reward:

$$r_3 = min\left(\frac{v_t - v_{\min}}{v_{\max} - v_{\min}}, 1\right) \tag{8}$$

where, v_t, $v_{\min}$ and $v_{\max}$ respectively represent the current speed, minimum speed, and maximum speed.

Constraint-related reward:

$$r_4 = -\exp\left(-\frac{(x-L)^2}{10L}\right) \tag{9}$$

where, x is the distance traveled by the CAV on the ramp merging area, and L is the length of the ramp merging area. As the CAV approaches the merging end, the penalty increases to avoid deadlock (too long waiting time in the merging area).

In addition, due to the presence of multiple lanes on the arterial road in a multi-lane scenario, lane changing behavior can affect vehicle driving. Therefore, r_5 is designed to ensure driving comfort and reduce frequent lane changes. r_5 is recorded as -1 when the vehicle changes lanes, otherwise 0.

Total reward can be defined as follows:

$$r_{i,t} = w_1 r_1 + w_2 r_2 + w_3 r_3 + w_4 r_4 + w_5 r_5 \tag{10}$$

where, w_1, w_2, w_3, w_4 and w_5 are weight coefficients of different reward components, which can be adjusted to balance during training process.

Low-level Control of CAV

Kinematics Bicycle Model. Utilized in this study is a conventional vehicle kinematics model capable of tracking and describing vehicular motion [23]. The kinematics bicycle model accepts inputs like the steering wheel angle and acceleration output from the PID control algorithm to detail vehicle movements and forecast states such as lateral position, longitudinal position, speed, and yaw angle [24]. Given its capacity to genuinely represent vehicle characteristics, and its efficiency, the kinematics bicycle model stands out among other vehicle kinetics models. Hence, it serves to delineate the lower-level control of CAVs in This paper.

PID Control. The PID control algorithm consists of three control algorithms: proportional, integral, and derivative [25]. During the control process, they cooperate with each other to regulate the errors between input and output. The core formula of PID control algorithm is as follows (Eq. 11):

$$U(t) = K_p \mathrm{err}(t) + K_i \int \mathrm{err}(t)dt + K_d \frac{\mathrm{derr}(t)}{dt} \tag{11}$$

where, K_p is the proportional coefficient, K_i is the integral coefficient, K_d is the differential coefficient, err(t) is the feedback error, $K_p \mathrm{err}(t)$ represents proportional control, $K_i \int \mathrm{err}(t)dt$ represents integral control and $K_d \frac{\mathrm{derr}(t)}{dt}$ represents differential control. Based on this, the control flow of the PID control algorithm is shown below (see Fig. 3).

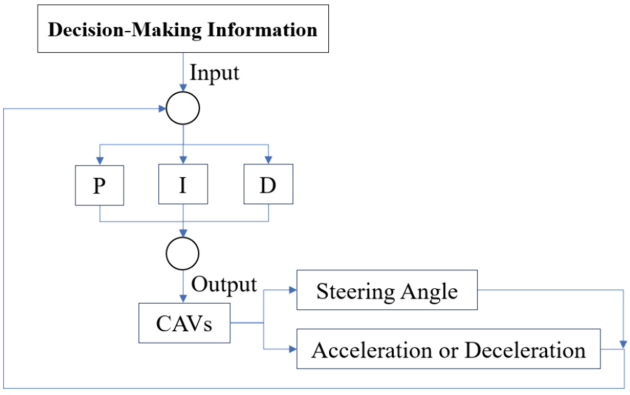

Fig. 3. PID Control-flow diagram.

2.4 Algorithm Design

In the collaborative decision-making of highway ramp merging that involves interactions among CAVs, this research develops an algorithm rooted in MARL for resolution. The training has two paradigms: centralized training and decentralized training. However, centralized training is difficult to consider the individual decision difference from

agents, and decentralized training is less adept at managing environmental instability while achieving globally optimal decisions. Therefore, to progress the collaboration of CAVs on highway ramps, the framework proposed adopts a Centralized Training and Decentralized Execution (CTDE) strategy All agents can leverage information of other agents to learn the parameters of the decision network. Hence, this study improves upon the performance and scenario adaptation of the classic CTDE Multi Agent Proximal Policy Optimization (MAPPO), proposing novel algorithms in the process.

Algorithm Performance Improvement (DDNRC-MAPPO). For developing the algorithm performance, we introduce deep dense network (DDN) to improve the sampling efficiency of agents in the environment, and change the policy network objective function of the algorithm from clipping objective function to rollback clipping objective function to improve the stability of the algorithm.

Deep Dense Network (DDN). MAPPO uses multilayer perceptron (MLP), and both the actor and critic networks contain two fully connected layers, with 128 hidden layer neurons. When faced with high-dimensional state space problems such as highway ramp merging decision-making, this network is difficult to achieve ideal learning results. If more layers are just simply added to MLP, the performance of the agent will only deteriorate. Deep dense networks connect state action pairs to each hidden layer of the network (See Fig. 4), using deeper features to better optimize the network while meeting data processing inequality.

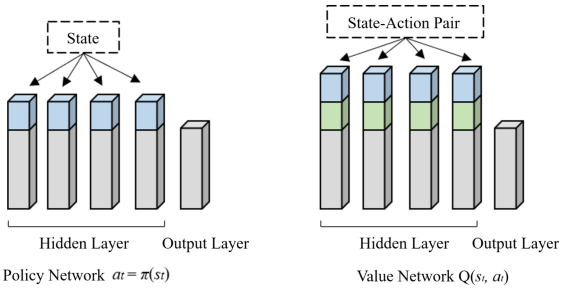

Fig. 4. Schematic diagram of Deep Dense Network.

Fallback Clipping Objective Function. MAPPO cannot strictly control the likelihood ratio as it attempts to do, so it still carries the risk of unstable performance. Therefore, we develop a new clipping function to support fallback operations, which adopts negative incentive measures to limit the differences between the new and old strategies.

Considering the original objective functions in MAPPO does not strictly limit the likelihood ratio within the pruning range, which may exceed $(1 - \varepsilon, 1 + \varepsilon)$. To address this issue, the improved fallback pruning objective function is proposed in the following form (see Fig. 5).

In DDNRC-MAPPO (see Fig. 6), the network of an individual agent has been modified by DDN. Firstly, the actor network interacts with the environment to obtain information. After training, the ratio of the new and old strategies is limited by a backtracking

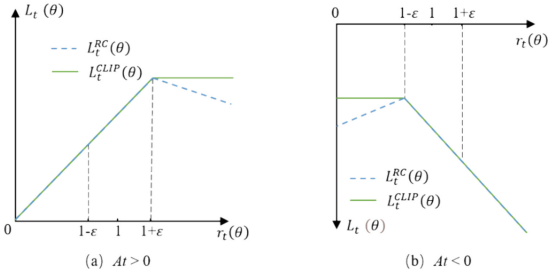

Fig. 5. A figure of improved clipping function.

pruning objective function. Finally, the critic network evaluates the actor network's strategy, which aims to implement strategy optimization.

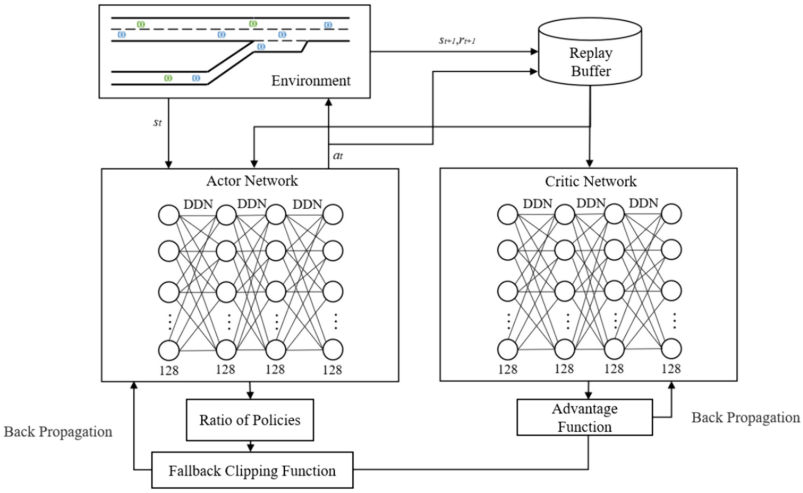

Fig. 6. A basic structure figure of DDNRC-MAPPO.

Scenario Adaptation Improvement (C-MAPPO). In Terms of adapting to ramp merging scenarios, this paper introduces the collision prediction optimization component to predict the driving trajectory of HDVs, which is compared with the trajectory of CAVs to optimize CAV actions. Based on this, a novel algorithm, C-MAPPO was proposed. The algorithm is divided into two operational steps: collision prediction and action optimization.

In the process of predicting a collision, if at any point at any time point $t(t = 1,2, ..., 8)$, the distance between two trajectories is less than the vehicle's width, it is determined that the vehicles are on a collision course. In light of this, optimizing the action selection is a necessity, which involves picking the best action from the available set to modify the present course. The rules for action change include the following two

points: firstly, if the target vehicle takes actions, the component will select the action from the available action set that minimizes the headway between the target vehicle and the colliding vehicle; Secondly, if the target vehicle takes left or right turns when colliding, the component will select the action from the available action set that minimizes the distance along the lane between the target vehicle and the colliding vehicle.

Comprehensive Improvement (DDNRCC-MAPPO). Combining the above algorithms with the improvement of performance and scenario adaptability, integrating the advantages of improving algorithm sampling efficiency and stability, as well as reducing vehicle collision rate, a comprehensive algorithm DDNRCC-MAPPO is proposed to achieve the best optimization effect.

2.5 Reward Mechanism Improvement

In the training process of MARL, the design of reward mechanism is directly related to the training effect and the cooperative decision-making efficiency of agents. In order to improve algorithm performance deeply, we optimized the reward mechanism by introducing local reward and curriculum learning mechanism.

Local Reward Mechanism. When it comes to ramp merging, an issue that involves cooperative MARL, the primary training objective is to maximize the global reward - the cumulative rewards of all agents. even though the global reward captures the total system rewards, it does have limitations. for instance, it doesn't provide specific rewards for each individual agent.

To tackle these challenges, we implement a local reward strategy that considers only the proximal vehicles surrounding our vehicle. The closest n vehicles ($n \leq 5$) within a 75 m stretch in our agent's longitudinal direction can be identified and the local reward computed as the mean reward of these n vehicles.

Curriculum Learning Mechanism. When using reinforcement learning to train agents, the rewards in the environment are sparse in most cases. To address the issue of sparse rewards, we adopt curriculum learning (CL). The concept of curriculum learning was first proposed by Bengio [26] as a training strategy that mimics human learning processes, advocating for agents to start learning from easy samples and gradually advance to complex samples and knowledge (See Fig. 7).

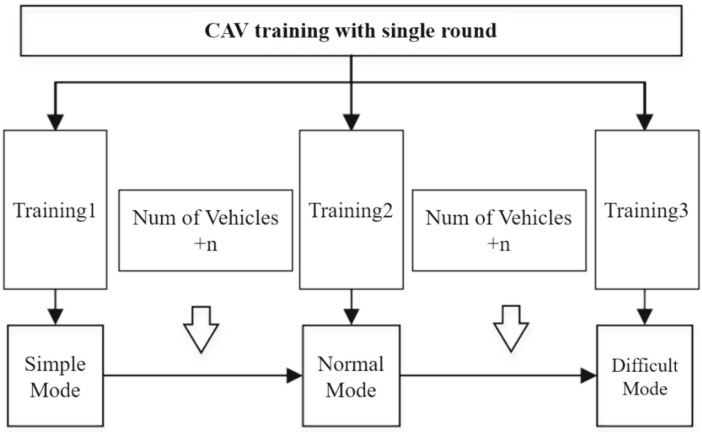

Fig. 7. Schematic diagram of Curriculum Learning.

2.6 Simulation Settings

In this section, simulation scenarios, control models, and algorithm hyper-parameters for highway ramp merging will be designed, and evaluation metrics for subsequent experimental results will be provided from multiple perspectives.

Parameters of HDV Model. In the simulation experiments, HDVs are driven by IDM and MOBIL for making longitudinal and lateral decision and control. The related parameters of IDM and MOBIL are set according to Table 1.

Table 1. Parameters setting of IDM and MOBIL.

Parameters	Value
IDM	
$a_{\max}$	6 m/s^2
$b_{\min}$	− 5 m/s^2
T_0	1.5 s
d_0	10 m
v_d	23–25 m/s
MOBIL	
p	0.1
b_s	2 m/s^2
Acceleration Gain	0.2 m/s^2

Hyper-Parameters of CAVs' High Level Decision-Making Model. The parameter design of the CAV high-level decision-making algorithm is shown in the Table 2 Below.

Table 2. Parameter settings of MARL algorithm.

Parameters	MAACKTR	MAPPO	DDNRC-MAPPO	C-MAPPO	DDNRCC-MAPPO
Actor Learning Rate	0.0005				
Critic Learning Rate	0.0005				
Optimization and Learning Rate	RMSprop, 0.00001				
Account Factor	0.99				
Replay Buffer Size	10000				
Batch Size	100				
Clipping factor ε	/	0.2	0.2	0.2	0.2
Fallback Clipping Factor α	/	/	0.3	/	0.3

Scenario Design. This Experiment was conducted on the Ubuntu 18.04, and all experiments are based on the highway ramp environment provided by highway env. In order to connect with the simulation software highway env, the experimental code is written in python, and the reinforcement learning framework uses PyTorch.

At the same time, we modify the density of traffic flow and set the number of lanes on the arterial road, completing the construction of the simulation scenario for the highway entrance ramp required for this experiment (Single-Lane & Low-Density, Single-Lane & Medium-Density, Single-Lane & High-Density, Double-Lane & Low-Density, Double-Lane & Medium-Density, and Double-Lane & High-Density).

Performance Evaluation Metrics. To give a comprehensive evaluation of the ramp merging vehicle collaborative decision model, six key performance indicators encompassing algorithmic and traffic performance are chosen. This entails three metrics assessing the learning performance of the proposed model and three additional metrics measuring the model's safety and efficiency. They are defined as follows: for collaborative performance:

(1) **Average Episode Reward (AER)**: With each training set at 20000 rounds, AER is the mean reward every 200 rounds.
(2) **Total Average Reward (TAR)**: Evaluating the algorithm's overall performance, TAR is calculated as the sum of rewards garnered from the 20000 rounds of training, averaged by dividing by 20000.
(3) **Average Step (AS_t)**: Each training round comprises of 100 steps. A higher average step size aligns with superior task execution by the agent, and generally a more efficient task completion rate, reflecting the efficacy and duration of task execution.

For traffic performance:

(1) **Average Speed (AS_p)**: A total of 20000 rounds of training will be conducted, and the average speed of every 200 rounds will be taken as the average speed, making it easy to observe the trend of speed changes.
(2) **Total Average Speed (TAS)**: Accumulate the speed obtained from 20000 rounds of training and divide it by 20000 to obtain the total average speed, which can analyze the overall performance of the algorithm.
(3) **Average Collision Rate (ACR)**: Collision can be determined by the step size of the turn, and the collision rate is obtained by dividing the number of turns in collision by the total number of turns of 20000.

With this basis, MAACKTR, MAPPO, DDNRC-MAPPO, C-MAPPO, and DDNRCC-MAPPO were trained through six simulation scenarios. Experimental results were analyzed and evaluated from the standpoint of policy algorithmic performance and traffic performance.

3 Results

In this section, we show the results about performance evaluation of 5 algorithms in six simulation scenarios from policy algorithm performance and traffic performance, and finally a comprehensive comparison among the algorithms is shown.

3.1 Policy Algorithm Performance Evaluation

Firstly, we analyze and evaluate AER. The experimental results of the single lane (in the first line) and two-lanes (in the second line) scenario on the arterial road are shown below (see Fig. 8). And three columns from left to right represent density from low to high respectively. By analyzing the AER of the ramp merging experiment with the scenarios on the arterial road, we can summarize the results as follows.

(1) The comparison between DDNRC-MAPPO proposed in this article and the baseline algorithm MAPPO demonstrates that DDNRC-MAPPO can help agents learn quickly in simple scenarios and significantly improve exploration efficiency, while the improvement is not significant in complex scenarios.
(2) By comparing C-MAPPO proposed in this article with the baseline algorithm MAPPO, it is found that C-MAPPO can significantly improve the reward value, and it can also be said that collisions are a key factor affecting the reward value, which indirectly confirms the role of C-MAPPO in reducing collisions.

(3) In the six scenarios, AER of the agent trained by DDNRCC-MAPPO proposed in the paper is much higher than the baseline algorithm, indicating that DDNRCC-MAPPO has achieved the goal of improving training effectiveness and excellent performance. However, deep dense networks and fallback pruning objective functions can only improve algorithm performance to a certain extent in low density scenarios with two lanes on the arterial road.

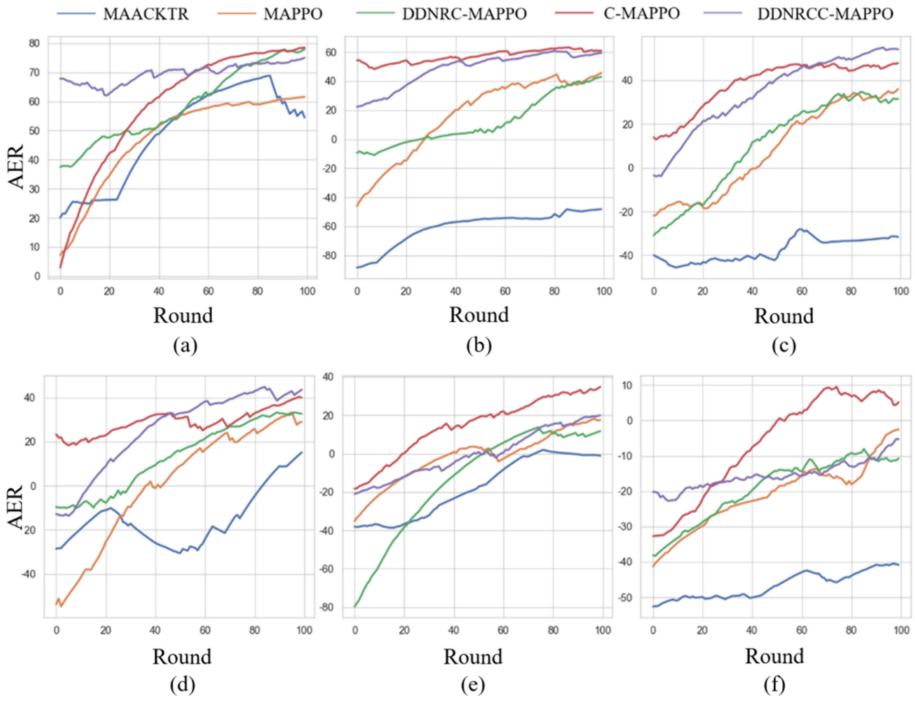

Fig. 8. Comparison of AER under Single Lane Arterial Road Scenarios.

The second policy performance evaluation indicator is AS_t. The relevant experimental results for the scenarios with single lane and two-lanes on the arterial road are shown in Fig. 9. The density from low to high is represented respectively in the 3 columns from left to right. The results can be described as follows.

(1) MAACKTR has little difference in training completion compared to other algorithms in simple scenarios, but in complex scenarios, its AS_t remains at a low value, making it difficult for agents to complete tasks successfully and resulting in poor training effectiveness.

(2) The performance difference between DDNRC-MAPPO and the baseline algorithm MAPPO in terms of training completion is not significant. In simple scenarios, agents can complete tasks smoothly, and in complex scenarios, agents can also achieve high task completion.

(3) AS_t of C-MAPPO and DDNRCC-MAPPO proposed in this paper is generally higher than the baseline algorithm during the training process, improving the task completion of the agent. In addition, the training effect of DDNRCC-MAPPO is slightly better.

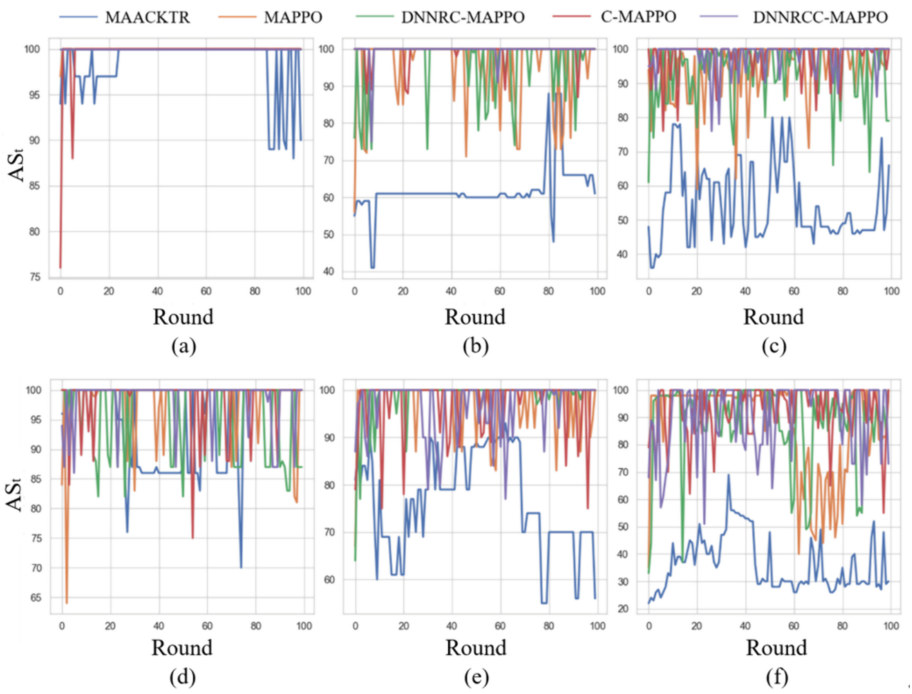

Fig. 9. Comparison of ASt under Single Lane Arterial Road Scenarios.

3.2 Traffic Performance Evaluation

In this section, we further validated algorithms' improvement in traffic performance, including the analysis and evaluation of ASp and ACR.

The experimental results of the single lane (in the first line) and two-lanes (in the second line) scenario on the arterial road are shown below (see Fig. 10). And three columns from left to right represent density from low to high respectively. Based on the analysis of ASp of the ramp merging experiments, the following conclusions can be drawn:

(1) MAACKTR algorithm trains agents with good ASp in low density scenarios, while in medium and high-density scenarios, low task completion results in low average speed reference value, indicating that the baseline algorithm is difficult to adapt to complex scenarios.

(2) The comparison between DDNRC-MAPPO and the baseline algorithm MAPPO shows that in low density scenarios, DDNRC-MAPPO significantly improves ASp of the agent. In medium and high-density scenarios, the training performance of the two algorithms is basically the same. It can be seen that DDNRC-MAPPO has a significant help in improving ASp performance in simple traffic scenarios.
(3) ASp of C-MAPPO and DDNRCC-MAPPO are basically the same in low and medium density scenarios, while in high-density scenarios, DDNRCC-MAPPO can help agents achieve higher ASp. We can see that among the five algorithms, DDNRCC-MAPPO has the best average speed performance.

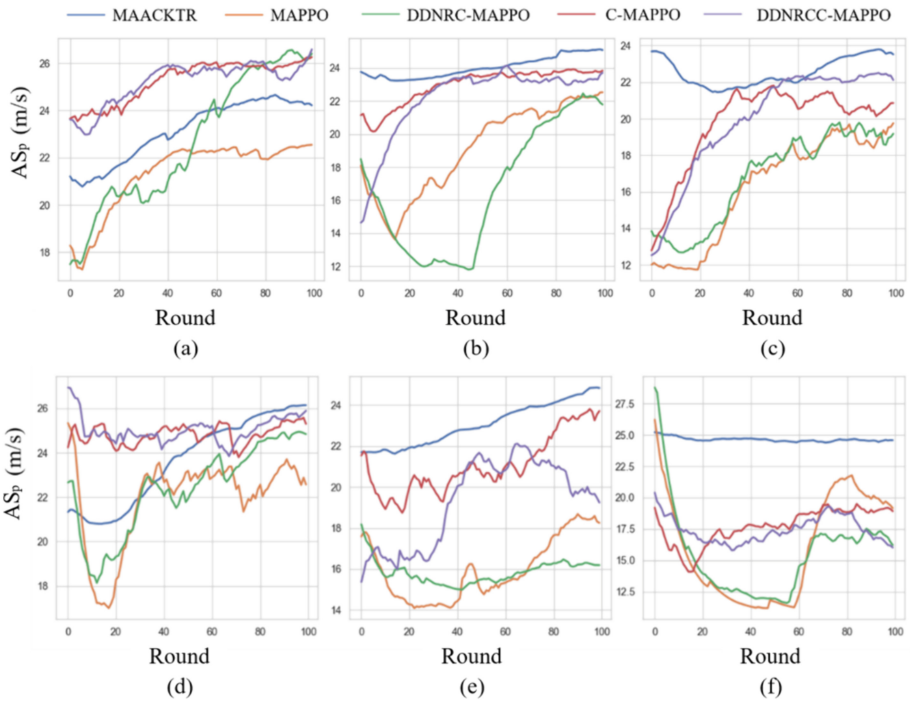

Fig. 10. Comparison of ASp under Single Lane Arterial Road Scenarios.

The ACR (Average Collision Rate) of various algorithms in six distinct scenarios is presented in Table 3. Upon analysis, it becomes evident that, with the exception of the two-lane medium density scenario, DDNRCC-MAPPO consistently exhibits the lowest ACR across all five scenarios. Following closely behind is C-MAPPO, with a slightly higher ACR compared to the former. DDNRC-MAPPO demonstrates the lowest ACR in the single lane low density and two-lane medium density scenarios, while slightly surpassing C-MAPPO and DDNRCC-MAPPO in other scenarios. Conversely, MAACKTR consistently yields the highest ACR in all six scenarios, resulting in poor performance.

Table 3. Comparison of ACR under different scenarios.

Algorithms/Scenarios	Single-Lane			Double-Lane		
	Low-Density	Medium-Density	High-Density	Low-Density	Medium-Density	High-Density
MAACKTR	0.06	0.69	0.92	0.15	0.58	0.98
MAPPO	0.01	0.09	0.24	0.07	0.08	0.39
DDNRC-MAPPO	0	0.09	0.20	0.11	0.05	0.37
C-MAPPO	0.01	0.03	0.08	0.04	0.07	0.19
DDNRCC-MAPPO	0	0.01	0.08	0.04	0.11	0.17

3.3 Comprehensive Performance Evaluation

This section comprehensively analyzes the performance of three improved algorithms and two baseline algorithms, and compares them from four aspects: safety, efficiency, robustness, and effectiveness (see Fig. 11).

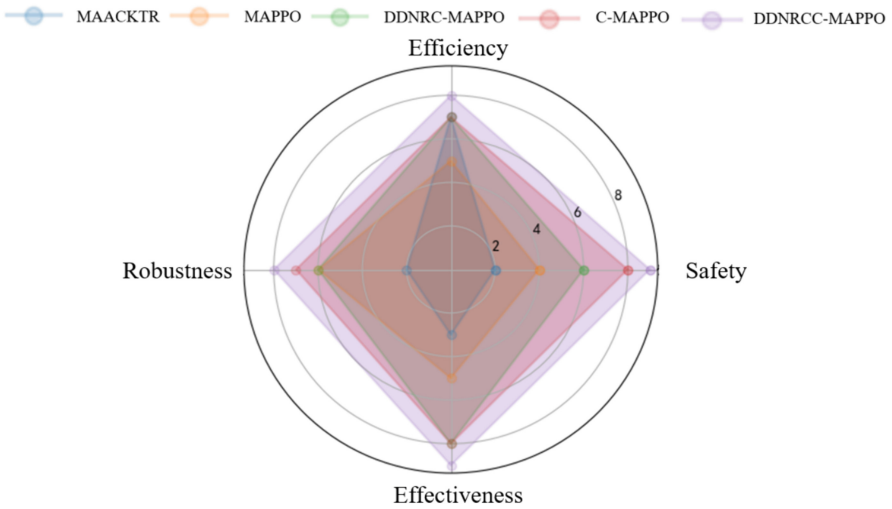

Fig. 11. Comprehensive comparison of algorithm performance.

From Fig. 11, we see that the three algorithms proposed in this paper have improved in terms of security, efficiency, robustness, and effectiveness compared to the baseline algorithms. Among them, DDNRCC-MAPPO has the highest score and the best performance in all four aspects.

4 Discussion

This section further discussed the performance of algorithms in typical scenarios based on the above experimental results and the improved reward mechanisms from 6 indicators.

For the two kinds of reward mechanisms, this experiment takes the improved algorithm DDNRCC-MAPPO as an example to train in a single lane medium density scenario. In this medium difficulty scenario, the impact of using the local reward mechanism and Curriculum Learning on training effectiveness is compared. The experimental results are shown in Fig. 12 and Table 4.

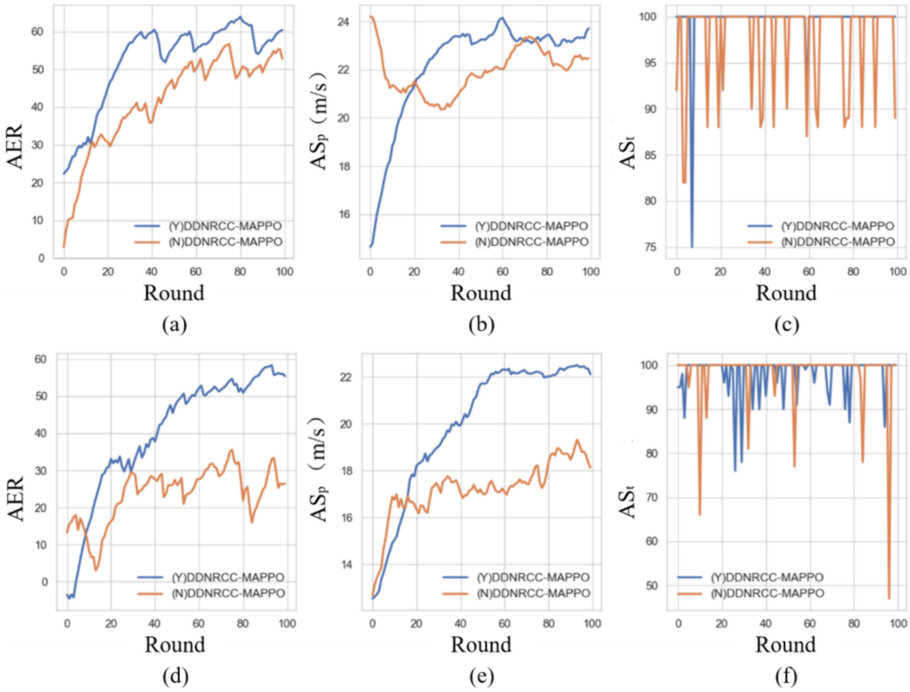

Fig. 12. Performance comparison of reward mechanisms before and after used.

In the 6 subgraphs of Fig. 12, the blue line represents the experimental results using the improved reward mechanisms, while the orange line represents the experimental results without using them. Except for whether using the reward mechanisms, all other settings are the same.

Figure 12 (a, d) show a comparison chart of AER, with the horizontal axis representing training rounds and the vertical axis representing AER. All curves of the two subgraphs show an upward trend, but the blue line is always higher than the orange line, indicating that the use of local reward mechanism and curriculum learning improves the exploration efficiency of the agent and can obtain higher AER.

Figure 12 (b, e) show1 a comparison chart of AS_p. In (b), the blue line in the figure shows an overall upward trend, with small fluctuations. In the later stage of training, it tends to converge and the AS_p is higher than the orange line. In (e), the orange line has a higher AS_p than the blue line in the first 20 rounds but shows a downward trend, then rises and tends to converge in the later training, However, its AS_p is always lower

than the blue line, indicating that curriculum learning can help the agent improve speed faster.

Figure 12 (c, f) show a comparison of AS_t. In (c), the orange line has 19 rounds with an AS_t of less than 100, while the blue line has only 2 rounds with an AS_t of less than 100. It can be seen that using the local reward mechanism can improve the task completion of the agent. And AS_t of algorithms using curriculum learning fluctuates between 76 and 100, while those without curriculum learning fluctuate between 45 and 100. In view of this, it can be seen that the curriculum learning mechanism can make agents complete tasks as much as possible.

Table 4. Performance comparison table of reward mechanisms before and after used.

Whether Using Local Reward Mechanism	TAR	AS_p	ACR
(Y)DDNRCC-MAPPO	55.06	23.01	0.01
(N)DDNRCC-MAPPO	45.97	21.79	0.07
Whether Using Curriculum Learning	TAR	AS_p	ACR
(Y)DDNRCC-MAPPO	45.63	20.86	0.08
(N)DDNRCC-MAPPO	24.85	17.71	0.09

In Table 4, (Y) DDNRCC-MAPPO represents algorithms that use novel reward mechanisms, and (N) DDNRCC-MAPPO represents algorithms that do not use novel reward mechanisms. It can be observed that after using local reward mechanism, TAR of the agent increased from 45.97 to 55.06, TAS increased from 21.79 m/s to 23.01 m/s, and ACR decreased from 7% to 1%. And with curriculum learning, TAR of the agent increased from 24.85 to 45.63, TAS increased from 17.71 m/s to 20.86 m/s, and ACR decreased from 9% to 8%.

Based on the above experimental analysis, it can be found that after using the local reward mechanism and curriculum learning, all evaluation indicators have been improved to varying degrees, indicating that the improvement can effectively help agents explore the environment and improve training effectiveness.

5 Conclusions

The gradual popularization of Connected Autonomous Vehicles (CAV) provides a new solution to the problem of highway ramp confluence. In recent years, there are optimization control, MPC control, single agent Reinforcement learning and other methods for the decision control of CAV, which improves the traffic performance to a certain extent, but insufficient consideration is given to the interactive decision-making between multiple vehicles in complex scenes. For this reason, this paper proposes a MARL based collaborative Decision model and strategy for highway ramp merging vehicles. Based on the MAPPO algorithm, the algorithm performance and adaptation to ramp merging scenarios are improved respectively, and the training mechanism is optimized. Then,

the strategy is tested and evaluated in a series of ramp merging scenarios with different number of lanes on the arterial road and vehicle density. The main research findings can be summarized as follows:

(1) A comprehensive framework was developed for decision-making and control of merging vehicles on highway ramps. This framework incorporates a decentralized partial observable Markov decision model, specifically designed to tackle the collaborative decision-making challenges faced by merging vehicles on highway ramps;
(2) Multiple enhanced versions of collaborative decision algorithms based on MAPPO were designed specifically for Connected Autonomous Vehicles (CAVs) during ramp merging;
(3) The use of local rewards and curriculum learning training mechanisms proved effective in training the proposed MARL models;
(4) Extensive simulation evaluations were carried out to assess the performance of collaborative decision-making strategies for merging vehicles on highway ramps.

For six simulation scenarios, namely Single-Lane & Low-Density, Single-Lane & Medium-Density, Single-Lane & High-Density, Double-Lane & Low-Density, Double-Lane & Medium-Density, and Double-Lane & High-Density, four indicators (AER, AS_t, AS_p and ACR) were used to assess both the algorithm performance and the traffic performance of the merging vehicle collaborative decision-making strategy. This was done from three aspects, and the results from employing baseline algorithms MAACKTR, MAPPO, and improved algorithms DDNRC-MAPPO, C-MAPPO, and DDNRCC-MAPPO were systematically compared. Such comparisons verified the efficiency of our proposed cooperative decision-making strategy, anchored on MARL for on-ramp merging.

References

1. Xin, W., Michalopoulos, P.G., Hourdakis, J., Lau, D.: Minnesota's new ramp control strategy: design overview and preliminary assessment [J]. Transp. Res. Rec. **1867**(1), 69–79 (2004)
2. Hadj-Salem, H., Blosseville, J.M., Papageorgiou, M.: ALINEA: a local feedback control law for on-ramp metering; a real-life study. In: Third International Conference on Road Traffic Control, 1990. (pp. 194-198). IET (1994)
3. He, X., Liu, H.X., Liu, X.: Optimal vehicle speed trajectory on a signalized arterial with consideration of queue. Trans. Res. Part C: Emerg. Technol. **61**, 106–120 (2015). https://doi.org/10.1016/j.trc.2015.11.001
4. Wen, J., Wang, S., Wu, C., Xiao, X., Lyu, N.: A longitudinal velocity CF-MPC model for connected and automated vehicle platooning. IEEE Trans. Int. Trans. Syst. **24**(6), 6463–6476 (2022)
5. Zhou, M., Yu, Y., Qu, X.: Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach. IEEE Trans. Intell. Transp. Syst. **21**(1), 433–443 (2020)
6. Chen, J., Zhou, Y., Chung, E., Ozbay, K.: CAV-based active congestion resolving for improving mainline traffic flow efficiency of a freeway on-ramp merging section. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE

7. Zhu, J., Tasic, I., & Qu, X.: Improving freeway merging efficiency via flow-level coordination of connected and autonomous vehicles. *arXiv preprint* arXiv:2108.01875 (2021)
8. Zhu, J., Tasic, I.: Safety analysis of freeway on-ramp merging with the presence of autonomous vehicles. Accid. Anal. Prev. **152**, 105966 (2021)
9. Hu, Z., Huang, J., Yang, Z., Zhong, Z.: Embedding robust constraint-following control in cooperative on-ramp merging. IEEE Trans. Veh. Technol. **70**(1), 133–145 (2021). https://doi.org/10.1109/TVT.2021.3049866
10. Wang, M., Hoogendoorn, S.P., Daamen, W., van Arem, B., Happee, R.: Game theoretic approach for predictive lane-changing and car-following control. Trans. Res. Part C: Emerg. Technol. **58**(SEP.PT.A), 73–92 (2015). https://doi.org/10.1016/j.trc.2015.07.009
11. Sabouni, E., Cassandras, C.G.: Optimal merging control of an autonomous vehicle in mixed traffic: an optimal index policy. IFAC-PapersOnLine **56**(2), 2353–2358 (2023)
12. Zhou, Y.: Trajectory planning strategies of connected automated vehicles for cooperative on-ramp merging and mainline facilitating maneuvers. Queensland University of Technology (2019)
13. Sun, Z., Huang, T., Zhang, P.: Cooperative decision-making for mixed traffic: a ramp merging example. Trans Res. Part C: Emerg. Technol. **120**, 102764 (2020)
14. Huang, T., Sun, Z.: Cooperative ramp merging for mixed traffic with connected automated vehicles and human-operated vehicles. IFAC-PapersOnLine **52**(24), 76–81 (2019). https://doi.org/10.1016/j.ifacol.2019.12.384
15. Gao, Z., Zhizhou, W., Hao, W., Long, K., Byon, Y.-J., Long, K.: Optimal trajectory planning of connected and automated vehicles at on-ramp merging area. IEEE Trans. Int. Trans. Syst. **23**(8), 12675–12687 (2022). https://doi.org/10.1109/TITS.2021.3116666
16. Le, V.-A., Malikopoulos, A.A.: A cooperative optimal control framework for connected and automated vehicles in mixed traffic using social value orientation. In: 2022 IEEE 61st Conference on Decision and Control (CDC), Cancun, Mexico, pp. 6272–6277 (2022), https://doi.org/10.1109/CDC51059.2022.9993337
17. Liao, X., Zhao, X., Wang, Z.: Game theory-based ramp merging for mixed traffic with unity-SUMO co-simulation. IEEE Trans. Syst., Man, Cybernet.: Syst. **52**(9), 5746–5757 (2022). https://doi.org/10.1109/TSMC.2021.3131431
18. Zhou, S., Zhuang, W., Yin, G., Liu, H., Qiu, C.: Cooperative on-ramp merging control of connected and automated vehicles: distributed multi-agent deep reinforcement learning approach. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, (pp. 402–408) (2022) https://doi.org/10.1109/ITSC55140.2022.9922173
19. Li, M., Li, Z., Wang, S., Zheng, S.: Enhancing cooperation of vehicle merging control in heavy traffic using communication-based soft actor-critic algorithm. IEEE Trans. Intell. Transp. Syst. **24**(6), 6491–6506 (2023). https://doi.org/10.1109/TITS.2022.3221450
20. Wang, S., Fujii, H., Yoshimura, S.: Generating merging strategies for connected autonomous vehicles based on spatiotemporal information extraction module and deep reinforcement learning. Phys. A: Stat. Mech. Appl. **607**, 128172 (2022)
21. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E **62**(2), 1805–1824 (2000). https://doi.org/10.1103/PhysRevE.62.1805
22. Kesting, A., Treiber, M., Helbing, D.: General lane-changing model MOBIL for car-following models. Trans. Res. Rec.: J. Trans. Res. Board **1999**(1), 86–94 (2007). https://doi.org/10.3141/1999-10
23. Polack, P., Altché, F., d'Andréa-Novel, B., de La Fortelle, A.: The kinematic bicycle model: a consistent model for planning feasible trajectories for autonomous vehicles?. In: 2017 IEEE intelligent vehicles symposium (IV) (pp. 812-818). IEEE. (2017) https://doi.org/10.1109/IVS.2017.7995816

24. Park, F.C., Lynch, K.M.: Modern robotics: mechanics, planning, and control. Cambridge University Press, Cambridge, UK (2017)
25. Kiong, T.K., Qing-Guo, W., Chieh, H.C., Hägglund, T.J.: Advances in PID control. Springer London, London (1999)
26. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning (pp. 41-48) (2009).