Wang *et al.*
Chinese Journal of Mechanical Engineering (2023) 36:80
https://doi.org/10.1186/s10033-023-00904-7

Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE

Open Access

A Combined Reinforcement Learning and Model Predictive Control for Car-Following Maneuver of Autonomous Vehicles

Liwen Wang[1], Shuo Yang[1,2], Kang Yuan[2,3], Yanjun Huang[1*] and Hong Chen[3]

Abstract

Model predictive control is widely used in the design of autonomous driving algorithms. However, its parameters are sensitive to dynamically varying driving conditions, making it difficult to be implemented into practice. As a result, this study presents a self-learning algorithm based on reinforcement learning to tune a model predictive controller. Specifically, the proposed algorithm is used to extract features of dynamic traffic scenes and adjust the weight coefficients of the model predictive controller. In this method, a risk threshold model is proposed to classify the risk level of the scenes based on the scene features, and aid in the design of the reinforcement learning reward function and ultimately improve the adaptability of the model predictive controller to real-world scenarios. The proposed algorithm is compared to a pure model predictive controller in car-following case. According to the results, the proposed method enables autonomous vehicles to adjust the priority of performance indices reasonably in different scenarios according to risk variations, showing a good scenario adaptability with safety guaranteed.

Keywords Model predictive control, Reinforcement learning, Autonomous vehicles

1 Introduction

As one of the development directions, autonomous driving is drawing more and more attention worldwide. Model predictive control (MPC) is able to deal with optimization problems with multiple objectives [1]. It iteratively solves an optimization problem over a finite horizon, to provide online optimal solutions subjected to constraints [2]. By constructing an optimization problem, the MPC algorithm ensures the car avoids collision and improves its performance in the dynamic environment [3]. According to Google Scholar, more than 7000 papers based on MPC have been published every year in the area of autonomous vehicles in the past three years.

MPC-based studies on autonomous vehicles mainly focus on planning or tracking the planned trajectory. A reasonable design of performance indices enables autonomous vehicles maintain an appropriate relative distance and velocity from surrounding vehicles, effectively alleviate traffic congestion, and reduce traffic accidents [4]. Jeong et al. [5] designed a MPC with fixed weights to improve trajectory and speed tracking performance by distributing control forces to multi-actuators. Ammour et al. [6] studied the trajectory planning of autonomous vehicles on the expressway based on a weight-fixed MPC such that a vehicle can improve safety by overtaking and lane changing. Wu et al. [7] designed a non-local controller based on MPC, which can attenuate the oscillation obviously and present a good riding comfort. Zhou et al. [8] developed an MPC based on the car following (CF) model with fixed weights. The objective function is designed based on the historical state data of the front

*Correspondence:
Yanjun Huang
yanjun_huang@tongji.edu.cn
[1] School of Automotive Studies, Tongji University, Shanghai 201804, China
[2] Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 201804, China
[3] College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China

vehicle, realizing a smooth response to the merged traffic. Sun et al. [9] proposed a hybrid MPC method for autonomous speed regulation, and it improved ride comfort by approximating the vehicle longitudinal dynamics as a two-mode discrete-time hybrid logic dynamics system. As one of the critical evaluation indices of vehicle performance, the energy-optimal speed trajectories are widely studied. Dollar et al. [10] applied a weight-fixed MPC to the longitudinal motion regulation in mixed traffic scenarios to realize the improvement of fuel economy, where objective function is designed using the instantaneous motion information of multiple preceding vehicles. Sun et al. [11] adopted a target switching MPC for global velocity tracking and adaptation, which greatly reduced the computation time for speed planning and saved 22.0% energy compared to human driving. As can be seen in existing studies, MPC controllers mostly adopt fixed weights, which can achieve a better performance in certain scenarios. However, driving scenarios in real world are uncertain, dynamical, and time-varying such that they cannot perform well in the real-world driving [12].

An excellent driving algorithm should be able to dynamically reach a balance among multiple performance indicators and adapt to different scenarios [13]. At present, there has been some researches on this problem [14–16], such as weight-adjustment or adaptive methods based on fuzzy control or personal driving data. Chang et al. [17] performed a real-time optimization of a weight matrix in MPC via fuzzy control, which improved the accuracy of tracking, ride comfort, and stability. Pang et al. [18] proposed an MPC that adaptively adjusts the weights of the cost function based on a fuzzy inference system, which significantly improved the MPC performance and control accuracy. Shivram et al. [19] used MPC based on fuzzy logic to improve the CF accuracy and ride comfort of independent vehicles. Tian et al. [20] proposed a coordinated tracking control strategy, they based on a fuzzy rule to adjust the weights of MPC, which finally improved the path tracking accuracy and stability of the vehicle at high speed with large curvature. Liu et al. [21] presented a shared steering mechanism based on MPC, so that the weight can be changed adaptively based on the consequence of risk assessment and predefined strategy to ensure driving safety. Liang et al. [22] proposed an adaptive multi-MPC scheme, which introduced a weighted adaptive mechanism based on rules to handle various driving conditions, especially some extreme cases. Rokonuzzaman et al. [23] proposed a longitudinal MPC controller for autonomous driving, which adaptively adjusts its weights based on a data-driven approach, so that the vehicle dynamics aspects such as speed, acceleration and jerk can be balanced. The existing weight adjustment methods with fault tolerance, reliability, and traceability advantages, can improve vehicle performance in several specific scenarios, but they generally are not able to cope well with dynamic scenarios.

Reinforcement learning (RL) as an auto-learning method, is able to encourage vehicles to explore under different scenarios with a reasonable reward function, through trial-and-error approach to accumulate experience and improve performance [24]. In fact, existing studies have extensively discussed the reward function design and vehicle performance optimization for diverse application scenarios [25–27]. It is suitable for dealing with the problem of scene changes and performance optimization [28]. However, RL has some shortcomings, for example, it is difficult to converge, and even if convergence is achieved, the effect of training is not always satisfactory [29]. In addition, as a safety-critical system, the vehicle cannot be tried randomly, but should always explore based on safety [30]. So just using RL methods is not enough to meet the requirements.

In this paper, we considered combining MPC with RL to automatically reach a balance among different performance index in dynamical scenarios. In this way, we can not only use MPC algorithm with hard constraints to ensure safety, but also extract complex features of dynamic scenarios as the basis of adaptive correction through RL. Therefore, the proposed combined strategy possesses the features to adaptively adjust controller parameters under different scenarios. Following is a summary of the main contributions in this paper.

1. This study proposes a weight-adjustment strategy for MPC based on RL, according to the state of surrounding environment to achieve a trade-off among safety, comfort, and energy saving of autonomous vehicles.
2. It summarizes the correlation between performance indices of speed control and the adaptive rules for MPC parameters to comprehensively improve the overall performance of autonomous vehicles.

This paper follows the following structure: Section 2 introduces the proposed combined RL and MPC strategy for autonomous vehicles. Section 3 defines the scenario risk assessment and vehicle performance analysis. Section 4 introduces MPC and RL algorithm. Finally, evaluation results of the proposed strategy for autonomous speed control are presented in Section 5.

2 Structure of the Combined Strategy

This section introduces the architecture of the combined RL and MPC strategy for autonomous vehicles and define four different scenarios depending on the risk level of the environment. As shown in Figure 1, the RL algorithm

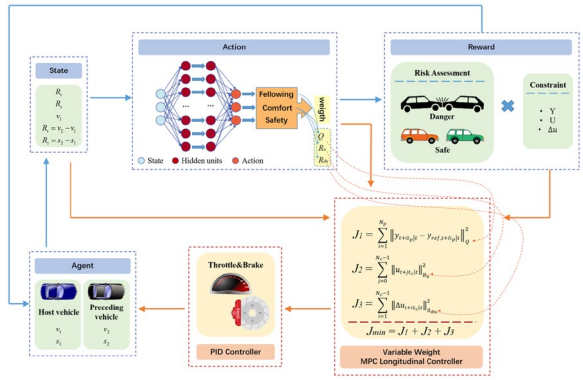

Figure 1 Combined RL and MPC strategy

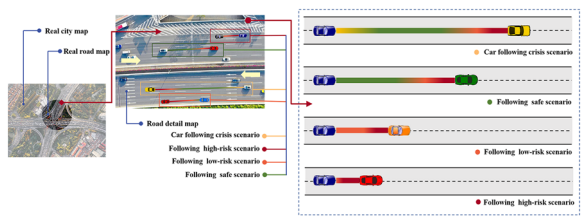

Figure 2 Dynamic scene and classification

takes the environment state in the current scene as input and outputs MPC variable weights. MPC calculates the optimal acceleration for lower layer controller. Furthermore, a risk threshold model is proposed for scenario risk assessment, so as to guide the reward function of RL, and MPC constraints are considered as well.

As shown in Figure 2, the scenarios of autonomous vehicles are complex and diverse. Different driving scenes have different characteristics, and the road conditions change dynamically. Even different vehicles on the same section of the road also drive differently. Four scenarios are shown on the right side in Figure 2, and the gradation of color is used to represent the transformation of scene risk degree caused by the change of the relative distance. Yellow, green, orange, and red represent the crisis scenario, the safety scenario, the low-risk scenario, and the high-risk scenario respectively. Under the crisis scenario, the vehicle should reduce ego-predecessor distance to ensure the CF task; under the safety scenario, the vehicle should keep the speed constant or accelerate slowly to reduce unnecessary jitters and pursue a better energy efficiency and ride comfort; under the low-risk scenario, due to the risk of collisions, much more emphasis should be placed on safety; in the high-risk scenario, the vehicle must slow down immediately to increase the distance between two vehicles and ensure that a collision does not occur. Scenario adaptability mentioned in this paper means that vehicles can autonomously adjust their behaviors based on different scenario risk levels to maximize vehicle performance as mentioned above.

3 Scenario Risk Assessment and Performance Analysis

It is necessary to analyze the risk degree and the priority requirements of vehicles in different scenarios to improve the adaptive adjustment ability of MPC with multiple weights. This paper proposes the risk threshold model (RTM), which can be used to evaluate the scene risk by analyzing the characteristics of the environment, classifying the scenes into four risk levels: the crisis scenario, the high-risk scenario, the low-risk scenario, and the safety scenario.

According to the RTM, the inputs are the relative distance and velocity of two vehicles in the CF maneuver, and the risk level of the scene is the output, as shown in Figure 3.

Based on the adjustment mechanism, some reference values are shown in Table 1.

When the relative distance exceeds the maximum following distance defined in the CF scenario, the vehicle does not meet the prerequisites of CF task such that the vehicle will be in the crisis scenario case. When the relative distance is between the safe driving distance and the maximum following distance, the autonomous vehicle can safely stop and the vehicle is in the safety scenario. When the relative distance is between the dangerous stopping distance and the safe stopping distance, where, when the relative velocity is less than the dangerous relative one, the stopping distance can be divided into the low-risk scenario and the safety scenario, and if the relative velocity is greater than the safe relative velocity, the vehicle will be able to stop instantly and it will be in the safety scenario. In cases that the relative distance is

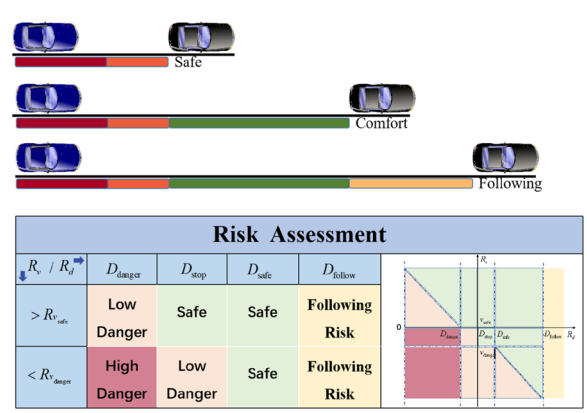

Figure 3 Risk assessment mechanism of the RTM

Table 1 Parameters and symbols of RTM

Parameters	Symbols	Value
D_{stop}	The average level of safe stopping (m)	20
D_{danger}	The lower limit of D_{stop} (m)	18
D_{safe}	The upper limit of D_{stop} (m)	22
D_{fellow}	The maximum following distance (m)	60
Rv_{safe}	The relative velocity that driver feels most comfortable (m/s)	0
Rv_{danger}	The relative velocity where there is a risk of collision (m/s)	−1

smaller than the dangerous stopping distance, and the relative velocity is less than the safety one, collision accidents will happen such that the vehicle is in the high-risk scenario, when the relative velocity is greater than the safe one, the preceding vehicle will always maintain the leading position, which is generally in the safety scenario. However, the state of the preceding vehicle is uncertain, the stopping distance of the ego is related to the relative velocity, which can be subdivided into the low-risk and the safety scenario.

During driving, vehicles are seeking for a balance among different performance, such as CF performance, safety, fuel economy, ride comfort under dynamic scenarios. However, there are conflict constraints for different performance. Exploring the correlation between performance and constraints is not only helpful to optimize vehicle performance and improve vehicle scene adaptability, but also lay a foundation for subsequent research on performance improvement and scene extension.

4 Combined RL and MPC Strategy

This section introduces the objective function and constraints of MPC controller and the RL algorithm for the adaptive adjustment of the weights in the objective function.

4.1 Model Predictive Controller

As the main algorithm of longitudinal following control, MPC is constructed based on the velocity and position information of the ego and the front vehicle, and outputs the acceleration that meets the constraints [31]. The longitudinal motion planning problem in this study is developed based on a longitudinal kinematic model as follows:

$$\dot{X} = v_X, \dot{V} = a_X, \tag{1}$$

where X, v_X and a_X represent the longitudinal displacement, velocity, and acceleration, respectively; the longitudinal displacement and velocity are denoted by the state variable and the output variable y, and the longitudinal acceleration is the control variable u:

$$x = \begin{bmatrix} X \\ v_X \end{bmatrix}, \quad y = \begin{bmatrix} X \\ v_X \end{bmatrix}, \quad u = a_X. \tag{2}$$

Then, the MPC longitudinal motion planning problem can be described as follows:

$$\begin{aligned} J_t(x(0), u_{t-1}, \Delta u, \varepsilon) &= \sum_{i=1}^{N_p} \left\| y_{t+it_p|t} - y_{\text{ref},t+it_p|t} \right\|_Q^2 \\ &+ \sum_{j=0}^{N_c-1} \left\| u_{t+jt_c|t} \right\|_{R_u}^2 + \sum_{i=1}^{N_c-1} \left\| \Delta u_{t+it_c|t} \right\|_{R_{du}}^2 + \rho\varepsilon^2, \\ &\min_{\Delta u, \varepsilon} J_t(x(0), u_{t-1}, \Delta u, \varepsilon), \end{aligned} \tag{3}$$

$$\begin{gathered} \text{s.t.}\, u_{\min} \le u(k) \le u_{\max}, k = 0, 1, \cdots, N_c - 1, \\ \Delta u_{\min} \le \Delta u(k) \le \Delta u_{\max}, k = 0, 1, \cdots, N_c - 1, \\ x_{\min} - \varepsilon 1_{n\times 1} \le x(k) \le x_{\max} - \varepsilon 1_{n\times 1}, k = 0, 1, \cdots, N_p, \\ y_{\min} - \varepsilon 1_{p\times 1} \le x(k) \le y_{\max} - \varepsilon 1_{p\times 1}, k = 0, 1, \cdots, N_p, \\ 0 \le \varepsilon(k) \le \varepsilon_{\max}, \end{gathered}$$

where,

$$\begin{gathered} \text{s.t.}\, u_{\min} \le u(k) \le u_{\max}, k = 0, 1, \cdots, N_c - 1, \\ \Delta u_{\min} \le \Delta u(k) \le \Delta u_{\max}, k = 0, 1, \cdots, N_c - 1, \\ x_{\min} - \varepsilon 1_{n\times 1} \le x(k) \le x_{\max} - \varepsilon 1_{n\times 1}, k = 0, 1, \cdots, N_p, \\ y_{\min} - \varepsilon 1_{p\times 1} \le x(k) \le y_{\max} - \varepsilon 1_{p\times 1}, k = 0, 1, \cdots, N_p, \\ 0 \le \varepsilon(k) \le \varepsilon_{\max}. \end{gathered}$$

The symbols of the relevant parameters are shown in Table 2.

The desired value of longitudinal position $y_{\text{ref},t+it_p|t}$ is jointly decided on the position of the preceding vehicle $X_{\text{f},t+it_p|t}$ and the safe stopping distance $D_{\text{safe},t}$, and is affected by the velocity $V_{\text{f},t+it_n|t}$ and acceleration $a_{\text{f},t+it_p|t}$ of the preceding vehicle.

In the objective function, the first item reflects the longitudinal CF safety requirements and the tracking ability to the desired values. The second reflects the fuel economy requirements, that is, the ability to suppress

Table 2 Parameters and symbols of MPC problems

Parameters	Symbols
t	Current moment
N_p	Prediction horizon
N_c	Control horizon
J_t	The objective function of the longitudinal motion planning
$x(0)$	The state vector at the current moment
u_{t-1}	The control vector at the previous moment
$y_{t+it_p\|t}$	The predictive output of longitudinal position and velocity corresponding to each predictive step in the prediction horizon at the current step
$y_{\mathrm{ref},t+it_p\|t}$	The desired output of longitudinal position and velocity of each predictive step in the prediction horizon
$u_{t+jt_c\|t}$	The output of acceleration corresponding to each predictive step in the control horizon at the current step
$\Delta u_{t+it_c\|t}$	The output of jerk corresponding to each predictive step in the control horizon at the current step
ε	Relaxation factor
Q, R_u, R_{du}, ρ	Weights of each optimization objectives
$1_{p\times 1}, 1_{n\times 1}$	A unit column vector of dimension n and p

excessive values of longitudinal acceleration. The third shows the ride comfort requirements, the ability to limit the excessive values of longitudinal jerk. The fourth item is used to prevent that the optimization problem will have no solution due to the error of the prediction model.

In this paper, we trained the weight of the first term in the MPC objective function by reinforcement learning, so as to dynamically adjust the expected longitudinal CF distance and improve the scene adaptability of vehicles.

4.2 Reinforcement Learning Algorithm

4.2.1 Action Space

The action $a_t \in \mathrm{A}$ is constituted by the CF weight coefficient Q, i.e., $a_t = [Q], Q \in [Q_{\min}, Q_{\max}]$, where $Q_{\min}, Q_{\max}$ correspond to the maximum and minimum values of the CF weight respectively.

4.2.2 State Space

The state $s_t \in S$ is composed of the relative distance, the relative velocity and the longitudinal velocity of the ego vehicle and the preceding vehicle, $s_t = \left[R_v, R_d, v_{\mathrm{ego}}\right]^{\mathrm{T}}$.

4.2.3 Reward Space

Four parts are used to define the reward as follows:

$$r(s,a) = r_{\mathrm{collision}} + r_U + r_{\Delta U} + r_D. \tag{4}$$

Collision reward $r_{\mathrm{collision}}$: Once a collision happens, the car will get a negative reward r_c, and r_c is a constant set to −10.

$$r_{\mathrm{collision}} = \begin{cases} 0, & \text{not collision}, \\ r_c, & \text{collsion}. \end{cases} \tag{5}$$

Acceleration reward r_U: If the vehicle satisfies the acceleration constraint of the MPC problem, it will get a reward r_U, r_U consists of four cases.

$$r_U = \begin{cases} r_{u1} + k_1 \times |u|, & \text{safe}, \\ r_{u2} + k_2 \times u, & \text{high danger}, \quad u \in [u_{\min}, u_{\max}], \\ r_{u3}, & \text{low danger}, \\ r_{u4}, & u \notin [u_{\min}, u_{\max}]. \end{cases} \tag{6}$$

1. When in the safety scenario, the reward is defined to travel at a lower acceleration, where r_{u1} and k_1 are constants set to 30 and −10;
2. When in the high-risk scenario, the reward encourages the vehicle to travel at a higher deceleration to avoid a dangerous situation. In addition, it penalizes the acceleration behavior to prevent the danger, where r_{u2} and k_2 are constants set to −10 and 1;
3. When in the low-risk scenario, it gets a reward r_{u3}, where r_{u3} is a constant set to 0;
4. Once the vehicle does not meet the acceleration constraint of the MPC problem, it will get a negative reward r_{u4} to punish the over-constrained behavior, where r_{u4} is a constant set to −100.

Jerk reward $r_{\Delta U}$: If the vehicle does not meet the jerk constraint of the MPC problem, it will get a negative reward $r_{\Delta u}$, where $r_{\Delta u}$ is a constant set to −30.

$$r_{\Delta U} = \begin{cases} 0, & \Delta u \in [\Delta u_{\min}, \Delta u_{\max}], \\ r_{\Delta u}, & \Delta u \notin [\Delta u_{\min}, \Delta u_{\max}]. \end{cases} \tag{7}$$

CF reward r_D: When the vehicle is in the crisis scenario, it will get a negative reward r_d. Its purpose is to

enable the autonomous vehicle to guarantee the basic CF task, where r_d is a constant set to -41.

$$r_D = \begin{cases} 0, & \text{not fellow risk,} \\ r_d, & \text{fellow risk.} \end{cases} \tag{8}$$

4.2.4 State Transition Probability

The vehicle in state s_t takes action a_t, and the state s_t transfers to state s_{t+1}. The probabilistic state transition mode is denoted as follows:

$$P_s^a = P[s = s_{t+1} | s = s_t, a = a_t]. \tag{9}$$

4.2.5 Value Function

In this paper, we choose a soft actor-critic (SAC) algorithm that optimizes random strategies in a non-strategic way to make the value function optimal [32].

First, introduce the definition of entropy, define x to be a stochastic variable with probability mass or density function. The entropy H of x is calculated from its distribution P according to:

$$H(P) = \underset{x \sim P}{E}[-\log P(x)]. \tag{10}$$

In entropy-regularized RL, the agent gets an additional bonus at each time step relative to the entropy of the policy at the same timestep. This changes the RL problem to:

$$\pi^* = \arg\max_{\pi} \underset{\tau \sim \pi}{E}\left[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot|s_t)))\right]. \tag{11}$$

Where, α is the entropy regularization coefficient, which particularly controls the explore-exploit tradeoff, with positive and negative correlation with exploration. γ is the discount factor, $\gamma \in [0, 1]$.

The value function is as follows:

$$V^{\pi}(s) = \underset{\tau \sim \pi}{E}\left[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot|s_t)))|s_0 = s\right]. \tag{12}$$

The Q function corresponds to:

$$Q^{\pi}(s, a) = \underset{\tau \sim \pi}{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) + \alpha \sum_{t=1}^{\infty} \gamma^t H(\pi(\cdot|s_t))|s_0 = s, a_0 = a\right]. \tag{13}$$

Combining the above definitions, we get:

$$V^{\pi}(s) = \underset{a \sim \pi}{E}[Q^{\pi}(s, a)] + \alpha H(\pi(\cdot|s)), \tag{14}$$

in addition, the bellman equation of $Q^{\pi}(s, a)$ is:

$$\begin{aligned} Q^{\pi}(s, a) &= \underset{\substack{s' \sim P \\ a' \sim \pi}}{E}\left[R(s, a, s') + \gamma\left(Q^{\pi}(s', a') + \alpha H(\pi(\cdot s'))\right)\right] \\ &= \underset{s' \sim P}{E}\left[R(s, a, s') + \gamma V^{\pi}(s')\right], \end{aligned} \tag{15}$$

rewrite it with the definition of entropy:

$$\begin{aligned} Q^{\pi}(s, a) &= \underset{\substack{s' \sim P \\ a' \sim \pi}}{E}\left[R(s, a, s') + \gamma\left(Q^{\pi}(s', a') + \alpha H(\pi(\cdot|s'))\right)\right] \\ &= \underset{\substack{s' \sim P \\ a' \sim \pi}}{E}\left[R(s, a, s') + \gamma\left(Q^{\pi}(s', a') - \alpha \log \pi(a'|s')\right)\right]. \end{aligned} \tag{16}$$

The right hand side is the expected value of the next action from the current policy as well as the next state from the replay buffer. Since it is an expectation, we can approximate it with samples:

$$\begin{aligned} Q^{\pi}(s, a) \approx r + \gamma\left(Q^{\pi}\left(s', \tilde{a}'\right) - \alpha \log \pi\left(\tilde{a}'|s'\right)\right), \\ \tilde{a}' \sim \pi(\cdot|s'). \end{aligned} \tag{17}$$

The loss function of Q-network in SAC is:

$$L(\phi_i, \mathcal{D}) = \underset{(s,a,r,s',d) \sim \mathcal{D}}{E}\left[\left(Q_{\phi_i}(s, a) - y(r, s', d)\right)^2\right], \tag{18}$$

where the target y is given by:

$$\begin{aligned} y(r, s', d) = r + \gamma(1 - d)\left(\min_{j=1,2} Q_{\phi_{t\arg,j}}(s', \tilde{a}') - \alpha \log \pi_{\theta}(\tilde{a}'|s')\right), \\ \tilde{a}' \sim \pi_{\theta}(\cdot|s'). \end{aligned} \tag{19}$$

The strategy should act to maximize the sum of expected future benefits and entropy in each state. That is, it should maximize V^{π}, which this paper expands out into:

$$\begin{aligned} V^{\pi}(s) &= \underset{a \sim \pi}{E}[Q^{\pi}(s, a)] + \alpha H(\pi(\cdot|s)) \\ &= \underset{a \sim \pi}{E}[Q^{\pi}(s, a) - \alpha \log \pi(a|s)]. \end{aligned} \tag{20}$$

The way we optimize the policy makes use of the reparameterization trick, in which a sample from $\pi_{\theta}(\cdot|s)$ is derived by calculating a deterministic function of policy parameters, state, and independent noise. To illustrate: We use a squashed Gaussian policy, which means that samples are obtained, according to

$$\tilde{a}_{\theta}(s, \xi) = \tanh(\mu_{\theta}(s) + \sigma_{\theta}(s) \odot \xi), \quad \xi \sim \mathcal{N}(0, I). \tag{21}$$

The re-parameterization technique allows us to rewrite the operational expectation into the noise expectation:

$$\underset{a\sim\pi_\theta}{E}\left[Q^{\pi_\theta}(s,a)-\alpha\log\pi_\theta(a|s)\right]$$
$$=\underset{\xi\sim\mathcal{N}}{E}\left[Q^{\pi_\theta}\left(s,\tilde{a}_\theta(s,\xi)\right)-\alpha\log\pi_\theta\left(\tilde{a}_\theta(s,\xi)|s\right)\right]. \tag{22}$$

To get the policy loss, the last step is that we need to substitute Q^{π_θ} with one of our function approximators, the policy is thus optimized according to Eq. (23):

$$\max_\theta \underset{\substack{s\sim\mathcal{N}\\ \xi\sim\mathcal{N}}}{E}\left[\min_{j=1,2}Q_{\phi_j}\left(s,\tilde{a}_\theta(s,\xi)\right)-\alpha\log\pi_\theta\left(\tilde{a}_\theta(s,\xi)|s\right)\right]. \tag{23}$$

5 Simulation Results

This section briefly introduces the configuration of the simulation environment and the design of training process, followed by the result demonstration and analysis. In this paper, a RL environment is built in Carla simulator (an open source software) to implement and test the proposed strategy. In order to verify the performance of the proposed combined controller with a dynamic scene adaptability, this paper compares the variable-weight MPC controller with the fixed-weight MPC controller under three different operating conditions. The scenarios involved in the training process can be classified into the four risk levels mentioned above. The relevant initial conditions and parameters used in the simulation are shown in Table 3.

The RL algorithm is trained with random seeds for 120000 iterations per evaluation, and Figure 4 shows the cumulative rewards for the average reward $\bar{r}$ at each step during the evaluation period. The maximum possible reward for each step is 30, and the reward decreases when the agent deviates from the expected acceleration and jerk. The results show that after 50000 training steps, the vehicle is able to learn how to perform better. As the training moves on, the average reward $\bar{r}$ continues to increase until around 100000 steps. During training, the set ends when (1) a collision occurs, (2) the relative distance is greater than the maximum CF distance, (3) the CF task ends or (4) the maximum number of iterative steps per set of 750 timesteps is reached.

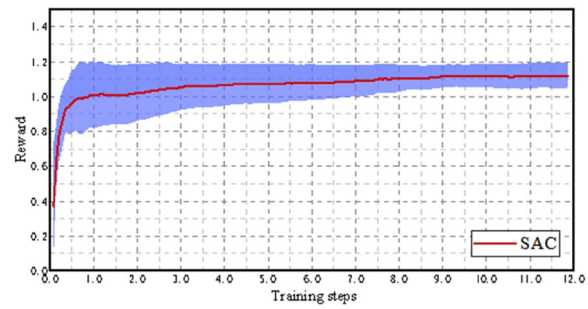

Figure 4 Learning curve

Then, the test conditions are defined. The initial speed of the ego is set to 0, and the initial position is fixed. The initial speed of the preceding vehicle is randomly set, and it successively experiences three driving conditions of acceleration, steady state and deceleration, the specific process is shown in Table 4. The D_{fellow} in this paper is 60 m, to extensively test the proposed CF algorithm performance, this paper set three working conditions with initial relative distances of 50 m, 60 m, and 70 m, respectively. The training results are shown in Figures 5, 6, 7, 8, 9, 10, 11, where the black curves indicate the following results of the fixed weight coefficient MPC controller and the red shows the results of the variable weight MPC controller with dynamic scene adaptability.

It is verified that this strategy can make the driving process basically within the following safety scenario through the adaptive adjustment of the vehicle, regardless of whether the initial relative distance is greater than D_{fellow}. Even in the emergency deceleration stage of the front vehicle, the ego vehicle is only unavoidably caught in the following danger scenario for a short time, and then immediately gets rid of the dilemma, as shown in Figure 5. Among them, the adjustment curve of weight

Table 3 Main simulation parameters

Parameters	Symbols	Value
System initial conditions	MPC_ρ	0.01
	Control horizon N_c	10
	Predicted horizon N_p	30
	Episode number N	120000
	Sampling time T (s)	0.1
	Track length (timestep)	750
System constrains	Input variable (m/s^2)	$-4\le u(k)\le 3$
	Velocity v (m/s^2)	$0\le v(k)\le 20$
	Acceleration a (m/s^2)	$-4\le a(k)\le 3$
	Vehicle jerk j (m/s^2)	$-3\le j(k)\le 3$
	Relaxation factor ε	$0\le\varepsilon\le 0.02$

Table 4 Driving process of the preceding vehicle

Time (s)	State	Acceleration (m/s^2)
0–75	Total time of an episode	–
0–11	Acceleration	1
31–34	Acceleration	2
59–63	Deceleration	−1.5
Others	At a constant speed	0

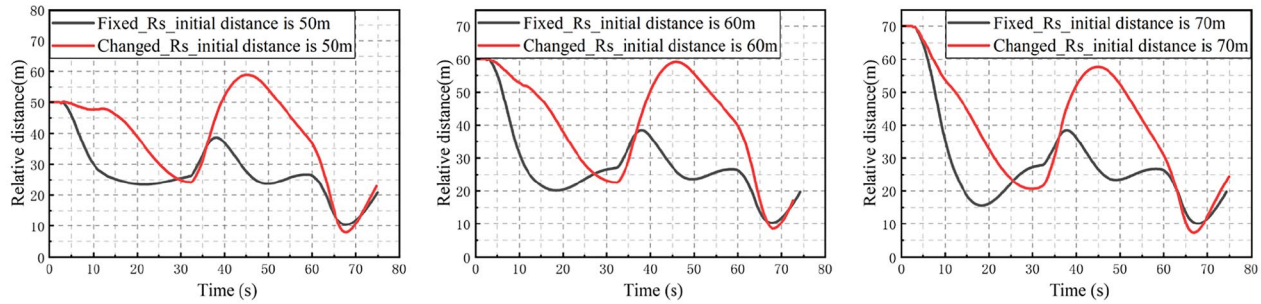

Figure 5 Relative distance

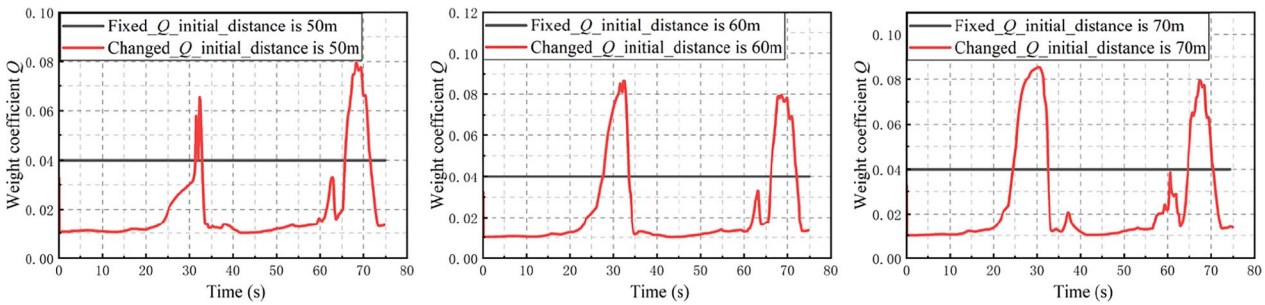

Figure 6 Weight coefficient *Q*

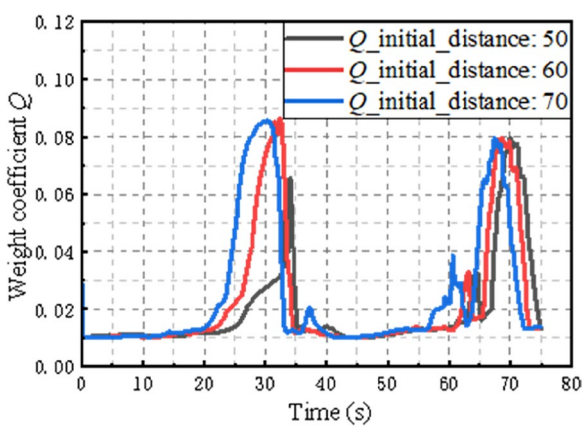

Figure 7 Variation of the weight coefficient *Q* under different initial distances

Q is shown in Figure 6, which can adjust its trend of increase or decrease by judging the state of environment. If the vehicle is maintaining the safe following distance, weight *Q* will reduce to improve the comfort and energy saving. As the possibility of the ego car getting into a following crisis situation increases when the front car accelerates, the *Q* value is gradually increased to ensure that the following distance is within the desired range. And the larger the initial relative distance is, the more likely it is to lead to CF task failure, so the growth trend of *Q* is positively correlated with the initial relative distance, as shown in Figure 7. In the low and high-risk scenario, the vehicle with this strategy rapidly increases the distance to the front vehicle through adjusting the weight *Q*, effectively avoiding the collision and reflecting the good safety.

Compared to the vehicle with conventional MPC controller, the vehicle under this strategy maintains a smaller relative velocity to the preceding vehicle for most of the time. As one of the causes of driver sight distance jitter, the reduction of relative velocity is necessary. In addition, the relative velocity reduction has little effect on the vehicle velocity and always satisfies the controller constraints. That is, the strategy can improve the comfort while ensuring the vehicle traffic efficiency, as shown in Figures 8 and 9.

In addition, as shown in Figures 10 and 11, the acceleration curves and jerk curves show that the adaptive adjustment strategy tends to follow the preceding vehicle with less acceleration in the safety scenario. It allows the vehicle to have the opportunity to pursue higher ride comfort and fuel economy while maintaining safety. Besides, this is in line with human habits and expectations, demonstrating its intelligence. When the vehicle is caught in a high-risk scenario due to sudden changes in environmental conditions, the strategy has better performance in terms of the state prediction and response speed. The

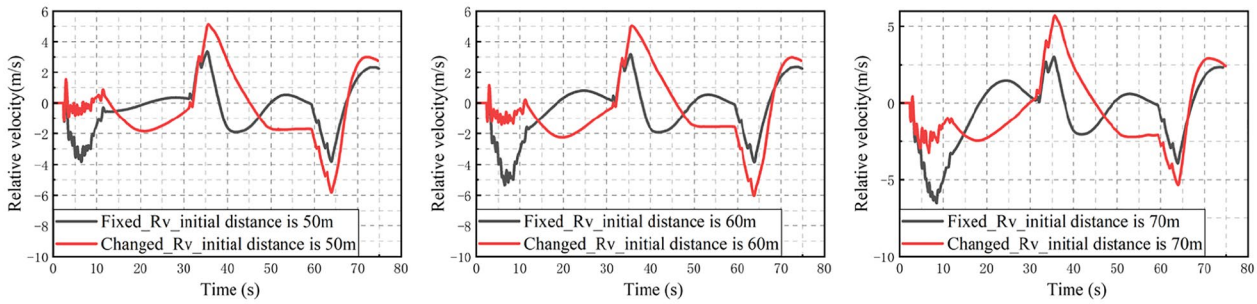

Figure 8 Relative velocity

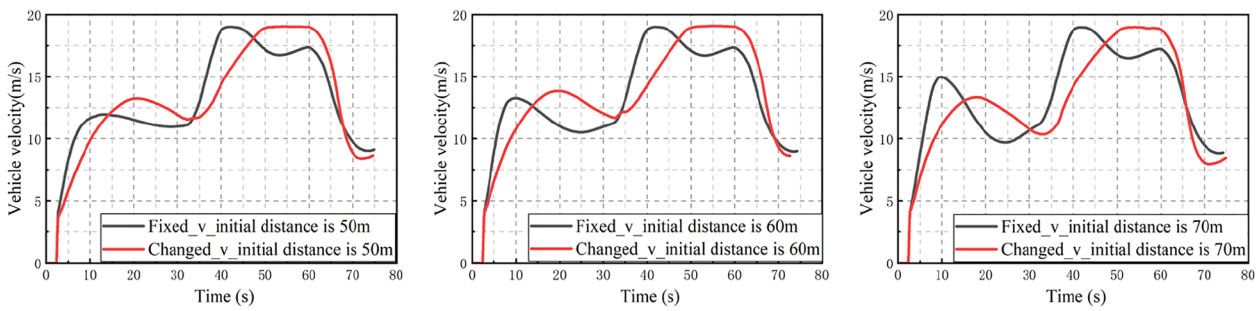

Figure 9 Vehicle velocity

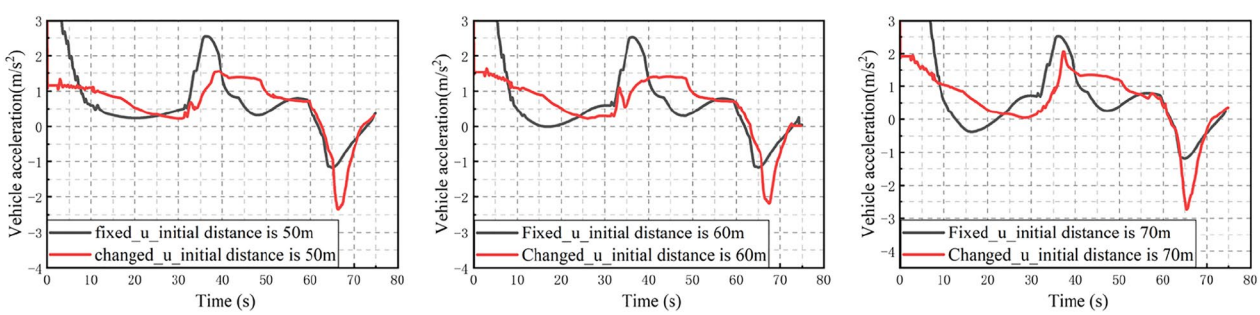

Figure 10 Vehicle acceleration

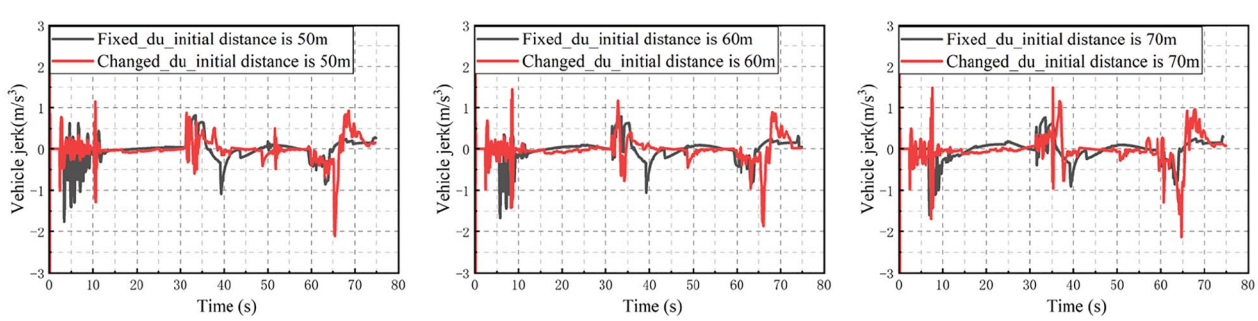

Figure 11 Vehicle jerk

vehicle with this strategy performed faster and decelerate more, and it is satisfying the controller constraint, that is the vehicle has good emergency handling capability to avoid the collision in a timely and effective manner. It indicates that the strategy has good adaptability in dynamic scenarios, which helps to improve the safety of the vehicle.

The training results show that the variable weight controller guided by the proposed combined strategy is able to adjust the vehicle performance priority in dynamic scenarios. Compared to the traditional controller with fixed weights, the proposed controller can adjust adaptively and has better performance, such as when in the crisis scenario, it can respond quickly and accelerate to keep up with the preceding vehicle. The probability of following the vehicle under the safety scenario is higher, and the riding comfort and fuel economy of the vehicle are effectively improved through weight adjustment. In dangerous scenarios such as the low-risk and the high-risk scenarios, the vehicle's emergency response ability is better, and collisions can be avoided by decelerating in time.

6 Conclusions

(1) The results of this paper have verified the effectiveness of using RL in combination with traditional MPC controller to construct an adaptive controller for autonomous vehicles driving in dynamic scene.
(2) The effectiveness of the proposed method was verified in the Carla environment with a high-fidelity vehicle model as the main control object. According to the results, when compared with the traditional controller, the autonomous vehicle based on this method could brake quickly in dangerous scenarios for the sake of safety, and it has a more stable acceleration performance and better ride comfort and fuel economy. Accordingly, ability of autonomous vehicles to trade-off safety, comfort and energy efficiency is significantly improved.
(3) A risk threshold model is developed to classify scenes based on feature information and guide the design of the RL reward function, which helps to accelerate the convergence process of RL and the probability of finding a more optimal solution.
(4) The study found that the adjustment of MPC weight coefficient has a direct impact on vehicle performance and the adjustment of tracking weight Q and its effect are described. To enhance the controller's scene adaptability, when designing the adjustment rules, it is necessary to fully consider the appropriate scene risk assessment method, reasonable safe following distance and safe stopping distance meeting the mechanical requirements.

There are still some problems to be further studied in the combined strategy of RL and MPC for autonomous. In the future, we plan to extend the adaptive adjustment method for vehicle lane change in complex environments.

Acknowledgements
The authors sincerely thank the anonymous reviewers for their valuable suggestions.

Authors' Contributions
LW conducted the research work, including the literature research, coding, and writing. SY and YK assisted with the coding and result analysis. YH and HC are supervisors who offered the original idea and coordinated the manuscript revision. All authors read and approved the final manuscript.

Authors' Information
Liwen Wang received the B.S. degree from *College of Automotive Engineering, Jilin University, Changchun, China,* in 2021. She is currently pursuing the M.S. degree at *School of Automotive Studies, Tongji University, Shanghai, China*. Her research interests include reinforcement learning, model predictive control and autonomous vehicle.
Shuo Yang received the B.S. and M.S. degrees from *College of Automotive Engineering, Jilin University, Changchun, China,* in 2017. He is currently pursuing the Ph.D. degree at *School of Automotive Studies, Tongji University, Shanghai, China*. His research interests include reinforcement learning, autonomous vehicle, intelligent transportation system and vehicle dynamics.
Kang Yuan received the B.S. and M.S. degrees in vehicle engineering from *Chongqing University, Chongqing, China,* in 2015 and 2019, respectively. He was also a visiting master at *Department of Mechanical and Mechatronics Engineering, University of Waterloo, Canada,* in 2018. He is currently pursuing the Ph.D. degree at *College of Electronic and Information Engineering, Tongji University, Shanghai, China*. His research interests include reinforcement learning, model predictive control and applications in decision-making and motion control for connected and autonomous vehicles.
Yanjun Huang is a professor at *School of Automotive Studies, Tongji University, Shanghai, China*. He received his PhD degree in 2016 from *Department of Mechanical and Mechatronics Engineering at University of Waterloo, Canada.* His research interest is mainly on autonomous driving and artificial intelligence in terms of decision-making and planning, motion control, human-machine cooperative driving.
Hong Chen received the B.S. and M.S. degrees in process control from *Zhejiang University, China,* in 1983 and 1986, respectively, and the Ph.D. degree in system dynamics and control engineering from *University of Stuttgart, Germany,* in 1997. In 1986, she joined *Jilin University of Technology, China*. From 1993 to 1997, she was a assistant research fellow at *Institute of System Dynamics and Control Engineering, University of Stuttgart, German.* Since 1999, she has been a professor at *Jilin University, China,* and hereafter a Tang Aoqing professor. Recently, she joined *Tongji University, China,* as a distinguished professor. Her current research interests include model predictive control, nonlinear control, artificial intelligence and applications in mechatronic systems e.g. automotive systems.

Funding
Supported by National Key R&D Program of China (Grant No. 2022YFB2502900), Fundamental Research Funds for the Central Universities of China, Science and Technology Commission of Shanghai Municipality of China (Grant No. 21ZR1465900), and Shanghai Gaofeng & Gaoyuan Project for University Academic Program Development of China.

Data Availability
Data will be made available on request.

Declarations

Competing Interests

The authors declare no competing financial interests.

Received: 20 December 2022 Revised: 31 May 2023 Accepted: 8 June 2023
Published online: 03 July 2023

References

[1] K Yang, X Tang, Y Qin, et al. Comparative study of trajectory tracking control for automated vehicles via model predictive control and robust H-infinity state feedback control. *Chinese Journal of Mechanical Engineering*, 2021, 34(1).
[2] J Zhou, H Zheng, J Wang, et al. Multiobjective optimization of lane-changing strategy for intelligent vehicles in complex driving environments. *IEEE Transactions on Vehicular Technology*, 2019, 69(2): 1291–1308.
[3] F Vitale, C Roncoli. An MPC-based task priority management approach for connected and automated vehicles reference tracking with obstacle avoidance. *2021 European Control Conference (ECC)*, Delft, Netherlands, 2021: 813–819.
[4] X Tang, K Yang, H Wang, et al. Driving environment uncertainty-aware motion planning for autonomous vehicles. *Chinese Journal of Mechanical Engineering*, 2022, 35(1).
[5] Y Jeong, S Yim. Model predictive control-based integrated path tracking and velocity control for autonomous vehicle with four-wheel independent steering and driving. *Electronics*, 2021, 10(22): 2812.
[6] M Ammour, R Orjuela, M Basset. An MPC combined decision making and trajectory planning for autonomous vehicle collision avoidance. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(12): 24805–24817.
[7] F Wu, A M Bayen. A hierarchical MPC approach to car-following via linearly constrained quadratic programming. *IEEE Control Systems Letters*, 2022, 7: 532–537.
[8] H Zhou, A Zhou, T Li, et al. Congestion-mitigating MPC design for adaptive cruise control based on Newell's car following model: History outperforms prediction. *Transportation Research Part C: Emerging Technologies*, 2022, 142: 103801.
[9] X Sun, Y Cai, S Wang, et al. Optimal control of intelligent vehicle longitudinal dynamics via hybrid model predictive control. *Robotics and Autonomous Systems*, 2019, 112: 190–200.
[10] R A Dollar, T G Molnár, A Vahidi, et al. MPC-based connected cruise control with multiple human predecessors. *2021 American Control Conference (ACC), IEEE*, 2021: 405–411.
[11] C Sun, J Leng, F Sun. A fastoptimal speed planning system in arterial roads for intelligent and connected vehicles. *IEEE Internet of Things Journal*, 2022, 9(20): 20295–20307.
[12] Q Sun, X Wang, G Yang, et al. Adaptive robust formation control of connected and autonomous vehicle swarm system based on constraint following. *IEEE Transactions on Cybernetics*, 2022.
[13] Y Zhang, M Xu, Y Qin, et al. MILE: Multi-objective integrated model predictive adaptive cruise control for intelligent vehicle. *IEEE Transactions on Industrial Informatics*, 2022.
[14] L Xiong, M Liu, X Yang, et al. Integrated path tracking for autonomous vehicle collision avoidance based on model predictive control with multi-constraints. *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), IEEE*, 2022: 554–561.
[15] J M Guirguis, S Hammad, S A Maged. Path tracking control based on an adaptive MPC to changing vehicle dynamics. *International Journal of Mechanical Engineering and Robotics Research*, 2022, 11(7).
[16] Y Li, J Fan, Y Liu, et al. Path planning and path tracking for autonomous vehicle based on MPC with adaptive dual-horizon-parameters. *International Journal of Automotive Technology*, 2022, 23(5): 1239–1253.
[17] G Chang, Q Suqin. An adaptive MPC trajectory tracking algorithm for autonomous vehicles. *2021 17th International Conference on Computational Intelligence and Security (CIS), IEEE*, 2021: 197–201.
[18] F Pang, M Luo, X Xu, et al. Path tracking control of an omni-directional service robot based on model predictive control of adaptive neural-fuzzy inference system. *Applied Sciences*, 2021, 11(2): 838.
[19] S Shivram, M Tajuddin, V Singhal, et al. Route tracking controller for self-directed vehicles based on an enhanced adaptive weight model predictive control. *2021 Innovations in Power and Advanced Computing Technologies (i-PACT), IEEE*, 2021: 1–7.
[20] Y Tian, Q Yao, P Hang, et al. Adaptive coordinated path tracking control strategy for autonomous vehicles with direct yaw moment control. *Chinese Journal of Mechanical Engineering*, 2022, 35(1).
[21] J Liu, H Guo, L Song, et al. Driver-automation shared steering control for highly automated vehicles. *Science China Information Sciences*, 2020, 63(9): 1–16.
[22] Y Liang, Y Li, A Khajepour, et al. Holistic adaptive multi-model predictive control for the path following of 4WID autonomous vehicles. *IEEE Transactions on Vehicular Technology*, 2020, 70(1): 69–81.
[23] M Rokonuzzaman, N Mohajer, S Mohamed, et al. A customisable longitudinal controller of autonomous vehicle using data-driven mpc. *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE*, 2021: 1367–1373.
[24] J Wu, Z Huang, C Lv. Uncertainty-aware model-based reinforcement learning: Methodology and application in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2022.
[25] N J Zakaria, M I Shapiai, N Wahid. A study of multiple reward function performances for vehicle collision avoidance systems applying the DQN algorithm in reinforcement learning. *In IOP Conference Series: Materials Science and Engineering*, 2021, 1176(1): 12–33.
[26] X L Tang, J X Chen, K Yang, et al. Visual detection and deep reinforcement learning-based car following and energy management for hybrid electric vehicles. *IEEE Transactions on Transportation Electrification*, 2022, 8(2): 2501–25153.
[27] Z Wang, H Huang, J Tang, et al. Velocity control in car-following behavior with autonomous vehicles using reinforcement learning. *Accident Analysis & Prevention*, 2022, 174: 106729.
[28] K Rezaee, P Yadmellat, S Chamorro. Motion planning for autonomous vehicles in the presence of uncertainty using reinforcement learning. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE*, 2021: 3506–3511.
[29] M J P Peixoto, A Azim. Improving environmental awareness for autonomous vehicles. *Applied Intelligence*, 2023, 53(2): 1842–1854.
[30] J Seo, J Lee, E Baek, et al. Safety-critical control with nonaffine control inputs via a relaxed control barrier function for an autonomous vehicle. *IEEE Robotics and Automation Letters*, 2022, 7(2): 1944–1951.
[31] K Yuan, H Shu, Y J Huang, et al. Mixed local motion planning and tracking control framework for autonomous vehicles based on model predictive control. *IET Intelligent Transport Systems*, 2019, 13(6): 950–959.
[32] Z Ahmed, N L Roux, M Norouzi, et al. Understanding the impact of entropy on policy optimization. *International Conference on Machine Learning. PMLR,* 2019: 151–160.