An enhanced eco-driving strategy based on reinforcement learning for connected electric vehicles: cooperative velocity and lane-changing control

Haitao Ding, Wei Li, Nan Xu and Jianwei Zhang
State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun, China

Abstract
Purpose – This study aims to propose an enhanced eco-driving strategy based on reinforcement learning (RL) to alleviate the mileage anxiety of electric vehicles (EVs) in the connected environment.

Design/methodology/approach – In this paper, an enhanced eco-driving control strategy based on an advanced RL algorithm in hybrid action space (EEDC-HRL) is proposed for connected EVs. The EEDC-HRL simultaneously controls longitudinal velocity and lateral lane-changing maneuvers to achieve more potential eco-driving. Moreover, this study redesigns an all-purpose and efficient-training reward function with the aim to achieve energy-saving on the premise of ensuring other driving performance.

Findings – To illustrate the performance for the EEDC-HRL, the controlled EV was trained and tested in various traffic flow states. The experimental results demonstrate that the proposed technique can effectively improve energy efficiency, without sacrificing travel efficiency, comfort, safety and lane-changing performance in different traffic flow states.

Originality/value – In light of the aforementioned discussion, the contributions of this paper are two-fold. An enhanced eco-driving strategy based an advanced RL algorithm in hybrid action space (EEDC-HRL) is proposed to jointly optimize longitudinal velocity and lateral lane-changing for connected EVs. A full-scale reward function consisting of multiple sub-rewards with a safety control constraint is redesigned to achieve eco-driving while ensuring other driving performance.

Keywords Ecological driving, Electric vehicles, Reinforcement learning in hybrid action space, Velocity and lane-changing control, Reward function

Paper type Research paper

1. Introduction
There is no doubt that electric vehicles (EVs) have been booming in recent years due to their environment-friendly characteristic and higher energy efficiency (Li et al., 2019; Deng et al., 2020). Nevertheless, additional challenges arise coupled with inherent advantages of green mobility. Notably, the range and charging are enormous practical problems for EVs: its range is shorter and the charging time is longer than traditional internal combustion engine vehicles which restricts the popularity of EVs (He et al., 2018). Nevertheless, on the bright side, the research studies on eco-driving of EVs have brought a great promise to solve this problem (Hardman et al., 2018; He and Wu, 2018; Afshar et al., 2021; Tran et al., 2021), especially under the condition of connected vehicles (Vahidi and Sciarretta, 2018; Olovsson et al., 2022; Shi et al., 2021).

1.1 Literature review
1.1.1 Eco-driving control strategies
A substantial amount of existing literature concentrated on the eco-driving of EVs by longitudinal control. Bertoni et al. (2017) proposed an algorithm of energy-saving cooperative adaptive cruise control, which uses a trajectory preview from the preceding vehicle and reduces inter-vehicular distance and smooths speed profile to minimize the energy consumption of autonomous EVs. Guo et al. (2021) designed traction control and brake control to track the desired acceleration and deceleration so as to accurately obtain the desired longitudinal motion and maximize the braking energy recovery. Kang et al. (2017) put forward a velocity control,
optimization system considering traffic lights, which prompts EVs to pass the green light immediately without delay. Similarly, Yu et al. (2019) presented a consensus and optimal speed advisory model for connected vehicle platoon at an isolated signalized intersection to enhance energy-efficiency and safety performance of the mixed traffic. The system can significantly cut down energy consumption without increasing travel time. Furthermore, Dong et al. (2021), based on the prediction of the queue ahead, proposed a hierarchical eco-approach control strategy. This method proves that even under the influence of the front queue, it can still achieve good energy-saving effect. On the other hand, a part of scholars are committed to achieving ecological functions through reasonable lateral movement. Xu et al. (2018) adopted a multi-layer control method to simultaneously optimize lane-changing stability and reduce energy consumption. The results show that the control system has good stability and energy saving, but there exists a query about real-time performance of the control system because of the complicated control algorithm and the huge amount of calculation. Tajedddin et al. (2019) developed a multi-lane adaptive cruise controller which computes instantaneous trip cost for each lane and selects the lane of lowest cost. As they used the exhaustive method to calculate the cost of each possible driving trajectory, the real-time performance is also of concern. Chen et al. (2020) jointly optimized lane-changing time and energy consumption of all cooperating vehicles by optimal control methods. Despite they have adopted some methods to speed up the solution, the amount of calculation is still a problem.

We find that most eco-driving research studies focus on either longitudinal or lateral control. This is not reasonable in real driving mode because both longitudinal velocity and lateral lane-changing are indispensable parts of driving. One obvious reason is that the controller composed of acceleration and lane change is pretty complex and easy to produce a huge computational burden, especially for lane change control.

1.1.2 Reinforcement learning

In recent years, reinforcement learning (RL) has shown its great advantages in computing speed and dealing with complex scenario tasks in the field of autonomous driving. Zhu et al. (2020), Li et al. (2021), Du et al. (2022), Wang et al. (2022) have confirmed that RL runs much faster than model predictive control (even more than 200 times faster), which holds a great promise for real-time implementations. Kendall et al. (2019) demonstrated the first application of RL to a full-sized autonomous driving in real-world driving experiments. In their study, a single monocular image was used as input of model to learn a policy for lane following in a handful of training episodes by a continuous RL algorithm. The experimental results illustrate that the RL algorithm can learn lane following with under only 30 min of training, which reveals the great potential of RL in practical application. Qu et al. (2020) reduced the electric energy consumption and improved transportation efficiency by dampen traffic oscillations using RL. Rezaee et al. (2019) proposed a hierarchical RL framework with a novel state-action space abstraction to control the vehicle to maintain the desired speed and ensure safety, which allows the trained model to be transferred from a simulation environment without dynamics to an environment with more real dynamics. Krasowski et al. (2020) extended RL with a safety layer which limits the action space into the sub-action of safe actions to address the safety problem of autonomous vehicles. Ye et al. (2020) used proximal policy optimization (PPO) to realize automatic lane-changing strategy. The results show that this method can learn and execute lane-changing actions safely, stably and efficiently. Owing to the aforementioned recent research progress, the great power of RL has been authenticated.

Additionally, in the process of driving, as the lane-changing duration is trivial relative to the whole travel time, the continuous lane-changing process will not make a substantial impact on the energy consumption of whole travel process. Therefore, the lane changing can be regarded as an instantaneous discrete action of lane selection, whereas acceleration is still implemented in continuous action. Accordingly, the whole action tuple of driving process is equivalent to a mixture of discrete and continuous actions. On the other hand, the RL algorithms based on hybrid action space have made remarkable progress to tackle the discrete–continuous hybrid action scenarios recently (Xiong et al., 2018; Fan et al., 2019). Thus, the RL methods in hybrid action space can effectively resolve the continuous acceleration and discrete lane-changing problem. It should be noted that although Guo et al. (2021) combined Deep Deterministic Policy Gradient (DDPG) for continuous action space and Deep Q Network (DQN) for discrete action space to control the longitudinal velocity and lateral lane-changing decision, the inherent mechanism of the simple integration is not explicit, and it may result in local optimum because the movements in two dimensions are not optimize jointly. Bai et al. (2022) also considered both longitudinal velocity and lateral lane-changing, but the application of DQN will cause drastic changes in acceleration and cannot guarantee global optimization. Differing from the aforementioned methods, this paper proposes an eco-driving model based on an advance RL algorithm in hybrid action space can jointly optimize longitudinal velocity and lateral lane-changing and yield to better performance.

1.2 Contribution

In light of the aforementioned literature review, our aim is three-fold. First, this paper is expected to find a reasonable solution to the problem on cooperative velocity and lane-changing control for eco-driving. Second, as many studies focus on promoting the target functions, while ignoring other driving characteristics, we are prone to redesign an full-scale reward function to meet the above requirements and ensure efficient training. In the field of eco-driving, to our best knowledge, our research is the first to comprehensively consider other driving performance while improving energy efficiency. Therefore, the major contribution and novelty of this paper lie within follows:

- An enhanced eco-driving strategy based on an advanced RL algorithm in hybrid action space (EEDC-HRL) is proposed to jointly optimize longitudinal velocity and lateral lane-changing for connected EVs.
- A full-scale reward function consisting of multiple sub-rewards with a safety control constraint is redesigned to achieve eco-driving and raise other performance to diverse extent including travel efficiency, comfort, safety, rational lane-changing maneuvers, while ensuring training efficiency.
The remainder of this paper is structured as follows. Section 2 presents the problem formulation, and Section 3 describes the eco-driving framework based on RL. In Section 4, a train of relevant experiments are carried out to estimate the performance of the proposed framework. The results are demonstrated in Section 5 together with the analysis, whereas Section 6 concludes this study along with the ideas for future work.

2. Problem formulation

This study aims to perform an eco-driving strategy based RL through cooperative velocity and lane-changing control in four-lane urban highway. Substantially, this theme is an optimal policy learning task for multiple objectives focusing on energy efficiency. There exist three key elements for the task:

1. energy consumption model;
2. state space and action space; and
3. full-scale and efficient-training reward function.

The above key elements are clearly introduced in this section, whereas the specific methodology and the establishment of system model will be discussed in next section.

2.1 Energy consumption model

Most energy consumption models (ECMs) of EVs are established by its powertrain model such as Vaz et al. (2015); Xu et al. (2019). However, for the purpose of ecological driving control, these ECMs are too sophisticated to be integrated into the reward function of RL. In this study, a simple and accurate energy consumption model (Galvin, 2017) for the particular vehicle derived from mathematical and engineering experience is applied to overcome the above problem. It only maps the combination of velocity and acceleration, which can be easily integrated into the reward function of RL.

Specifically, a Mitsubishi electric vehicle is used as the ego vehicle in this study, which was conducted the dynamometer test with up to 36 different test cycles to obtain the corresponding data of speed, acceleration and battery power demand. And the data were carried out multivariate regression analyses to match the best formula between the three variables. The adjusted value generated from the multivariate regression gains the highest (0.9703) in the form of equation (1), coined the expression of the energy consumption model in this paper:

\[ P = 1281VA + 840.4V - 55.312V^2 + 1.67V^3 \]  

Where \( P \) indicates power, \( V \) indicates velocity, \( A \) denotes acceleration (\( A \) takes negative values when the electric vehicle decelerates). As all experiments take into account the energy recovered by braking, the aforementioned equation also triggers the braking energy recovery.

2.2 State space and action space

State space stores the state information that the agent of RL interact with the environment. Note that in the connected conditions, the relevant information of surrounding vehicles can be directly collected by connected vehicle technologies (e.g. vehicle-to-vehicle). To explore ideal eco-driving through longitudinal velocity and lateral lane-changing movements, the state space should provide sufficient information for the agent learning; thus, it is necessary to consider the pertinent states containing the ego vehicle and all surrounding vehicles among the four lanes, as following: the velocity of the ego vehicle, the acceleration of the ego vehicle, the relative velocity between the ego vehicle and all leader vehicles among the four lanes, the relative distance between the ego vehicle and all leader vehicles among the four lanes, the relative velocity between the ego vehicle and all following vehicles among the four lanes, the relative distance between the ego vehicle and all following vehicles among the four lanes; 18 state variables in total.

Action space contains all possible action that can be executed by the agent in the environment. Especially, the object of this study is the hybrid action space consisting of both continuous acceleration and discrete lane-changing. The continuous acceleration is bounded by the maximum acceleration and deceleration (3 and \(-3\) m/s\(^2\), respectively, in this study), whereas discrete lane-changing includes three variables \((-1, 0, 1)\), in which \(-1\) represents turning right, \(0\) indicates keeping the current lane and \(1\) denotes turning left.

2.3 Reward function with a safety constraint

To hold a full-scale and efficient-training reward function, the trade-off is to redesign multiple corresponding sub-rewards with a safety control constraint.

2.3.1 Multiple sub-reward functions

- Economy: one of the indicators reflecting the economy of EVs is the electric energy consumed per kilometer. We notice from equation (1) that:

\[ \frac{E}{S} = \frac{P}{S} \cdot \frac{T}{V} = \frac{P}{V} = \frac{1281A + 840.4 - 55.312V + 1.67V^2}{V} \]  

Where:

- \( E \) = energy;
- \( S \) = distance; and
- \( T \) = time.

The derivation of the aforementioned equation (2) just meets the requirements of this study. Moreover, the visualization of equation (2) is depicted in Figure 1, which illustrates that reasonable speed and acceleration can achieve considerable energy-saving effect.

To make the energy consumption value as small as possible, we set the sub-reward function to the reciprocal of the above formula:

\[ r_1 = \frac{1}{1281A + 840.4 - 55.312V + 1.67V^2} \]  

- Travel efficiency: velocity, a simple and effective indicator, can be neatly used to represent the transport efficiency (He et al., 2020; Jan et al., 2020).

\[ r_2 = V \]
• Comfort: jerk, the change rate of acceleration, is regarded as a general indicator to evaluate comfort (Guo et al., 2021; Ye et al., 2020).

\[
r_3 = -\text{jerk}^2 = -\left(\frac{a(t) - a(t-1)}{\Delta t}\right)^2
\]

where \(a(t)\) denotes the acceleration in time \(t\), \(a(t-1)\) is the acceleration in time \(t-1\), and \(\Delta t\) indicates the sample interval.

• Lane-changing performance: to avoid aggressive lane-changing maneuvers, which would disturb traffic operation seriously (Park et al., 2019), it is essential to add a penalty discount when lane-changing maneuvers trigger:

\[
P' = -\frac{V}{t_\Delta}
\]

where \(t_\Delta\) indicates the elapsed time from the last lane-changing maneuver.

Thus, the multi-modal reward \(R\) is defined as:

\[
R = v_1r_1 + v_2r_2 + v_3r_3 + v_4P'
\]

where \(v_1, v_2, v_3, v_4\) describes the weight factors for respective sub-reward function.

2.3.2 Safety control constraint

Though safety is the first priority of autonomous driving, it is an enormous problem how to achieve safe driving simultaneously considering energy efficiency, comfort, travel efficiency and lane-changing performance. We have ever attempted to regard safety as a sub-reward function, but it tended to cause mutual interference with other sub-reward functions, and it is still impossible to completely avoid collision even after convergence, as a result of the soft constraint nature of reward function. On the worse side, it engenders an adverse impact on the convergence and speed of training, i.e. training efficiency. To tackle the problem and reach the full-scale and efficient-training goal, we ultimately treat the safety as control hard constraint as a previous study (Zhu et al., 2020): once the relative distance between the ego vehicle and the lead vehicle is less than the safe distance, the controlled vehicle will brake in the maximum deceleration without being controlled by the EEDC-HRL, that is:

\[
a(t) = \begin{cases} 
-3m/s^2 & d < d_{safe} \\
\text{RL model output} & \text{otherwise}
\end{cases}
\]

where, \(a(t)\) is the acceleration of the ego vehicle, \(d\) denotes the relative distance.

To determine the safe distance, a classic stop distance model (Wilson et al., 1997) was used:

\[
d_{safe} = \frac{v_\epsilon T'^2}{2a_{max}} - \frac{v_\epsilon^2}{2a_{max}}
\]

where \(T'\) stands for the driver’s reaction time (defined as 1 s in this paper), \(a_{max}\) denotes the maximum absolute deceleration value, \(v_\epsilon\) and \(v_l\) are the speed of the ego vehicle and the lead vehicle respectively.

3. Eco-driving framework based on reinforcement learning

In this section, an eco-driving framework based on RL in hybrid action space is established to handle the optimal policy learning task raised in Section 2.

3.1 Reinforcement learning in hybrid action space

Classic RL algorithms only can deploy either discrete action space or continuous action space, such as DQN (Mnih et al., 2013) for discrete action space, DDPG (Silver et al., 2014) for continuous action space and PPO (Schulman et al., 2017) for continuous or...
discrete action space. In recent years, RL algorithms based on hybrid action space have been proposed to deal with the scenario with discrete-continuous hybrid action space. To be specific, DDPG with a parametrized action space (PA-DDPG) allows DDPG to simultaneously output continuous and discrete actions (Hausknecht and Stone, 2015). Xiong et al. (2018) proposed a parametrized deep Q-network (P-DQN) framework through combining DQN and DDPG seamlessly. Based on PPO, Fan et al. (2019) creatively designed a hybrid architecture of actor-critic algorithm for hybrid action space (named H-PPO) because PPO is capable of learning stochastic policies in continuous action spaces or discrete action spaces. H-PPO is composed of multiple parallel sub-actor networks, which decompose the structured action space into simpler action spaces, along with a critic network as an estimator of the state-value function $V(s)$ to guide the training of all sub-actor networks. The multiple parallel sub-actor networks are divided into discrete actor network and continuous actor network, in which discrete actor networks learn stochastic policies $\pi_{\theta_d}$ to perform discrete action and continuous actor networks learn stochastic policies $\pi_{\theta_c}$ to perform continuous action. All actor networks share the first few layers to encode the state information and updates stochastic policies with the advantage function provided by the critic network. Different from PPO learn general stochastic policy $\pi_{\theta}$ by equation (10), H-PPO’s discrete policy $\pi_{\theta_d}$ and continuous policy $\pi_{\theta_c}$ are updated separately by minimizing their respective clipped surrogate objective, i.e. equations (11) and (12). Moreover, H-PPO is experimentally analyzed in four environments containing hybrid action space and compared with PA-DDPG and P-DQN. The experimental results show that H-PPO exhibits better stability, higher convergence values and lower variance than the other two algorithms in these environments [refer (Fan et al., 2019) for the specific process]. Thus, the H-PPO is adopted as the specific algorithm model of RL in this study. Algorithm 1 describes the pseudocode of the H-PPO implementation:

$$L_{CLIP}^\theta = \hat{E}_i \left[ \min \left( r_i(\theta) A_i, \text{clip}(r_i(\theta, 1 - 1 + )A_i) \right) \right]$$ (10)

$$L_2^\text{CLIP} (\theta_d) = \hat{E}_i \left[ \min \left( r^{\text{c}}(\theta_d) A_i, \text{clip}(r^{\text{c}}(\theta_d, 1 - 1 + )A_i) \right) \right]$$ (11)

$$L_4^\text{CLIP} (\theta_c) = \hat{E}_i \left[ \min \left( r^{\text{c}}(\theta_c) A_i, \text{clip}(r^{\text{c}}(\theta_c, 1 - 1 + )A_i) \right) \right]$$ (12)

Algorithm 1 Hybrid Proximal policy optimization (H-PPO)
1: initialize discrete policy parameters $\theta_{d_k}$, continuous policy parameters $\theta_c$, sequentially from internal nodes to leaf nodes and value function parameters $\phi$;
2: for each $k \in [0, n]$ do
3: Collect set of trajectories $D_k=\{\tau_i\}$ by running discrete policy $\pi_{\theta_d}$ and continuous policy $\pi_{\theta_c}$ in the environment;
4: Compute rewards-to-go $\hat{R}_i$;
5: Compute advantage estimates $\hat{A}_i$ (using any method of advantage estimation) based on the current value function $V_{\phi k}$;
6: Update the policies by maximizing the two kinds of clipped surrogate objectives in the previous sequence:

$$\theta_{d_{k+1}} = \arg \max_{\theta_d} \frac{1}{|D_k|T} \sum_{t \in D_k} \sum_{i=0}^T \min \left( \frac{\pi_{\theta_d}(a_i|s_i)}{\pi_{\theta_d}(a_i|s_i)} A_i, \text{clip}(r_i(\theta_d), 1 - \varepsilon, 1 + \varepsilon)A_i \right)$$

7: Fit value function by regression on mean-squared error:
$$\phi_{k+1} = \arg \min_{\theta} \frac{1}{|D_k|T} \sum_{t \in D_k} \sum_{i=0}^T \left( V_{\phi}(s_i) - \hat{R}_i \right)^2$$
8: end for

3.2 System architecture
The system architecture based on H-PPO is composed of two components as shown in Figure 2: a RL model and a traffic environment which is simulated in SUMO, a realistic urban traffic simulation software (Lopez et al., 2018). These two independent parts interact through TraCI, a bridge between SUMO and other control algorithms (Wegener et al., 2008). Specifically, the state information described in Section 2.2 from traffic simulation environment is respectively imported to the state encoding network and the critic network of RL model. The two actor networks share the state encoding network and generate corresponding stochastic continuous policy or stochastic discrete policy to output acceleration or lane-changing actions, whereas the critic network estimates these policies by state-value function. In addition, the reward represented in Section 2.3, is used to update all network parameters to maximize the return.

3.3 Neural network and hyperparameters
There possess two kinds of neural networks in the system: the discrete and continuous actor networks are designated for policy generation, alone with the single critic network for policy improvement. For the critic network, the input layer is the 18 state variables neurons, and the output layer is the state-value function. For the two actor networks, the input layer are both the 18 state variables neurons. The output layer of the continuous actor network is the mean and variable from Gaussian distribution, with a tanh activation function, which can map numerical values to the range $[-1, 1]$; thus, it enables bound the outputted accelerations between $-3$ and $3$ m/s² by multiplying three, whereas the output layer of discrete actor applies the softmax distribution to select discrete values for lane-changing. For the hidden layers, two-layer fully connected neural network with 1,024 neurons is adopted for all neural networks. Other number of nodes and layer had been tested. The result shows the more nodes under the same number of layers, the better the performance. However, when the number of nodes exceeds 1,024, the performance improvement is not obvious and running speed would be slower. When three hidden layers are selected, the performance is not improved, instead the running speed is affected. Thus, the chosen hidden layer architecture is sufficiently matched for our problem. Moreover, the Rectified Linear Unit activation function is used in the hidden layers, which contributes to facilitate the convergence of network parameter optimization (Krizhevsky et al., 2012).

Experimentally chosen hyperparameters of the H-PPO are listed in Table 1. Including hidden layers, the hyperparameters in the study were assigned by grid-search and “trial and error”
approach. We conclude that the H-PPO model is not sensitive to a substantial share of hyperparameters, which is consistent with the characteristic of the PPO (Schulman et al., 2017), except for the learning rate, where too large or too small values can cause performance degradation. When the learning rate is chosen as 0.005, the performance reaches a sweet spot.

4. Numerical experiments

In this section, a package of experiments are conducted for the above system architecture to widely evaluate the performance of the proposed control model. Specifically, we engage in the model training and corresponding testing design of RL model in three different traffic states which are generated by SUMO and introduce two comparative models in order to validate the performance of the EEDC-HRL model.

4.1 Setting of different traffic states

First, the scientific establishment of different traffic flow states is a basis for comprehensive and synthetic performance evaluation of the eco-driving system. The traffic flow fundamental diagram has been regarded as the foundation of traffic flow theory, which addresses the relationship among three fundamental parameters of traffic flow: traffic flow (vehs/h), speed (km/h), and traffic density (vehs/km). As Greenshields proposed the seminal Greenshields model (Greenshields et al., 1935), extensive traffic flow models had been followed up, which were systematically summarized by Qu et al. (2017). Here, the Greenberg model (Greenberg, 1959), one of the most classic traffic flow models is adopted in this paper, as it demonstrates the relationship between traffic flow and traffic density ratio:

\[ q = ck \ln \left( \frac{k_j}{k} \right) \]  
(13)

where \( q \) indicates the traffic flow, \( k \) presents the traffic density, \( c \) is a constant and \( k_j \) denotes the jam density when vehicle’s speed is 0. The value of \( k_j \) can be computed by:

\[ k_j = \frac{1}{l_{veh} + h \cdot n_{lane}} \]  
(14)

where \( l_{veh} \) is vehicle’s average length, \( h \) denotes the minimum headway when vehicles stop and \( n_{lane} \) denotes the number of lanes.

![Figure 2 Eco-driving system architecture based on RL in hybrid action space](image-url)
It is not laborious to deduce that when \( \frac{k}{k_j} = \frac{1}{\lambda} \), the traffic flow reaches its maximum value \( q_{m} \) which represents the most efficient operation point of transportation. Let \( k_m \) be the corresponding traffic density when the traffic flow reaches \( q_{m} \) as shown in Figure 3. Based on the aforementioned definition, we set \( k_1 = 0.2k_m \), \( k_2 = 0.8k_m \) and \( k_3 = 2k_m \) as the free traffic flow state, normal traffic flow state and congested traffic flow state, respectively. The relevant parameters used to calculate traffic flow can be found in Table 2.

### 4.2 Model training and corresponding testing design

Regarding training and testing phase, the 2,070 random traffic events for above each traffic flow state were grouped into a training data set and a held-out testing data set. The traffic events were randomly executed from 1,380 episodes during the training process, and the testing was repeated for other 690 episodes. An episode means a traffic events in this study. Through multiple running, the average reward values of the three traffic states with respect to the training episodes are depicted in Figure 4, where the bold line and the shaded areas represent the mean and the standard deviation, respectively.

As shown in Figure 4, the reward values of the three curves gradually ascend and converge, which proves the training of the EEDC-HRL model has achieved considerable effects in different environments. It is worth mentioning that, the difference of the reward values illustrates the difference of comprehensive performance in the three traffic flow states.

### 4.3 Comparative model

To quantify the effect of the proposed model, we introduced a classical car-following model: intelligent driver model (IDM) (Treiber et al., 2000) as baseline and a state-of-art energy-efficient electric driving model (E3DM) (Lu et al., 2019) as benchmark to compare the eco-driving performance. The function of IDM is that it has a safe driving mechanism to prevent vehicle collision (Li et al., 2021), whereas E3DM can generate smooth acceleration and efficient regenerative braking through adjusting its speed-dependent spacing to result in energy-efficient driving. Particularly, the lane-changing movements of the two comparative models is controlled by the rule-based lane-changing model of SUMO (Erdmann, 2015), which works to realistically implement usual function of lane-changing, such as obtaining higher speed, departing the dead lane, turning to the target lane and so on [refer to (dos Santos and Wolf, 2019; Dong et al., 2021; Silgu et al., 2021)for more details]. These methods are summarized as Table 3.

### 5. Results and analysis

After the establishment and training of the model above, the experimental results and the analyses based on testing data are presented in this section. The IDM and E3DM would be used for comparison to demonstrate various performance.

---

**Table 2** Relevant parameters used to calculate traffic flow

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l_{veh} )</td>
<td>5</td>
<td>Vehicle’s average length</td>
<td>m</td>
</tr>
<tr>
<td>( h )</td>
<td>3</td>
<td>Minimum headway</td>
<td>m</td>
</tr>
<tr>
<td>( k_j = \frac{1}{l_{veh} + h} n_{lane} )</td>
<td>500</td>
<td>The jam density</td>
<td>vehs/km</td>
</tr>
<tr>
<td>( k_m )</td>
<td>183.94</td>
<td>The density when the traffic flow reaches its maximum value</td>
<td>vehs/km</td>
</tr>
<tr>
<td>( k_1 = 0.2k_m )</td>
<td>36.79</td>
<td>The density as free traffic state</td>
<td>vehs/km</td>
</tr>
<tr>
<td>( k_2 = 0.7k_m )</td>
<td>128.76</td>
<td>The density as normal traffic state</td>
<td>vehs/km</td>
</tr>
<tr>
<td>( k_3 = 2k_m )</td>
<td>367.88</td>
<td>The density as congested traffic state</td>
<td>vehs/km</td>
</tr>
</tbody>
</table>
5.1 Evaluating metrics of relevant performance

First, the evaluating metrics of performance need to be defined. The economy of EVs is measured by energy consumption of each kilometer after the vehicle drive through the whole journey. The access of the energy consumption of each moment is based on the equation (1), and the whole energy consumption is computed by:

\[ E = \sum_{i=1}^{N} P_i \cdot \Delta t \]  

(15)

where \( E \) is energy consumption, \( P_i \) is the power of \( i \) sampling moment, \( \Delta t \) is the sampling interval, \( N \) is the final sampling time. Then energy consumption of each kilometer is calculated by:

\[ \frac{E}{S} = \frac{\sum_{i=1}^{N} P_i \cdot \Delta t}{S} \]  

(16)

In addition, travel efficiency and comfort are evaluated by average speed and the absolute value of jerk, respectively, and headway is used to illustrate safety, which are consistent with the evaluation methods of other related studies (Li et al., 2021; Du et al., 2022; Srisomboon and Lee, 2021).

5.2 Experimental results

Figure 5 presents the average values of each metric based on all tested data, and samples of the same number from IDM and E3DM are also counted in Figure 5 for comparison. Additionally, the percentage increase of EEDC-HRL and E3DM for each performance compared with IDM can be seen in Figure 6.

Figure 5(a) exhibits the economy of the three models, where the energy consumption of EEDC-HRL, E3DM and IDM are 12.15–12.58–13.41 kWh/100 km, 11.61–11.99–12.50 kWh/100 km, 16.67–16.84–17.21 kWh/100 km in three traffic flow states, respectively. Thus, it indicates that the EEDC-HRL model generates better energy-saving potential. In terms of travel efficiency, EEDC-HRL is on a par with IDM, whereas E3DM exhibits slower than IDM as shown in Figure 5(b). Besides, as for safety, there is almost no difference for average headway between EEDC-HRL, E3DM, but they are higher than IDM as shown in Figure 5(c). The smallest jerk of EEDC-HRL corresponding to the best comfort is illustrated in Figure 5(d).

It can be summarized from Figure 6 that compared with IDM, the EEDC-HRL model is capable of effectively improving energy-saving potential of EVs by 9.39%, 7.10%, 3.14% in three traffic states, respectively, without sacrificing the transportation efficiency, comfort and safety. Compared with E3DM, EEDC-HRL raises energy-saving potential by 3.42%, 3.22%, 0.10% and the performance of other aspects is equal to or slightly better than that of E3DM. The detailed data distribution for EEDC-HRL and E3DM is presented in Appendix, which conduces to more subtle observation. For lane-changing performance, the average number of lane changing in different traffic states for the three methods is counted in Table 4. It is obvious that the EEDC-HRL produces less intense lane-changing frequency than other two methods to effectively avoid aggressive lane-changing maneuvers. In addition, we tested the computational efficiency of IDM, E3DM and EEDC-HRL based all traffic events of the whole testing phase. The results demonstrate that the average running time of the EEDC-HRL (0.88 s) is shorter than that of the E3DM (2.13 s) and IDM (1.95 s), which is attribute to the great advantages of RL in computing speed and dealing with complex scenario tasks.

5.3 Analysis through randomly sampling events

To understand the detailed reasons of the above results, we need to analyze them from specific testing events. Thus, one traffic
Figure 5 Average values of each metric for EEDC-HRL based on all tested data

![Graphs showing average values of metric for different traffic states: energy consumption, speed, headway, and jerk.](image)

Notes: E3DM and IDM are also presented for comparison: (a) energy consumption; (b) speed; (c) headway; (d) jerk

Figure 6 Percentage increase of EEDC-HRL compared with (a) IDM, (b) E3DM

![Percentage increase graph for different traffic states showing the improvement of EEDC-HRL.](image)

Table 4 Average number of lane changing in different traffic states for the three methods

<table>
<thead>
<tr>
<th>Traffic states</th>
<th>Free</th>
<th>Normal</th>
<th>Congested</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDM</td>
<td>1.43</td>
<td>2.85</td>
<td>0.21</td>
</tr>
<tr>
<td>E3DM</td>
<td>1.14</td>
<td>2.11</td>
<td>0.15</td>
</tr>
<tr>
<td>EEDC-HRL</td>
<td>0.86</td>
<td>1.44</td>
<td>0.11</td>
</tr>
</tbody>
</table>

event is randomly extracted from the testing events for each traffic state, respectively. Figures 7–9 shows the velocity, acceleration, jerk, energy consumption, headway and lane index in each time step of the three sampled traffic events. According to these sampled events, there exist three-fold to discuss:

1. The lane-changing performance: For the free traffic flow in Figure 7, as there is a better lane-changing moment by training and learning of the proposed objectives, the
**Figure 7** The sampled event in free traffic states

**Figure 8** The sampled event in normal traffic states
fluctuation of velocity and acceleration of EEDC-HRL perform more lower than E³DM and IDM which helps to produce better economy on the whole. The reason is that dampening erratic acceleration patterns could reduce the energy consumption of a given journey significantly (Qu et al., 2020; Galvin, 2017). Instead, the velocity and acceleration of IDM generate larger range of change because of improper lane-changing moment. For the normal traffic flow in Figure 8, more intense lane-changing actions were carried out by IDM and E³DM, which also induces more unstable oscillations in velocity and acceleration. The EEDC-HRL performs more rational lane-changing actions, so that the fluctuation of velocity and acceleration is not so violent.

2 The longitudinal acceleration performance: From the three sampling events in Figures 7–9, the velocity and acceleration of E³DM perform more smoothly than IDM, whereas EEDC-HRL still exists the overall marginal lead compared with E³DM, especially as can be seen from the Figure 9. It is noted that there never occurs any lane-changing behavior in the sampled event of congested traffic states because of traffic jam, so the whole driving process can be deemed as the car-following driving only controlled by longitudinal acceleration. Thus, the above behaviors result in better energy-efficient performance for the EEDC-HRL model.

3 The other performance: It can be clearly observed that EEDC-HRL and E³DM produce smaller jerk and larger headway, which corresponds to better comfort and safety. For travel efficiency, EEDC-HRL is almost equivalent to IDM and marginally higher than E³DM.

In summary, the reasons we analyze from randomly sampling events are as follows. Through the training and learning of the proposed goals, the vehicle controlled by EEDC-HRL model enables perform lane-changing action at a more rational moment; its lane-changing action is less intense. In other words, its lane-changing action is not so frequent; its longitudinal control performs more smoothly. Consequently, the fluctuation of velocity and acceleration is alleviated to produce better energy efficiency.

6. Conclusion

In conclusion, this paper is devoted to using a cooperative velocity and lane-changing control to achieve the purpose of eco-driving for EVs and ensure other performance. We propose an eco-driving model based on RL control in hybrid action space (EEDC-HRL) and redesign a full-scale reward function to balance economy, travel efficiency, comfort and lane-changing maneuvers, with a safety control constraint to ensure safety. To better integrate with the reward function, a simple and accurate energy consumption model is applied. Afterward, the eco-driving model is trained and tested in three traffic states: free traffic flow, normal traffic flow and congested traffic flow. The experimental results show that as its better lateral and longitudinal control performance, velocity and acceleration controlled by the EEDC-HRL model run more smoothly. It helps the electric vehicle to trigger considerable energy-efficient potential on average 9.39%, 7.10% and 3.14% in different traffic conditions.
traffic environments and exceeds E1DM. Besides, compared with two comparative methods, the other performance of EEDC-HRL is not inferior to that of them. Therefore, the proposed EEDC-HRL model is of considerable value in the field of energy-efficiency for EVs.

In the future work, we will consider more complex scenarios such as mandatory lane-changing and intersections to comprehensively verify the performance of the proposed algorithm. Moreover, we will be devoted to platoon-based eco-driving research, because platoon at small inter-vehicle distances can reduce aerodynamic drag, which brings greater energy-efficient potential. Meanwhile, considering the more complex traffic environment of urban roads, traffic lights will be added into the traffic model. Furthermore, the traffic scene for more macroscopic level will be considered, where the controlled vehicle or platoon can choose a more energy-saving trajectory.

References


Enhanced eco-driving strategy
Haitao Ding et al.


Appendix. The detailed data distribution of each performance metric from testing data

See Figure A1-A3.

Figure A1  The data distribution of each performance indicator in the free traffic flow
Figure A2  The data distribution of each performance indicator in the normal traffic flow
Figure A3  The data distribution of each performance indicator in the congested traffic flow

Corresponding author
Wei Li can be contacted at: liw19@mails.jlu.edu.cn