World Electric Vehicle Journal

MDPI

Article

TD3-Based EMS Using Action Mask and Considering Battery Aging for Hybrid Electric Dump Trucks

Jinchuan Mo [1], Rong Yang [1], Song Zhang [2], Yongjian Zhou [1] and Wei Huang [1,*]

[1] School of Mechanical Engineering, Guangxi University, Nanning 530004, China
[2] Guangxi Yuchai Machinery Company Limited, Yulin 537000, China
* Correspondence: huangwei@gxu.edu.cn

Abstract: The hybrid electric dump truck is equipped with multiple power sources, and each powertrain component is controlled by an energy management strategy (EMS) to split the demanded power. This study proposes an EMS based on deep reinforcement learning (DRL) algorithm to extend the battery life and reduced total usage cost for the vehicle, namely the twin delayed deep deterministic policy gradient (TD3) based EMS. Firstly, the vehicle model is constructed and the optimization objective function, including battery aging cost and fuel consumption cost, is designed. Secondly, the TD3-based EMS is used for continuous action control of ICE power based on vehicle state, and the action mask is applied to filter out invalid actions. Thirdly, the simulations of the EMSs are trained under the CHTC-D driving cycle and C-WTVC driving cycle. The results show that the action mask improves the convergence efficiency of the strategies, and the proposed TD3-based EMS outperforms the deep deterministic policy gradient (DDPG) based EMS. Meanwhile, the battery life is extended by 36.17% under CHTC-D and 35.49% under C-WTVC, and the total usage cost is reduced by 4.30% and 2.49% when the EMS considers battery aging. In summary, the proposed TD3-based EMS can extend the battery life and reduce usage cost, and provides a method to solve the optimization problem for the EMS of hybrid power systems.

Keywords: energy management strategy; hybrid electric dump truck; deep reinforcement learning; battery aging; action mask

Citation: Mo, J.; Yang, R.; Zhang, S.; Zhou, Y.; Huang, W. TD3-Based EMS Using Action Mask and Considering Battery Aging for Hybrid Electric Dump Trucks. *World Electr. Veh. J.* **2023**, *14*, 74. https://doi.org/10.3390/wevj14030074

Academic Editors: Joeri Van Mierlo, Danial Karimi and Amin Hajizadeh

Received: 16 December 2022
Revised: 4 February 2023
Accepted: 15 March 2023
Published: 17 March 2023

1. Introduction

Hybrid electric dump trucks significantly reduce carbon emissions and fuel consumption compared with internal combustion engine (ICE) vehicles in the transportation industry. The hybrid power system improves the operating efficiency of the ICE by an electric motor, and it solves the contradiction between long-endurance mileage and low energy usage [1]. The energy management strategy (EMS) distributes the demand power among various powertrain components, so the energy-saving capability of the hybrid electric dump truck is further enhanced [2].

An excellent EMS plays an influential role in reducing the fuel consumption of hybrid electric vehicles (HEV). Researchers have proposed a large number of EMSs which can be classified into three categories: rule-based EMSs, optimization-based EMSs, and learning-based EMSs [3]. Rule-based EMSs have been applied to real-time control in the automotive industry to make HEVs more fuel efficient than ICE vehicles. However, the EMSs require engineering experience to calibrate them and are poorly adapted to different driving cycles [4]. Consequently, the optimization-based EMSs use optimization algorithms to achieve optimal control of HEVs has been substantially encouraged, which includes dynamic programming (DP) [5], equivalent consumption minimization strategy (ECMS) [6], model predictive control (MPC) [7] and so on. DP-based EMS searches for the global optimal fuel consumption of the vehicles by knowing the velocity of the driving cycle in advance, so it can only be solved offline and cannot be applied to real-time control.

Moreover, the DP-based EMS is commonly used as a benchmark for other EMSs [8]. ECMS and MPC are real-time optimization-based EMSs. However, the computation of ECMS and MPC grows exponentially with constraints, model accuracy, and additional issues [9]. The thriving of artificial intelligence has caught the attention of researchers who have proposed learning-based EMSs for hybrid electric vehicles [10–13]. Among them, deep reinforcement learning (DRL) incorporates neural networks, so the DRL has the perceptual capability of deep learning to receive continuous states and has the decision capability of reinforcement learning (RL) [14–17]. Moreover, numerous DRL-based EMSs are proposed to achieve continuous action control to eliminate discrete errors [18–21]. A large number of studies on EMS have been conducted by numerous researchers and a brief review of previous studies on EMS is given in Table 1.

Table 1. A brief review of previous studies on EMS.

Ref.	Author	Year	Categories	Continuous		Main Topic
				State	Action	
[4]	Padmarajan et al.	2016	Rule-based			System structure and strategy
[5]	Zhou et al.	2018	Optimization-based			Improvement of DP-based EMS for different HEVs
[7]	East et al.	2022	Optimization-based			Scenario MPC for data-based EMS
[13]	Liu et al.	2015	Learning-based			Reinforcement learning of adaptive EMS
[15]	Wu et al.	2018	Learning-based		x	Continuous RL-based EMS
[16]	Han et al.	2019	Learning-based		x	DDQL-based EMS avoids falling into policy value overestimation
[17]	Li et al.	2019	Learning-based		x	EMS with terrain information
[18]	Tan et al.	2019	Learning-based	x	x	Continuous state and action spaces
[19]	Wu et al.	2019	Learning-based	x	x	Continuous control and traffic information
[21]	Li et al.	2022	Learning-based	x	x	SAC-AET-based EMS to improve the control effects

Table 1 presents a summary of previous research work. Most EMSs only optimize fuel economy and focus on discovering the best strategy to minimize fuel consumption. This study takes a proactive approach to extend battery life and reduce total usage cost from an energy management optimization perspective, proposes a DRL-based EMS for power-split hybrid electric dump trucks, namely the twin delayed deep deterministic policy gradient algorithm (TD3), and uses the action mask technology to avoid unsafe exploration. Moreover, the main contributions of this study can be summarized as follows.

(1) A TD3-based EMS is proposed to extend the battery life and reduce the total usage cost. Because battery aging affects vehicle range, costly battery replacements are required when battery life terminates.
(2) Most of EMSs ignore safety issues during the exploration stage such as the MG1 overloading, which cause serious problems in automotive control and is unacceptable in industrial applications. Then action masks are used to eliminate invalid actions that exceed the physical limits and improve the training efficiency of the policy.
(3) The TD3 algorithm can reduce the overestimation bias of DDPG, thus the TD3 algorithm is applied as an EMS for hybrid electric dump trucks and trained by the self-learning capability of DRL. Finally, a comparison with DDPG-based EMS is presented.

(4) The reward function that includes battery aging cost and fuel consumption cost is designed to extend the battery life and reduce fuel consumption.

The rest of the study is organized as follows: In Section 2, the vehicle model is formulated and the optimization problem is presented. In Section 3, the TD3-based EMS is introduced and designed. In Section 4, some results and analysis are described. Finally, some conclusions are given in Section 5.

2. Vehicle Modeling and Optimization Problem

2.1. Vehicle Model

A 31-ton hybrid electric dump truck is used as the research object, combined with the power-split hybrid system of planetary gear (PG), as shown in Figure 1. The system consists of an ICE, battery, generator (MG1), and drive motor (MG2), and the vehicle structure parameters and the main parameters of the power and transmission system are shown in Table 2.

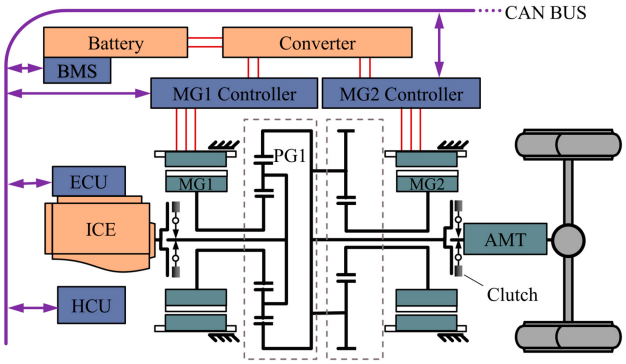

Figure 1. Power-split hybrid electric dump truck powertrain.

Table 2. The parameters of the vehicle structure and the main parameters of the power and transmission systems.

Parts	Parameter Name	Value
Vehicle	Gross weight (kg)	31,000
	Dimension (mm)	9662 × 2495 × 3450
	Dimension of cargo box (mm)	6800 × 2350 × 1500
	Drive form	8 × 4
	Drag coefficient	0.56
	Frontal area (m^2)	8.24
	Rolling resistance coefficient	$0.0041 + 0.0000256v$
ICE	Max. power (kW)	243
	Max. torque (Nm)	1400
	Max. speed (rpm)	2200
MG1	Max. power (kW)	110
	Max. torque (Nm)	340
	Max. speed (rpm)	7500
MG2	Max. power (kW)	196
	Max. torque (Nm)	375
	Max. speed (rpm)	15,000
	Gear ratio	6.7
Transmission	Transmission ratio of PG	4.4
	AMT gears ratio	6.3/2.1/1/0.86
	Final drive ratio	5.1
Battery	Capacity (Ah)	70
	Voltage (V)	576

The resistance of the vehicle is determined by rolling resistance, air resistance, slope resistance, and acceleration resistance [22]. The overall resistance for the vehicle is defined as

$$F_{req} = (fmg\cos i + \frac{C_d A v^2}{21.15} + mg\sin i + \delta m\frac{\mathrm{d}v}{\mathrm{d}t}) \tag{1}$$

where f is the rolling resistance coefficient, m is the gross weight, g is the gravity coefficient, C_d is the aerodynamic drag coefficient, A is the frontal area, v is the vehicle velocity, i is the angle of the road slope, and δ is the rotational mass conversion coefficient.

The required power for the hybrid vehicle is provided by the ICE and the battery with the following equation

$$P_{req} = (P_{ICE} + P_{batt}\eta_{batt})\eta_i \tag{2}$$

where P_{req} is the vehicle requirement power, P_{ICE} is the ICE power, P_{batt} is the battery power, η_{batt} is the battery efficiency, and η_i is the transmission efficiency.

The planetary gear is the mechanism that realizes the power distribution. ICE is attached to the planetary frame, MG1 is attached to the solar gear, and MG2 is attached to the gear ring. The relation between rotation speed and the torque of each component in the system can be expressed as

$$\begin{cases} \omega_{MG1} = (1+k_1)\omega_{ICE} - k_1\omega_{R1} \\ \omega_{MG2} = k_2\omega_{R1} \\ T_{MG1} = -\frac{T_{ICE}}{1+k_1} \\ T_{R1} = \frac{k_1 T_{ICE}}{1+k_1} + k_2 T_{MG2} \end{cases} \tag{3}$$

where the ω_{ICE}, ω_{MG1}, ω_{MG2}, ω_{R1} are the rotation speeds of the ICE, MG1, MG2 and planetary gear ring respectively, the T_{ICE}, T_{MG1}, T_{MG2}, T_{R1} are the torques of the ICE, MG1, MG2 and planetary gear ring respectively, k_1 is the transmission ratio of the PG, and k_2 is the gear ratio of the MG2.

ICE is a quasi-static model. The fuel consumption between the ICE rotation speed and the ICE torque is obtained by using the ICE bench test method. The map of ICE fuel consumption is shown in Figure 2.

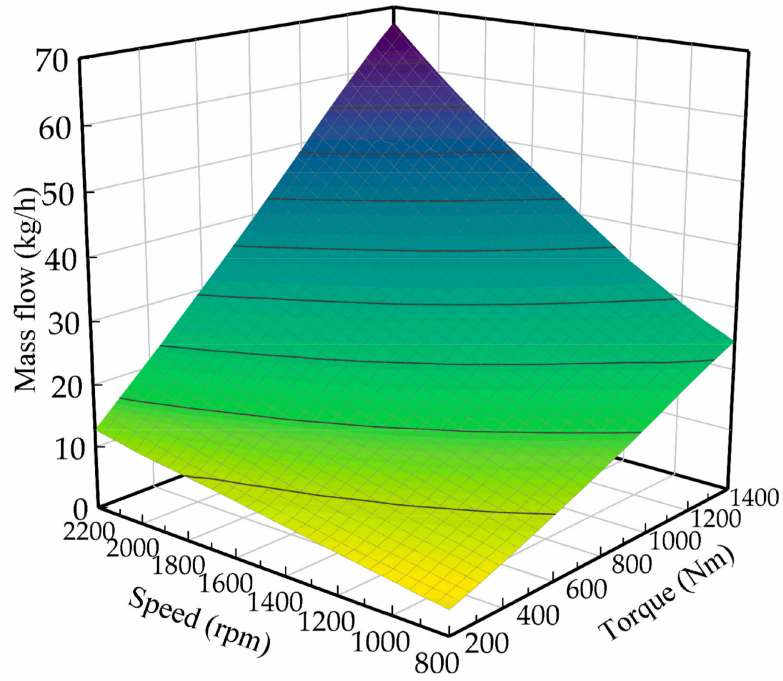

Figure 2. Map of ICE fuel consumption.

The instantaneous fuel consumption is obtained using methods such as data interpolation, and the fuel consumption relation is given by

$$Fuel = \int_0^T m_{fuel}(T_{ICE}, \omega_{ICE})\mathrm{d}t \tag{4}$$

where $Fuel$ is total fuel consumption, m_{fuel} is instantaneous fuel consumption function, T is total time.

The motor is a quasi-static model and the efficiency is derived from the rotational velocity and torque of the motor. The efficiency of the motor is shown in Figure 3. The motor can be operated in two modes, drive mode and generation mode, and the power of the motor can be described as

$$P_{motor\_req} = \begin{cases} T_m\omega_m/\eta_m, \text{(drive mode)} \\ T_g\omega_g\eta_g, \text{(generation mode)} \end{cases} \tag{5}$$

where P_{motor_req} is the motor requirement power, T_m, ω_m, η_m is the torque, rotation speed and efficiency of drive mode respectively, T_g, ω_g, η_g is the torque, rotation speed and efficiency of generation mode respectively.

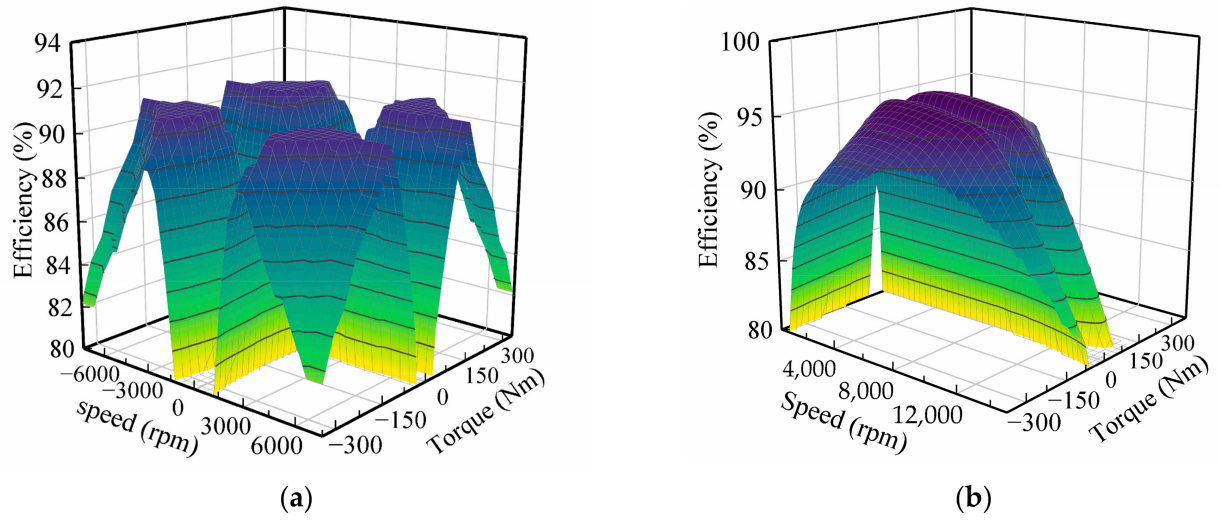

Figure 3. Efficiency of the motor. (**a**) MG1. (**b**) MG2.

The battery is modeled on an equivalent circuit model, and the battery current is related to the open circuit voltage and internal resistance of the battery by the following relation equation.

$$I_{batt} = \frac{U_{oc} - \sqrt{U_{oc}^2 - 4R_{batt}P_{batt}}}{2R_{batt}} \tag{6}$$

where I_{batt} is the battery current, U_{oc} is the open circuit voltage, R_{batt} is the internal resistance of the battery. The state of charge (SOC) is defined as

$$SOC(t+1) = SOC(t) - \frac{I_{batt}}{Q_{batt}}\Delta t \tag{7}$$

where SOC is the state of charge, Q_{batt} is the nominal battery capacity.

2.2. Battery Aging Model and Optimization Problem for EMS

This study is mainly focused on the control problem. A semi-empirical model of battery aging was used. The model takes into account the physical chemistry of Li-ion batteries, performs an aging test on Li-ion batteries, and finally fits the data to obtain a set of equations

describing the capacity loss of Li-ion batteries. The American scholars Wang John et al. [23] have established a representative semi-empirical model for Li-ion battery aging.

$$Q_{loss} = B \cdot \exp(\frac{-E_a}{R_{\text{gas}} \cdot T_k})(Ah)^z \tag{8}$$

where Q_{loss} is the percentage of Li-ion battery capacity loss, B is the pre-exponential factor, E_a is the activation energy, R_{gas} is the gas constant, R_{gas} = 8.314 J/(mol·K), T_k is the battery temperature, Ah is the battery Ah-throughput, z is the power law factor.

There are three main factors affecting battery aging: current rate, battery temperature, and depth of discharge (DOD). Replace DOD with SOC, and the battery aging model is built with the following equation [24].

$$\begin{gathered} Q_{loss} = (\alpha \cdot SOC + \beta) \exp(\frac{-31700+163.3I_c}{R_{\text{gas}} \cdot T_k})(Ah)^{0.57} \\ \alpha = \begin{cases} 1287.6, & SOC \leq 0.45 \\ 1385.5, & SOC > 0.45 \end{cases} \\ \beta = \begin{cases} 6356.3, & SOC \leq 0.45 \\ 4193.2, & SOC > 0.45 \end{cases} \end{gathered} \tag{9}$$

where α and β are the coefficients, I_c is the current rate.

Battery temperature has a significant impact on battery degradation, with battery operating temperatures typically ranging from −20 to 60 °C. Too high or too low a temperature can affect the performance of the battery and eventually lead to a decrease in capacity. However, the EMS in this study is trained under the same driving cycles, therefore, the battery temperature is not taken into account and is set to 25 °C. The above equation is a semi-empirical model to calculate the battery capacity loss. While facing the optimization problem, it is necessary to set the objective function that incorporates the battery aging model. Typically, we consider battery life termination when the battery loses 20% of its nominal capacity, and the Ah-throughput is given by

$$Ah(SOC, I_c) = [20/(\alpha \cdot SOC + \beta) \cdot \exp(\frac{-31700 + 163.3I_c}{8.314 \times 298.15})]^{\frac{1}{0.57}} \tag{10}$$

The total number of battery operating cycles before the end of life can be calculated as

$$N(SOC, I_c) = \frac{3600Ah(SOC, I_c)}{Q_{batt}} \tag{11}$$

The total Ah-throughput includes charged and discharged, so the state of health (SOH) of the battery can be described as

$$\frac{\mathrm{d}SOH(t)}{\mathrm{d}t} = -\frac{|I(t)|}{2N(SOC, I_c) \cdot Q_{batt}} \tag{12}$$

To account for fuel consumption cost and battery aging cost, the following total cost objective function is established.

$$J = \int_0^T w_1 \cdot fuel(t) + w_2 \cdot \mathrm{d}SOH(t)\mathrm{d}t \tag{13}$$

where the first term is the instantaneous fuel consumption cost and the second term is the battery aging cost, w_1 is the diesel oil price, which is set to 7.2 CNY/L, w_2 is the Li-ion battery price, which is set to 1700 CNY/kWh, CNY is Chinese Yuan.

In this study, the optimization of energy management strategies for hybrid electric dump trucks by minimizing the objective function.

3. Method and Design of TD3-Based EMS

3.1. Reinforcement Learning

The schematic diagram of the reinforcement learning principle is shown in Figure 4. The agent chooses an action based on the state of the environment, the action acts on the environment, and the environment feeds back the reward. The agent is guided to select a more appropriate action next time by iterating the policy function in this way.

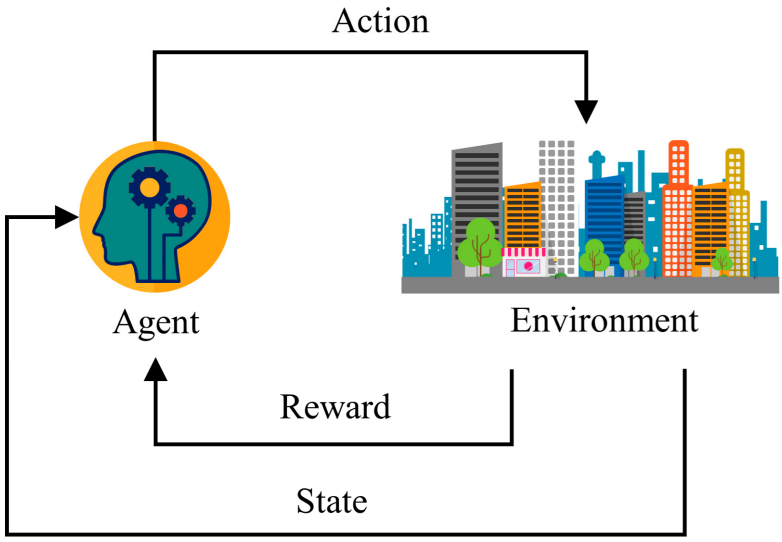

Figure 4. Schematic diagram of reinforcement learning principle.

The reinforcement learning must follow the Markov decision process (MDP). The MDP has four elements: the state of the environment, the action of the agent, the transition probability of the state space, and the reward function. In MDP, the reinforcement learning algorithm continuously updates the policy function by interacting with the environment cyclically and based on the reward values fed by the environment. A policy function is a mapping from states to actions, and its performance can be evaluated by the state-value function $V(s)$ or the Q-value function $Q(s,a)$, so reinforcement learning algorithms usually find the optimal policy function π^* by iterating over the optimal value function to maximize the expected reward [25]. Its expected total reward is as follows

$$R_t = \sum_{t=k}^{\infty} \gamma^{t-k} r(s_t, a_t) \tag{14}$$

where γ is the discount factor, s is the state, a is the action, t is the time, and $r(s_t,a_t)$ is the reward at each moment.

To find the optimal policy function π^*, many reinforcement learning algorithms use the Q-value function $Q_\pi(s_t,a_t)$ to evaluate the policy function. The Bellman equation is given by

$$Q_\pi(s_t, a_t) = r(s_t, a_t) + \gamma \mathrm{E}_\pi[Q_\pi(s_{t+1}, a_{t+1})] \tag{15}$$

where E_π is the expectation.

The policy function π^* can be reversed by a = argmax $Q^*_\pi(st,a')$ when the optimal Q-value function $Q^*_\pi(s_t,a_t)$ is known. The $Q^*_\pi(s_t,a_t)$ of the Q-learning algorithm can be solved by temporal difference. However, the Q-learning algorithm faces the "dimensional disaster" when the dimensionality of state and action is large, and deep reinforcement learning using neural networks can solve this problem.

3.2. TD3 Algorithm

The TD3 algorithm is an algorithm in deep reinforcement learning, and the TD3 algorithm uses the actor-critic (AC) framework to approximate the policy function and

the Q-value function through neural networks called actor and critic, respectively. The actor network is the action output of the TD3 algorithm and is used to interact with the environment, while the critic network scores the policy through feedback rewards from the environment. With this clever learning approach, the TD3 algorithm can handle problems with continuous action spaces.

The TD3 algorithm draws on the experience of the DDQN algorithm and improves the DDPG algorithm by integrating the dual critic network technology on the AC framework [26]. One critic network and one target critic network are added to the four neural networks of the DDPG algorithm: actor network, critic network, target actor network, and target critic network. By using two sets of critic networks, the smaller target Q-value is taken into account in calculation, thus suppressing the network overestimation problem. The critic network is updated by minimizing the loss function.

$$L(\theta) = \mathrm{E}[\sum_{i=1}^{2} (y_t - Q_{\theta_i}(s_t, a_t))^2] \tag{16}$$

where y_t is the Q-target value, calculated by temporal difference, and $Q_{\theta'}(s_{t+1},a_{t+1})$ uses the smaller of the two sets of target critic networks.

$$y_t = r(s_t, a_t) + \gamma \mathrm{min} Q_{\theta_i'}(s_{t+1}, a_{t+1}) \tag{17}$$

$$a_{t+1} \sim \pi_{\phi'}(s_{t+1}) \tag{18}$$

where $Q_{\theta'}$ is the Q-target value and $\pi_{\Phi'}$ is the target actor policy.

The updating method of the actor network is a deterministic strategy gradient algorithm.

$$\nabla_\phi J(\phi) = \mathrm{E}_{s\sim p\pi}[\nabla_a Q_\theta(s,a)|_{a=\pi(s)} \nabla_\phi \pi_\phi(s)] \tag{19}$$

Soft update method is used for the TD3 algorithm to ensure more stable training [27].

$$\begin{cases} \theta' = \tau\theta + (1-\tau)\theta' \\ \phi' = \tau\phi + (1-^{\tau})\phi' \end{cases} \tag{20}$$

where τ is the soft update coefficient and has a value range of $0 \leq \tau \leq 1$, θ is the critic network, ϕ is the actor network, θ' is the target critic network, ϕ' is the target actor network.

The TD3 algorithm gives two improvements to address the problem of high variance: (1) delayed policy updates, where the actor network is updated asynchronously with the critic network. (2) target policy smoothing regularization, where noise is added to the action output from the target actor network.

$$a' \leftarrow \pi_{\phi'}(s') + \varepsilon,\ \varepsilon \sim \mathrm{clip}(\mathcal{N}(0,\sigma), -c, c) \tag{21}$$

where a' is the next action, s' is the next state, ε is the noise, N is the normal distribution, σ is the standard deviation, c is the range of noise.

The actor and critic networks are trained using empirical replay. This means that the TD3 algorithm deposits samples <s, a, r, s', d> consisting of state s, action a, reward value r, next state s', and whether the round is over d into an experience buffer, which is obtained by interacting with the environment, and then mini-batches of samples are randomly sampled from the experience buffer to train the neural networks. This approach can improve sample utilization and attenuate correlations between training samples.

In summary, the TD3 algorithm has the advantage of avoiding Q-value overestimation. It can improve training efficiency and stability. The pseudo-code of the TD3 algorithm is shown in Algorithm 1.

Algorithm 1: TD3

initialization:
  critic networks $Q_{\theta 1}$, $Q_{\theta 2}$ with random parameters θ_1, θ_2,
  actor network π_φ with random parameters φ,
  target critic networks $\theta'_1 \leftarrow \theta_1$, $\theta'_2 \leftarrow \theta_2$,
  actor network $\phi' \leftarrow \phi$
  replay buffer $\mathcal{B}$
for $t = 1$ **to** T **do**
  observe state s choose action with exploration noise $a \sim \pi_\varphi(s) + \varepsilon$,
  $\varepsilon \sim \mathcal{N}(0, \sigma)$ and observe reward r and new state s',
  store transition tuple (s, a, r, s', d) in $\mathcal{B}$
  randomly sample a mini-batch of N transitions $\{(s, a, r, s', d)\}$ from $\mathcal{B}$
  $a' \leftarrow \pi_{\phi'}(s') + \varepsilon$, $\varepsilon \sim \mathrm{clip}(\mathcal{N}(0, \sigma), -c, c)$
  $y_t = r(s_t, a_t) + \gamma \min_{i=1,2} Q_{\theta'_i}(s_{t+1}, at + 1)$
  update critic networks $\theta_i \leftarrow \mathrm{argmin}_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(s, a))^2$
  if t mod d **then**
    update φ by deterministic policy gradient:
    $\nabla_\phi J(\phi) = N^{-1} \sum \nabla_a Q_{\theta_1}(s, a)\big|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$
    Update target networks by soft update:
    $\theta' \leftarrow \tau\theta + (1 - \tau)\theta'$
    $\phi' \leftarrow \tau\phi + (1 - \tau)\phi'$
  end if
end for

3.3. Action Mask

Action mask is a common method used to avoid invalid actions [28]. DRL is an algorithm that performs complex tasks by trial-and-error methods. The agent needs to explore the action space. In practice, unlimited exploration can cause serious safety problems in the case of vehicle control, which can damage the vehicle. There are often many constraints, and some actions are not selectable in some states. There are invalid actions in the action space. Soft constraints are commonly used to set a large negative reward is set for the invalid action to guide the agent to avoid such actions. However, this approach does not avoid unsafe actions and makes the model take a long time to train and is prone to scatter. Therefore, a hard-constrained approach was chosen in this study by setting the action mask. A schematic diagram of the action mask is shown in Figure 5. It adds a masking process after the output layer of the neural network to filter out invalid actions and output the available actions, so that the agent cannot select invalid actions.

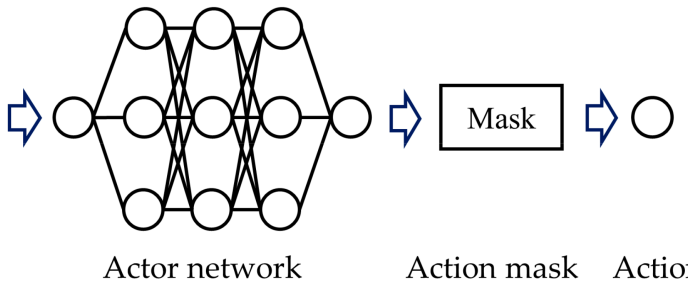

Figure 5. Schematic diagram of the action mask.

It is necessary to design an action mask to filter out invalid actions during the DRL training. It can avoid unsafe issues, reduce unnecessary exploration, and accelerate the training efficiency of the algorithm. In a power-split hybrid system, the generator may overshoot the revs or overshoot the torque when the engine is at high power, damaging the powertrain. The action mask is designed as follows

Firstly, the ICE power is discrete.

$$P_{ICE} = \{0, 1, 2, \ldots, 243 \text{ kW}\} \tag{22}$$

Secondly, at different times t, all actions are traversed in advance to prevent the generator from operating outside the constraints and to obtain the allowed range of engine operating power.

$$P_{ICE}(t) = \left[P_{ICE}^{\min}(t), P_{ICE}^{\max}(t)\right] \tag{23}$$

Thirdly, the actor network outputs the action that is clipped by using the clip function so that the action is within the allowed range.

$$P_{ICE}(t) = P_{ICE}(t).\text{clip}\left[P_{ICE}^{\min}(t), P_{ICE}^{\max}(t)\right] \tag{24}$$

Both the actor network and the target actor network apply action masks to improve the deep reinforcement learning capabilities. The clip method can be used for algorithms such as DDPG and TD3 that output continuous actions based on the AC framework and deterministic strategies. It can set the Q-value of the invalid action to negative infinity on the Q-value function vector for Q-value-based algorithms such as DQN and DDQN, so the invalid actions are not selected by argmax Q-value [29]. For algorithms such as proximal policy optimization, the action mask can be implemented by tuning the probability of an invalid action to be sampled to zero [30].

3.4. Design of TD3-Based EMS

In this study, the TD3-based EMS of hybrid electric dump trucks is taken as the research object. The definition of state, action, and reward in the TD3-based EMS is as follows.

The vehicle velocity, vehicle acceleration, and battery SOC are defined as states to accurately represent the vehicle's state. The vehicle motion state is reflected by the vehicle velocity and vehicle acceleration, and the vehicle energy management state is reflected by the battery SOC.

$$s = \{v, acc, SOC\} \tag{25}$$

where *acc* is the vehicle acceleration.

The power of the ICE is defined as the action, while the optimal operating line (OOL) of the ICE is applied to reduce the action variables, in order to simplify the power system model and improve the convergence efficiency of DRL. The OOL is shown by the red solid line in Figure 6, where each power corresponds to the associated speed and associated torque.

$$a = \{P_{ICE} | P_{ICE} \in [0, 243\text{kW}]\} \tag{26}$$

The reward function includes the instantaneous fuel consumption cost, the battery aging cost, and the sustainability penalty of the SOC. The optimal strategy is searched by maximizing the reward function.

$$r = -\left[w_1 \cdot fuel(t) + w_2 \cdot \Delta Q_{loss}(t) + w_3 \cdot (SOC(t) - SOC_{init})^2\right] \tag{27}$$

where the first term is the instantaneous fuel consumption cost, the second term is the battery aging cost, the third term is the sustainability penalty of the SOC, w_3 is the SOC sustainability penalty coefficient and SOC_{init} is the initial SOC.

Since the TD3 algorithm is based on the AC framework and applied to continuous state space and continuous action space, it is combined with the neural networks. The structure of the actor and critic network is shown in Figure 7. There are three neurons in the input layer and one neuron in the output layer for the actor network. There are four neurons in the input layer and one neuron in the output layer for the critic network. There are three hidden layers with 512, 256, and 128 neurons for the actor and critic networks. All networks consist of fully connected layers with ReLU activation function for hidden

layers and Tanh activation function for output layers. The target actor network and target critic network have the same structure as the actor and critic networks, respectively, and the Adam optimizer is used to update the networks.

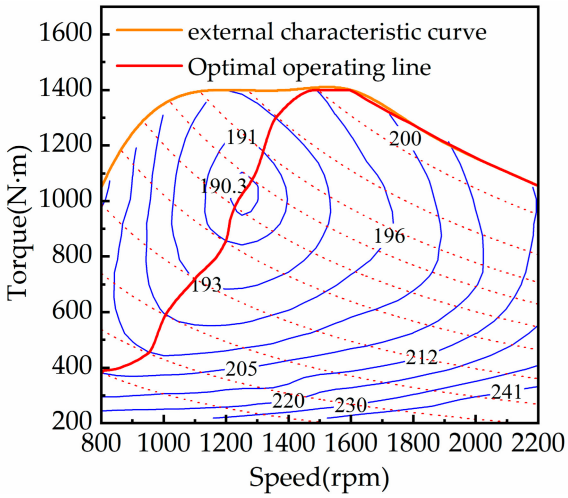

Figure 6. The optimal operating line of the engine.

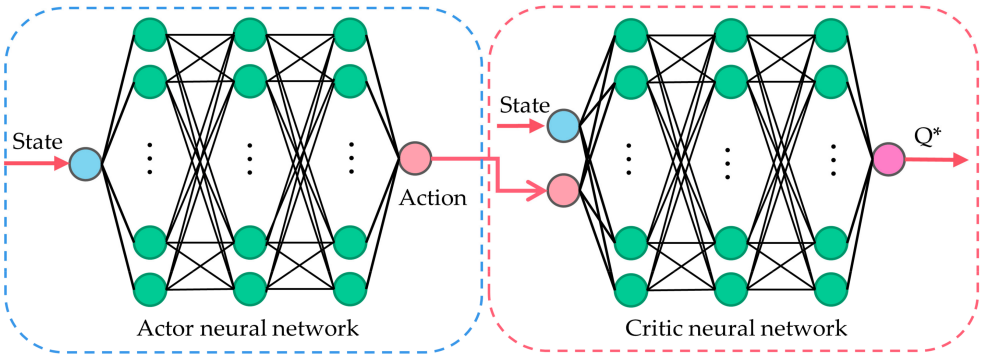

Figure 7. The structure of the neural networks.

The relevant hyper-parameters are listed in Table 3. The learning rate affects how fast the policy learns. Too large will prevent the policy from searching for the optimal solution, and too small will make the policy learn inefficiently. The discount factor in reinforcement learning is the discount of the expected future reward at the current moment and regulates the effect of the future reward on the agent. Because of the exponential relationship of the discount factor, when the discount factor takes a larger value, γ^n is larger, the longer the steps the agent considers in the future, and the training difficulty increases. When the discount factor takes a smaller value, γ^n is smaller, the agent focuses more on the current reward, and the training difficulty decreases. The discount factor is as large as possible in order to allow the agent to consider as much of the global regression as possible, provided the algorithm can converge, so it takes the value of 0.99. The mini-batch size is the number of samples for a single training session, and the experience buffer size is the maximum capacity to record the training experience, and the earliest samples are removed if the training experience exceeds the maximum capacity.

The overall control framework of this study is illustrated in Figure 8. The actor network in the agent outputs action based on the state of the vehicle, action acts on the vehicle after action mask, the ICE power is controlled, vehicle feedback the reward of fuel consumption, battery aging and SOC maintenance cost, the experience buffer stores the samples and mini-batch is sampled to training the neural networks.

Table 3. Hyper-parameters for TD3-Based EMS.

Parameters	Value
Actor network learning rate	0.0001
Critic network learning rate	0.0002
Discount factor	0.99
Mini-batch size	256
Experience buffer size	1×10^6

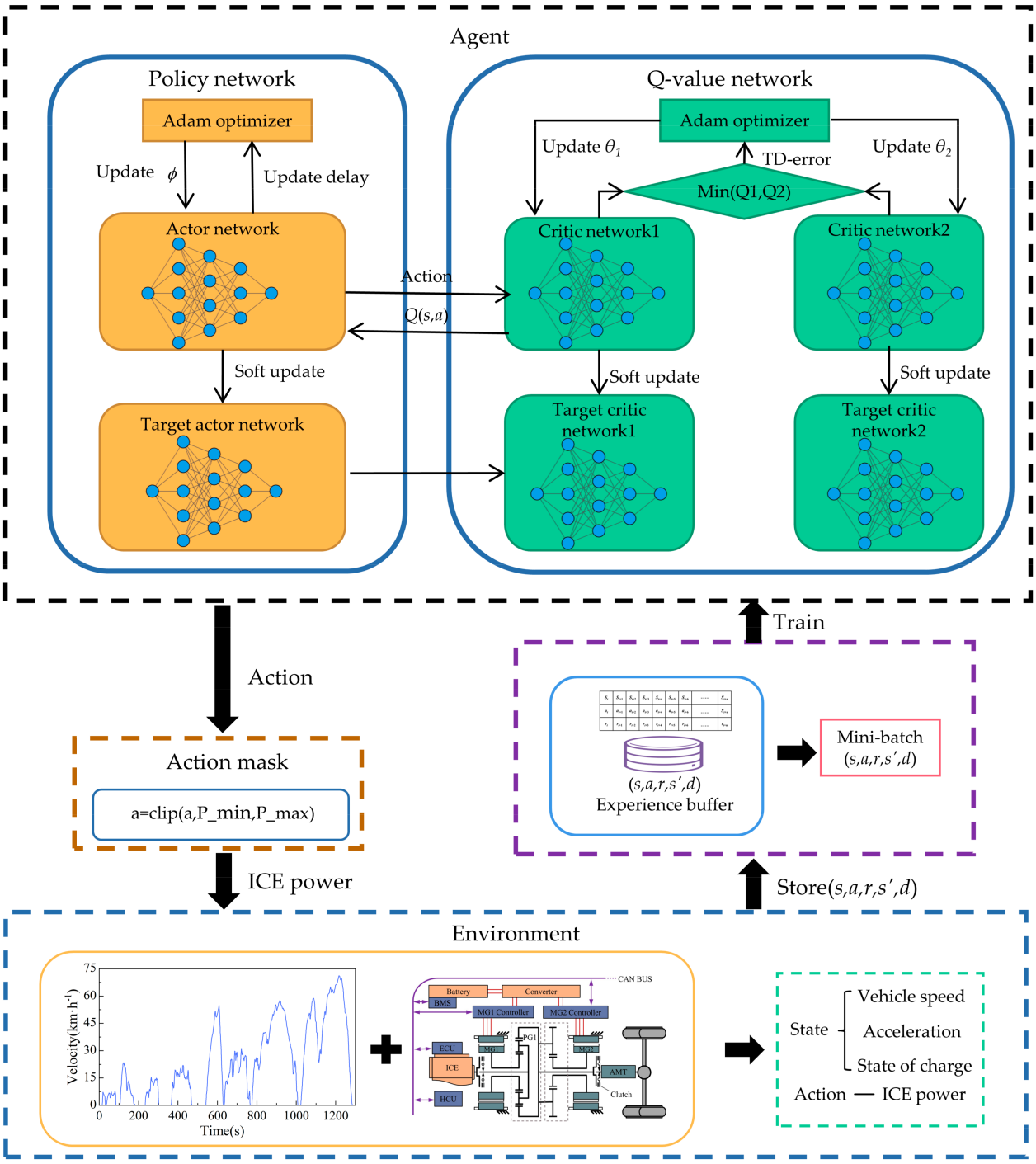

Figure 8. Overall control framework.

4. Results

In Section, EMSs are trained under the China Heavy-duty Commercial Vehicle Test Cycle for Dump Truck (CHTC-D) driving cycle to verify the effectiveness of the proposed EMS. The velocity of the CHTC-D driving cycle is shown in Figure 9. The simulation model is built in Python, and the DRL framework is programmed with Python using the PyTorch package. First, the convergence efficiency is compared between EMSs with and without action masks. Second, it compared and analyzed the battery capacity loss and fuel economy under DDPG-based EMSs and TD3-based EMSs with and without considering battery aging. Finally, it is simulated on the China-World Transient Vehicle Cycle (C-WTVC) driving cycle to verify the applicability of the TD3-based EMS.

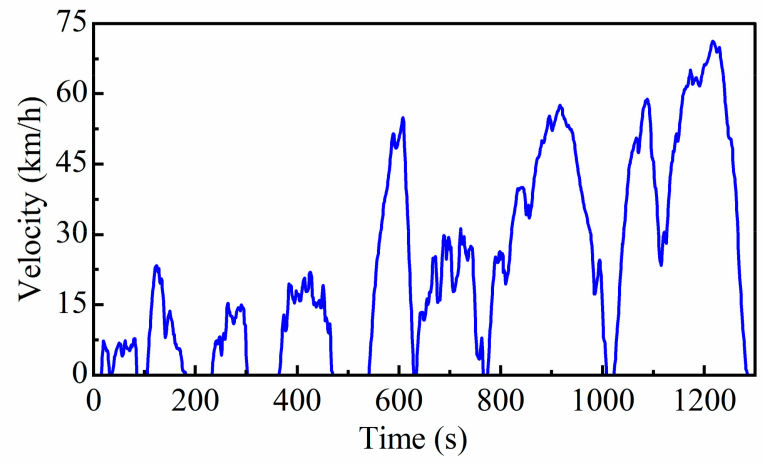

Figure 9. Velocity of the CHTC-D driving cycle.

4.1. The Impact of Action Mask

The action mask filters out invalid actions that cause the MG1 to overshoot in speed and torque and cause damage to the powertrain. At the same time, action masks can reduce useless exploration and improve learning efficiency. In iterative learning, the optimal EMS can be searched by a trial-and-error approach based on the DRL-based EMS. The learning process ends when the reward remains stable and converges. To achieve the global optimization objective, the number of episodes of the learning process is defined to be 500 for all EMSs. Figure 10 shows the reward for the training process and these results for the same driving cycle (CHTC-D) in online testing. It can be found that the reward maintains stable convergence, and the DRL-based EMSs search for the optimal policy through their self-learning capability. Also, it can be seen that the strategies without action masks converge slower than the ones with action mask, proving that the action mask can improve the training efficiency of the DRL-based EMSs.

All EMSs initially have low rewards because the neural network parameters are randomly initialized. In the beginning, EMSs are learned without any prior experience that facilitates energy optimization. The reward then floats up and down as the DRL-based EMSs keep exploring, with different results between great and poor. The overall effect is upward.

The ICE power of the DRL-based EMS for one of the episodes is shown in Figure 11. The maximum allowed power per second for ICE is the solid red line. Due to the exploration of TD3-based EMS without action masks, the ICE power exceeds the allowed range several times. This condition can damage electrical systems or cause unsafe accidents in industrial applications and therefore needs to be avoided. The ICE power never exceeds the maximum limit because the TD3-based EMS with action mask technique can filter out invalid actions, which demonstrates the excellent reliability of the proposed action mask approach. Both TD3-based EMSs and DDPG-based EMSs studied later use AM techniques to enable policies to avoid invalid actions while improving training efficiency.

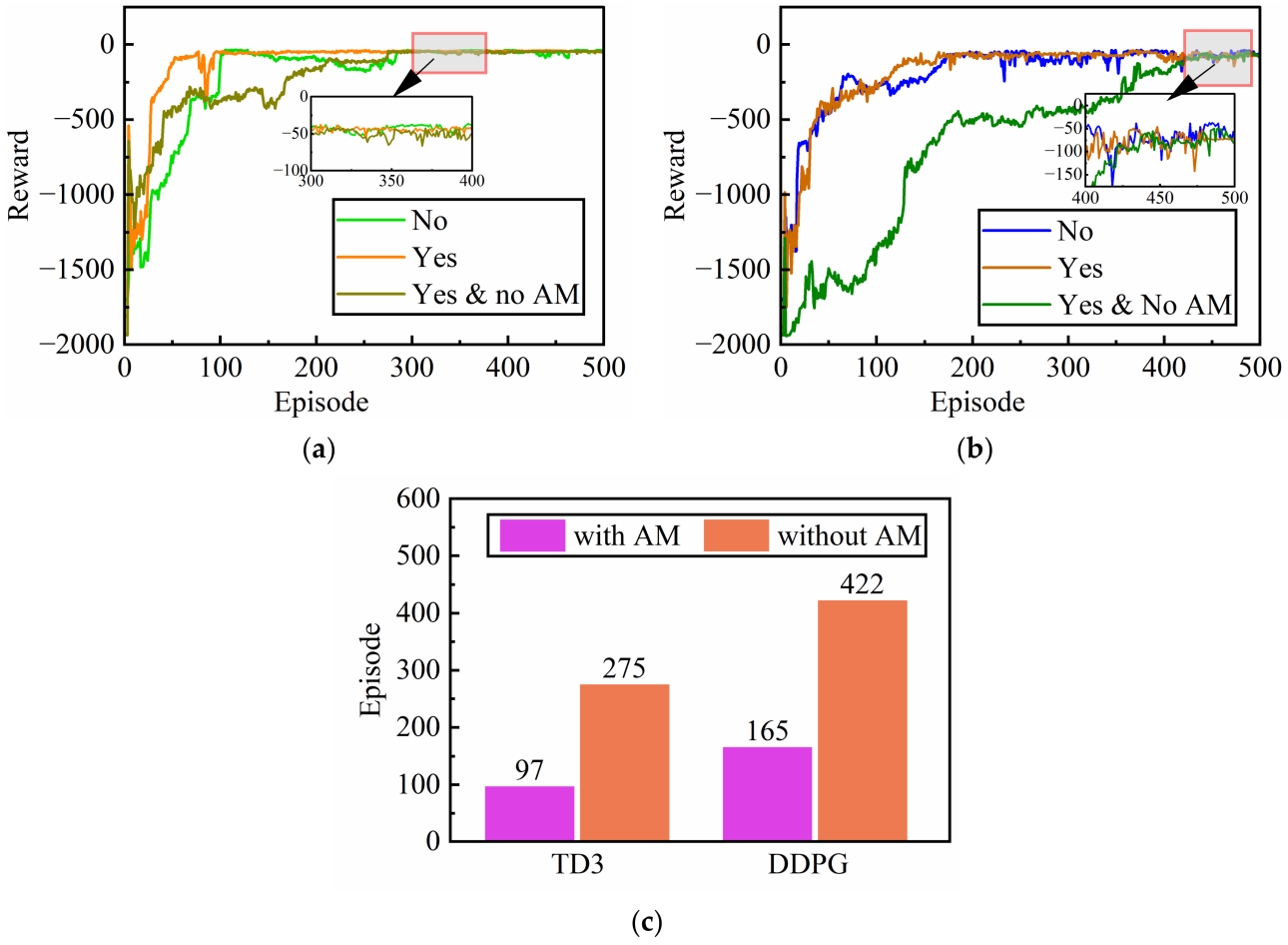

Figure 10. Rewards for the episode. (**a**) TD3-based EMSs. (**b**) DDPG-based EMSs. (**c**) Convergence speed for different EMSs.

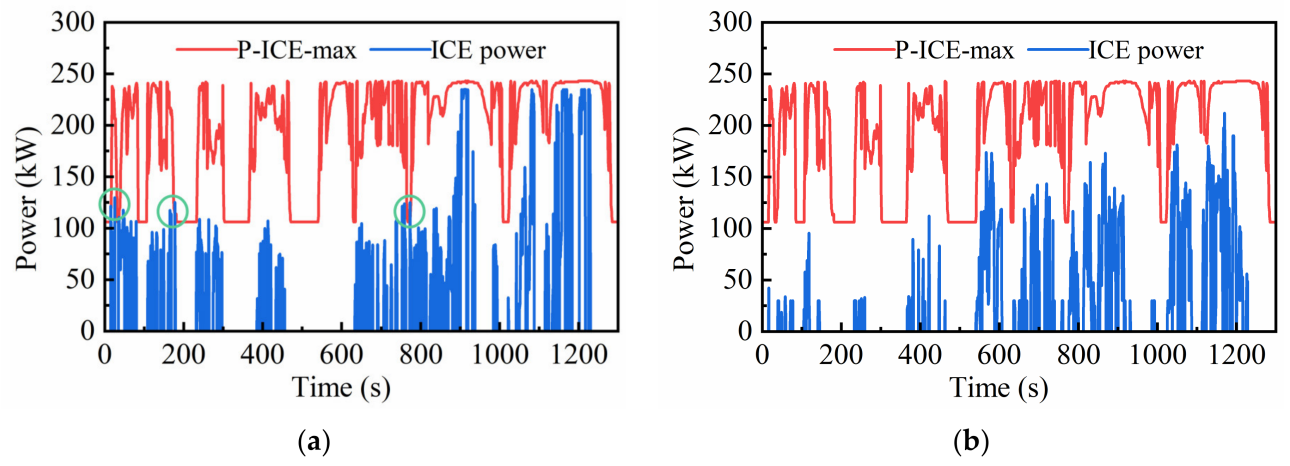

Figure 11. The ICE power of the DRL-based EMS for one of the episodes. (**a**) TD3-based EMS without action mask. (**b**) TD3-based EMS with action mask.

4.2. Battery Capacity Loss and Fuel Consumption

Energy management strategies for non-plug-in hybrid electric dump trucks require battery SOC sustainment capabilities to ensure that the strategy does not run out of battery energy by using only battery power. The SOC of the TD3-based EMSs and the DDPG-based EMSs under the CHTC-D driving cycle is shown in Figure 12. It can be seen that the SOC

of all strategies returns to around the initial value of 0.6, which indicates that all strategies can achieve battery energy balance. At the same time, the SOC fluctuates in a small range of 0.57–0.61 and the battery can operate in shallow cycles, which helps to increase the charging and discharging efficiency and reliability of the battery.

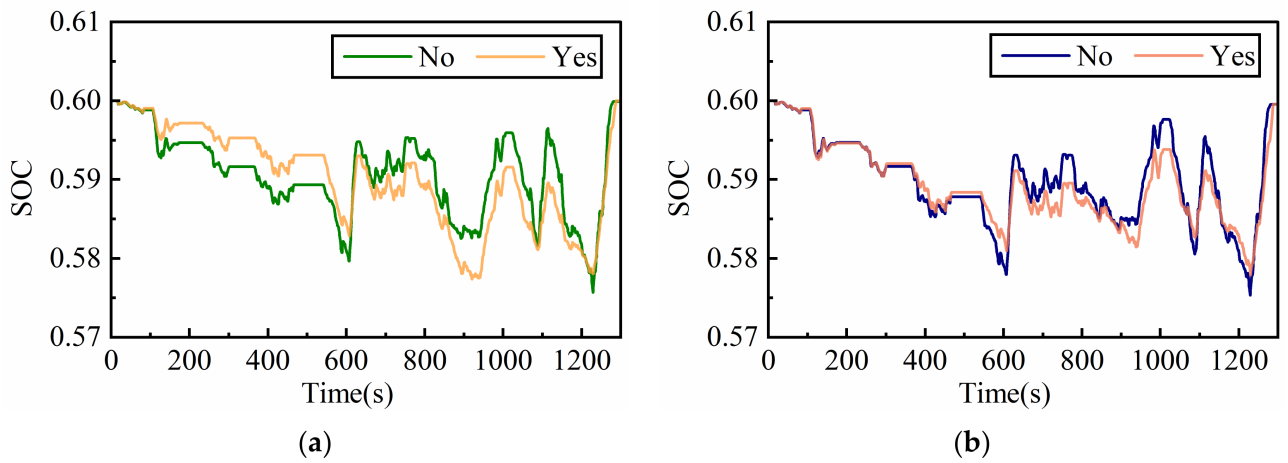

Figure 12. SOC of the hybrid electric dump truck under the CHTC-D driving cycle. (**a**) TD3-based EMSs. (**b**) DDPG-based EMSs.

The loss of battery capacity is shown in Figure 13. The battery capacity loss is present at all EMSs and increases more when do not consider battery aging. The largest loss of battery capacity is reported for the DDPG-based EMS without considering battery aging, with a maximum capacity loss of 0.0434% per 100 km. The smallest battery capacity loss is achieved by the TD3-based EMS considering battery aging, with a minimum capacity loss of 0.0270% per 100 km. These results demonstrate that EMS which takes into account battery aging slows down battery capacity loss.

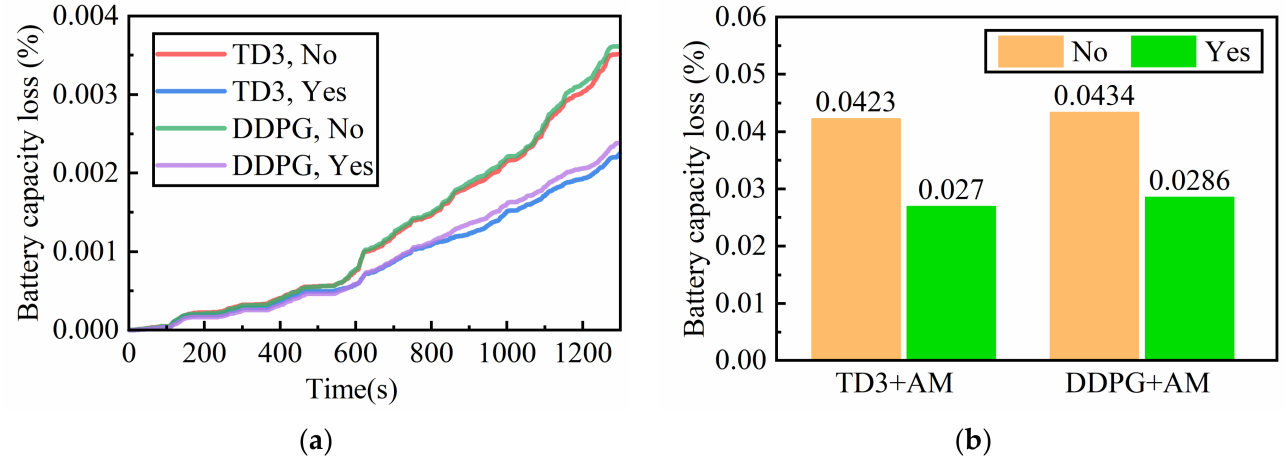

Figure 13. The loss of battery capacity. (**a**) Battery capacity loss at each time. (**b**) Total battery capacity loss.

The battery power of different EMSs is shown in Figure 14. One DDPG-based EMS, which does not consider battery aging, leads to maximum battery capacity loss, while the other TD3-based EMS, which considers battery aging, leads to minimum battery capacity loss. It can be seen that the maximum battery power of 163 kW for DDPG-based EMS and 100 kW for TD3-based EMS. The TD3-based EMS, which takes into account battery aging, takes a smaller battery power load than the DDPG-based EMS, which does not take into account battery aging. As a result, a smaller battery load reduces the battery capacity loss and extends battery life.

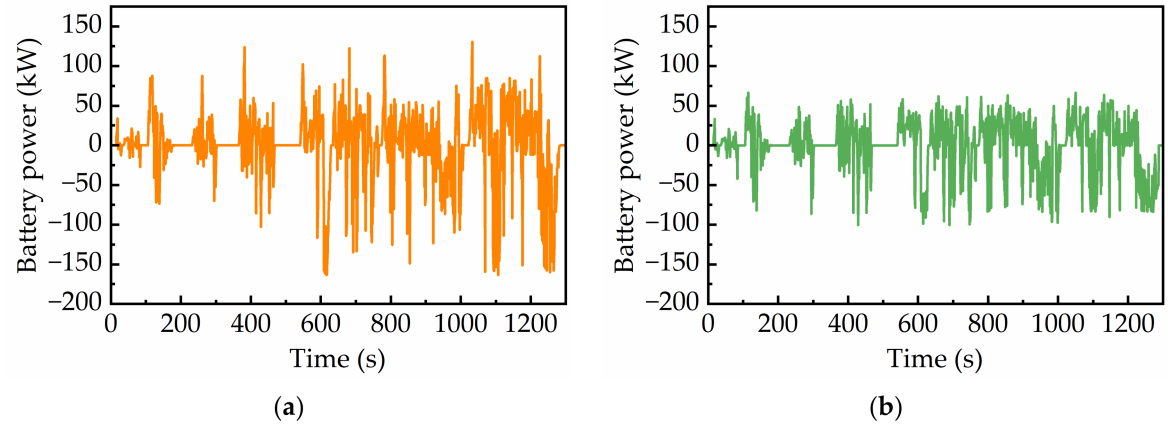

Figure 14. The battery power of different EMSs. (**a**) DDPG-based EMS, No. (**b**) TD3-based EMS, Yes.

The distribution of operating points on the battery for different EMSs is shown in Figure 15. It can be seen that the battery current rate (C-rate) of the two strategies without considering battery aging is as high as four, and the C-rate of numerous operating points is higher than three. For EMSs that take into account battery aging, there are a few cases where the battery C-rate is higher than three. The results show that the battery current can be controlled by EMSs that take into account the battery aging, which can help to slow down the battery capacity loss by keeping the battery operating at a lower C-rate. An EMS that takes into account the battery aging results in a battery that operates in the low-load region and extends the battery life.

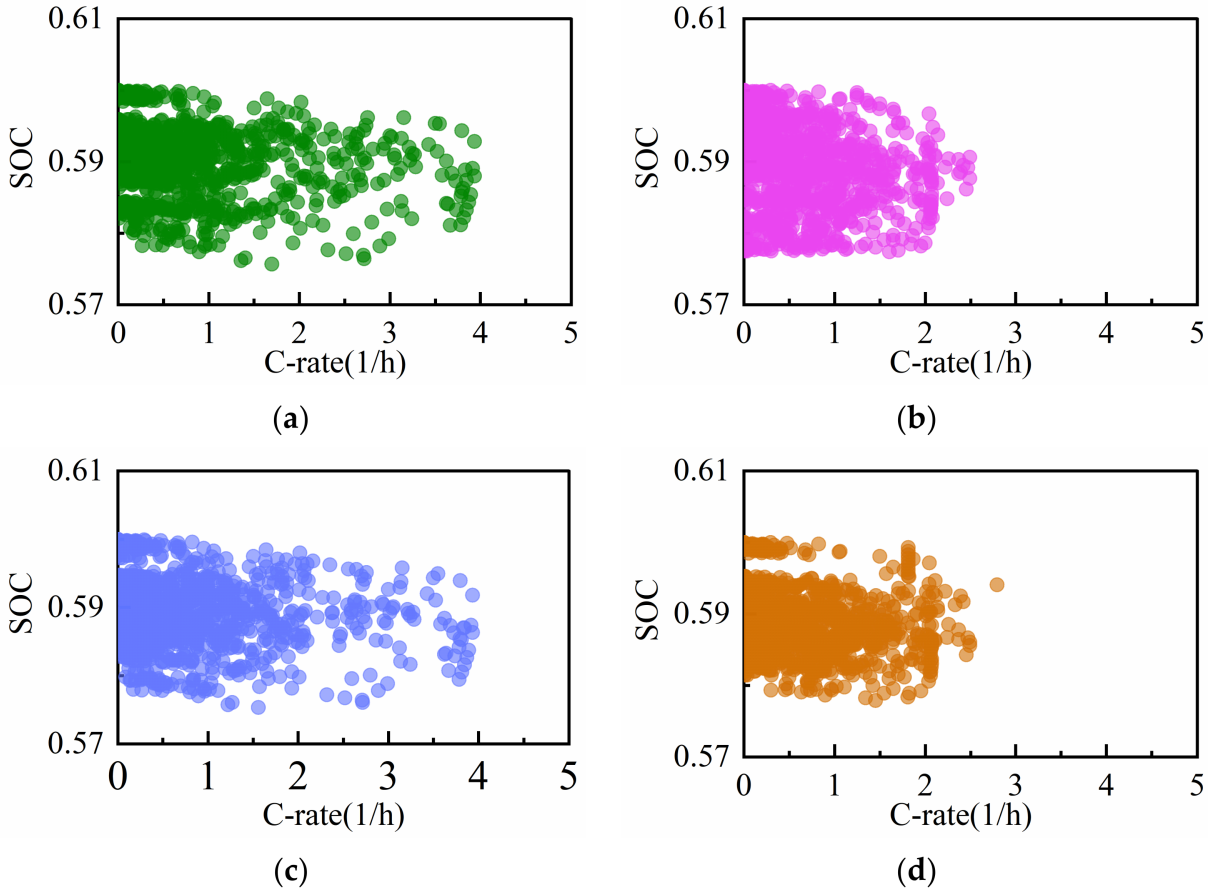

Figure 15. Distribution of operating points on the battery for different EMSs. (**a**) TD3-based EMS, No. (**b**) TD3-based EMS, Yes. (**c**) DDPG-based EMS, No. (**d**) DDPG-based EMS, Yes.

Since the fuel consumption of a vehicle is the main objective of studying energy management strategies, the fuel consumption of the ICE is shown in Figure 16. The lowest fuel consumption is achieved by the DDPG-based EMS, which consumes 24.74 L/100 km without considering battery aging. The maximum fuel consumption is achieved by the DDPG-based EMS considering battery aging with a consumption of 25.27 L/100 km. Moreover, the fuel consumption of TD3-based EMS is 25.06 L/100 km without considering battery aging, and 25.27 L/100 km without considering battery aging. The results show that fuel consumption will increase a little when considering battery aging.

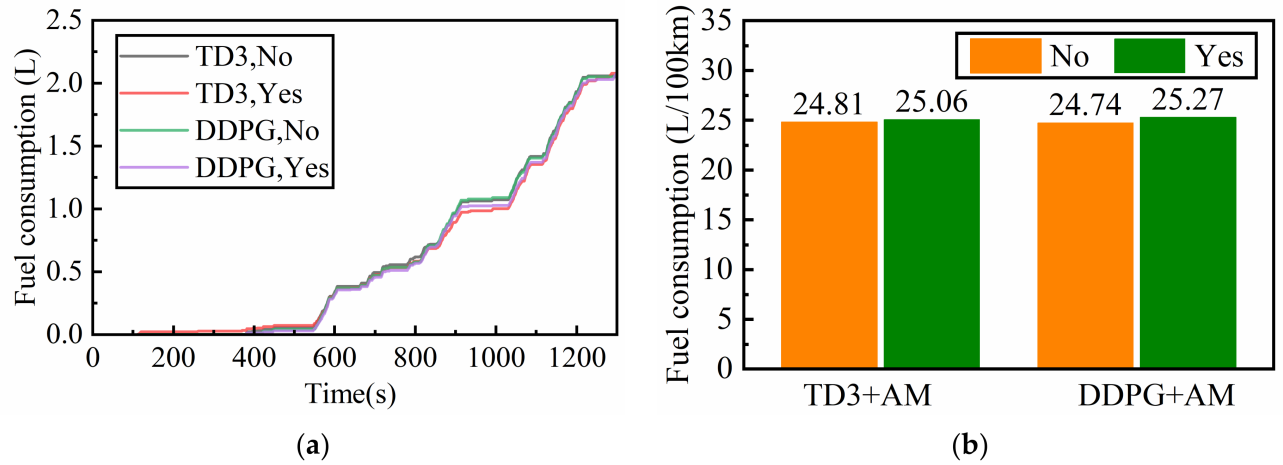

Figure 16. The fuel consumption of the ICE. (**a**) Fuel consumption at each time. (**b**) Fuel consumption for 100 km.

The distribution of engine operating points can reasonably reflect the performance of the strategy. As shown in Figure 17, the distribution of ICE power for different EMSs corresponds to the brake-specific fuel consumption (BSFC) curve of the optimal operating line of the ICE, and the red curve shows the BFSC for each power. Most of the ICE operating points are located in the lower regions of the BSFC, especially in the two EMSs where battery aging is not considered. It indicates that the DRL-based EMSs learn to maximize the efficiency of ICE. However, to mitigate the loss of battery capacity, EMSs considering battery aging operate ICE in higher BFSC region by optimizing the operating point of the battery, resulting in an increase in ICE fuel consumption. At the same time, the engine operating point in the lower power region is more significant when considering battery aging. That is bad for fuel economy, but it extends battery life.

The fuel consumption and battery capacity loss for different EMSs under the CHTC-D driving cycle are shown in Table 4. The two EMSs with the lowest fuel consumption, without considering battery aging, are 24.81 L/100 km for the TD3-based EMS and 24.74 L/100 km for the DDPG-based EMS, resulting in the largest battery capacity loss of 0.0423% and 0.0434%. When the EMSs take into account battery aging, the fuel consumption of the TD3-based EMS increased to 25.06 L/100 km and the battery capacity loss decreased to 0.0270%, the fuel consumption of DDPG-based EMSs increased to 25.27 L/100 km and the battery capacity loss decreased by 36.17% to 0.0286%. It can be seen that when the EMS considers battery aging, it mitigates the loss of battery capacity but also leads to an increase in fuel consumption. There is a small increase in fuel consumption, but a large improvement in battery capacity loss when the EMS considers battery aging.

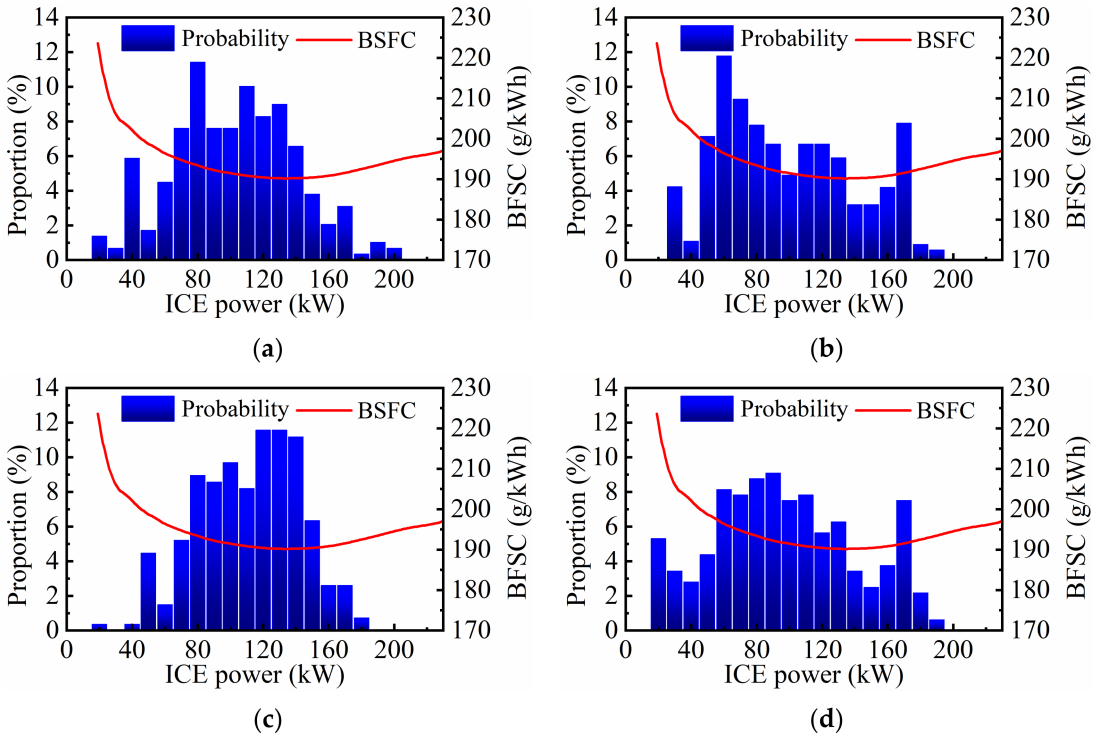

Figure 17. The distribution of ICE power. (**a**) TD3-based EMS, No. (**b**) TD3-based EMS, Yes. (**c**) DDPG-based EMS, No. (**d**) DDPG-based EMS, Yes.

Table 4. Fuel consumption, battery capacity loss, and cost under CHTC-D driving cycle.

EMS	Consider Battery Aging	F.C. (L/100 km)	F.C. Cost (CNY)	Battery Capacity Loss (%)	Battery Aging Cost (CNY)	Total Cost (CNY)	Performance
TD3-based	No	24.81	178.63	0.0423	28.99	207.62	95.82%
	Yes	25.06	180.43	0.0270	18.51	198.94	100%
DDPG-based	No	24.74	178.13	0.0434	29.75	207.88	95.70%
	Yes	25.27	181.94	0.0286	19.60	201.54	98.71%

Note: F.C. is the fuel consumption.

The cost per 100 km for different EMSs is given in Table 4. It can be seen that the DDPG-based EMS without considering the battery aging has the highest cost of 207.88 CNY, and the battery cost accounts for 14.31% of the total cost. When battery aging is considered, the TD3-based EMS has the lowest cost, with a 4.30% reduction to 198.94 CNY and a battery cost of 9.30% of the total cost. At the same time, the TD3-based EMS outperforms the DDPG-based EMS in the results when considering battery aging since the TD3 algorithm is an improvement of the DDPG algorithm, which reduces the overestimation bias of values in DDPG networks. The results demonstrate that the TD3-based EMS, considering the battery, can extend the battery life while the fuel consumption is slightly increased.

Also, it is trained on the road cycle parts and high-speed cycle parts of the China-World Transient Vehicle Cycle (C-WTVC) driving cycle to verify the applicability of the TD3-based EMS. The velocity of the C-WTVC is shown in Figure 18. The C-WTVC driving cycle, lasts for 900 s, with a top velocity of 87.8 km/h.

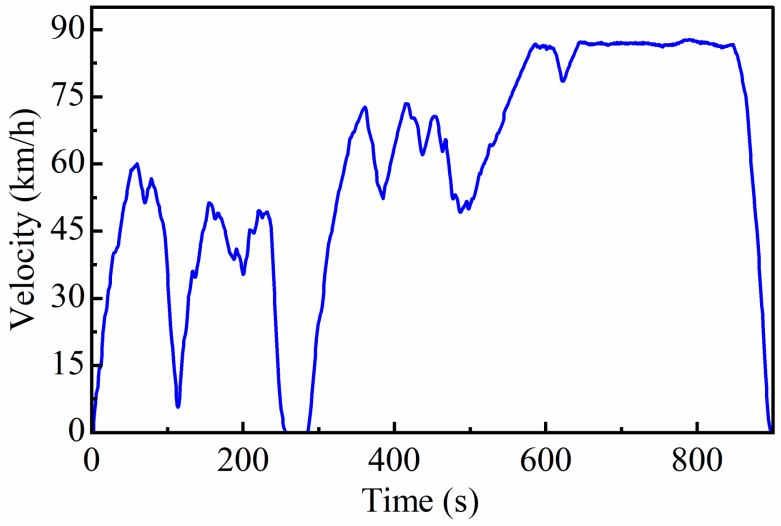

Figure 18. Velocity of the C-WTVC driving cycle.

The reward and SOC for the C-WTVC driving cycle are shown in Figure 19. It can be seen that when the reward remains stable and converges, it means that the TD3-based EMS is trained successfully and the optimal EMS is explored. The SOC also returns to its initial value of around 0.6, which indicates the adaptability of the TD3-based EMS that can be applied to different driving cycles.

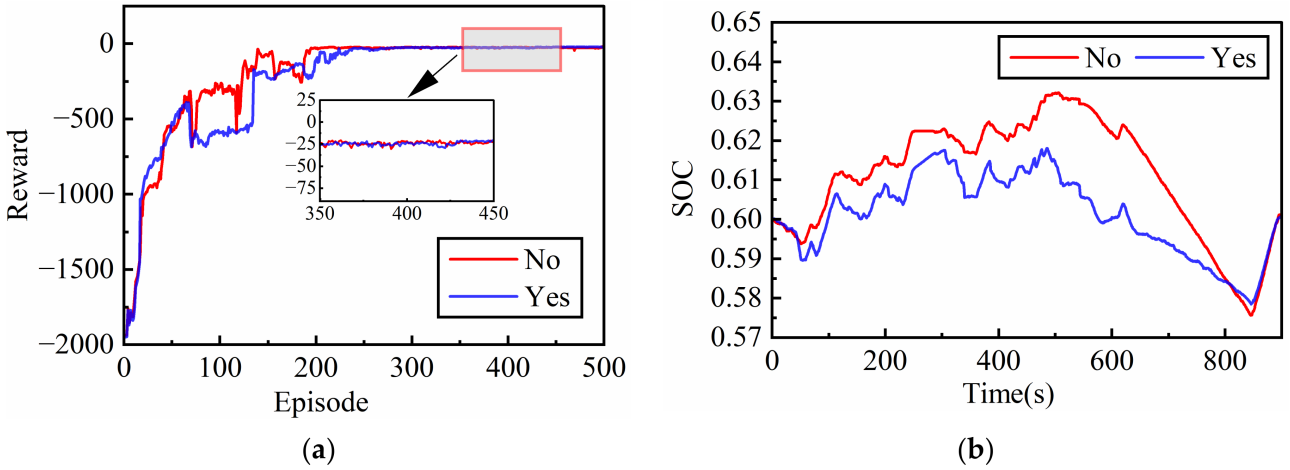

Figure 19. Reward and SOC. (**a**) Rewards for the episode. (**b**) SOC under the C-WTVC driving cycle.

The fuel consumption, battery capacity loss, and cost under the C-WTVC driving cycle are shown in Table 5. It can be seen that the TD3-based EMS without considering the battery aging has a cost of 237.09 CNY, and the battery cost accounts for 8.47% of the total cost. When battery aging is taken into account, the cost of the TD3-based EMS is 231.18 CNY and the cost of the battery is 5.60% of the total cost. Moreover, the battery life is extended by 35.49% when considering battery aging. Due to the high-velocity bias of the C-WTVC driving cycle, EMS prefers to use ICE in high-speed situations, resulting in high fuel consumption. Compared to the CHTC-D driving cycle, the battery is not used as often, resulting in relatively less battery capacity loss and lower battery cost as a percentage of the total cost. However, the EMS considering battery aging can still improve battery life and reduce the total cost.

Table 5. Fuel consumption, battery capacity loss, and cost under C-WTVC driving cycle.

EMS	Consider Battery Aging	F.C. (L/100 km)	F.C. Cost (CNY)	Battery Capacity Loss (%)	Battery Aging Cost (CNY)	Total Cost (CNY)	Performance
TD3-based	No	30.14	217.01	0.0293	20.08	237.09	97.51%
	Yes	30.31	218.23	0.0189	12.95	231.18	100%

5. Conclusions

In this study, a DRL-based EMS, namely the TD3-based EMS, is proposed to extend the battery life and reduced usage cost for hybrid electric dump trucks. The TD3-based EMS utilizes neural networks and the AC framework for continuous action control with continuous state space. An optimized objective function including battery aging cost and fuel consumption coat is established, and the TD3-based EMS is designed in this study. Finally, the EMSs are simulated and analyzed under the CHTC-D driving cycle and C-WTVC driving cycle.

The results show that the TD3-based EMS has a strong ability to adapt to the energy management problem of hybrid electric dump trucks through the self-learning capability of DRL. EMSs with action masks filter out invalid actions in the exploration stage, which avoids unsafe exploration and improves training efficiency. It extends battery life and slightly increases fuel consumption when the DRL-based EMSs consider battery aging. Moreover, TD3-based EMS performs better than DDPG-based EMS. Finally, the best-performing strategy is the TD3-based EMS considering battery aging. The proposed EMS is compared to the TD3-based EMS without considering battery aging. The battery life is extended by 36.17% under CHTC-D and 35.49% under C-WTVC, and the total cost is reduced by 4.30% and 2.49% when the EMS considers battery aging. In addition, TD3-based EMS can output actions based on the real-time state of the vehicle, and action masks can avoid invalid actions, which is suitable for industrial applications.

There are two things that should be done in the future. One is to improve learning efficiency and the other is to test on real vehicles. The contribution of this study is that TD3-based EMS using an action mask and considering battery aging, the action mask can filter out invalid actions and the EMS considering battery aging reduces the total usage cost and extend battery life. The TD3-based EMS gives a reference for real-time applications.

Author Contributions: Conceptualization, J.M. and W.H.; methodology, J.M. and R.Y.; software, J.M.; validation, J.M., S.Z. and W.H.; formal analysis, J.M. and Y.Z.; investigation, Y.Z.; resources, S.Z.; data curation, J.M.; writing—original draft preparation, J.M.; writing—review and editing, R.Y. and W.H.; visualization, J.M. and Y.Z; supervision, W.H.; project administration, W.H.; funding acquisition, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangxi Science and Technology Plan (AA22068062 and AA22068061).

Data Availability Statement: Not applicable.

Conflicts of Interest: Song Zhang is employee of Guangxi Yuchai Machinery Company Limited. The paper reflects the views of the scientists, and not the company.

References

1. Ali, A.; Söffker, D. Towards Optimal Power Management of Hybrid Electric Vehicles in Real-Time: A Review on Methods, Challenges, and State-Of-The-Art Solutions. *Energies* **2018**, *11*, 476. [CrossRef]
2. Saiteja, P.; Ashok, B. Critical Review on Structural Architecture, Energy Control Strategies and Development Process towards Optimal Energy Management in Hybrid Vehicles. *Renew. Sust. Energ. Rev.* **2022**, *157*, 112038. [CrossRef]
3. Tran, D.; Vafaeipour, M.; El Baghdadi, M.; Barrero, R.; Van Mierlo, J.; Hegazy, O. Thorough State-of-the-Art Analysis of Electric and Hybrid Vehicle Powertrains: Topologies and Integrated Energy Management Strategies. *Renew. Sust. Energ. Rev.* **2020**, *119*, 109596. [CrossRef]

4. Padmarajan, B.; McGordon, A.; Jennings, P. Blended Rule-Based Energy Management for PHEV: System Structure and Strategy. *IEEE Trans. Veh. Technol.* **2016**, *65*, 8757–8762. [CrossRef]
5. Zhou, W.; Yang, L.; Cai, Y.; Ying, T. Dynamic Programming for New Energy Vehicles Based on Their Work Modes Part I: Electric Vehicles and Hybrid Electric Vehicles. *J. Power Sources* **2018**, *406*, 151–166. [CrossRef]
6. Rezaei, A.; Burl, J.; Zhou, B.; Rezaei, M. A New Real-Time Optimal Energy Management Strategy for Parallel Hybrid Electric Vehicles. *IEEE Trans. Control Syst. Technol.* **2019**, *27*, 830–837. [CrossRef]
7. East, S.; Cannon, M. Scenario Model Predictive Control for Data-Based Energy Management in Plug-In Hybrid Electric Vehicles. *IEEE Trans. Control Syst. Technol.* **2022**, *30*, 2522–2533. [CrossRef]
8. Yu, P.; Li, M.; Wang, Y.; Chen, Z. Fuel Cell Hybrid Electric Vehicles: A Review of Topologies and Energy Management Strategies. *World Electr. Veh. J.* **2022**, *13*, 172. [CrossRef]
9. Zhang, F.; Wang, L.; Coskun, S.; Pang, H.; Cui, Y.; Xi, J. Energy Management Strategies for Hybrid Electric Vehicles: Review, Classification, Comparison, and Outlook. *Energies* **2020**, *13*, 3352. [CrossRef]
10. Hu, Y.; Li, W.; Xu, K.; Zahid, T.; Qin, F.; Li, C. Energy Management Strategy for a Hybrid Electric Vehicle Based on Deep Reinforcement Learning. *Appl. Sci.* **2018**, *8*, 187. [CrossRef]
11. Zou, Y.; Liu, T.; Liu, D.; Sun, F. Reinforcement Learning-Based Real-Time Energy Management for a Hybrid Tracked Vehicle. *Appl. Energy* **2016**, *171*, 372–382. [CrossRef]
12. Xiong, R.; Cao, J.; Yu, Q. Reinforcement Learning-Based Real-Time Power Management for Hybrid Energy Storage System in the Plug-in Hybrid Electric Vehicle. *Appl. Energy* **2018**, *211*, 538–548. [CrossRef]
13. Liu, T.; Zou, Y.; Liu, D.; Sun, F. Reinforcement Learning of Adaptive Energy Management with Transition Probability for a Hybrid Electric Tracked Vehicle. *IEEE Trans. Ind. Electron.* **2015**, *62*, 7837–7846. [CrossRef]
14. Li, Y.; He, H.; Peng, J.; Wang, H. Deep Reinforcement Learning-Based Energy Management for a Series Hybrid Electric Vehicle Enabled by History Cumulative Trip Information. *IEEE Trans. Veh. Technol.* **2019**, *68*, 7416–7430. [CrossRef]
15. Wu, J.; He, H.; Peng, J.; Li, Y.; Li, Z. Continuous Reinforcement Learning of Energy Management with Deep Q Network for a Power Split Hybrid Electric Bus. *Appl. Energy* **2018**, *222*, 799–811. [CrossRef]
16. Han, X.; He, H.; Wu, J.; Peng, J.; Li, Y. Energy Management Based on Reinforcement Learning with Double Deep Q-Learning for a Hybrid Electric Tracked Vehicle. *Appl. Energy* **2019**, *254*, 113708. [CrossRef]
17. Li, Y.; He, H.; Khajepour, A.; Wang, H.; Peng, J. Energy Management for a Power-Split Hybrid Electric Bus via Deep Reinforcement Learning with Terrain Information. *Appl. Energy* **2019**, *255*, 113762. [CrossRef]
18. Tan, H.; Zhang, H.; Peng, J.; Jiang, Z.; Wu, Y. Energy Management of Hybrid Electric Bus Based on Deep Reinforcement Learning in Continuous State and Action Space. *Energy Conv. Manag.* **2019**, *195*, 548–560. [CrossRef]
19. 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. *Appl. Energy* **2019**, *247*, 454–466. [CrossRef]
20. Zhou, J.; Xue, S.; Xue, Y.; Liao, Y.; Liu, J.; Zhao, W. A Novel Energy Management Strategy of Hybrid Electric Vehicle via an Improved TD3 Deep Reinforcement Learning. *Energy* **2021**, *224*, 120118. [CrossRef]
21. Li, T.; Cui, W.; Cui, N. Soft Actor-Critic Algorithm-Based Energy Management Strategy for Plug-In Hybrid Electric Vehicle. *World Electr. Veh. J.* **2022**, *13*, 193. [CrossRef]
22. Cheng, Y.; Xu, G.; Chen, Q. Research on Energy Management Strategy of Electric Vehicle Hybrid System Based on Reinforcement Learning. *Electronics* **2022**, *11*, 1933. [CrossRef]
23. Wang, J.; Liu, P.; Hicks-Garner, J.; Sherman, E.; Soukiazian, S.; Verbrugge, M.; Tataria, H.; Musser, J.; Finamore, P. Cycle-Life Model for Graphite-LiFePO4 Cells. *J. Power Sources* **2011**, *196*, 3942–3948. [CrossRef]
24. Tang, L.; Rizzoni, G.; Onori, S. Energy management strategy for HEVs including battery aging optimization. *IEEE Trans. Transp. Electrif.* **2015**, *1*, 211–222. [CrossRef]
25. Xu, D.; Cui, Y.; Ye, J.; Cha, S.W.; Li, A.; Zheng, C. A Soft Actor-Critic-Based Energy Management Strategy for Electric Vehicles with Hybrid Energy Storage Systems. *J. Power Sources* **2022**, *524*, 231099. [CrossRef]
26. Fujimoto, S.; Hoof, H.; Meger, D. Addressing Function Approximation Error in Actor-Critic Methods. In Proceedings of the PMLR/35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.
27. Zhou, G.; Huang, F.; Liu, W.; Zhao, C.; Xiang, Y.; Wei, H. Comprehensive Control Strategy of Fuel Consumption and Emissions Incorporating the Catalyst Temperature for PHEVs Based on DRL. *Energies* **2022**, *15*, 7523. [CrossRef]
28. Nam, H.; Kim, Y.; Bae, J.; Lee, J. GateRL: Automated Circuit Design Framework of CMOS Logic Gates Using Reinforcement Learning. *Electronics* **2021**, *10*, 1032. [CrossRef]
29. Wu, Y.; Tseng, B.; Rasmussen, C. Improving Sample-Efficiency in Reinforcement Learning for Dialogue Systems by Using Trainable-Action-Mask. In Proceedings of the ICASSP/2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 1 May 2020; pp. 8024–8028.
30. Tang, C.; Liu, C.; Chen, W.; You, S.D. Implementing Action Mask in Proximal Policy Optimization (PPO) Algorithm. *ICT Express* **2020**, *6*, 200–203. [CrossRef]