# Collaborative adaptive cruise control and energy management strategy for extended-range electric logistics van platoon

| | |
|---|---|
| Journal: | *Part D: Journal of Automobile Engineering* |
| Manuscript ID | JAUTO-22-0550.R1 |
| Manuscript Type: | Original Article |
| Date Submitted by the Author: | 15-Mar-2023 |
| Complete List of Authors: | Wang, Gang; Nanjing University of Science and Technology, School of Mechanical Engineering<br>Wang, HongLiang; Nanjing University of Science and Technology School of Mechanical Engineering,<br>Pi, Dawei; Nanjing University of Science and Technology School of Mechanical Engineering<br>Sun, Xiaowang; Nanjing University of Science and Technology School of Mechanical Engineering<br>Wang, Xianhui; Nanjing University of Science and Technology |
| Keywords: | Extended-range electric logistics van platoon, Cooperative adaptive cruise control, Distributed model predictive control, Energy management strategy, Multi-agent deep deterministic policy gradient |
| Abstract: | This paper improves the economy of the extended-range electric logistics van (ERELV) platoon from two aspects of cooperative adaptive cruise control (CACC) and energy management strategy (EMS). Based on the vehicle-to-everything (V2X) communication, to improve the economy of heterogeneous vehicle platoon CACC system, a distributed model predictive controller (DMPC) with stability, comfort, and the economy as optimization goals are designed. The sufficient conditions for the asymptotic stability of the vehicle platoon closed-loop system are obtained by Lyapunov stability analysis. The multi-agent deep reinforcement learning (MADRL) algorithm is used to solve the EMS of the ERELV platoon. Under the framework of centralized Training Distributed Execution (CTDE), the experience of all agents can be obtained during training, and the actions can be output only according to their local observation states during execution. The simulation results |

| | |
|---|---|
| | show that the designed ecological cooperative adaptive cruise control (Eco-CACC) effectively balances the stability and economy of a heterogeneous vehicle platoon. Taking dynamic programming (DP) as the benchmark, compared with the single-agent algorithm, EMS based on a multi-agent deep deterministic strategy gradient (MADDPG) algorithm can achieve a near-optimal solution while significantly improving the learning efficiency. |
| | |

# Collaborative adaptive cruise control and energy management strategy for extended-range electric logistics van platoon

Gang Wang[a], Hongliang Wang[a,*], Dawei Pi[a], Xiaowang Sun[a], Xianhui Wang[a]

[a]School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

Submitted to

**Journal of Automobile Engineering**

*Corresponding Author: School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China. Email address: whl343@163.com

# Collaborative adaptive cruise control and energy management strategy for extended-range electric logistics van platoon

Gang Wang, Hongliang Wang*, Dawei Pi, Xiaowang Sun, Xianhui Wang

**Abstract**: This paper improves the economy of the extended-range electric logistics van (ERELV) platoon from two aspects of cooperative adaptive cruise control (CACC) and energy management strategy (EMS). Based on the vehicle-to-everything (V2X) communication, to improve the economy of heterogeneous vehicle platoon CACC system, a distributed model predictive controller (DMPC) with stability, comfort, and the economy as optimization goals are designed. The sufficient conditions for the asymptotic stability of the vehicle platoon closed-loop system are obtained by Lyapunov stability analysis. The multi-agent deep reinforcement learning (MADRL) algorithm is used to solve the EMS of the ERELV platoon. Under the framework of centralized Training Distributed Execution (CTDE), the experience of all agents can be obtained during training, and the actions can be output only according to their local observation states during execution. The simulation results show that the designed ecological cooperative adaptive cruise control (Eco-CACC) effectively balances the stability and economy of a heterogeneous vehicle platoon. Taking dynamic programming (DP) as the benchmark, compared with the single-agent algorithm, EMS based on a multi-agent deep deterministic strategy gradient (MADDPG) algorithm can achieve a near-optimal solution while significantly improving the learning efficiency.

**Keywords:** Extended-range electric logistics van platoon; Cooperative adaptive cruise control; Distributed model predictive control; Energy management strategy; Multi-agent deep deterministic policy gradient

## 1 Introduction

The global cargo volume and turnover are increasing year by year, and the emission problem is also becoming more serious. With the continuous development of intelligent transportation systems (ITS) and intelligent connected vehicle (ICV) technology, platoon driving will be one of the main driving modes in the future. CACC combines V2X and adaptive cruise control systems to change the control object from a single vehicle to a platoon of vehicles.[1] CACC enables vehicle platoon to operate fully automatically in the ITS without driver intervention, improving platooning efficiency and safety.[2]

Compared with hybrid logistics vehicles and pure electric logistics vehicles, the structural characteristics of ERELV make it have better fuel economy and longer mileage, so it is more suitable for logistics vehicle transportation scenarios.[3] The engine of ERELV is used only to generate electricity, providing the power needed to drive the motor and battery. The engine can always operate in the high efficiency zone, thus reducing fuel consumption.[4] The EMS of ERELV is used to balance the power output of the engine and battery to meet the vehicle's driving demand and improve fuel economy at the same time.

### *1.1 Eco-Cooperative Adaptive Cruise Control*

ICV can obtain the status information of surrounding vehicles and road status information through vehicle-to-vehicle (V2V) communication and vehicle-to-infrastructure (V2I) communication.[5] Eco-CACC takes economy as the optimization goal and improves economy on the basis of achieving stability and safety of vehicle platoon. The research directions of Eco-CACC can be divided into two categories: V2I-based macro decision making and V2V-based micro control.[6]

Eco-CACC based on V2I takes the leading vehicle (LV) in the platoon as the research object, takes the road information as the constraint condition, and obtains the optimal speed by solving the optimal control problem. Qiu S. et al.[7] designed a model prediction controller (MPC) combined with signal phase and timing (SPAT) information to predict the optimal velocity profile over a finite time horizon for a vehicle platoon. Ma F. et al.[8] also used SPAT information to obtain the optimal speed for the vehicle platoon through an improved DP algorithm, making the vehicle platoon better suitable for urban roads.

Eco-CACC based on V2V takes the following vehicle (FV) in the platoon as the research object, takes the state information of other vehicles in the platoon as the constraint condition, and achieves a safer and more economical driving goal by solving the optimal control.[9] A distributed economic model predictive control algorithm is proposed in [10], achieving overall platoon fuel economy by solving two open-loop optimization problems. Pi D. et al.[11] used DMPC to achieve multi-objective optimal control of electric vehicle platoon, combined with regenerative braking to achieve economical driving of platoon.

### *1.2 EMS for connected hybrid electric vehicle*

Fig.1 shows the basic framework of EMS based on deep reinforcement learning (DRL). The environment model includes hybrid power system and driving environment, and the agent module contains learning algorithm. Through the interaction between agent and environment to optimize network parameters, agent will learn to output optimal actions to the environment to maximize the cumulative reward.

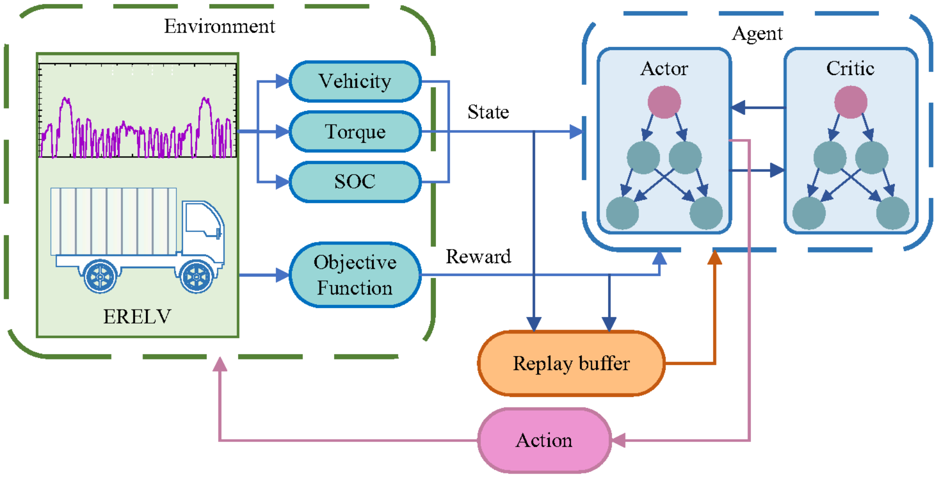


Fig.1. Structure of EMS based on DRL

Hybrid electric vehicle (HEV) can provide the required road information and speed information to EMS through V2X, which makes EMS more adaptive and robust. Referring to the classification scheme of intelligent power management systems for electric vehicles in [12], this paper divides the EMS of intelligent connected HEV into two cases: predictive-EMS (P-EMS) and cognitive-EMS (C-EMS).

The P-EMS method is based on the prediction algorithm to improve the real-time performance of EMS, using historical driving data or real-time road condition information to predict the driving state, usually using the optimal control algorithm as the EMS.[12] Since the running state of the vehicle has the characteristics of time series, Jamali H. et al.[13] used the long short-term memory algorithm to predict the engine state and vehicle speed information and used it to improve the hysteresis of rule-based EMS control. In [14], a neural network algorithm is used to provide the required vehicle speed information for the prediction time domain of MPC. In [15], the Gray model and Dijkstra algorithm are combined to plan the optimal path, and the Deep Q-Network (DQN) algorithm is adopted as the EMS of HEV. In [16], The real-time distance between the two vehicles is obtained by visual algorithm, under the premise of keeping a reasonable following distance, the multi-objective control of the hybrid power system is realized by using the EMS based on DQN.

C-EMS is mainly based on different reinforcement learning (RL) algorithms, and the algorithm framework generally includes an offline layer and an online layer. The offline task is trained by the RL algorithm to obtain the optimal control actions of the vehicle in different operating states, to maximize the accumulated reward value.[17] When running online, the corresponding control actions are output according to the actual running state of the vehicle.[18] Since the heavy training is decoupled to the offline layer, the online computation requirement is alleviated and the real-time performance of C-EMS is improved.[12] In [19], the battery characteristics and brake specific fuel consumption (BSFC) are embedded into a deep deterministic policy gradient (DDPG), which speeds up the learning process by reducing the action space of control variables, and achieves better fuel economy. A more efficient DRL algorithm can meet the requirements of real-time, optimization and applicability of EMS. Some studies use distributed DRL as EMS to realize multi-thread parallel computing on multi-core CPU, which greatly improves the learning efficiency.[20,21] Aiming at the commonalities between EMS of different types of HEVs, Some studies applied a deep transfer reinforcement learning algorithm to realize the cross-type knowledge transfer between DRL-based EMS.[22,23] In [24], use the generative adversarial imitation learning algorithm to improve the optimization performance of DRL to improve its applicability.

### *1.3 Contribution and article structure*

Since the accuracy of prediction and the computing power of the controller will affect the real-time applicability of P-EMS, and the actual driving situation is very complex, it will increase the instantaneous error of the prediction. Moreover, C-EMS has the disadvantages of long training time and low efficiency, and it is necessary to further improve the convergence speed. The current EMS research on platoon is mainly P-EMS, and there is no mutual influence between the EMS of platoon nodes.

Based on the above studies, this paper proposes an overall framework as shown in Fig.2, and makes the following contributions to the Eco-CACC and EMS of vehicle platoon: 1) using the predecessor leader following (PLF) communication topology, based on the DMPC algorithm, an Eco-CACC controller with the optimization goals of stability, comfort and energy-saving is designed, and the asymptotic stability of the algorithm is analyzed; 2) construct the EMS of the ERELV platoon as a partially observable Markov decision process (POMDP), and describe the EMS of the ERELV platoon as a fully cooperative multi-agent problem, where multiple agents jointly explore the optimal actions under different vehicle states; 3) the MADRL algorithm is used to solve the EMS of multiple ERELV, combined with the CTDE network framework of the MADDPG algorithm, and the learning rate is accelerated by learning the experience of other vehicle nodes in the platoon.

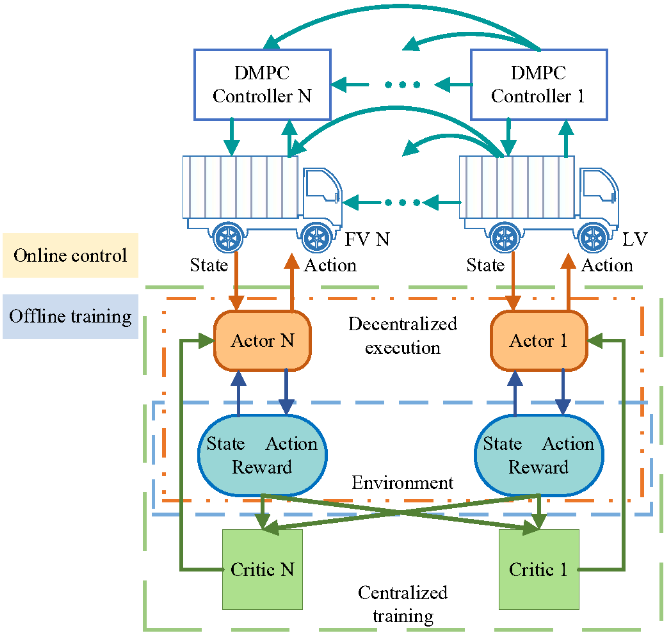


Fig.2. Eco-CACC and MADRL EMS overall control framework

The rest of this paper is as follows. Section 2 establishes the nonlinear vehicle platoon model and the dynamic system model of ERELV. Section 3 designs the Eco-CACC controller of the vehicle platoon based on the DMPC algorithm. Section 4 uses the MADDPG algorithm to solve the EMS of the ERELV platoon. Section 5 simulates and analyzes the effect of Eco-CACC multi-objective optimization control, and verifies the performance of the designed MADDPG EMS. The last section is the conclusion of this paper.

## 2 Modeling of platoon and ERELV

According to the research on the communication topology in [9], the vehicle platoon with the PLF communication structure has the best stability, and its topology is shown in Fig.3. The FV in the platoon can obtain the vehicle information of the preceding vehicle (PV) and the LV.

The powertrain of ERELV is shown in Fig.4, where the black and blue lines represent the transfer of mechanical and electrical energy. The power system consists of a power battery, power auxiliary unit (APU), drive motor, and control system. APU consists of an internal combustion engine (ICE) and integrated starter and generator (ISG).

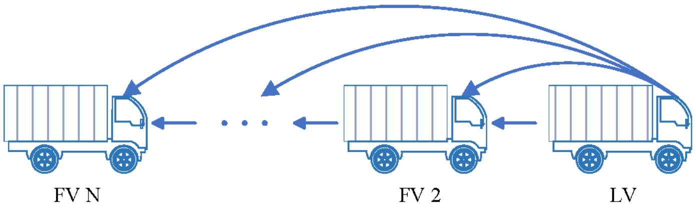


Fig.3. PLF communication topology of vehicle platoon

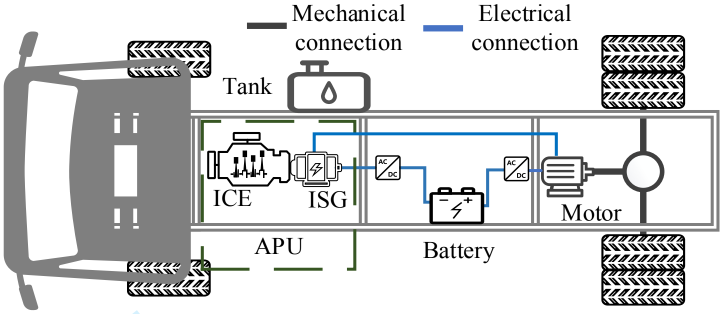


Fig.4. Powertrain structure of ERELV

## *2.1 Nonlinear platoon model*

The discrete nonlinear dynamic model of the vehicle node-$i$ in the platoon is given in Eq. (1).

$$\begin{cases} s_i(t+1) = s_i(t) + v_i(t)\Delta t \\ v_i(t+1) = v_i(t) + \left( \dfrac{\eta_i}{r_i} T_i(t) - C_{A,i} v_i^2(t) - m_i g f_i \right) \dfrac{\Delta t}{m_i} \quad i \in N \\ T_i(t+1) = T_i(t) - \dfrac{T_i(t)}{\tau_i}\Delta t + \dfrac{u_i(t)}{\tau_i}\Delta t \end{cases} \tag{1}$$

Where $N$ is the number of vehicles in the platoon, $s_i(t)$ is the displacement, $v_i(t)$ is the speed, $\Delta t$ is the discrete time step, $C_{A,i}$ is the coefficient of aerodynamic drag, $m_i$ is the mass of the vehicles in the platoon, $f_i$ is the rolling resistance coefficient, $T_i(t)$ is the actual driving/braking torque, $u_i(t)$ is the desired driving/braking torque, $\tau_i$ is the vehicle longitudinal dynamics time-delay constant, $r_i$ is the wheel radius, $\eta_i$ is the efficiency of the mechanical transmission system.

The state variables of a single-vehicle node are the position, speed, and wheel torque of the vehicle, denoted as $x_i(t)=\left[s_i(t),v_i(t),T_i(t)\right]^T$, The control variable is the wheel torque, $u_i(t)=T_i(t)$.

Eq. (1) can be further simplified to Eq. (2).

$$\begin{gathered} x_i(t+1)=A_i\left(x_i(t)\right)+B_i u_i(t) \\ y_i(t)=C_i x_i(t) \end{gathered} \tag{2}$$

Where $A_i=\left[s_i(t)+v_i(t)\Delta t, v_i(t)+\left(\eta_i T_i(t)/r_i-C_{A,i}v_i^2(t)-m_i g f_i\right)\Delta t/m_i, T_i(t)-T_i(t)\Delta t/\tau_i\right]^T$, $B_i=\left[0,0,\Delta t/\tau_i\right]^T$, $C_i=\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}$.

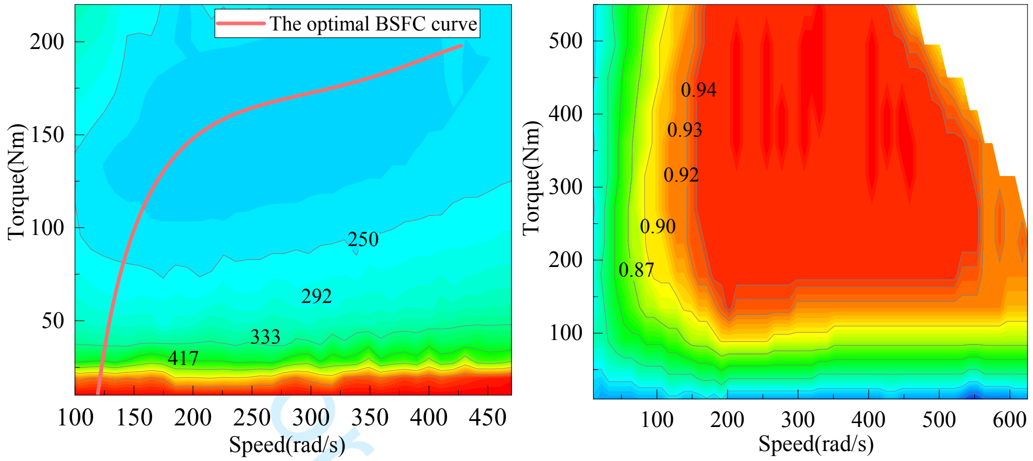
$X(t)=\left[x_1^T(t),x_2^T(t),\cdots,x_N^T(t)\right]^T$, $Y(t)=\left[y_1^T(t),y_2^T(t),\cdots,y_N^T(t)\right]^T$ and $U(t)=\left[u_1^T(t),\cdots,u_N^T(t)\right]^T$ represent the state, output, and input vectors of all nodes in the platoon at time *t*. Then the discrete nonlinear platoon dynamics model can be expressed as Eq. (3).

$$\begin{gathered} X(t+1)=A\left(X(t)\right)+BU(t) \\ Y(t)=CX(t) \end{gathered} \tag{3}$$

Where $A(t)=\left[A_1^T(x_1), A_2^T(x_2),\cdots,A_N^T(x_N)\right]^T$, $B=diag\left\{B_1,B_2,\cdots,B_N\right\}$, $C=diag\left\{C_1,C_2,\cdots,C_N\right\}$. The configuration and information flow topology characteristics of the platoon can be reflected by controlling the input vector.

### *2.2 Powertrain system model of ERELV*

ERELV has two working modes: charge depleting (CD) and charge sustaining (CS). In CD mode, the vehicle is powered by the battery, and the APU is turned off. In SD mode, the power from the APU can power the battery and can drive the motor through the power converter.[3]



(a) Engine (b) Drive motor

Fig.5. Efficiency map of the (a) engine and (b) drive motor

The engine in the powertrain is modeled by the efficiency map as shown in Fig.5. In [19], the operating point of the engine is fixed on the optimal BSFC curve, instead of searching globally in the map graph. This constraint reduces the space for motion exploration and improves training efficiency while achieving better fuel economy.

The energy consumption of the drive motor is considered in the optimization objective of Eco-CACC, and Eq. (4) gives the calculation method of the motor power.

$$P_i(k|t)=\begin{cases}\dfrac{T_i(k|t)v_i(k|t)}{r_i\eta_i\eta_d} & T_i(k|t)\geq 0\\[2ex] \dfrac{T_i(k|t)v_i(k|t)}{r_i}\eta_i\eta_b & T_i(k|t)<0\end{cases} \tag{4}$$

Where $\eta_d$ is the motor drive efficiency, $\eta_b$ is the braking efficiency of the motor, their values are obtained from the drive motor efficiency map in Fig.5. $P_i(k|t)$ is the drive motor power of vehicle node $i$ at time $t$, $k \in \left[0,1,2,\cdots,N_{p-1},N_p\right]$, $N_p$ is the prediction time domain of DMPC.

Eq. (5) is the equivalent circuit model of the battery, which represents the relationship between the current, voltage, and power of the battery.

$$I\left(t\right) = -\frac{V_{oc}\left(SOC\right) - \sqrt{V_{oc}^{2}\left(SOC\right) - 4P_bR_b\left(SOC\right)}}{2Q_bR_b\left(SOC\right)} \tag{5}$$

Where $SOC$ is the state of charge, $I\left(t\right)$ is the current, $V_{oc}$ and $R_b$ are the open-circuit voltage and internal resistance of the battery, respectively, both of which are related to the SOC of the battery, $P_b$ is the output power of the battery, and $Q_b$ is the nominal capacity of the battery.

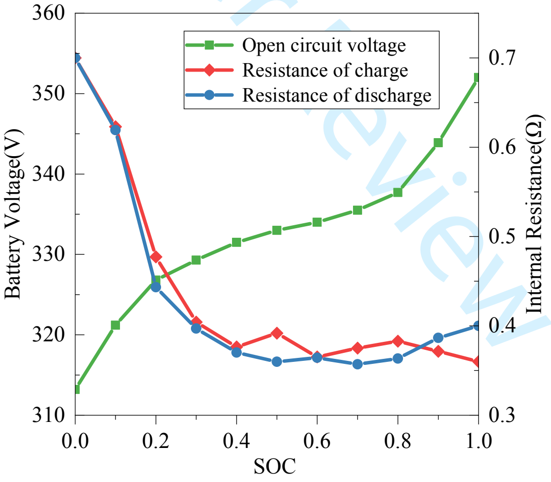


Fig.6. Battery characteristic

The curves of the open-circuit voltage and internal resistance of the battery with SOC are shown in Fig.6. When SOC is between 0.4 and 0.85, the internal resistance is relatively small, and the open-circuit voltage of the battery also changes little.

During the actual operation of ERELV, the power system components need to meet the constraints in Eq. (6).

$$\begin{cases} SOC_{\min} \le SOC(t) \le SOC_{\max} \\ P_{b,\min} \le P_b(t) \le P_{b,\max} \\ T_{x,\min} \le T_x(t) \le T_{x,\max} \\ \omega_{x,\min} \le \omega_x(t) \le \omega_{x,\max}, x = eng, mot \end{cases} \tag{6}$$

Where $P_b$ is the output power of the battery, $T$ and $\omega$ are the torque and speed of the components $x$ (engine and motor).

## 3 Eco-CACC based on DMPC

DMPC can couple multiple ICVs to each other through the information topology structure, and the overall control objectives are achieved through the coordinated control of each vehicle.[28]

### *3.1 DMPC Controller Design*

The control variable transfer flow of Eco-CACC based on DMPC is shown in Fig.7. A single-point optimization problem is defined on each vehicle node in the platoon, and the model prediction optimization problem of all nodes needs to be solved at each optimization moment. Using the domain node information obtained by the PLF communication topology, each sub-problem is optimized to obtain the control input of the vehicle.[11]

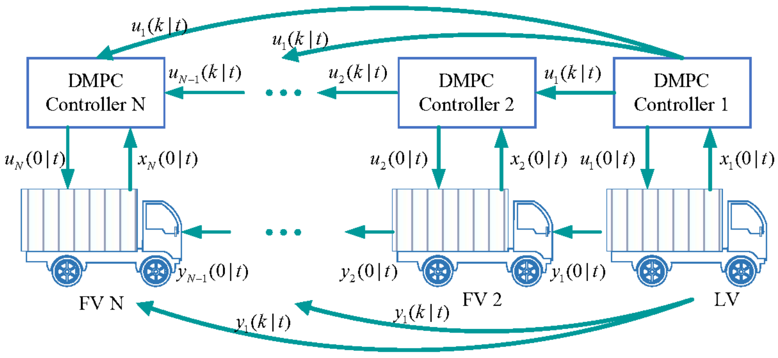


Fig.7. Eco-CACC controls variable transfer flow

The position and velocity of the LV are denoted by $s_0(t)$ and $v_0(t)$, respectively. $u_{i,des}(t)= T_{i,des}(t)$ and $x_{i,des}(t)=\left[s_{i,des}(t),v_{i,des}(t),T_{i,des}(t)\right]^T$ represent the desired input and desired state of the node in the platoon. $s_{i,des}(t)=s_0(t)-id_0$, $d_0$ is the expected following distance between the two vehicles. $v_{i,des}(t)=v_0(t)$. $T_{i,des}(t)=h_i(v_0)$, $h_i(v_0)=r_i\left(C_{A,i}v_0^2+m_i g f_i\right)\big/\eta_i$ is the vehicle driving resistance equation.

For each vehicle node $i\in N$, $A=\left[a_{i,j}\right]\in$ ¡ $^{N\times N}$ is an adjacency matrix, representing the communication direction between nodes, $a_{i,j}=1(=0)$ indicates that node $i$ can (cannot) receive the information sent by node $j$. Define $\mathsf{F}=\left\{j\left|a_{i,j}=1, j\in N\right.\right\}$ as the set of domain vehicle information that node $i$ can obtain. $P_i=1(=0)$ indicates that node $i$ can (cannot) obtain LV information, $\mathsf{P}_i$ is the set of LV information that node $i$ can receive, $\mathsf{P}_i=\{0\}$ when $P_i=1$, $\mathsf{P}_i=\varnothing$ when $P_i=0$, $\mathsf{I}_i=\mathsf{F}_i\,\mathsf{U}\,\mathsf{P}_i$ is the set of all information that vehicle node $i$ can receive.

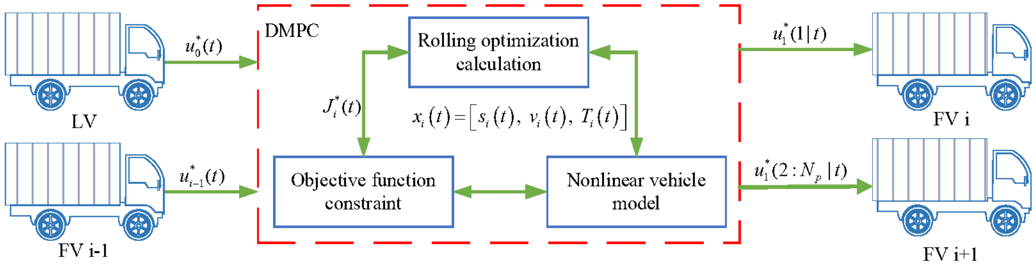


Fig.8. Vehicle node control variable information transfer flow

In each prediction time domain $\left[t,t+N_p\right]$ of the nonlinear MPC problem, define $u_i^p\left(k\left|t\right.\right)$ as the vehicle state trajectory sequence for optimal control, $u_i^*\left(k\left|t\right.\right)$ is the optimal predictive control input sequence obtained by rolling optimization, $u_i^a\left(k\left|t\right.\right)$ is the control input sequence passed to other vehicles, $k\in\left[0,1,2,\cdots,N_{p-1},N_p\right]$. The corresponding three output sequences are defined as prediction output sequence $y_i^p\left(k\left|t\right.\right)$, optimal prediction output sequence $y_i^*\left(k\left|t\right.\right)$ and hypothesis output sequence $y_i^a\left(k\left|t\right.\right)$.

The single-node control variable information transfer of DMPC is shown in Fig.8. Eq. (7) is the designed multiple optimization objective functions, $J_{1,i}$ and $J_{2,i}$ are the following error cost functions, $J_{3,i}$ is the comfort cost function, $J_{4,i}$ is the stability cost function, and $J_{5,i}$ is the economic cost function.

$$
\begin{aligned}
&J_{1,i}(k|t)=\left\|Q_i\left(y_i^p(k|t)-y_{i,des}(k|t)\right)\right\|_2\\
&J_{2,i}(k|t)=\sum_{j\in\mathrm{F}_i}\left\|G_i\left(y_i^p(k|t)-y_j^a(k|t)-d_{j,i}\right)\right\|_2\\
&J_{3,i}(k|t)=\left\|R_i\left(u_i^p(k|t)-h_i(v_i^p(k|t))\right)\right\|_2\\
&J_{4,i}(k|t)=\left\|F_i\left(y_i^p(k|t)-y_i^a(k|t)\right)\right\|_2\\
&J_{5,i}(k|t)=\left\|W_iP_i(k|t)\right\|_2
\end{aligned}
\tag{7}
$$

Where $Q_i$, $G_i$, $R_i$, $F_i$, $W_i$ are symmetric non-negative definite weight matrices, $Q_i$ is the error coefficient matrix of the FV and the LV, $G_i$ is the error coefficient matrix of the assumed trajectory of the node *i* and its neighboring nodes, $R_i$ reflects the penalty on acceleration and deceleration behavior, $F_i$ is the communication stability weight coefficient matrix of the FV, $W_i$ is the weight coefficient matrix of the vehicle's energy consumption.

The prediction optimization problem of vehicle node *i* can be formulated as Eq. (8).

$$
\begin{gathered}
\min_{U_i} J_i\left(y_i^p(:|t),u_i^p(:|t),y_i^a(:|t),y_{-i}^a(:|t)\right)=\\
\sum_{k=0}^{N_P-1}\left\{J_{1,i}(k|t)+J_{2,i}(k|t)+J_{3,i}(k|t)+J_{4,i}(k|t)+J_{5,i}(k|t)\right\}\\
st:\ \ x_i'^p(k+1|t)=A_i\left(x_i^p(k|t)\right)+B_iu_i^p(k|t)\\
y_i^p(k|t)=C_ix_i^p(k|t)\\
x_i^p(0|t)=x_i(t)\\
T_{\min}\le u_i^p(k|t)\le T_{\max}\\
y_i^p(N_p|t)=\frac{1}{|\mathrm{I}_i|}\sum_{j\in\mathrm{I}_i}\left(y_j^a(N_p|t)-d_{j,i}\right)\\
T_i^p(N_p|t)=h_i\left(v_i^p(N_p|t)\right)
\end{gathered}
\tag{8}
$$

Where $U_i = \left[ u_i^p\left(0|t\right), u_i^p\left(1|t\right), \cdots, u_i^p\left(N_p - 1|t\right) \right]^T$ is the control sequence to be optimized. The constraints in Eq. (8) include: dynamic constraints in the prediction time domain, amplitude constraints of the control input, prediction terminal equation constraints, and acceleration and deceleration constraints. $|\mathbb{I}_i|$ is the number of elements in the set, $\mathbb{I}_i = \mathbb{F}_i \cup \mathbb{P}_i$. $d_{j,i} = \left[ -(j-i)d_0, 0 \right]^T$ is the deviation of the following distance between nodes.

### 3.2 Asymptotic stability analysis

The optimal cost function value of node *i* at time *t* can be obtained as Eq. (9).

$$\begin{aligned} J_i^*\left(t\right) &= J_i^*\left(y_i^*\left(:|t\right), u_i^*\left(:|t\right), y_i^a\left(:|t\right), y_{-i}^*\left(:|t\right)\right) \\ &= \sum_{k=0}^{N_p-1} \left\{ J_{1,i}\left(k|t\right) + J_{2,i}\left(k|t\right) + J_{3,i}\left(k|t\right) + J_{4,i}\left(k|t\right) + J_{5,i}\left(k|t\right) \right\} \end{aligned} \tag{9}$$

For the optimization problem *P* of vehicle node *i* in DMPC, when the predicted terminal of the node is consistent with the expected state, the value of the single-step iteration of the node cost function is Eq. (10).

$$J_i^*\left(t+1\right) - J_i^*\left(t\right) \le -l_i\left(y_i^*\left(0|t\right), u_i^*\left(0|t\right), y_i^a\left(0|t\right), y_{-i}^a\left(0|t\right)\right) + \varepsilon_i \tag{10}$$

Where $\varepsilon_i$ is defined as Eq. (11).

$$\varepsilon_i = \sum_{k=1}^{N_p-1} \left\{ \sum_{j \in \mathbb{F}_i} \left\| G_i\left(y^j\left(k|t\right) - y_i^p\left(k|t\right)\right) \right\|_2 - \left\| F_i\left(y_i^*\left(k|t\right) - y_i^a\left(k|t\right)\right) \right\|_2 \right\} \tag{11}$$

According to the Lyapunov stability principle, the overall optimization objective function of the vehicle platoon is used as the candidate function. For the optimization problem of DMPC, when the predicted terminal of the node is consistent with the expected state, the single-step iterative decline value of the objective function of all nodes in the platoon is Eq. (12).

$$J_{\Sigma}^{*}(t+1)-J_{\Sigma}^{*}(t)\le -\sum_{i=1}^{N} l_i\left(y_i^{*}(1|t),u_i^{*}(0|t),y_i^{a}(1|t),y_{-i}^{a}(1|t)\right)+\sum_{k=1}^{N_p-1}\varepsilon_{\Sigma}(k) \tag{12}$$

Where $J_{\Sigma}^{*}(t)=\sum_{i=1}^{N} J_i^{*}(t)$ is the sum of the optimal values of all node vehicles in the platoon at time $t$, $\varepsilon_{\Sigma}(k)$ is defined as Eq. (13).

$$\varepsilon_{\Sigma}(k)=\sum_{i=1}^{N}\left[\sum_{j\in \mathrm{O}_i}\left\|G_i\left(y_i^{*}(k|t)-y_i^{a}(k|t)\right)\right\|_2-\left\|F_i\left(y_i^{*}(k|t)-y_i^{a}(k|t)\right)\right\|_2\right] \tag{13}$$

Where $\mathrm{O}_i$ is the set of nodes that can receive the information status of node $i$.

For Eq. (13), according to the Lyapunov stability principle, when $J_{\Sigma}^{*}(t+1)\le J_{\Sigma}^{*}(t)$, the asymptotic stability of the platoon can be achieved, the Eq. (14) needs to be satisfied.

$$\sum_{j\in \mathrm{O}_i}\left\|G_j\left(y_i^{*}(k|t)-y_i^{a}(k|t)\right)\right\|_2<\left\|F_i\left(y_i^{*}(k|t)-y_i^{a}(k|t)\right)\right\|_2 \tag{14}$$

Eq. (14) can be satisfied by artificially setting $G_i$ and $F_i$ values, so that the asymptotic stability of the heterogeneous nonlinear vehicle platoon can be achieved.

## 4 EMS for ERELV platoon

In this paper, the ERELV platoon is described as a multi-agent system (MAS), which relies on multiple agents working together to achieve a common goal to improve the optimization rate.[25]

### *4.1 MADDPG algorithm*

The state transition of ERELV has the Markov property and generates a corresponding energy consumption at each time step.[3] Since each agent cannot know the complete environment state, the EMS of the ERELV platoon can be modeled as a partially observable POMDP.[27] The task types applied by MADRL algorithm are divided into fully cooperative, fully competitive and hybrid types.[25] For the EMS of the ERELV platoon, each vehicle node needs to jointly explore the optimal control behavior under different vehicle states, so it is a fully cooperative MADRL problem.

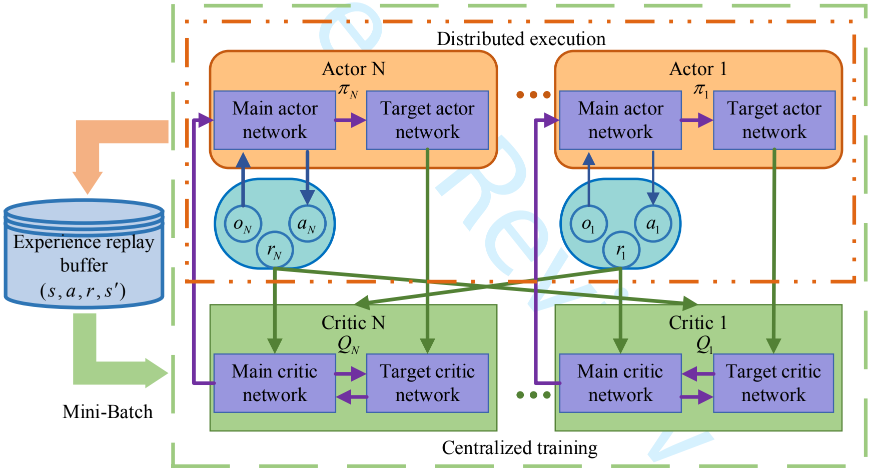


Fig.9. MADDPG algorithm structure

The MADDPG algorithm pioneered the CTDE framework shown in Fig.9, each agent has a critic network and an actor network, MADRL can improve the learning efficiency by learning the experience of other vehicles in the platoon, achieve the overall fuel economy.[26] The actor network input is the local state information of the agent, while the critic network input includes the action and observation state information of all agents.[29]

The actor network uses random exploration to find the optimal reward policy, and the input of the network is the local observation $o_i$ of the agent *i*, the output action is $a_i=\pi_i\left(o_i,\theta_i\right)$, to change the policy $\pi_i\left(o_i,\theta_i\right)$ by updating the parameter $\theta_i$ to find the optimal policy. The cumulative expected reward of the *i*-th agent is $J\left(\theta_i\right)=E_{s:\rho^{\pi},a_i:\pi_{\theta_i}}\left[\sum_{t=0}^{\infty}\gamma^t r_{t,i}\right]$, that the goal of the actor network is to maximize this expectation. For a random policy, the policy gradient is calculated as:

$$\nabla_{\theta_i}J\left(\theta_i\right)=E_{s:\rho^{\pi},a_i:\pi_i}\left[\nabla_{\theta_i}\log\pi_i\left(a_i\middle|o_i\right)Q_i^{\pi}\left(x,a_1,\cdots,a_N\right)\right] \tag{15}$$

Where $Q_i^{\pi}\left(x,a_1,\cdots,a_N\right)$ is the centralized action function of the *i*-th agent, which is used to represent the value of joint actions, $x=\left\{o_1,\cdots,o_N\right\}$ is the set of observations for all agents. Since each agent can independently learn its centralized action function $Q_i^{\pi}$, each agent has a different reward function, so it is possible to complete cooperative or competitive tasks.[30]

The action-value function of the centralized critic network is $Q_i^{\mu}$. The update method is similar to the traditional TD-error algorithm. The parameter $\theta_i$ is updated by minimizing the loss function $L\left(\theta_i\right)$. The gradient update method is as follows:

$$\begin{gathered}L\left(\theta_i\right)=E\left[\left(Q_i^{\mu}\left(x,a_1,\cdots,a_N\right)-y\right)^2\right]\\ y=r_i+\gamma Q_i^{\mu'}\left(x',a_1',\cdots,a_N'\right)\Big|_{a_j'=\mu_j'\left(o_j\right)}\end{gathered} \tag{16}$$

Where $\mu'=\left\{\mu_{\theta_i'},\cdots,\mu_{\theta_N'}\right\}$ is the strategy set of the target network, and $Q_i^{\mu'}$ is the action-value function, which is used to evaluate the pros and cons of the action in the subsequent steps, derived from the output of the target value network. The function $y$ is the cumulative future average return of the agent $i$ in the target actor network, the output of the value network is $Q_i^{\mu}$, the loss function $L(\theta_i)$ is the square of the difference between $Q_i^{\mu}$ and $y$, and the gradient descent method is used to minimize the objective function, thereby updating the network parameters $\theta_i$.

### *4.2 Multi-agent EMS based on MADDPG*

The principle of interaction between each agent and environmental information based on the MADDPG algorithm is shown in Fig.10. The MADDPG algorithm adopts the CTDE learning framework. The critic network of each agent can obtain the observation $(o_1, o_2, \cdots, o_N)$ and action $(a_1, a_2, \cdots, a_N)$ of all agents, and the actor network only executes the action $a_i$ according to the local observation $o_i$.

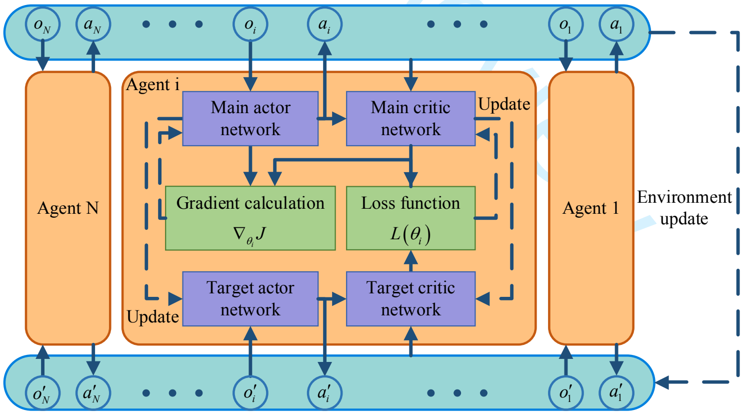


Fig.10. Interaction between agent and environment

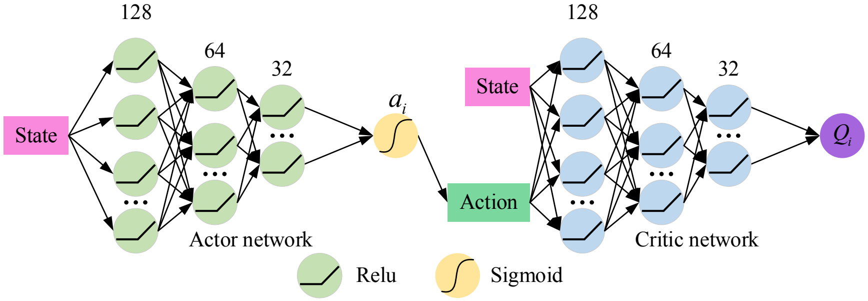


Fig.11. Architecture of actor-critic network.

The structure of Critic network and Actor network is shown in Fig.11, there are three fully connected hidden layers with the number of nodes of 128-64-32, and each layer is connected with Relu activation function. The output layer of the Actor network is processed by the Sigmoid activation function to generate actions. The hyperparameters of the network are shown in Table 1.

Table 1. Hyperparameters of agent.

| Hyperparameters | Values |
|---|---|
| Discount factor | 0.99 |
| Learning rate of Actor | 0.001 |
| Learning rate of Critic | 0.01 |
| Memory capacity | 10000 |
| Replay buffer size | 128 |
| Soft target update | 0.01 |

Eq.(17) is the observation, action, and reward of each agent. The state variable $o_i$ includes the desired torque demand $T_i$, the vehicle speed $v_i$ and the state of charge $SOC_i$. $T_i$ and $v_i$ are obtained by the DMPC controller. Action $a_i$ is the output power $P_i$ of the APU. $r_i$ is the reward value, $m_{fuel,i}$ is the fuel consumption in the simulation step. $SOC_{ref}$ is the reference value of the electric charge, which can maintain the power of the battery near the optimal state. $\alpha$ is used to balance fuel consumption and maintain electricity.

The key of the multi-objective reward function is the adjustment of parameters, which aims to meet the battery requirements and improve the fuel economy. Different weights represent different optimization effects, which are studied in [19].

$$\begin{cases} o_i = \{T_i, V_i, SOC_i\} \\ a_i = \{P_i\} \\ r_i = -\left(m_{fuel,i} + \alpha[SOC_{ref} - SOC_i]^2\right) \end{cases} \tag{17}$$

## 5 Simulation and Results Analysis

In this section, the proposed algorithm is simulated and verified under typical standard driving conditions FTP75, WLTC and NEDC, and the data set of agent training is also these three standard conditions. The research object is a heterogeneous ERELV platoon with the same powertrain structure. The parameters of the vehicle platoon are shown in Table 2.

Table 2. Parameters of ERELV platoon

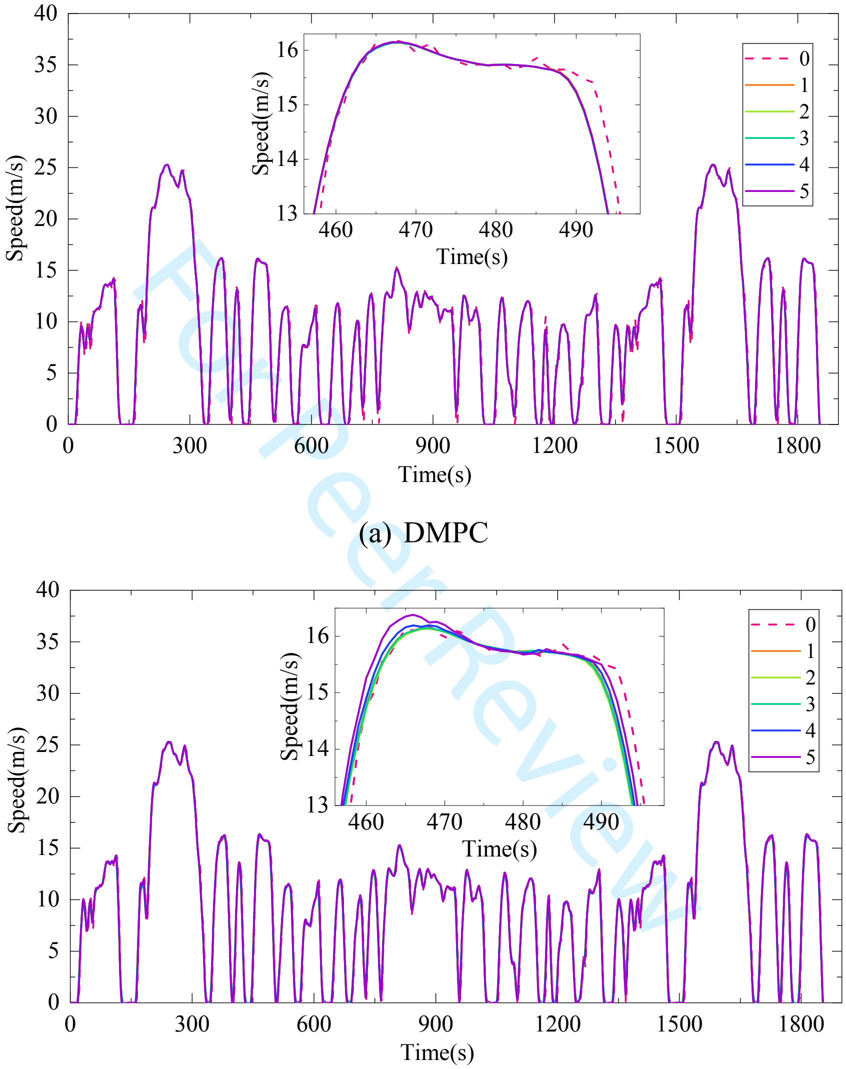
| Symbol | Parameters | Unit | Value |
|---|---|---|---|
| Vehicle | $m_{1,2,3,4,5}$ | kg | 4495、4200、3900、3600、3350 |
| | $f_i$ | -- | 0.15 |
| | $C_D$ | -- | 0.6 |
| | $A$ | $m^2$ | 5.1 |
| | $d_0$ | m | 10 |
| Transmission | $r_i$ | m | 0.376 |

| | | | |
|---|---|---|---|
| | $\tau_i$ | s | 0.75 |
| | $\eta_i$ | -- | 0.95 |
| | Reduction ratio | -- | 6.15 |
| Traction motor | Maximum power | kw | 100 |
| | Maximum torque | Nm | 850 |
| | Maximum speed | rpm | 6000 |
| APU | Rated power | kw | 60 |
| | Maximum torque | Nm | 265 |
| | Maximum speed | rpm | 5500 |
| Battery | Capacity | kWh | 26.25 |
| | Voltage | V | 350 |
| | $SOC_{ref}$ | -- | 0.55 |
| Cost function weight | $Q_i / G_i / R_i / F_i / W_i$ | -- | 10/5/10/8/10 |

### *5.1 Eco-CACC algorithm optimizes performance*

Aiming at the multi-objective optimization performance of Eco-CACC, the indicators proposed in [11] are used for evaluation and compared with the method proposed in [7]. The speed following performance, vehicle distance keeping performance, comfort performance, and energy-saving performance are used as evaluation indicators to evaluate the control effect of different algorithms on vehicle platoon. $\delta_{\max}(v_{l,i}) = |v_l - v_i|$ is the maximum speed deviation between the FV and the LV during driving, which is used to evaluate the speed following performance. The maximum distance deviation $\delta_{\max}(d_{i,i+1})$ between the FV is used to evaluate the distance keeping performance. The absolute value $a_{\max} = |a_i|$ of the maximum acceleration is used to represent the driving comfort of the vehicle. The average energy consumption $P_{mean} = \sum_{i=1}^{N} P_i \Big/ N$ is used to evaluate the energy-saving effect of different algorithms in the whole driving cycle.

Fig.12 shows the speed following effect of Eco-CACC based on DMPC and MPC under FTP75. As can be seen from the figure, each FV controlled by DMPC can follow the speed of LV well, and the speed following error is smaller than that of MPC.



(a) DMPC

(b) MPC

Fig.12. Platoon speed curves of (a) DMPC and (b) MPC algorithms under FTP75

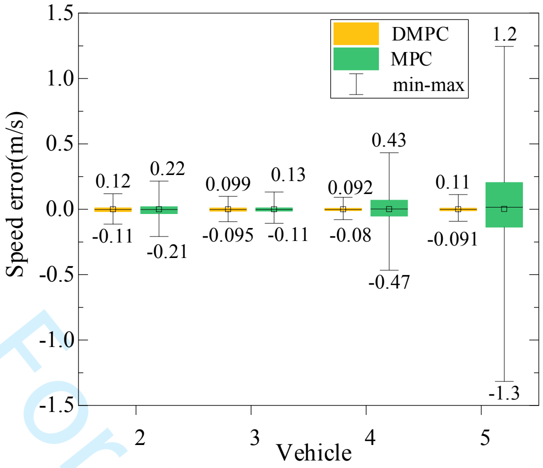
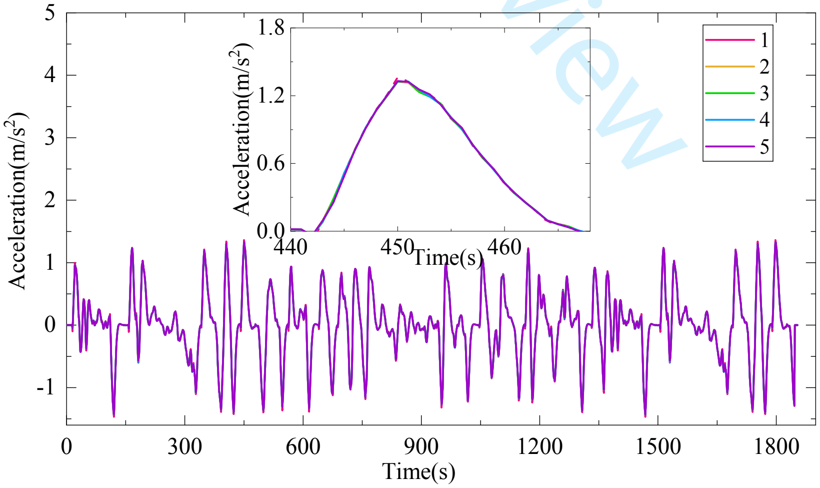
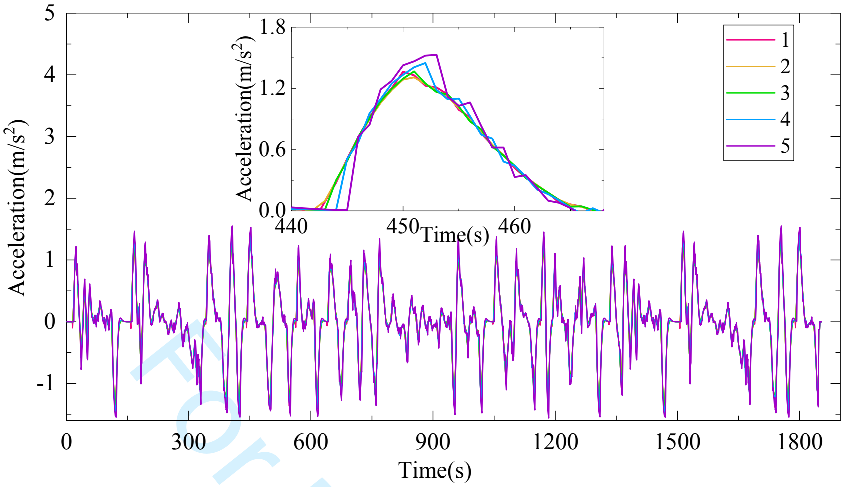


Fig.13. Box plot of vehicle speed error between LV and FV *i* under FTP75

Fig.13 shows the speed error between the FV and the LV under different algorithms. The maximum speed error of DMPC and MPC are 0.12 m/s and 1.3 m/s, respectively. Since the FV based on the MPC algorithm only accepts the state information of the PV, the speed error of the FV will gradually enlarge.



(a) DMPC



(b) MPC

Fig.14. Platoon acceleration curves of (a) DMPC and (b) MPC algorithms under FTP75

The acceleration of each FV in the platoon is too high or changes frequently, which will affect the driving comfort of the FV and increase energy consumption. Fig.14 is the acceleration curve of platoon vehicles controlled by DMPC and MPC under FTP75, the absolute maximum acceleration values of the DMPC algorithm and MPC algorithm are 1.46 m/s$^2$ and 1.56 m/s$^2$, respectively. The designed Eco-CACC based on DMPC has little change in acceleration and deceleration, showing good driving comfort and energy-saving. Under the MPC algorithm, since the error is gradually amplified, the further behind the FV, the greater the acceleration and deceleration.

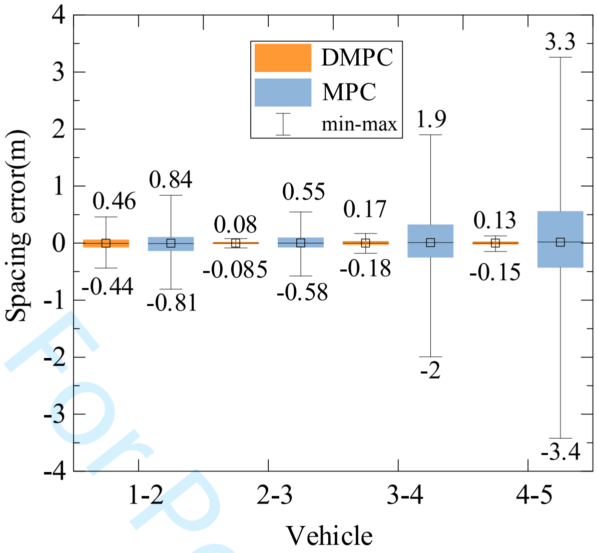


Fig.15. Box plot of spacing error of vehicles in platoon under FTP75

The deviation between the vehicle distance and the ideal distance reflects the stability of the vehicle platoon. The box plot in Fig.15 shows the distribution of the distance deviation range for different control methods. For the vehicle platoon using the MPC method, the further back the vehicles in the platoon are, the larger the spacing error and the lower the stability. And DMPC makes the fluctuation range of the spacing error of the FVs relatively concentrated, and the stability of the platoon is significantly improved.

Fig.16 shows the energy consumption of different algorithms under FTP75. In the case of considering regenerative braking, DMPC saves an average of 1% power consumption compared to MPC, and the energy consumption of vehicles at each node in the platoon is also relatively reduced.

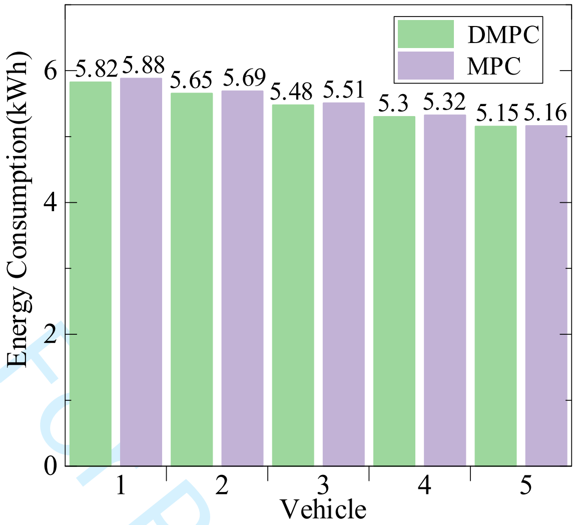


Fig.16. Energy consumption of vehicles in platoon under FTP75

Comparing the control effects of Eco-CACC under different algorithms, under multiple driving cycles, Table 3 summarizes the evaluation indicators of the speed following, driving comfort, distance keeping, and energy-saving of the platoon. Under various indicators, the control effect of DMPC is better than that of MPC.

Table 3. Comparison of different algorithms

| Algorithm | Driving cycle | $\delta_{\max}(v_{l,i})$ (m/s) | $a_{\max}$ (m/s$^2$) | $\delta_{\max}(d_{i,i+1})$ (m) | $P_{mean}$ (kWh) |
|---|---|---|---|---|---|
| DMPC | FTP75 | 1.4 | 1.46 | 0.46 | 5.48 |
| | WLTC | 1.6 | 1.48 | 0.44 | 11.62 |
| | NEDC | 1.2 | 1.37 | 0.41 | 4.33 |
| MPC | FTP75 | 1.9 | 1.56 | 3.41 | 5.51 |
| | WLTC | 2.2 | 1.65 | 3.12 | 11.67 |
| | NEDC | 1.6 | 1.48 | 2.72 | 4.36 |

### 5.2 Comparative analysis of EMS

Since the multi-agent EMS is a fully cooperative type of MADRL, the EMS of each agent is the same, so the control effect of the first vehicle is used for analysis. Taking DP-based EMS as a benchmark, the performance of DDPG-based and MADDPG-based EMS on optimization performance and convergence rate is compared.

During the training iteration of the DRL algorithm, the change in the average reward value remains stable to indicate that the training is complete. The average reward value trajectories of DDPG- and MADDPG-based EMS over 100 episodes under the FTP75 driving cycle are shown in Fig.17. The MADDPG-based EMS converged to a steady state after the 12th episode, while the DDPG-based EMS only reached a steady state after the 24th episode. The average reward value of the DDPG algorithm at the beginning is greater than that of the MADDPG algorithm. Since it needs to accumulate enough experience to start training, MADDPG uses the experience of multiple agents, so the accumulation speed is relatively fast. In the MADDPG algorithm, different agents make different actions for the same state, which saves the agent time to explore the action space. These advantages make the MADDPG-based EMS converge faster and the training results are relatively stable.

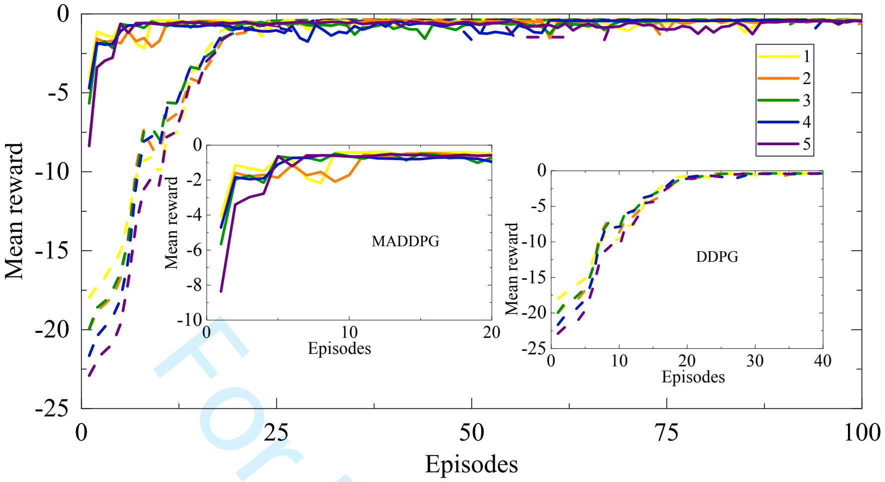


Fig.17. Convergence curves for different algorithms under FTP75

The DP algorithm can consider all the states of the vehicle, so it can realize the global optimal EMS, and use it to evaluate the global optimization characteristics of EMS based on DDPG and MADDPG.[31] The initial value of SOC is set to 0.65, and the end value is 0.55.

The SOC trajectories of the first vehicle in the platoon are selected for comparison. Under the FTP75 driving cycle, the SOC trajectories of the three different algorithms are shown in Fig.18. The three algorithms have similar SOC trajectories, which are related to the characteristics of the driving cycles selected for training. The DP-based EMS can obtain the global optimal fuel economy, so the SOC can be strictly limited within the preset range, while the DRL-based EMS only considers the instantaneous SOC value.[18] There is a deviation between the final SOC values of the three algorithms, and the deviation is converted into fuel consumption.

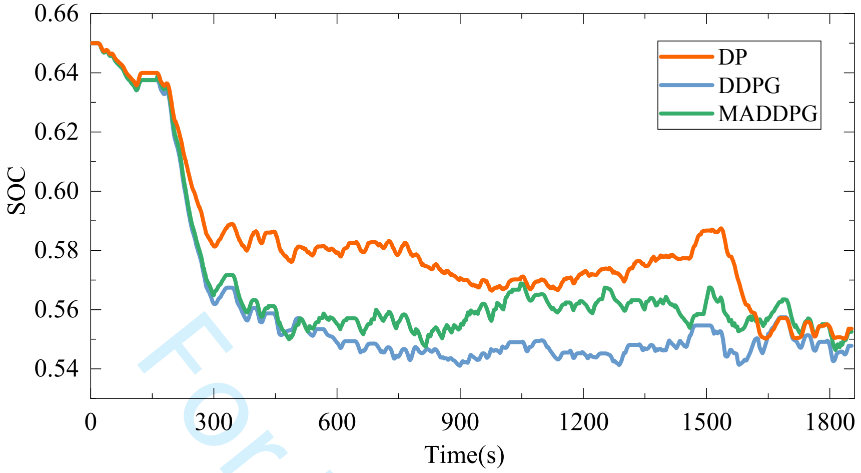
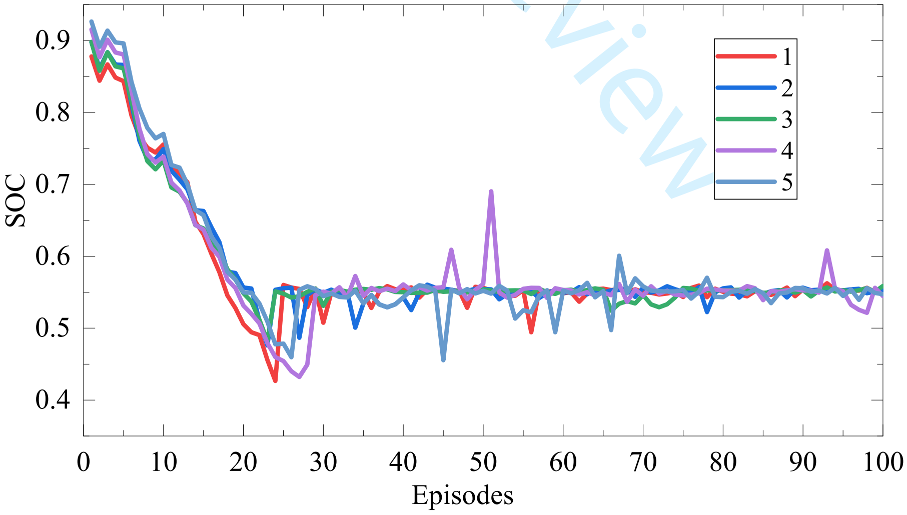
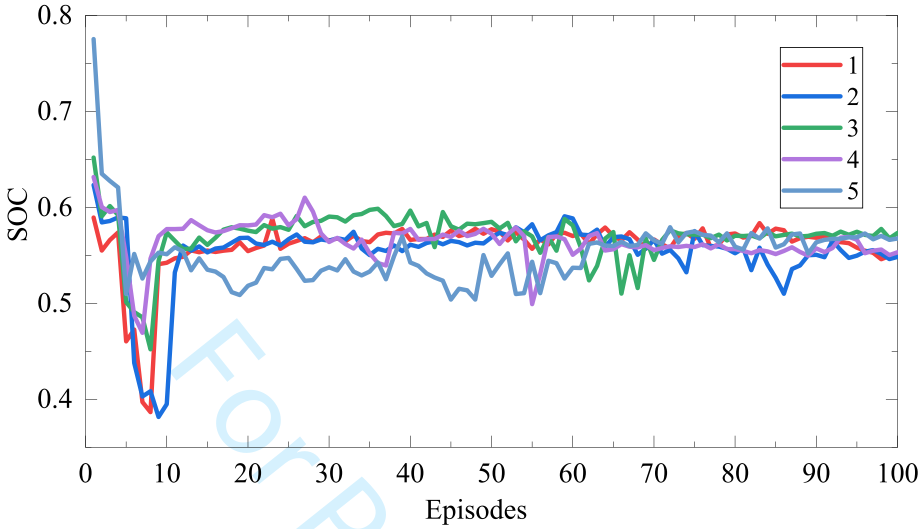


Fig.18. SOC trajectories for different algorithms under FTP75

Fig.19 shows the final SOC values of EMS based on the two algorithms. The final SOC values of both algorithms are near $SOC_{ref}$, indicating the constraint performance of the reward function, and the convergence speed of MADDPG algorithm is fast.



(a) DDPG



(b) MADDPG

Fig.19. SOC final valuefor different algorithms under FTP75

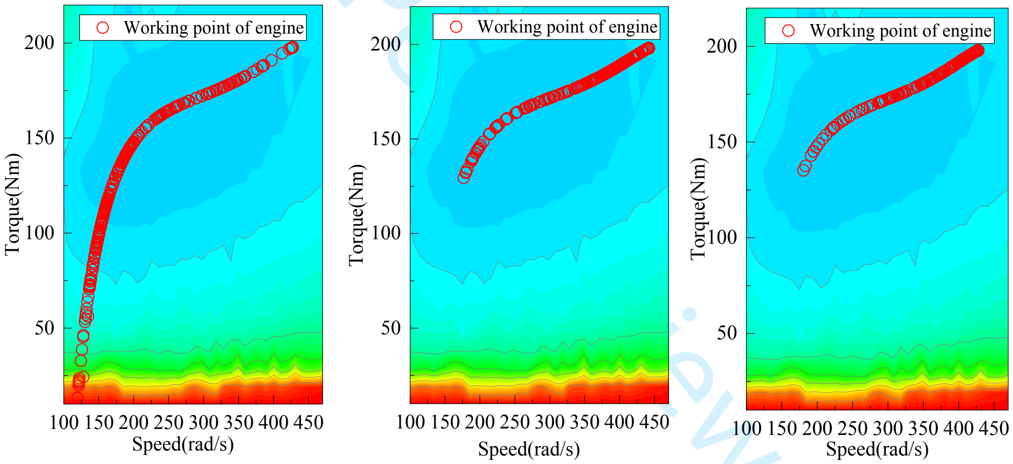
The control effects of the three EMS are summarized in Table 4. The difference in fuel consumption per 100 kilometers between the two DRL-based EMS is small. This is because MADDPG is extended based on DDPG, and its advantage lies in its fast convergence speed. Compared with DP-based EMS, DDPG-based and MADDPG-based EMS achieved 95.1% and 95.6% of the optimal fuel consumption, and both achieved good fuel economy.

Table 4. EMS Comparison of Different Algorithms

| Algorithm | DP | DDPG | MADDPG |
|---|---|---|---|
| Fuel consumption (L/100 km) | 9.01 | 9.47 | 9.42 |
| Terminal SOC | 0.554 | 0.548 | 0.553 |
| Fuel economy (%) | 100 | 95.1 | 95.6 |
| Average convergence rate (Episode) | -- | 24 | 12 |

The engine operating point distribution of the first vehicle in the platoon is shown in

Fig.20. (a), (b), and (c) is the engine operating points of the EMS based on DP, DDPG, and MADDPG algorithms. The engine operating points of DP-based EMS are most widely distributed, but some are located in the low power consumption area. This is because the DP algorithm needs to perform power distribution on a global basis to achieve global optimal control. In contrast, EMS based on DDPG and MADDPG only considers the instantaneous working state, so the engine operating points are concentrated in the fuel economy area, and the distribution of engine operating points is relatively similar.



(a) DP (b) DDPG (c) MADDPG

Fig.20. Engine operating point

Under FTP75 operating conditions, the distribution of engine operating points under DP, DDPG and MADDPG control algorithms is shown in Fig.21. Since DP algorithm can accurately calculate the engine output power of each step from a global perspective, its engine operating points are mainly concentrated in the interval [0-45]. DDPG and

MADDPG algorithms mainly optimize control according to the reward function of each step and pay more attention to the current reward value, so the engine operating points are mainly distributed in the interval [65-85], which is the same as the situation in [19].

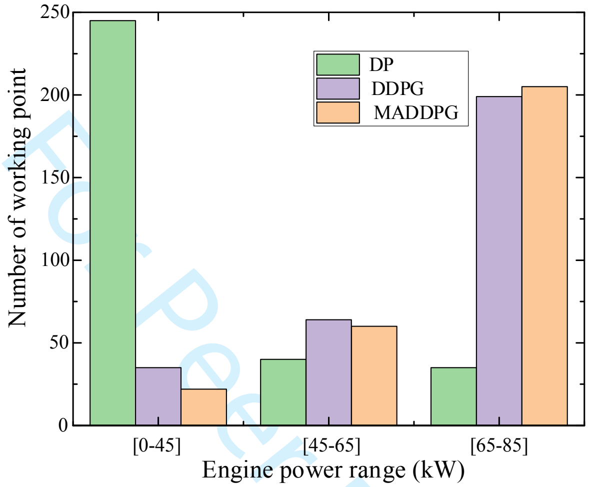


Fig.21. The distribution of engine operating points with different algorithms

Table 5 summarizes the performance of three different EMS. The higher number of episodes in NEDC is due to the short duration of NEDC and relatively few training times per episode. It can be seen that under different driving cycles, thanks to the advantages of complete cooperation among multiple agents, the convergence speed of EMS based on MADDPG is twice that of DDPG, and the optimization effect is roughly equal to that of DP algorithm.

Table 5. EMS comparison under different driving cycles

| Algorithm | Driving cycle | Fuel consumption (L/100 km) | Fuel consumption of DP (L/100 km) | Fuel economy (%) | Average convergence rate (episodes) |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| DDPG | FTP75 | 9.47 | 9.01 | 95.1 | 24 |
| | WLTC | 14.25 | 13.48 | 94.6 | 29 |
| | NEDC | 9.21 | 8.78 | 95.3 | 42 |
| MADDPG | FTP75 | 9.42 | 9.01 | 95.6 | 12 |
| | WLTC | 14.21 | 13.48 | 94.9 | 16 |
| | NEDC | 9.23 | 8.78 | 95.3 | 24 |

## 6 Concluding remarks and future work

In this paper, we improve the fuel economy of ERELV platoon from CACC and EMS. Under the PLF communication topology, a DMPC-based Eco-CACC controller is designed to optimize comfort and economy based on achieving stability. Compared with MPC, DMPC has better safety performance, improved driving comfort, and relatively better economic performance. MADRL is used to solve the EMS problem of the ERELV platoon. Under the CTDE framework of the MADDPG algorithm, the experience of all agents is used for learning during offline training, and the actions are output according to local observations during online control. Compared with the DDPG algorithm, the final training results of MADDPG are similar, but the training convergence speed of MADDPG is twice as fast. Compared with the DP, the MADDPG can obtain an approximate optimal solution. The MADRL EMS proposed in this paper can significantly improve the learning efficiency while achieving a near-optimal solution.

The simulation of the vehicle platoon EMS based on MARL proposed in this paper is completed on one computer, but in the actual application scenario, it is completed by multiple computing units. In the calculation process, they also need to communicate with each other, and its feasibility should be verified in the future.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 52172369).

## References

1. Shladover SE, Nowakowski C, Lu X-Y, Ferlis R. Cooperative Adaptive Cruise Control:Definitions and Operating Concepts. Transportation Research Record. 2015;2489(1):145-52.
2. Vahidi A, Sciarretta A. Energy saving potentials of connected and automated vehicles. Transportation Research Part C: Emerging Technologies. 2018;95:822-43.
3. Li J, Wang Y, Chen J, Zhang X. Study on energy management strategy and dynamic modeling for auxiliary power units in range-extended electric vehicles. Applied Energy. 2017;194:363-75.
4. Xiao B, Walker PD, Zhou S, Yang W, Zhang N. A Power Consumption and Total Cost of Ownership Analysis of Extended Range System for a Logistics Van. IEEE Transactions on Transportation Electrification. 2022;8(1):72-81.
5. Tang, X., Yang, K., Wang, H. et al. Driving Environment Uncertainty-Aware Motion Planning for Autonomous Vehicles. Chin. J. Mech. Eng. 35, 120 (2022).
6. Nie Z, Farzaneh H. Real-time dynamic predictive cruise control for enhancing eco-driving of electric vehicles, considering traffic constraints and signal phase and timing (SPaT) information, using artificial-neural-network-based energy consumption model. Energy. 2022;241:122888.

7. Qiu S, Qiu L, Qian L, Pisu P. Hierarchical energy management control strategies for connected hybrid electric vehicles considering efficiencies feedback. Simulation Modelling Practice and Theory. 2019;90:1-15.

8. Ma F, Yang Y, Wang J, Li X, Wu G, Zhao Y, et al. Eco-driving-based cooperative adaptive cruise control of connected vehicles platoon at signalized intersections. Transportation Research Part D: Transport and Environment. 2021;92:102746.

9. Zheng Y, Li SE, Li K, Borrelli F, Hedrick JK. Distributed Model Predictive Control for Heterogeneous Vehicle Platoons Under Unidirectional Topologies. IEEE Transactions on Control Systems Technology. 2017;25(3):899-910.

10. Bian Y, Du C, Hu M, Li SE, Liu H, Li C. Fuel Economy Optimization for Platooning Vehicle Swarms via Distributed Economic Model Predictive Control. IEEE Transactions on Automation Science and Engineering. 2021:1-13.

11. Pi D, Xue P, Xie B, Wang H, Tang X, Hu X. A Platoon Control Method Based on DMPC for Connected Energy-Saving Electric Vehicles. IEEE Transactions on Transportation Electrification. 2022;8(3):3219-35.

12. Ali AM, Moulik B. On the Role of Intelligent Power Management Strategies for Electrified Vehicles: A Review of Predictive and Cognitive Methods. IEEE Transactions on Transportation Electrification. 2022;8(1):368-83.

13. Jamali H, Wang Y, Yang Y, Habibi S, Emadi A, editors. Rule-Based Energy Management Strategy for a Power-Split Hybrid Electric Vehicle with LSTM Network Prediction Model. 2021 IEEE Energy Conversion Congress and Exposition (ECCE); 2021 10-14 Oct. 2021.

14. Zhang F, Hu X, Liu T, Xu K, Duan Z, Pang H. Computationally Efficient Energy Management for Hybrid Electric Vehicles Using Model Predictive Control and Vehicle-to-Vehicle

Communication. IEEE Transactions on Vehicular Technology. 2021;70(1):237-50.

15. Xiong S, Zhang Y, Wu C, Chen Z, Peng J, Zhang M. Energy management strategy of intelligent plug-in split hybrid electric vehicle based on deep reinforcement learning with optimized path planning algorithm. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering. 2021;235(14):3287-98.

16. X. Tang, J. Chen, K. Yang, M. Toyoda, T. Liu and X. Hu, "Visual Detection and Deep Reinforcement Learning-Based Car Following and Energy Management for Hybrid Electric Vehicles," in IEEE Transactions on Transportation Electrification, vol. 8, no. 2, pp. 2501-2515, June 2022.

17. Du G, Zou Y, Zhang X, Kong Z, Wu J, He D. Intelligent energy management for hybrid electric tracked vehicles using online reinforcement learning. Applied Energy. 2019;251:113388.

18. Bo L, Han L, Xiang C, Liu H, Ma T. A real-time energy management strategy for off-road hybrid electric vehicles based on the expected SARSA. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering. 2022:09544070221079173.

19. Lian R, Peng J, Wu Y, Tan H, Zhang H. Rule-interposing deep reinforcement learning based energy management strategy for power-split hybrid electric vehicle. Energy. 2020;197:117297.

20. Tang X, Chen J, Liu T, Qin Y, Cao D. Distributed Deep Reinforcement Learning-Based Energy and Emission Management Strategy for Hybrid Electric Vehicles. IEEE Transactions on Vehicular Technology. 2021;70(10):9922-34.

21. Zhou J, Xue Y, Xu D, Li C, Zhao W. Self-learning energy management strategy for hybrid electric vehicle via curiosity-inspired asynchronous deep reinforcement learning. Energy. 2022;242:122548.

22. Lian R, Tan H, Peng J, Li Q, Wu Y. Cross-Type Transfer for Deep Reinforcement Learning Based Hybrid Electric Vehicle Energy Management. IEEE Transactions on Vehicular Technology. 2020;69(8):8367-80.

23. He H, Wang Y, Li J, Dou J, Lian R, Li Y. An Improved Energy Management Strategy for Hybrid Electric Vehicles Integrating Multistates of Vehicle-Traffic Information. IEEE Transactions on Transportation Electrification. 2021;7(3):1161-72.

24. Wang K, Yang R, Huang W, Mo J, Zhang S. Deep reinforcement learning-based energy management strategies for energy-efficient driving of hybrid electric buses. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering. 2022:09544070221103392.

25. Zhang K, Yang Z, Başar T. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In: Vamvoudakis KG, Wan Y, Lewis FL, Cansever D, editors. Handbook of Reinforcement Learning and Control. Cham: Springer International Publishing; 2021. p. 321-84.

26. Wu Y, Tan H, Peng J, Zhang H, He H. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus. Applied Energy. 2019;247:454-66.

27. Huang B, Liu X, Wang S, Pan L, Chang V. Multi-agent reinforcement learning for cost-aware collaborative task execution in energy-harvesting D2D networks. Computer Networks. 2021;195:108176.

28. Balador A, Bazzi A, Hernandez-Jayo U, de la Iglesia I, Ahmadvand H. A survey on vehicular communication for cooperative truck platooning application. Vehicular Communications. 2022;35:100460.

29. Lowe R, Wu YI, Tamar A, Harb J, Pieter Abbeel O, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems. 2017;30.

30. Li J, Yu T, Zhang X. Coordinated load frequency control of multi-area integrated energy system using multi-agent deep reinforcement learning. Applied Energy. 2022;306:117900.

31. Xu N, Kong Y, Yan J, Zhang Y, Sui Y, Ju H, et al. Global optimization energy management for multi-energy source vehicles based on “Information layer - Physical layer - Energy layer - Dynamic programming” (IPE-DP). Applied Energy. 2022;312:118668.