This is a repository copy of *Power allocation strategy for urban rail HESS based on deep reinforcement learning sequential decision optimization*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/194308/

Version: Accepted Version

Article:

Wang, X., Luo, Y., Qin, B. et al. (1 more author) (2023) Power allocation strategy for urban rail HESS based on deep reinforcement learning sequential decision optimization. IEEE Transactions on Transportation Electrification, 9 (2). pp. 2693-2710. ISSN 2332-7782

https://doi.org/10.1109/tte.2022.3227900

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works. Reproduced in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Power allocation strategy for urban rail HESS based on deep reinforcement learning sequential decision optimization

XIN WANG[1,2], YINGBING LUO[1,2], BIN QIN*[1,2], and LINGZHONG GUO[3], *Member, IEEE*

(1 School of Electrical & Information Engineering, Hunan University of Technology, Zhuzhou, Hunan, 412007, China)

(2 Hunan Engineering Research Center of Electric Drive and Regenerative Energy Storage Utilization, Zhuzhou, Hunan, 412007, China)

(3 Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, S1 3JD, UK)

*Corresponding author: BIN QIN (qinbin99p@163.com)

Abstract: A hybrid energy storage system (HESS) is adopted to tackle the traction network voltage fluctuation problem caused by high power and large energy demand during the starting and braking of urban rail trains. The system is composed of on-board ultracapacitors and ground lithium batteries, aiming to smooth out the power fluctuation to realize "peak-shaving and valley-filling". Based on deep reinforcement learning (DRL) online sequence decision, a dynamic power allocation strategy is proposed to improve the energy-saving and voltage stabilization of DC traction networks as well as HESS life protection. Furthermore, to enhance the DRL's efficiency under time-varying operating conditions, an annealing bias - priority experience replay twin delayed deep deterministic policy gradient algorithm (A-TD3) is proposed to train the replay buffer in DRL. The online learning and optimization strategy is implemented via the mechanism of "trial and error" and "feedback" of the agent. RT-LAB semi-physical real-time simulation systems are adopted to verify the effectiveness of the proposed strategy. Compared with the traditional rule-based control, filter-based control and DRL method, the results show that the proposed method converges faster and is more energy saving and stable while effectively protecting the HESS.

Keywords: Regenerative braking energy; Hybrid energy storage system (HESS); Power dynamic allocation; Deep reinforcement learning; HESS protection

1 Introduction

In recent years, urban rail trains have become a major means of transporting people and goods in the urban as well as suburban areas, and as such contribute greatly to the reduction of traffic and environmental pollution such as CO_2 emission [1-3]. With the steady progress of urbanization and the growth of population worldwide, currently there is a fast development of urban rail transport systems, which leads to high increase of energy consumption [4]. Due to the huge energy consumption and high operating voltage requirements of urban rail lines [5], recovering excess braking energy in the form of energy storage system (ESS) becomes an effective means to achieve energy savings and ensure the voltage safety of the traction network [6][7]. As the ESS with a single type of device cannot meet the dual demands of high power and large energy, a HESS is often used to suppress the sharp rise in traction network voltage [8][9]. In the urban rail transit, the hybrid energy storage system (HESS) consisting of on-board ultracapacitor/supercapacitor and ground lithium battery mainly uses the "peak-shaving and valley-filling" strategy for the traction power network to achieve energy saving and voltage stabilization [10-12]. Given that the energy allocation among each ESS depends on their control strategies, how to improve the braking energy utilization and achieve a better voltage stabilization and energy saving effectiveness is one of the key technical issues in HESS management strategies that need to be addressed.

In rule-based HESS energy management strategies, it is often to use the given voltage or power as reference, for example, rule-based PI control [13] and filter-based control [14][15]. The former has the advantages of being easy to implement and deal with the power constraints of the battery and ultracapacitor while the latter is able to readily achieve power allocation through frequency decomposition. However, neither of these methods takes into consideration the system optimization, because the main problem is traction energy-saving and network stabilization. Fuzzy control strategies [16][17] can design fuzzy rules to reach above requirements, where the design is not an easy task. A power threshold strategy with power quality index constraint is proposed in [18], which effectively reduce the system-side negative sequence current and improving the average power factor, but the ESS protection problem [19] may need consider under frequent braking conditions.

Another class of HESS power management strategies is optimization-based methods. A model predictive control (MPC) method combined neural networks, was proposed in [20]. The proposed method can effectively lower the requirements of prediction accuracy and step size, and at the same time control the ESSs (State of Charge) SOC within the desired range, but the method may have high real-time control requirement. In [21], a dynamic programming method was used to optimize the offline HESS power allocation, which can guide the online control optimization. The method requires to predict the state transition probability in advance, which hinders the use in practice.

From the above discussions, it can be concluded that although the recent progress of the HESS management strategies, there remains some challenges in the energy-saving and voltage stabilization of DC traction networks of rails. The main challenge comes from the uncertainties in the system caused by many factors such as voltage fluctuations, the degradation of the battery, and driving patterns etc. The first two factors are related to the designing strategies, which can be used in real-time to minimize the power consumption, optimize power distribution between different power elements and prolonging battery/ultracapacitor life under these uncertain environments. It is worth pointing out that the driving patterns in urban rail transit are normally set in advance, therefore not the focus of this paper. Reinforcement learning (RL) method has been shown capable of guaranteeing real-time and robust performance in decision making [22], which has the potential to provide a robust and optimal solution to the power allocation problem. Furthermore, taking into consideration the randomness of the state of the rail transit system under studied, such as the speed and acceleration of the train, the power consumptions in different driving modes, the states of the charge of the ultracapacitor and battery, the problem in this study perfectly fits Markov Decision Process (MDP) framework, which is the basis of the reinforcement learning (RL).

In the urban rail power supply environment, the cycle characteristics of battery have great randomness. For example, the internal resistance of a battery can be significantly increased as the temperature drops or the battery degrades, resulting in an increase in the discharge current and discharge rate [23]. This parameters variation could significantly affect the performance of controller. Therefore, it is necessary to design an adaptive controller that can adapt to battery decay. Reinforcement learning (RL) has been presented to implement and optimize the energy management strategies recently. It provides a mathematical framework for discovering or learning strategies that map situations onto actions with the goal of maximizing a reward function [22]. In contrast to prior work, which requires deterministic or stochastic knowledge of underlying systems, the advantage of RL is that it can adapt to complex and changing environments and carry out strategy learning without knowing accurate system models and the state transition probability. The RL for energy management has been studied by some researchers [24][25]. In [24], a RL-based real time management strategy for electric vehicles (EV) was proposed and simulation verification of the improved RL under different loads was performed. Moreover, a novel approach was proposed combining a rule-based controller with RL to achieve an adaptive optimal solution. For hybrid electric vehicles (HEV), a RL algorithm was proposed to minimize the HEV fuel consumption over any driving cycle without prior knowledge of the cycle [25]. The results from these studies indicate that the RL-based energy management strategy can considerably improve fuel and battery efficiency and allow real-time implementation. Motivated by these RL-based studies for energy management, we aim to apply DRL techniques [26] for the HESS power allocation optimization in urban rail transit system in this study.

Currently, the twin-delayed deep deterministic policy gradient (TD3) is the more popular algorithm in DRL [27][28], which has been successfully applied in many different areas [29-34]. Given the time-varying characteristics in urban rail transit systems, we propose a modified TD3 algorithm by introducing a priority experience sampling strategy, termed A-TD3 (Annealing bias-priority experience replay twin delayed deep deterministic policy gradient algorithm). The proposed method is used to train the replay buffer in TD3 for the purpose of improving the training efficiency and accuracy. The online learning and optimization strategy will be implemented via the mechanism of "trial and error" and "feedback" of the agents. To the best of our knowledge, combing TD3 with annealing bias - priority experience replay in the power allocation strategies for urban rail HESS in real time taking into consideration onboard capacitor life protection has not been investigated. The main contributions of the paper are as follows:

(1) To enhance the performance of the online optimization and adaptive power compensation of RL algorithms in urban rail power supply environment, an improved TD3 algorithm, namely A-TD3, is proposed, which adopts an annealing bias - priority empirical sampling strategy. Compared with the previous methods, it can effectively save energy, stabilize voltage, protect HESS

(2) The proposed A-TD3 algorithm overcomes the limitations of traditional DDPG and TD3 algorithms by determining different priority probabilities through a combination of priority experience replay and annealing bias - importance sampling, which can eliminate the deviation under the distribution change and has the advantages of fast convergence and being not easy to fall into local optima.

(3) An off-line training - online optimization - online sequential decision method is designed to solve the stochastic control problem under initial conditions and save computing resources. The proposed strategy is validated on a RTLAB hardware-in-the-loop real-time simulation system and the experimental results verify its effectiveness.

The rest of the paper is organized as follows: Section 2 describes in detail the metro rail power supply model and the proposed control strategy. In Section 3, the A-TD3 is presented. In Section 4 the proposed control strategy is verified by simulations and semi-physical real-time simulation experiments. Finally, conclusions are drawn in Section 5.

2 Traction power supply structure

2.1 System components

The structure of the main power supply system of the urban rail studied is shown in Figure 1.

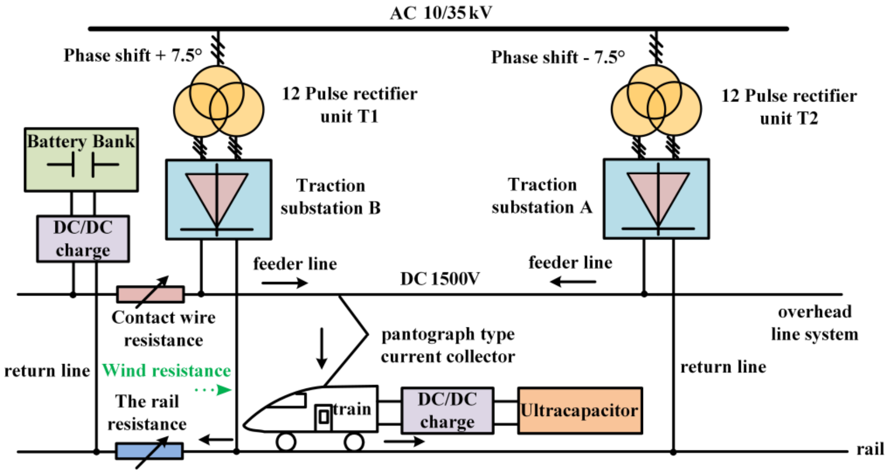


Figure 1 The urban rail power supply system

The traction network is stepped down from 110kV / 220kV power system to 10kV / 35kV AC, and then obtains 1500V DC bus voltage by rectifier, which provides power supply in both direction of the trains. In order to reduce the ripple coefficient of the output voltage and the harmonic current of the traction power supply system, 24-pulse rectifier is used to suppress the generation of harmonics. The 24-pulse rectifier is operated in parallel by two sets of 12-pulse rectifier units, where the windings on the high voltage side of the transformer are phase-shifted by ±7.5°. The train uses the rotor magnetic field vector control to realize the SVPWM control of the inverter, thereby controlling the operation of the train. The parameters of the train are listed in Table 1.

Table 1 Parameters of the train

Parameters	Conditions/Values
Train formation	3M1T
Rated load	AW0: 129.9t, AW2: 204.3t, AW3: 228.3t
Highest operation speed v	80km/h
Average acceleration a_c	⩾1.0m/s^2 (v~[0,40km/h])
	⩾0.6m/s^2 (v~[0,80km/h])

Basic resistance per unit mass of train (N/t)	r=1.1064+0.0295v+0.000248v^2
Running resistance F_m (N/t)	F_m=2340.9+62.42v+0.525v^2
Rated voltage of DC traction network	1500V
Voltage standard range	1000~1800V
Wheel diameter (semi-wear condition)	805mm
Transmission ratio	6.68

The HESS, which is composed of on-board ultracapacitor pack and ground battery pack, is connected to the DC traction network through bi-directional DC/DC converter. When the train is electrically braked, the traction motor acts as a generator to deliver energy to the traction network while the HESS absorbs the excess braking energy; when the train starts to accelerate, the HESS provides the required energy for the traction motor together with the traction network.

2.2 Energy management strategy for HESS

The energy management strategy for the HESS is shown in Figure 2. The main sections are as follows: i) the coordinated control of permanent magnet traction system (PMSM) and HESS; ii) the dynamic power allocation strategy based on A-TD3 sequential decision optimization. The system will ultimately achieve optimal energy saving and voltage stabilization between the permanent magnet traction system and the HESS.

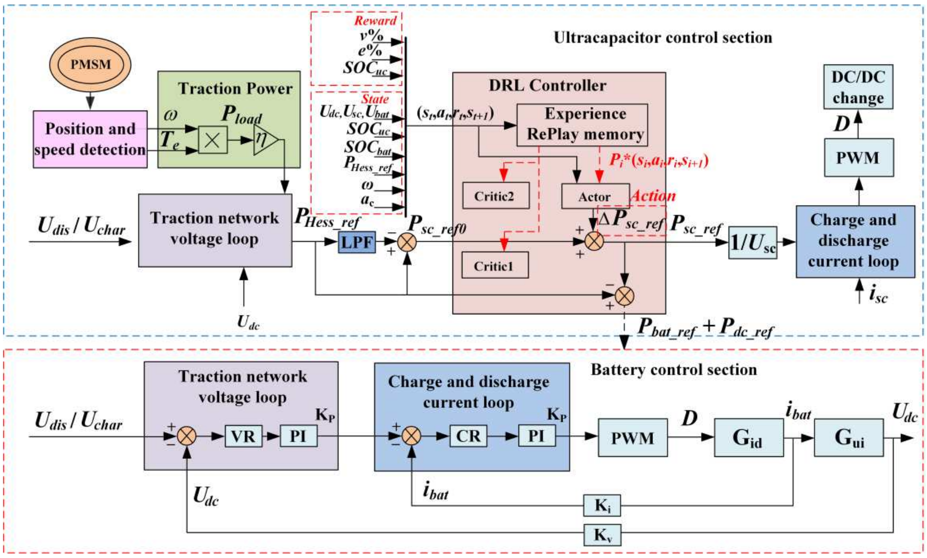


Figure 2 Control strategy of HESS

In the HESS, the basic controller of on-board ultracapacitor is the filter-based traction power feedforward control [36][37] and the basic controller of ground battery is the rule-based voltage PI control. The former is mainly making use of their fast response characteristics to accomplish rapid acceleration of the traction system and braking energy exchange while the latter is stabilizing the DC traction voltage and providing the auxiliary power supply. The control of ultracapacitor will be detailed in section 2.3. In the battery control, the error between the given charging and discharging voltages threshold U_{char} / U_{dis} and the real-time feedback traction network voltage U_{dc} will be fed into a PI controller and compared with the feedback battery current $i_{bat.}$ The duty cycle D of the driving BDC switch device is finally obtained through PWM control by a second PI controller, more details as shown in [35]. Therefore, the control rules as follows: 1) when U_{dc} is lower than U_{dis}, the HESS discharging; 2) when U_{dc} is higher than U_{char}, the HESS charging.

2.3 *Coordinated control strategy between PMSM and HESS*

The coordinated control strategy of the HESS based on traction power feedforward is shown in Figure 3. The output action ΔP_{sc_ref} from A-TD3 sequential decision is used to adjust the given ultracapacitor power in real time, which can realize the dynamic power allocation between the HESSs. Finally, the given ultracapacitor power P_{sc_ref}, the given on-board battery power P_{bat_ref}, and the grid P_{dc_ref} are obtained. In Figure 3, i_{sc_ref} is the current allocated to the on-board ultracapacitor. i_{sc} is the output current of ultracapacitor. The energy allocation of each ESS is obtained through the coordinated control of the actual train traction power.

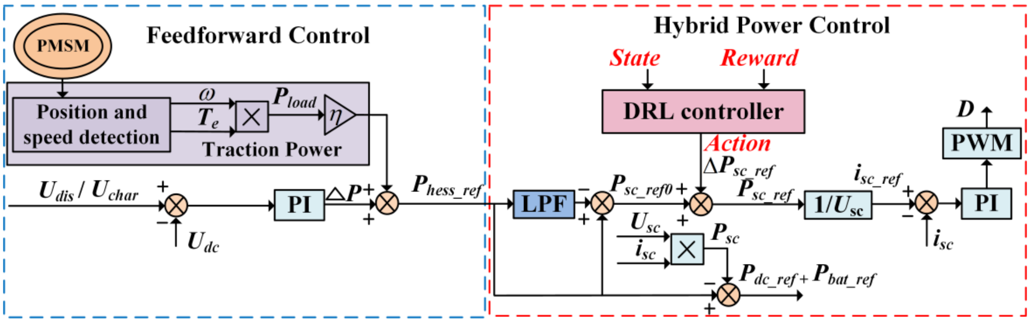


Figure 3 Coordinated control strategy between PMSM and HESS

3 Improved deep reinforcement learning (DRL)

3.1 Reinforcement learning (RL)

For the nonconvex nonlinear optimization model, the gradient-based optimization algorithm cannot guarantee the global optimal while the gradientless heuristic algorithm is slow in searching the solution. The goal of RL is to find an optimal strategy to maximize the expected return value [38], and automatically learn the global optimal decision through trial and error and feedback based on real-time action and environmental information feedback.

RL does not need labelled data, and its learning control process is based on continuous-times Markov Decision Process (MDP) by taking into consideration both Action information (Action) and Environment information (Environment). The interaction between Actions and Environment leads to the transition of the Environment from one state to another while a Reward is given based on the new state of the Environment. RL can evaluate the quality of an action sequence according to the reward information. Because of the feedback delay of reward and the real-time change of environmental information, RL constantly analyzes and learns useful information from the reward sequence.

In the actor-critic framework, which integrates a value function estimation algorithm and a policy search algorithm, the policy network (actor) performs network updates via deterministic policy gradients [38]:

$$\nabla_{\phi} j(\phi) = E_{s-p_{\tau}}[\nabla_{a} Q^{\pi}(s,a)\,|_{a=\pi(s)}\, \nabla_{\phi}\pi_{\phi}(s)] \tag{1}$$

Where $Q^{\pi}(s,a) = E_{s-p_{\tau}}[R_t \mid s,a]$ is the action value function, which represents the expected return value after action a is taken in state s under the policy π. It takes advantage of the algorithm in continuous action space and changes the stochastic policy into a deterministic policy, as shown in Eq. (2) [39]:

$$a = \pi[s_t \mid \theta^{\pi}] \tag{2}$$

3.2 Annealing bias- priority experience replay-based TD3

TD3 is a DRL algorithm based on the actor-critic framework [28], which is developed on the basis of DDPG and can solve the problem of *Q* overestimation in actor-critic framework algorithm. In TD3, the experience replay mechanism can eliminate the correlation between data samples and at the same time improve the utilization rate of the sample. Note that the experience in the buffer is history data (s_t, a_t, r, s_{t+1}), which is basically sampled uniformly during learning. Some generated neighboring samples through the Agent experience of interaction with the environment are strongly correlated with the data itself. Moreover, different data's contribution to the gradient learning may be different. All these will lead to low learning efficiency and even over-fitting.

Traditional experience priority replay is based on $|TD\text{ - }error|$ method [40], which is the difference between the current action value function *Q*' and the target value function *Q* in sequential difference, as shown in Eq. (3). It can effectively improve the utilization rate of samples. Training goal is to make $|TD\text{ - }error|$ expectations as small as possible, therefore it often uses its value to determine the prioritization $rank_i$ and the priority indicator p_i, as shown in Eq. (4):

$$|TD-error| = r(s_t,a_t)+\gamma\max_{a_{t+1}} Q'(s_{t+1},a_{t+1})-Q(s_t,a_t) \tag{3}$$

$$p_i = |\frac{1}{rank_i}| \tag{4}$$

Where $rank_i$ indicates the ranking starts from the i^{th} experience to the smallest.

Thus, the sampling probability is determined as:

$$P(i) = |\frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}}| \tag{5}$$

Where α is the variable priority factor which does not alter the monotonicity priority, but rather adjusts the priority of |TD-error|. α=0 indicates uniform random sampling whereas α=1 indicates greedy strategy sampling. $\alpha \in [0,1]$ and k is batch quantity.

However, this method is liable to an excessively high frequency of access to those experiences with a relatively high |TD-error|, which may easily cause overfitting due to lack of diversity of samples.

At the same time, the priority experience replay also changes the status distribution and bias is bound to be introduced, so in this paper bias annealing-importance sampling [40] is introduced to eliminate the bias and also reduce the gradient magnitude. The importance sampling weight ω is expressed as follows:

$$\omega_i = (\frac{1}{N} \cdot \frac{1}{P(i)})^{\beta} \tag{6}$$

In addition, the range of weight ω_{i} is standardized for the sake of training stability, such as $\omega_i / \max_i \omega_i$. Finally, the sampling probability is:

$$P(i)^* = \frac{\omega_i}{\max_i \omega_i} \cdot P(i) = (\frac{1}{N} \cdot \frac{1}{P(i)})^{\beta} \cdot P(i) \tag{7}$$

In the above two equations, N is the number of samples in the experience buffer, and β is the annealing factor, which is used to determine the influence of priority experience replay on the convergence results. When β=0, it represents full importance sampling; When β=1, ω_{i} and $P(i)$ are exactly canceled, which means that the influence of experience replays on the convergence results is completely eliminated. At this time, uniform random sampling is then used. At the beginning and end of the actual training, β anneals linearly from the initial value 0 to 1. α and β have an interactive effect, and both parameters simultaneously determine the priority.

When initializing the network model, all samples in the replay buffer are initialized with an immediate reward value of 0. During training, experience samples in the replay buffer are selected with the probability of $P(\mathrm{t})^*$. The algorithm training process and network block diagram of ATD3 are shown in Appendix. Table 1 and Figure 4, respectively.

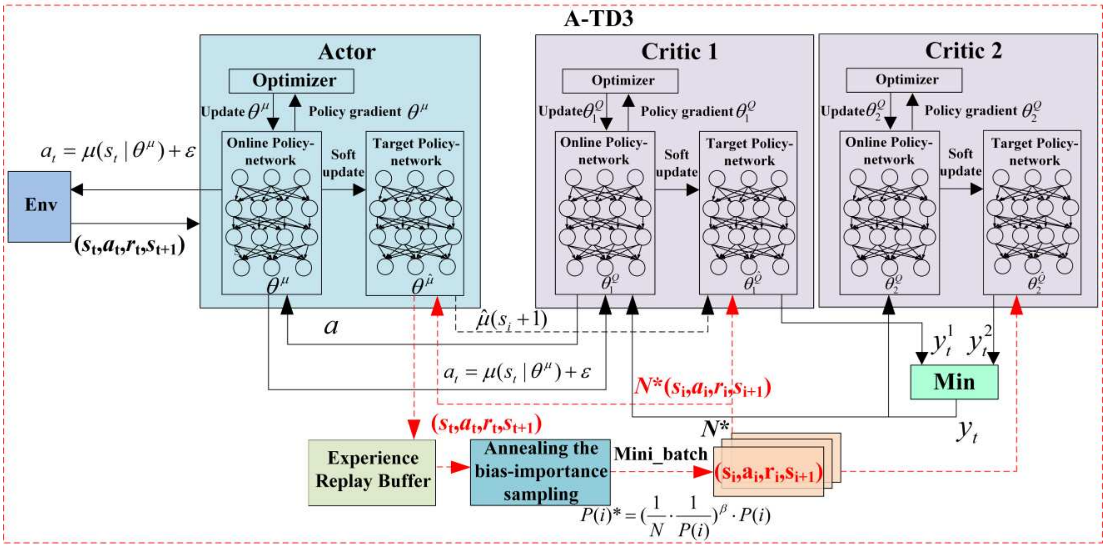


Figure 4 The schematic diagram of the proposed A-TD3 algorithm

3.3 PMSM power supply environment and state characterization

In the process of A-TD3 execution, at each step after the action is applied to the environment, the environment will feedback the updated state to the A-TD3 Agent. The DRL interactive environment in the proposed method is shown in Figure 5.

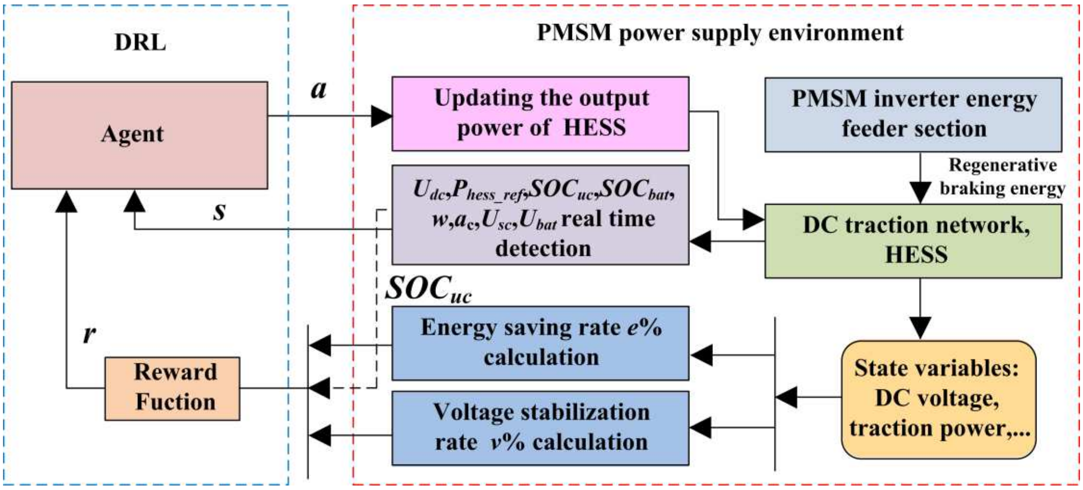

DRL
Agent
a
s
r
Reward Fuction
PMSM power supply environment
Updating the output power of HESS
PMSM inverter energy feeder section
Regenerative braking energy
U_{dc},P_{hess_ref},SOC_{uc},SOC_{bat}, w,a_c,U_{sc},U_{bat} real time detection
DC traction network, HESS
SOC_{uc}
Energy saving rate e% calculation
Voltage stabilization rate v% calculation
State variables: DC voltage, traction power,...

Figure 5 Agent-based decision making for permanent magnet traction power supply environment

In the proposed method, the HESS energy management system is considered as an Agent for learning and decision-making, and the permanent magnet traction power supply system is considered as the environment in which the Agent is located. The Agent senses the environment and its state changes, performing specific action that affects the state of the environment and cause the environment to generate the corresponding reward signals and adjust the policy based on the obtained feedback signal to maximize the cumulative reward over the time period considered. Due to the characteristics of instantaneous high power and short-term large energy when breaking the train, it is necessary to consider the life protection of the energy storage components, i.e., ultracapacitor and battery while realizing voltage stabilization and energy saving. Therefore, the DC traction voltage U_{dc}, the given hybrid energy storage power P_{hess_ref}, the SOC_{uc} of the ultracapacitor, the Voltage U_{sc} of the ultracapacitor, the SOC_{bat} of the battery, the Voltage U_{bat} of the battery, the motor speed ω and the acceleration a_c of the train are selected as the state of the environment that the Agent is to observe. It follows that the state space S of the system with n subsystems is expressed as:

$$\begin{aligned} S = [&U_{dc1}, P_{hess\_ref1}, SOC_{uc1}, U_{uc1}, SOC_{bat1}, U_{bat1}, \omega_1, a_{c1}, \ldots, \\ &U_{dcn}, P_{hess\_refn}, SOC_{ucn}, U_{ucn}, SOC_{batn}, U_{batn}, \omega_n, a_{cn}] \end{aligned} \tag{8}$$

3.4 Continuous action space selection and action execution

The redistribution action of the permanent magnet traction power selected by A-TD3 from the action space should be able to fully cover the feasible region of the HESS. In order to avoid the failure of A-TD3 to learn the global optimal decision of the HESS power allocation, the continuous action space should not be selected too small. However, if the selected action space is too large, the training efficiency of A-TD3 will deteriorate. In this study, the Agent action is selected as the given power adjustment value of the ultracapacitor ΔP_{sc_ref} and the policy function π is the mapping from the state space S to action space A (π : $S \rightarrow A$), that selects actions based on the observations from the environment. Considering the power fluctuation range of the ultracapacitor, the power demand adjustment is carried out in a continuous manner and the continuous action space A is obtained as shown in Eq. (9):

$$\begin{cases} A = [\Delta P_{sc\_ref1}, \Delta P_{sc\_ref2}, \ldots, \Delta P_{sc\_ref\mathrm{n}}] \\ |\Delta P_{sc\_ref}| < P_{hess\_ref} - P_{sc\_ref0} \end{cases} \tag{9}$$

Where P_{hess_ref} is the given HESS power, P_{sc_ref0} is the initial value of the given ultracapacitor power.

3.5 Reward function design

In each decision cycle, the Agent takes the action *a* under the state *s*, that is to select the proper given power to allocate the power for the HESS. The environment will then move to the next state. A reward *r* will be given to this action *a*, which represents the feedback from the environment to the Agent's action, and the learning goal of the Agent is to obtain the maximum cumulative reward. The variance of the reward value distribution should not be too large, otherwise it will increase the training time of A-TD3 and is easy to fall into local optima. On the other hand, if the variance of the reward value distribution is too small, A-TD3 may not be able to learn effectively. In order to improve the effect of energy saving, voltage regulation, and HESS protection, the reward *r* in this study is divided into two parts:

(1) Energy saving and voltage regulation part: r_1 is selected as the weighted sum of the coefficient of energy saving *e%* and the coefficient of voltage stabilizing *v%* within time step ΔT, as shown in Eq. (10):

$$r_1 = \mathrm{m}ax[\mu \cdot e\% + \lambda \cdot v\%] \tag{10}$$

Where μ is the weighting coefficient of *e%*, λ is the weighting coefficient of *v%*. The coefficient of energy saving *e%* is defined as the percentage of the change of the total output energy of the substation after the installation of the HESS in the total output energy of the substation without the energy storage system, and coefficient of voltage stabilizing *v%* is evaluated by the integration of the portion of the DC traction voltage above/below the limit [41], as shown in Eq. (11) and (12), respectively.

$$e\% = (1 - \frac{\int_0^T u_{dc}^{hess} i_{dc}^{hess} dt}{\int_0^T u_{dc}^{nohess} i_{dc}^{nohess} dt}) \tag{11}$$

Where $u_{dc}^{hess} / u_{dc}^{nohess}$ are the voltages of the DC traction network with/without the HESS installed, i_{dc}^{hess} $/ i_{dc}^{nohess}$ are the currents of DC traction network with/without the HESS installed.

$$v\% = (1 - \frac{\int_0^T \Delta h(u_{dc}^{hess} - u_{dc}^{refh}) + \Delta l(u_{dc}^{refl} - u_{dc}^{hess})dt}{\int_0^T \Delta h(u_{dc}^{nohess} - u_{dc}^{refh}) + \Delta l(u_{dc}^{refl} - u_{dc}^{nohess})dt}) \tag{12}$$

Where $u_{dc}^{refh} / u_{dc}^{refl}$ are the safety upper limit/safety lower limit of the DC traction network voltage, and $\Delta h/\Delta l$ are the times when the DC traction voltage exceeds the safety upper limit/lower limit under the operating conditions.

(2) HESS protection part: i) in ultracapacitor protection part, by setting the SOC_{uc} safety range to achieve ultracapacitor overcharge and overdischarge protection, and it is expected that SOC_{uc} can be kept within the safety range [0.15, 0.85]. ii) In battery protection part, the battery lifetime degradation is another important factor need to be taken into consideration in the HESS power allocation. The degradation of the battery is mainly characterized by the gradual decrease of the actual discharge capacity and discharge time as the increase of operation cycles. Combined with the actual operating environment of the urban rail systems and according to the C_r model of lithium iron phosphate battery capacity decay rate established by Swierczynski et al [42], it can be found that at 30 ℃, the capacity decay rates of the battery after 500, 1000, 1500, and 2000 cycles at a depth of discharge of 50% are about 5%, 8%, 11%, and 12.5%, respectively. Therefore, it is necessary to reduce the allocated power of the battery according to the capacity decay accordingly. In addition, due to uncertain factors such as differences in cell consistency and/or temperature, there could be rapid capacity decay caused by sudden battery discharge, which could shorten the battery life rapidly and reduce the system efficiency. Therefore, we take both the gradual decay and the sudden drop of the battery capacity into consideration by considering the following supplementary reward function term:

$$\begin{aligned} r_2 = \mathrm{m}ax[\sigma(0.15 \le SOC_{uc} \le 0.85) - \eta[(SOC_{uc-char} \ge 0.85) or (SOC_{uc-dis} \le 0.15)] \\ + \mathrm{m}ax[-\rho * C_r - \psi(C_{bat} - 1)] \end{aligned} \tag{13}$$

Where $SOC_{uc\text{-}char}$ and $SOC_{uc\text{-}dis}$ represent SOC values of the ultracapacitor in the state of charge/discharge; σ is the weighting coefficient of the full utilization of the ultracapacitor; η is the (penalty) weighting coefficient of the overcharge and overdischarge of the ultracapacitor; C_r is the capacity decay rate, which is related to the depth of battery discharge; C_{bat} is the discharge rate of the battery; ρ and ψ are the corresponding reward weights, respectively. When the capacity of battery undergoes gradual or sudden decline, the actual discharge rate and depth of discharge are reduced by gradually reducing the power allocated to the battery, so as to achieve the purpose of protecting the battery. Finally, the overall reward function is designed as follows:

$$r = \max(r_1 + r_2) \tag{14}$$

In the reward function, the reward is given as a single scalar value at each time instant. The Agent's goal is to maximize the accumulated reward, which is used to measure the level of progress in learning. In general, in order to stimulate the Agent to learn, the reward coefficients should be small while the penalty punishment coefficient should be large.

3.6 Off-line training - online optimization - online sequential decision making

The proposed sequential decision optimization framework for DRL is shown in Figure 6.

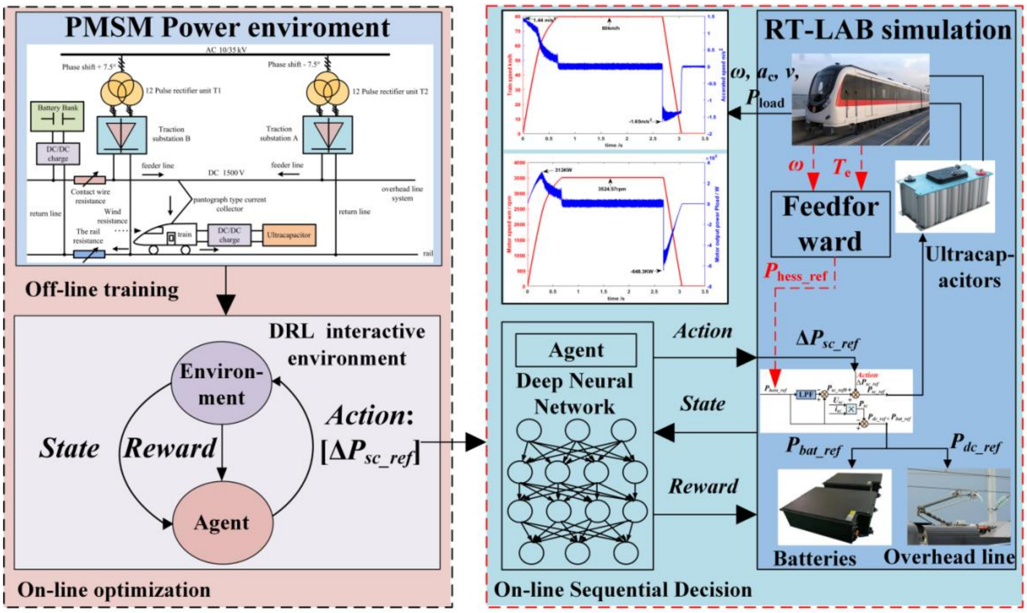


Figure 6 A-TD3 sequential decision optimization framework

(1) Offline model training - online optimization. Offline training is firstly performed on a simple urban rail plant model, taking the advantages of DRL training without accurate model and providing a good initial strategy for Agent to realize pre-training, which greatly saves the training time. A simulation platform of the rail transit traction power supply is then used as the environment for the

Agent. Preliminary control of the HESS is conducted through initial energy rule control to achieve a preliminary stable control effect (The on-board ultracapacitor is controlled by filter-based traction power feedforward control and the ground battery is controlled by rule-based PI control). At the same time, DRL is used as the supervision and supplement of the power allocation rule controller, and the guided exploration is used to make the Agent learn and progress quickly, which improves the dynamic adaptive ability of the controller and ensures the stability and the guided exploration effect under the condition of guaranteeing the benchmark control performance. The control block diagram is shown in Figure 7. In addition, the running speed of the train in each round is randomly changed within the safe operating range when the Agent is trained to improve the stability in the changing environment, particularly the gradual or abrupt change of the environment as well as achieve the stable application of the control strategy in the hardware-in-the-loop system.

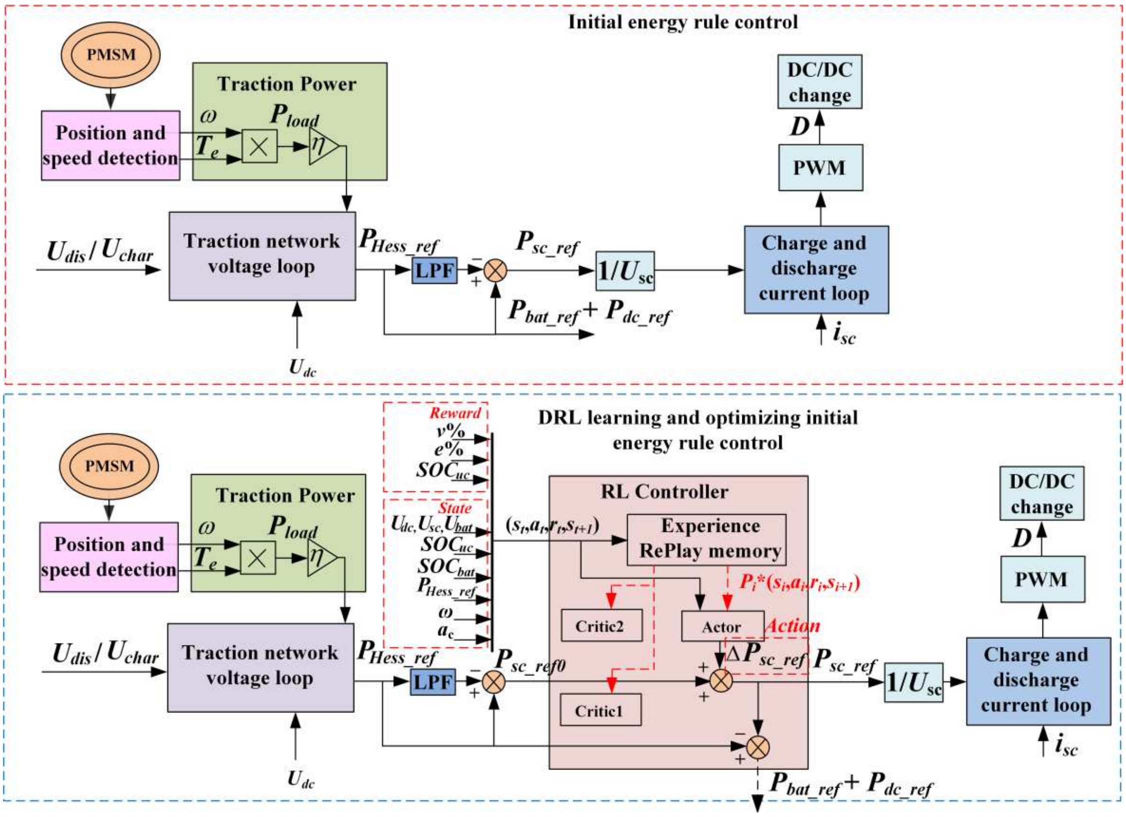


Figure 7 Initial energy rule control and DRL optimization

(2) Online sequential decision. RT-LAB semi-physical real-time simulation system is used to simulate real-time operating conditions. The Agent makes decisions and learns based on the real-time system state information. The combination of online training and online decision greatly shortens the

Agent exploration stage in the experiment and improves the learning efficiency of energy management algorithm. On the other hand, through the Agent offline training-online learning and decision making, online optimization of the control performance can be achieved in the presence of model errors and parameter changes in the system.

4. Simulation and results

4.1 Hyperparameter selection

The weighting coefficients in the reward function and the hyperparameter design in online learning are shown in Table 2. In order to minimize the impact of the selection of hyperparameters on the performance of different algorithms, the selected hyperparameters are general, and the hyperparameters of DDPG, TD3 and ATD3 are guaranteed to be consistent. Taking episodes in each training round as an example, the Agent takes a batch of 512 samples from the experience buffer (size 2e^6) for training through the Annealing bias - priority mechanism in the experience buffer, and improves the optimization efficiency by replacing random sampling strategy. In the training, the Agent adopts Gaussian noise model to smooth the update of the target strategy and an additional noise variance attenuation rate of α=1e^{-4} is added to increase the exploration capability at later stage. Finally, a discount factor γ=0.995 is introduced to increase the long-term awards.

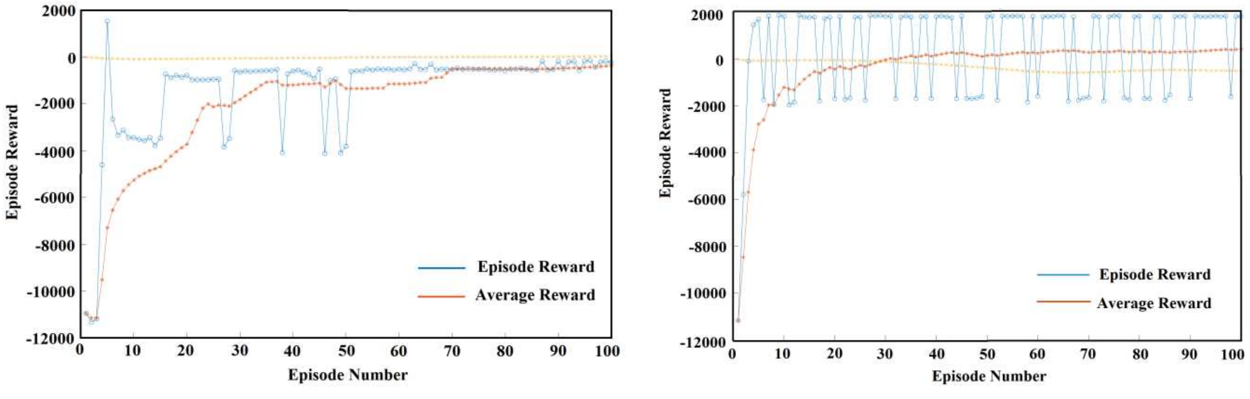
Table 2 Hyperparameter settings

symbol	parameter	The numerical	symbol	parameter	The numerical
-	Experience Buffer Length	2e^6	α	Initial value of variable priority factor	0.5
-	Mini Batch Size	512	β	Initial value of annealing factor	0
γ	Discount Factor	0.995	λ	weight coefficient 1	2
Var	Exploration / Policy Noise Variance	0.1	μ	weight coefficient 2	2

-	Exploration / Policy Noise Variance Decay Rate	1e^{-4}	σ	weight coefficient 3	10
-	Delay update frequency	2	η	weight coefficient 4	50
-	Critic learn Rate	0.0001	ρ	weight coefficient 5	10
-	Actor learn Rate	0.001	ψ	weight coefficient 6	50

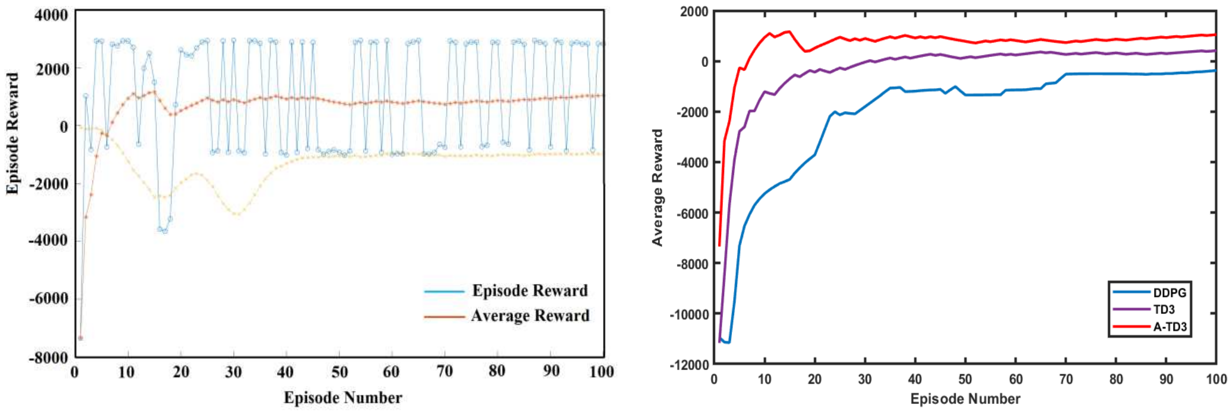
4.2 Online training

In order to verify the effectiveness of the proposed A-TD3 power dynamic allocation algorithm, TD3 algorithm and DDPG algorithm were trained online in the Matlab/Simulink simulation-based urban rail power supply model for the purpose of comparison. In order to save computing resources, we scaled the running time of the simulation model according to the actual working conditions at a time ratio of 1:42.85. Therefore, the running time in each episode round was set to be 3.5s until the training converges, and verification was still carried out according to the real running time in the RT-LAB real-time simulation system. The learning curve of each algorithm is shown in Figure 8 and the training times are shown in Table 3.



a. DDPG

b. TD3



c. A-TD3

d. Comparison

Figure 8 Agent learning process

Table 3 Training time

Symbol	Episodes to converge	Time of convergence (s)
DDPG	70	$1.7e^4$
TD3	60	$1.47e^4$
A-TD3	50	$1.16e^4$

It can be observed that the A-TD3 algorithm can make the system converge to a stable state more quickly and the average reward is much higher than the other two algorithms, improving the training efficiency and optimization accuracy of the agent.

4.3 Simulation result analysis

Based on the analysis in the previous section, a traction network and urban rail transit model was established. Since urban rail has the characteristics of uniform power distribution and uniform car speed, a permanent magnet traction motor can be selected to simulate the running process of the train under the condition of saving computing resources as much as possible (achieved by scaling overall traction system power level to a motor power level)[43],[44]. In this simplified simulation, the train departed from station A to station B at 0s, reached a maximum speed of 80km/h at 0.67s, where the traction motor speed was about 3524.57rpm, and braked and decelerated into station B at 2.68s. A “constant acceleration-constant power traction-idling-braking” mode of operation was adopted. The running state curve of train traction and motor traction are shown in Figure 9 and Figure 10, respectively.

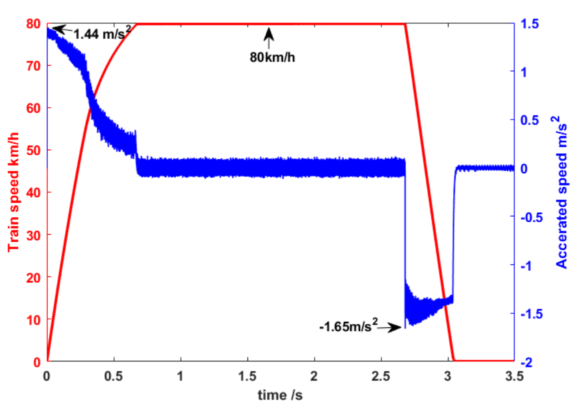


Figure 9 Train running state

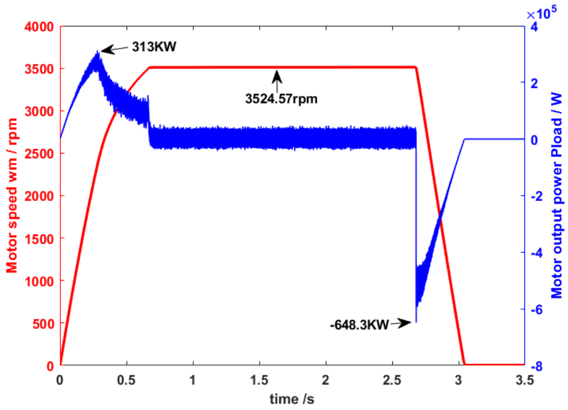


Figure 10 Motor traction state

From Figure 10, it can be determined that the highest traction motor startup power demand is about 313 kW and the highest braking power demand is about -648.3 kW. Considering that the lithium battery is the energy storage element, and it is not suitable to bear large power, its rated output power should be taken as the average power load demand power. Combined with the working range of low-voltage side voltage of DC/DC converter, which is generally 200V-600V, the rated voltage and total capacity of the battery were chosen as 495V and 40Ah. Since the ultracapacitor is assembled on the train and can only be used by one train, the capacity configuration can be directly adopted. In addition, considering the large flow of the trains on the platform and the frequent start of the braking of the trains on and off the platform, the capacity of the lithium battery pack on the platform should be appropriately increased while maintaining the rated voltage. Therefore, according to the actual operating energy range of urban rail transit, the capacity indexes of ultracapacitor and lithium battery were selected. Control parameters are shown in Table 4, the HESS capacity in the simulation has been reduced according to the actual operation conditions.

Table 4 HESS simulation control parameters

Parameter	Value	Parameter	Value
Working temperature	-20°C~45°C	Ultracapacitor charge and discharge cycle life	1 million times

Rated voltage of lithium battery cell	2.7V	Ultracapacitor monomer rated voltage	2.7V
Rated capacity of lithium battery cell	20Ah	Rated capacity of ultracapacitor monomer	3000F
Number of lithium battery in series	185	Number of ultracapacitor in series	223
Number of lithium battery in parallel	6	Number of ultracapacitor in parallel	5
Lithium battery initial capacity SOC	85%	Ultracapacitor initial capacity SOC	100%
Battery safe capacity range	20%~85%	Ultracapacitor safe capacity range	15%~100%
Rated voltage of battery module	495V	Rated voltage of ultracapacitor module	594V
Discharge threshold U_{dis}	1480V	Charge threshold U_{char}	1520V

(1) Simulation analysis of voltage fluctuations range: A comparison of the DC traction currents under different control strategies is shown in Figure 11 and Table 5.

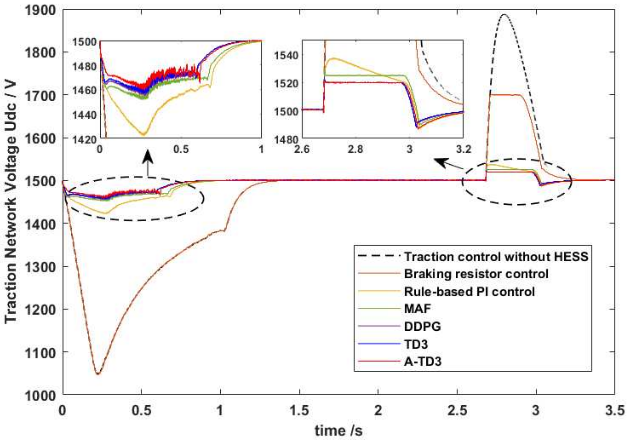


Figure 11 Traction network voltage comparison under different control strategies

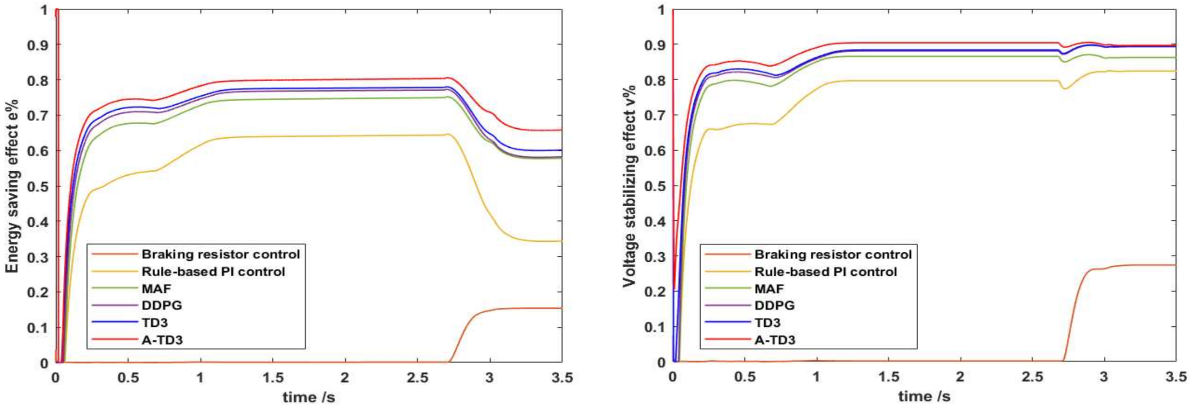
Table 5 Traction network voltage under different control strategies

Energy storing device	Lowest Traction Network Voltage (V)	Traction Network Voltage Peak (V)	Traction Network Voltage Fluctuation Range (V)

Traction control without HESS	1046	1887	1046~1887
Braking resistor control	1046	1700	1046~1700
Rule-based PI control	1422	1537	1422~1537
MAF	1452	1525	1452~1525
DDPG	1455	1520	1455~1520
TD3	1457	1520	1457~1520
A-TD3	1461	1520	1461~1520

From Figure 11 and Table 5, it can be observed that when the HESS is not used, the train uses electric braking to feedback regenerative energy, resulting in sharp fluctuation of DC traction voltage. At the moment of train start, the voltage drops to 1046V, and when the train brakes, the voltage rises to 1887V, this indicates that most of the regenerative braking energy is not recycled and is likely to cause safety hazards as well. When the braking resistor control is used, the remaining regenerative braking energy is consumed by the brake resistor, the traction voltage rises to 1700V, which keeps voltage fluctuations within the safe limits, but it causes a lot of energy dissipations. When both the on-board ultracapacitor and the ground battery is controlled by rule-based PI control, the traction voltage drops to 1422V for train start but recovers relatively slow, which cannot meet the rapid need of the peak-time traction power. At the breaking stage, the maximum value of the traction voltage is 1537V, which is not satisfactory. Therefore, rule-based PI control allocates power demand according to pre-set principles, and the quantitative control objective is not fundamentally clear, which makes it difficult to give full play to the characteristics of HESS. By adopting the traditional filter-based control (such as moving average filter (MAF)) in the HESS, the traction power is divided into high-frequency power component and low-frequency power component by MAF, which are assigned to the on-board ultracapacitor and ground battery respectively. The power compensation is realized for the DC traction network voltage under the condition of giving full play to the energy storage characteristics of the ultracapacitor and batter. Although the traction voltage can drop to 1452V, compared with rule-based PI control, the MAF method provides a faster response and a traction network voltage up to 1525V.

For the traditional DDPG algorithm, the convergent voltage stabilizing effect is better than that of MAF, and the traction voltage sag and increase range are relatively smaller, stable at 1455-1520V. For TD3, the addition of operations such as the prevention of the biased estimation of Q value, the delayed update of strategy and the smoothing of update of value function not only speed up the training convergence, but also improve the control performance. Finally, the voltage fluctuation of traction network is kept within 1457V-1520V. Furthermore, A-TD3 enables the agents to learn knowledge and experience that are more suitable for working conditions in complex and changing urban rail power supply scenarios through experience replay of priority, so as to better and faster complete knowledge interaction between the agents and the environment, which achieves the effect of fast convergence and high reward. The DC traction voltage is stabilized at 1461V~1520V, which is the best voltage stabilization performance compared with the other four algorithms.



a. *e%*

b. *v%*

Figure 12 Energy saving and voltage stabilization effect

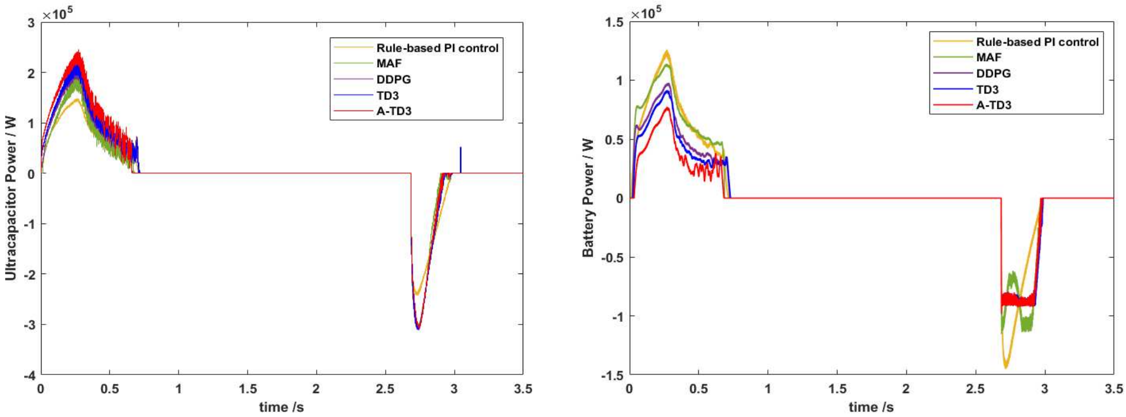
Table 6 *e%* and *v%* under different control strategies

Energy storing device	*e%*	*v%*
Traction control without HESS	0	0

Braking resistor control	15.3	27.4
Rule-based control	34.3	82.4
MAF	57.6	86.2
DDPG	58.1	89.3
TD3	60	89.4
A-TD3	65.6	89.6

Figure 12 and Table 6 represent the variation of energy saving parameters *e%* and voltage stabilization parameters *v%*. Compared with case where the HESS is not used, the rule-based filtering control algorithm and RL algorithm can realize the recovery and reuse of braking energy well, which greatly improves the effect of energy saving and voltage regulation. Compared with the traditional filtering algorithm MAF, the RL algorithm has better energy saving and voltage stabilizing effect. The A-TD3 performs even better, which the coefficient of energy saving *e%* is increased by 5.6% and the coefficient of voltage stabilizing *v%* is increased by 0.2% compared with the original TD3 methods.

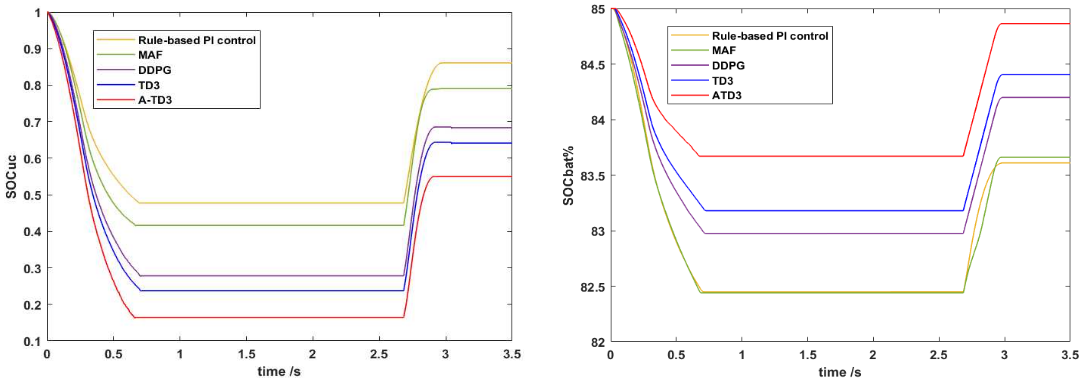
(2) Simulation analysis of HESS state variation: Figures. 13 and 14, 15 as well as Table 7 show the power and state changes of the HESS, among which the on-board ultracapacitor is mainly used to facilitate the fast exchange of permanent magnet traction power while the ground battery is used to assist in power supply and stabilize the traction voltage.



a. P_{sc}

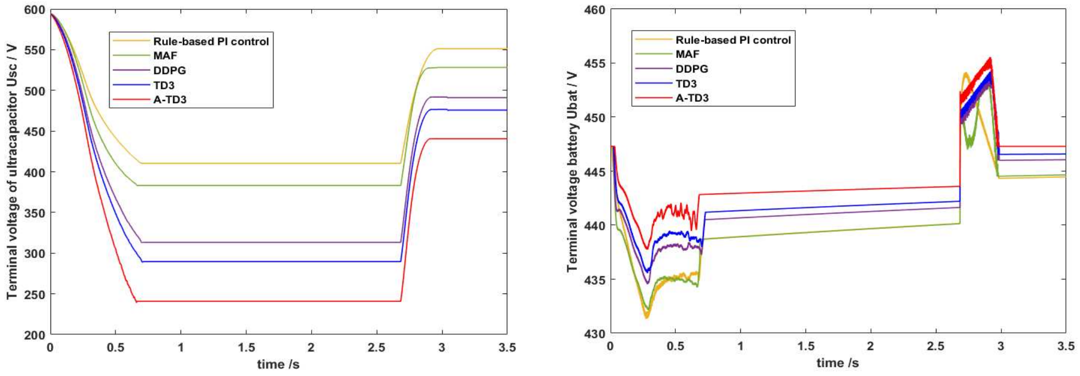
b. P_{bat}

Figure 13 Variation of given power of HESS



a. *SOC_{uc}* b. *SOC_{bat}*

Figure 14 Variation of SOC



a. U_{sc} b. U_{bat}

Figure 15 terminal voltage of HESS

Table 7 HESS power and state variation range

Energy storing device	P_{sc}(KW)	P_{bat}(KW)	SOC_{uc}%	SOC_{bat}%	U_{sc}(V)	U_{bat}(V)
Rule-based PI control	-241.2 ~ 146.5	-144.8 ~ 126	47.7% ~ 86.1%	82.45% ~ 83.6%	410.4 ~ 551.1	431.4 ~ 454
MAF	-304.8 ~ 194.8	-115.4 ~ 114	41.6% ~ 79%	82.44% ~ 83.66%	383.2~ 528.2	432.2 ~ 453.7
DDPG	-309.5 ~ 216.1	-98.5 ~ 97.8	27.8% ~ 68.3%	82.97% ~ 84.19%	313.3 ~ 491	434.5 ~ 453.4
TD3	-310.3 ~ 225.1	-96.3 ~ 91.5	23.7% ~ 64.1%	84.17% ~ 84.4%	289.5 ~ 475.7	435.7 ~ 454.2
A-TD3	-306.5 ~ 243.4	-96.5 ~ 77.3	16.4% ~ 55%	83.67% ~ 84.86%	240.8 ~ 440.5	437.8 ~ 455.4

It can be observed that the ultracapacitor under the control strategy of the rule-based PI control, MAF, or DDPG algorithms are not fully discharged. This is caused by the uneven power allocation of the ultracapacitor-battery-traction network, which will make it difficult to achieve a good energy-saving and voltage-stabilizing effect. In the case of the rule-based PI control, the train braking makes the ultracapacitor SOC_{uc} eventually rise to 86.1%, breaking the safety range [0.15~0.85], SOC_{uc} controlled by MAF, DDPG and TD3 also has broken through this range, which will greatly affect the life of the capacitor; and in the case of the MAF, DDPG, TD3 algorithms, the battery output power is very high, which indicates that the on-board ultracapacitor has not achieved a good energy saving effect. Compared with the other four algorithms, the proposed A-TD3 can make the ultracapacitor reach a better charge and discharge depth within the safety range and reduce battery output. In summary, the proposed control strategy can achieve optimal energy saving, voltage stabilization and ultracapacitor overcharge / overdischarge protection while fully completing the permanent magnet traction power exchange.

(3) Simulation analysis of battery cycle lifetime decay: In order to test the adaptive ability of the proposed DRL algorithm to cope with the dynamic changes of the environment, the ground battery is taken as the main protection object. It can be seen from Section 3.5 that when the initial capacity of the battery is 85%, the remaining capacities after 500, 1000, 1500, and 2000 cycles of discharge (gradual decay) correspond to 80.75%, 78.2%, 75.65%, and 74.375%, respectively. At the times of 0.2s~1s, the sudden capacity decline of the battery was simulated, and the corresponding sudden discharge currents were 40A, 50A, 65A, and 80A, respectively. The proposed A-TD3 method was used to detect the operating conditions of the battery in real time, dynamically allocate the given battery power, and respond to the two conditions of gradual battery attenuation and sudden attenuation in time. In the simulation, the experiments with 500, 1000, 1500, and 2000 charge-discharge cycles (1 cycle of the battery is composed of many charging and discharging processes, that is, the battery is reduced from 85% of the initial capacity to 20% of the minimum capacity after multiple discharges and charges, and then the system is cut out and charged to the initial capacity process, we regard this process as 1 cycle work of the battery) in the built urban rail operating environment were conducted and the MAF method was used for the purpose of comparison. Different from battery, in order to maximize the advantages of the ultracapacitor, we design a cycle process of the on-board ultracapacitor as one start

and break of the train between stations. The simulation results are shown in Figure 16 and Figure 17, respectively.

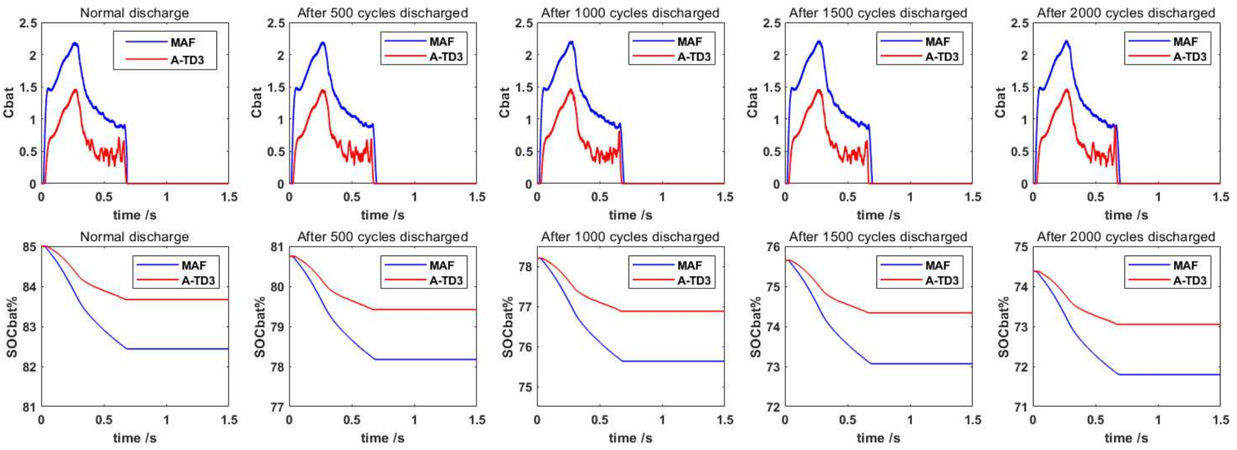


Figure 16 Battery discharge state after different discharge cycles

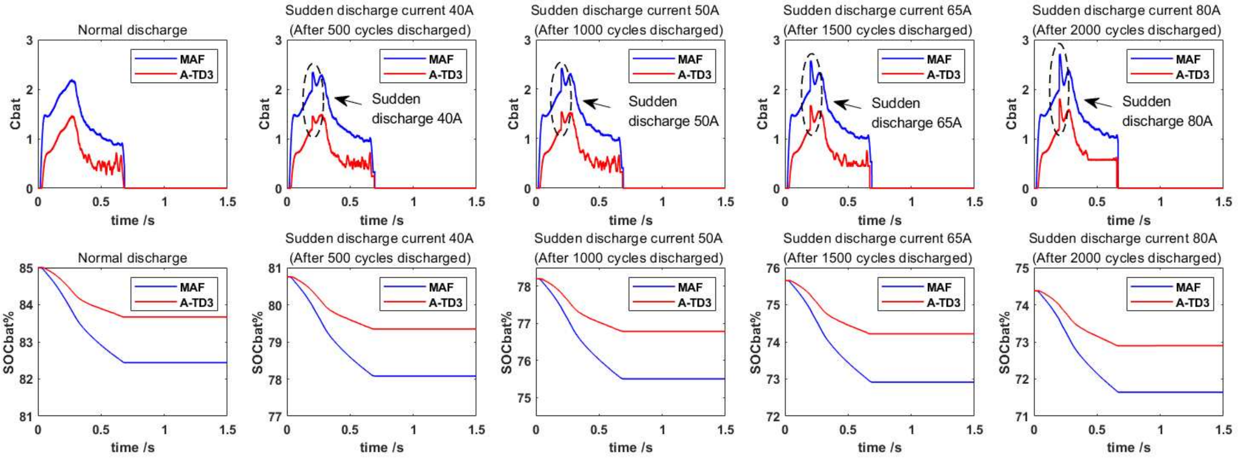
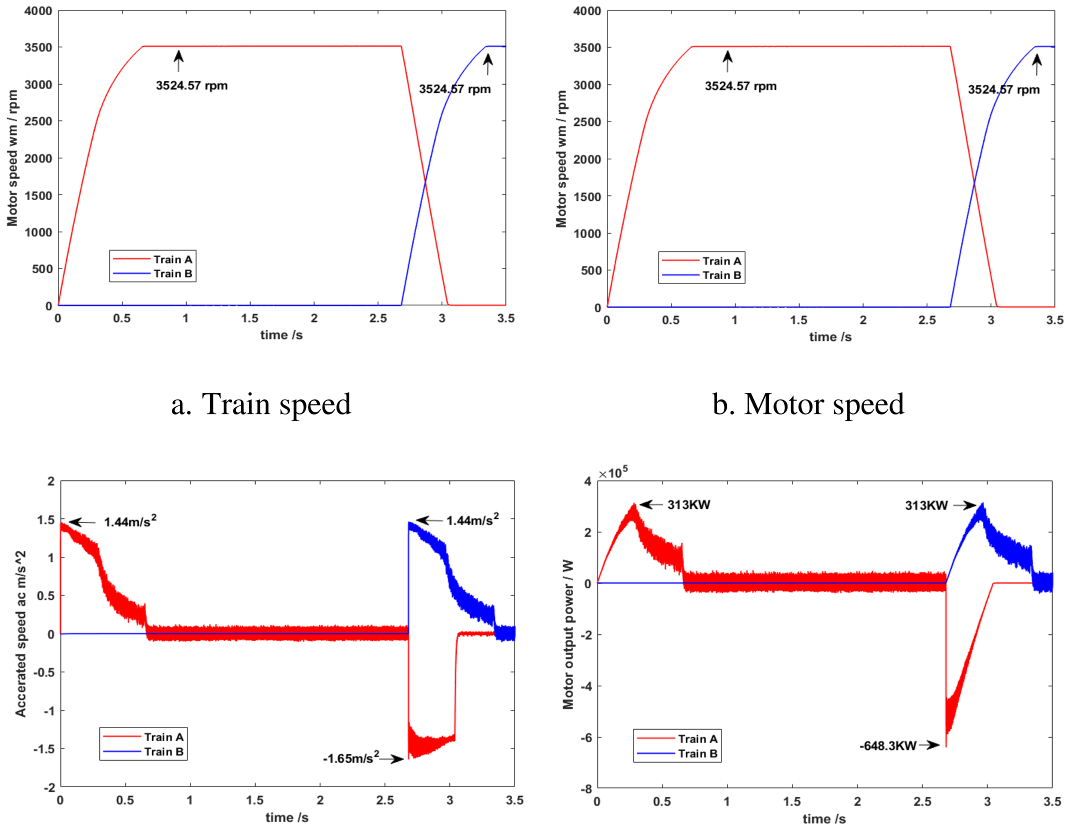


Figure 17 Discharge state of the battery under different sudden decay

It can be observed from Figure 16 that the total battery capacity SOC will gradually decrease with the increase of the number of cycles but the discharge current is still relatively stable. However, the sudden discharge current of the battery will have a greater impact on its life. In Figure17, the Agent can adaptively adjust the battery output power at the moment of sudden change of battery discharge current to reduce the battery discharge rate C_{bat} and the peak discharge current, thereby reducing the impact of instantaneous sudden change on the battery life. The method can timely respond to the sudden change of the urban rail power supply environment and has better adaptive ability.

(4) Simulation analysis of two-train energy interaction: In real urban rail line, the train has certain energy interaction with the adjacent train when stopping at the platform, which aggravates the voltage fluctuation of the traction network. In order to study the interaction conditions of multiple trains and verify that the proposed method still has a good voltage regulation effect, a down train is added in the simulation. Suppose that the battery is located in the platform, assisting in power supply to all the trains through the DC traction network. There are on-board ultracapacitor for every individual train, mainly used to facilitate the fast exchange of permanent magnet traction power. In this simulation, the train A departed from station A to station B at 0s, reached a maximum speed of 80km/h at 0.67s, where the traction motor speed was about 3524.57r/min, and braked and decelerated into station B at 2.68s. At the same time, the train B accelerated and departed from station B. The running state curves of the two trains are shown in Figure 18.



a. Train speed

b. Motor speed

c. Accelerated speed

d. Motor output power

Figure 18 The running state curve of train

It can be seen from Figure 18 that it is difficult to achieve a reasonable transfer of braking energy and starting energy at 2.68s even if two trains act at the same time (train A brakes and train B starts), because the braking process is faster than starting and the resulting instantaneous braking power is higher, so HESS is required to coordinate the process. After adding HESS and the proposed method, the DC voltage fluctuations are shown in Figure 19 and Table 8.

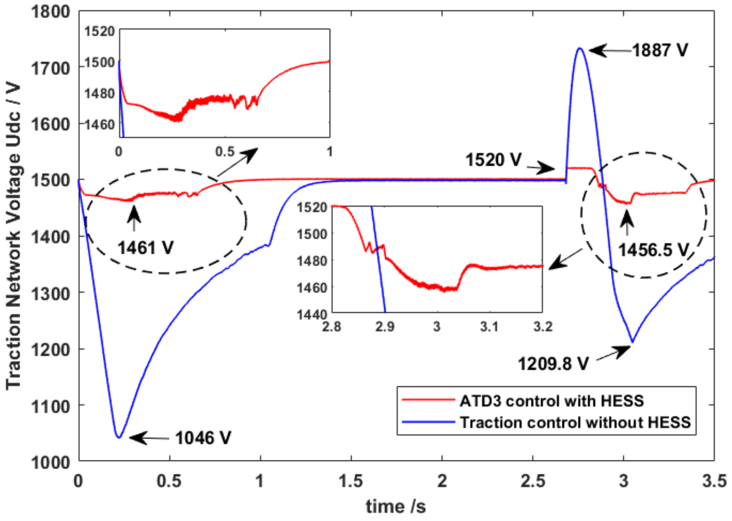


Figure 19 Traction network voltage

Table 8 Traction network voltage under different control strategies

Energy storing device	Lowest Traction Network Voltage (V)	Traction Network Voltage Peak (V)	Traction Network Voltage Fluctuation Range (V)
Traction control without HESS	1046	1887	1046~1887
A-TD3	1456.5	1520	1456.5~1520

It can be seen from Figure 19 and Table 8 that the DC voltage fluctuates sharply without HESS, and the time for voltage stabilization is slow. The voltage fluctuation range can be kept between 1456.5V and 1520V after adding HESS and the proposed ATD3 method, which can effectively suppress DC voltage fluctuation caused by the train energy interaction. The state change of HESS is shown in Figure 20.

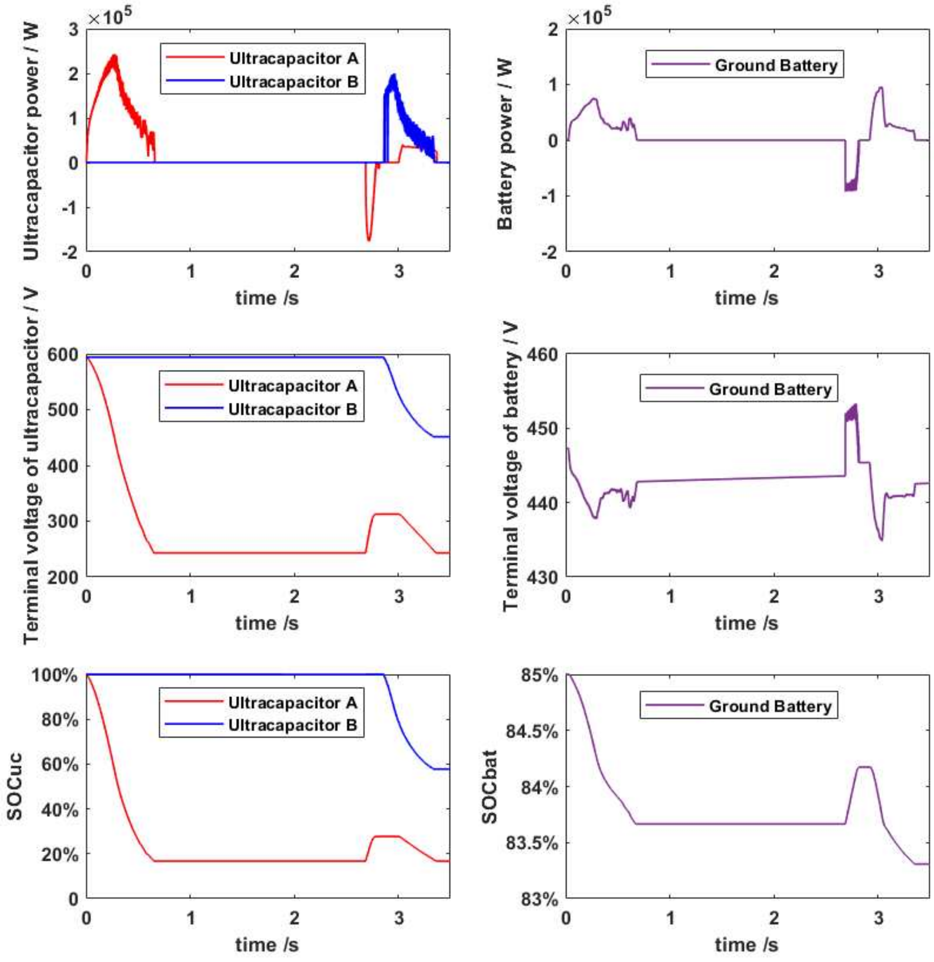


Figure 20 State of HESS

Table 9 HESS power and state variation range

Energy storage systems	Power (KW)	Voltage (V)	SOC%
Ultracapacitor A	-176.5 ~ 243.4	243 ~ 312.6	16.7% ~ 27.7%
Ultracapacitor B	0 ~ 199.7	451.5 ~ 594	57.7% ~ 100%
Ground Battery	-92.3 ~ 95.7	434.87 ~ 453.3	83.3% ~ 84.2%

It can be seen from Figure 20 and Table 10 that a reasonable energy allocation among ESS can still be achieved during energy interaction period between the two trains with the proposed method, so as to prevent the ground battery from taking on excessive power peaks. In addition, the ultracapacitor of train A can also be prevented from overcharging and over discharging with the proposed method, and their charge states vary from 16.7% to 27.7%, which is kept within the safety range [0.15, 0.85].

4.4 RT-LAB real-time simulation

To verify the effectiveness of the proposed control strategy, the RT-LAB semi-physical real-time simulation system was used, as shown in Figure 21. The DSP including the learned agents performs the control algorithm part while the RT-LAB mainly simulate the permanent magnet traction system with the HESS.

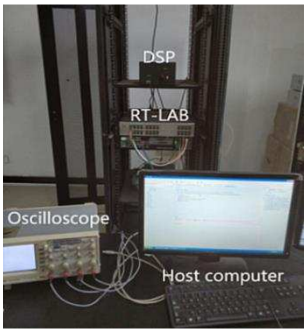
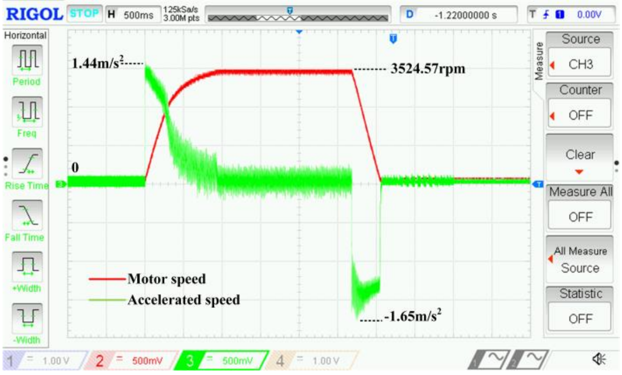


Figure 21 RT-LAB real-time simulation system

The simulation was based on the reference data from a real-world rail network. The train departed from station A with a starting acceleration of 1.44m/s^2, reached a maximum speed of 80km/h after acceleration, then ran at a constant speed, and decelerated at a braking acceleration of 1.65m/s^2. and finally stopped at Station B. The traction motor speed is shown in Figure 22. Output power of the HESS is shown in Figure 23.



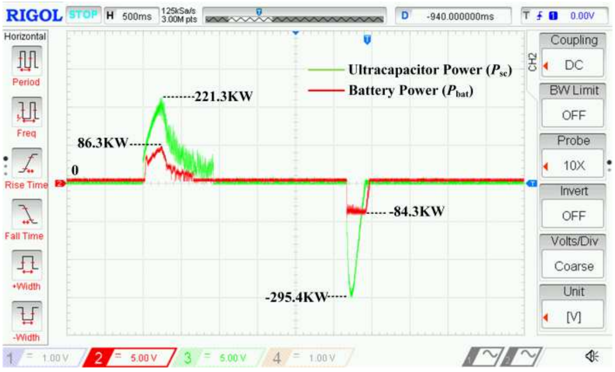


Figure 22 Train operating parameters

Figure 23 Variations of the HESS power

The controller in the HESS implemented the power allocation through the real-time information of the train running state. At the train traction/braking moment, a high frequency power given value was assigned to increase the output power of the ultracapacitor and therefore realize the energy exchange between the permanent magnet traction and the ultracapacitor, undertaking the task of high frequency power requirement at peak time. At the same time, the battery at the platform undertook low-frequency power to smooth the fluctuation of the traction network.

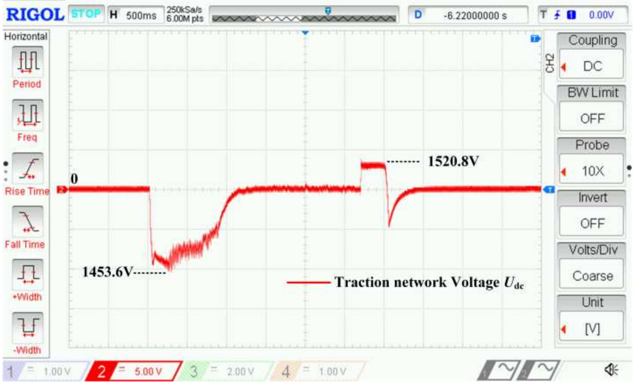
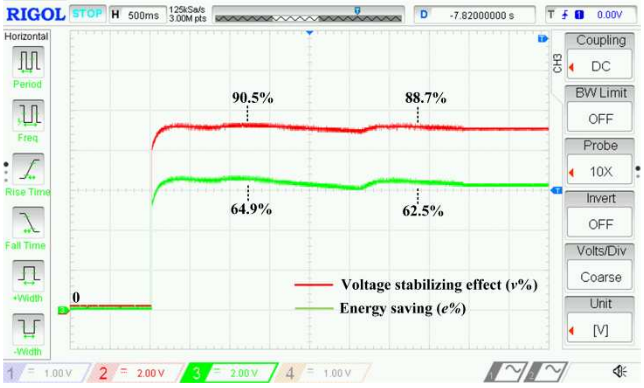


Figure 24 Voltage variation of DC traction network

Figure 24 shows the voltage variation of DC traction network. When the train accelerates at 0s, the traction voltage decreases to about 1453.6V, then rises rapidly and returns to stability. At this point, the output of the HESS realizes "valley filling". The train brakes and decelerates, and the HESS cuts off the sudden "peak" by recovering energy, making the voltage stable at about 1520.8V, which shows excellent voltage stabilization effect. Eventually, the regenerative energy recovery is realized while the DC voltage fluctuation is effectively stabilized.



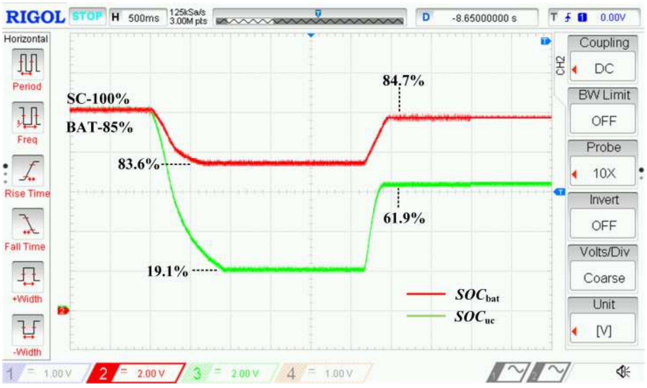


Figure 25 Energy saving and voltage stabilization effect

Figure 26 Changes of HESS SOC

Table 10 e%, v% and SOC changes

Energy storing device	*e%*	*v%*	*SOC*$_{\mathrm{uc}}$*%*	*SOC*$_{\mathrm{bat}}$*%*
A-TD3	62.5	88.7	19.1 ~ 61.9	83.6 ~ 84.7

Figure 25 shows the variation of coefficient of energy saving *e%* and coefficient of voltage stabilizing *v%*, and Table 10 shows the *e%*, *v%* rate and SOC. The initial SOC of the ultracapacitor is 100%, and the initial SOC of the battery is 85%. It can be observed that the *e%* under the proposed control strategy can reach up to 62.5%, and the *v%* reaches 88.7%, with excellent energy-saving and voltage stabilization effects. As can be seen from Figure 26, under the proposed control strategy, the adaptive working condition change can achieve more accurate online dynamic power distribution of the HESS and effectively prevent the overcharge and overdischarge of the ultracapacitor under the traction acceleration and braking deceleration of the train. During discharge, the SOC_{uc} of ultracapacitor charge state drops to a minimum of 19.1% and rises to a maximum of 61.9% during charging, so that the SOC_{uc} is effectively kept within the safe range [0.15, 0.85], which effectively prolongs the life of on-board ultracapacitor.

The experimental results show that under the proposed control strategy, the advantages of each energy storage element can be fully utilized to improve the effect of energy saving and voltage stabilization. In the same power supply section, the coordinated control and dynamic power distribution among the HESS are realized according to the DRL on-line sequential decision-making method, which can effectively stabilize the DC traction voltage fluctuations and protect the on-board ultracapacitor.

5. Conclusions

In this study, considering the severe voltage fluctuations of traction networks and unbalanced charge-discharge behavior between each ESSs caused by starting and braking of urban rail trains, a HESS power dynamic allocation strategy has been proposed, which is based on annealing bias -

priority experience replay with double-delay depth deterministic policy gradient algorithm (A-TD3). The main conclusions can be drawn as follows:

(1) By training the experience buffer in TD3 with priority probabilistic experience replay, in conjunction with the proposed off-line training - online optimization - online sequential decision making method, the training efficiency and the correctness of the agents can be greatly improved.

(2) Compared with the traditional filtering algorithms and DRL algorithms, the proposed control strategy can effectively stabilize the DC traction voltage fluctuation, make the "peak-shaving and valley-filling" effect more significant, and balance the utilization of each ESS, which effectively prevents overcharge and overdischarge of ultracapacitor and battery, taking on the excessive power demands.

(3) Due to paper length and time, we have conducted a case analysis based on the typical interaction conditions of two trains. The simulation shows that the proposed method is helpful for the energy interaction process of dual trains, such as realizing DC voltage stability and HESS protection. However, the energy interaction and scheduling of multiple trains is a relatively complex problem, which often needs to be combined with the optimization of the operating schedule, energy scheduling, the full life cycle calculation of energy storage, and even the economic cost of equipment, etc. In the future, we will combine more complex operating conditions with energy scheduling and control optimization.

(4) The real-time simulation experiments on RT-LAB demonstrate that the proposed method is suitable for real-time urban rail applications, and the results show that coefficient of energy saving reaches 62.5% and the coefficient of voltage stabilizing reaches 88.7%, which realizes the overall optimal dynamic performance and economic benefits.

(5) A limitation of the current study is the use of the semi-physical simulation platform, where the modeling error could affect the system performance when implementing in real world system. The future work will focus on the application to metro projects.

ACKNOWLEDGMENT

Financial support was provided in part by the National Natural Science Foundation of China (62033014, 61673166), Natural Science Foundation of Hunan Province (2021JJ50006, 2022JJ50074).

REFERENCES

[1] Shen, S., Wei, H., Li, W. “Study of trackside photovoltaic power integration into the traction power system of suburban elevated urban rail transit line,” Applied Energy, 2020, 260(15), pp.114177.

[2] González-Gil, A., Palacin, R., Batty, P., et al. “A systems approach to reduce urban rail energy consumption,” Energy Convers. Manag, 2014, 80, pp.509–524.

[3] Popescu, M., Bitoleanu, A., et al. “A review of the energy efficiency improvement in DC railway systems,” Energies, 2019, 12(6), pp.1092.

[4] Bao, X. “Urban Rail Transit Present Situation and Future Development Trends in China: Overall Analysis Based on National Policies and Strategic Plans in 2016–2020”, Urban Rail Transit, 2018, 4(1), pp.1–12.

[5]Lin, S., Huang, D., Wang, A., et al. “Research on the regeneration braking energy feedback system of urban rail transit,” IEEE Transactions on Venhicular Technology, 2019, 68(8), pp.7329-7339.

[6] Liu, R., Xu, L., Liu, F., et al. “A Novel Architecture of Urban Rail Transit Based on Hybrid Energy Storage Systems Using Droop Control,” 2018 IEEE International Conference on Electrical Systems for Aircraft, Railway, Ship Propulsion and Road Vehicles & International Transportation Electrification Conference (ESARS-ITEC), 2018, pp.1-6.

[7] Khodaparastan, M., Mohamed, A, A., Brandauer, W. “Recuperation of Regenerative Braking Energy in Electric Rail Transit Systems,” IEEE Transactions on Intelligent Transportation Systems, 2019, 20(8), pp.2831-2847.

[8] Gao, Z., Fang, J., Zhang, Y., et al. “Control of urban rail transit equipped with ground-based supercapacitor for energy saving and reduction of power peak demand,” International Journal of Electrical Power & Energy Systems, 2015, 67(May), pp. 439-447.

[9] Liu, Y., Yang, Z., Lin, F., et al. “Energy loss analysis of the stationary battery-supercapacitor HESS,” 2019 IEEE Vehicle Power and Propulsion Conference (VPPC), 2019, pp. 1-6.

[10] Qin, Q., Zhang, J., Li, Y., et al. “Research on time-phased control strategy of urban rail ground hybrid energy storage device based on train operation status,” Transactions of China Electrotechnical Society, 2019 (in Chinese), 34(S2), pp.318-327.

[11] Huang, X., Liao, Q., Li, Q. et al. “Power management in co-phase traction power supply system with super capacitor energy storage for electrified railways,” Railway Engineering Science, 2020, 28(1), pp.85–96.

[12] Graber, G., Galdi, V., Calderaro, V., et al. “Sizing and energy management of on-board hybrid energy storage systems in urban rail transit,” 2016 International Conference on Electrical Systems for Aircraft, Railway, Ship Propulsion and Road Vehicles & International Transportation Electrification Conference (ESARS-ITEC), 2016, pp.1-6.

[13] Zahedi, B., Norum, L. E., et al. “Modeling and Simulation of All-Electric Ships With Low-Voltage DC Hybrid Power Systems,” IEEE Transactions on Power Electronics, 2013, 20(10), pp.4525–4537.

[14] Liang, J., Li, Q., J. O. Lindtjørn. et al. “Frequency dependent DC voltage droop control for hybrid energy storage in DC microgrids,” IEEE Power & Energy Society General Meeting, 2015, pp.1–5.

[15] Manandhar. U, Tummuru. N. R, Kollimalla. S. K, et al, “Validation of Faster Joint Control Strategy for Battery- and Supercapacitor-Based Energy Storage System,” IEEE Transactions on Industrial Electronics, 2018, 65(4), pp. 3286-3295.

[16] Gao, C., Zhao, J., Wu, J., et al. “Optimal fuzzy logic based energy management strategy of battery/supercapacitor hybrid energy storage system for electric vehicles,” World Congress on Intelligent Control and Automation. 2016, pp.98-102.

[17] Herrera, V., Milo, A., Gaztañaga, H., et al. “Adaptive energy management strategy and optimal sizing applied on a battery-supercapacitor based tramway,” Applied Energy, 2016, 169(1), pp.831–845.

[18] Deng, W., Dai, C., Han, C., et al. “Back-to-back hybrid energy storage system of electric railway and its control method considering regenerative braking energy recovery and power quality improvement,” in Proceedings of the CSEE, 2019, 39(10), pp.2914-2924.

[19] Fang, S., Gou, B., Wang, Y., et al. “Optimal Hierarchical Management of Shipboard Multibattery Energy Storage System Using a Data-Driven Degradation Model,” IEEE Transactions on Transportation Electrification, 2019, 5(4), pp. 1306-1318.

[20] Jia, Z., Jiang, J., Lin, H., et al. “A Real-time MPC-based Energy Management of Hybrid Energy Storage System in Urban Rail Vehicles,” Energy Procedia, 2018, 152(10):526-531.

[21] Wang, Y., Yang, Z., Lin, F., et al. “Research on collaborative optimization of energy interactive management strategy and capacity allocation for tram hybrid system,” Transactions of China Electrotechnical Society, 2019, 34(08), pp.1780-1788.

[22] Sutton, R. S. and Barto. A. G. “Reinforcement Learning: An Introduction,” The MIT Press, Cambridge, MA, 1998.

[23] Hannan MA, Lipu M, Hussain A, Mohamed A. “A review of lithium-ion battery state of charge estimation and management system in electric vehicle applications: challenges and recommendations,” Renew Sustain Energy Rev 2017, 78(10), pp.834-54.

[24] Abdelhedi, R,. Lahyani, A,. Chiheb, A., et al. “Reinforcement learning-based power sharing between batteries and supercapacitors in electric vehicles,” 2018 IEEE International Conference on Industrial Technology (ICIT), 2018, pp.2072-2077.

[25] Lin, X., Wang, Y., Bogdan, P., et al. “Reinforcement learning based power management for hybrid electric vehicles,” 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2014, pp.33-38.

[26] Ma, Z., Qian, H., Zhang, T., et al. “Deep deterministic policy gradient based energy management strategy for hybrid electric tracked vehicle with online updating mechanism,” IEEE Access, 2021, 9, pp.7280–7292.

[27] Tiong, T., Saad, I., Teo, K., et al. “Deep reinforcement learning with robust deep deterministic policy gradient,” 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), 2020, pp.1-5.

[28] Fujimoto, S., van Hoof, H., Meger, D. “Addressing function approximation error in actor-critic methods,” arXiv:1802.09477, 2018.

[29] Hou, Y., Hong, H., Sun, Z., et al. “The Control Method of Twin Delayed Deep Deterministic Policy Gradient with Rebirth Mechanism to Multi-DOF Manipulator,” Electronics. 2021; 10(7), pp.870.

[30] Dankwa, S., Zheng, W. “Modeling a Continuous Locomotion Behavior of an Intelligent Agent Using Deep Reinforcement Technique,” 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET), 2019, pp.172-175.

[31] Woo, J. H., Wu, L., Park, J. B., Roh, J. H. “Real-Time Optimal Power Flow Using Twin Delayed Deep Deterministic Policy Gradient Algorithm,” IEEE Access, 2020, 8, pp.213611-213618.

[32] Cui, Q., Kim, G., Weng, Y. “Twin-Delayed Deep Deterministic Policy Gradient for Low-Frequency Oscillation Damping Control,” Energies, 2021, 14(20), pp.6695.

[33] Ye, Y., Qiu, D., Wang, H., et al. “Real-Time Autonomous Residential Demand Response Management Based on Twin Delayed Deep Deterministic Policy Gradient Learning,” Energies, 2021, 14(3), pp.531.

[34] Perera, A.T.D., Kamalaruban, P. “Applications of reinforcement learning in energy systems,” Renewable and Sustainable Energy Reviews, 2021, 137, pp.110618.

[35] Wei, W., Shaojun, X., Ting, H., et al. “Design and control of a combined BDC based energy storage system for regenerative applications”, 2014 IEEE Conference and Expo Transportation Electrification Asia-Pacific (ITEC Asia-Pacific), 2014, pp.1-4.

[36] Wang, X., Luo, Y., Qin, B., et al. “Hybrid energy management strategy based on dynamic setting and coordinated control for urban rail train with PMSM,” IET Renewable Power Generation. 2021, pp.1–13.

[37] Wang, X., Luo, Y., Qin, B., et al. “Power dynamic allocation strategy for urban rail hybrid energy storage system based on iterative learning control,” Energy. 2022, 245, pp.123263.

[38] Hailong, Z., Jiankun, P., Huachun, T., et al., “A Deep Reinforcement Learning Based Energy Management Framework with Lagrangian Relaxation for Plug-in Hybrid Electric Vehicle,” IEEE Transactions on Transportation Electrification, 2021, 7(3), pp.1146-1160.

[39] Saehong, P., Andrea, P., Michael, W., et al. “Reinforcement learning-based fast charging control strategy for Li-ion batteries,” 2020 IEEE Conference on Control Technology and Applications (CCTA), 2020, pp.100-107.

[40] Schaul, T., Quan, J., Antonoglou, I., et al. “Prioritized Experience Replay,” Computer Science, 2015.

[41] Yang, Z., Zhu, F., Lin, F. “Deep-Reinforcement-learning-based Energy Management Strategy for Supercapacitor Energy Storage Systems in Urban Rail Transit,” IEEE Transactions on Intelligent Transportation Systems, 2021, 22(2), pp.1150-1160.

[42] Swierczynski, M., Stroe, D., Stan, A., et al. “Lifetime Estimation of the Nanophosphate LiFePO4/C Battery Chemistry Used in Fully Electric Vehicles,” IEEE Transactions on Industry Applications, 2015, 51(4), pp.3453-3461.

[43] Zhong, Z., Yang, Z., Fang, X., et al. “Hierarchical Optimization of an On-Board Supercapacitor Energy Storage System Considering Train Electric Braking Characteristics and System Loss,” IEEE Transactions on Vehicular Technology, 2020, 69(3), pp. 2576-2587.

[44] Guilherme, C., Moraes, D., Brockveld, S., et al. “Power Conversion Technologies for a Hybrid Energy Storage System in Diesel-Electric Locomotives,” IEEE Transactions on Industrial Electronics, 2021, 68(10), pp. 9081-9091.