buildings

MDPI

Article

Comparative Analysis of Reinforcement Learning Approaches for Multi-Objective Optimization in Residential Hybrid Energy Systems

Yang Xu [1], Yanxue Li [2,*] and Weijun Gao [2,3]

[1] School of Information and Control Engineering, Qingdao University of Technology, 777 Jialingjiang Road, Qingdao 266520, China; 901020170087@qut.edu.cn
[2] Innovation Institute for Sustainable Maritime Architecture Research and Technology, Qingdao University of Technology, Fushun Road 11, Qingdao 266033, China; gaoweijun@me.com
[3] Faculty of Environmental Engineering, The University of Kitakyushu, Kitakyushu 808-0135, Japan
* Correspondence: liyanxue@qut.edu.cn

Abstract: The rapid expansion of renewable energy in buildings has been expedited by technological advancements and government policies. However, including highly permeable intermittent renewables and energy storage presents significant challenges for traditional home energy management systems (HEMSs). Deep reinforcement learning (DRL) is regarded as the most efficient approach for tackling these problems because of its robust nonlinear fitting capacity and capability to operate without a predefined model. This paper presents a DRL control method intended to lower energy expenses and elevate renewable energy usage by optimizing the actions of the battery and heat pump in HEMS. We propose four DRL algorithms and thoroughly assess their performance. In pursuit of this objective, we also devise a new reward function for multi-objective optimization and an interactive environment grounded in expert experience. The results demonstrate that the TD3 algorithm excels in cost savings and PV self-consumption. Compared to the baseline model, the TD3 model achieved a 13.79% reduction in operating costs and a 5.07% increase in PV self-consumption. Additionally, we explored the impact of the feed-in tariff (FiT) on TD3's performance, revealing its resilience even when the FiT decreases. This comparison provides insights into algorithm selection for specific applications, promoting the development of DRL-driven energy management solutions.

Keywords: deep reinforcement learning; building energy management; photovoltaics; heat pump; battery storage

check for updates

Citation: Xu, Y.; Li, Y.; Gao, W. Comparative Analysis of Reinforcement Learning Approaches for Multi-Objective Optimization in Residential Hybrid Energy Systems. *Buildings* **2024**, *14*, 2645. https://doi.org/10.3390/buildings14092645

Academic Editor: Francesco Nocera

Received: 19 July 2024
Revised: 22 August 2024
Accepted: 25 August 2024
Published: 26 August 2024

1. Introduction

1.1. Background

Currently, the operation of buildings is responsible for 26% of the world's energy-related emissions, of which 8% originate directly from the buildings and 18% derive indirectly from the electricity and heat produced for them [1]. Consequently, the deep decarbonization of building energy systems is poised to contribute significantly to the low-carbon transition of the energy system [2]. Prior research indicates that investments in distributed renewable energy sources (RESs) have noteworthy prospects for mitigating carbon emissions from the construction sector [3,4].

In light of the global carbon neutrality strategy, there has been a remarkable increase in the utilization of RESs to combat climate change [5]. Among the various RES options available, distributed photovoltaic (PV) systems have gained considerable attention due to their cost-effectiveness and ease of deployment, establishing themselves among the most widespread types of renewable energy [6,7]. However, the widespread adoption of distributed PV systems presents significant challenges in balancing the energy supply and demand within the grid [8]. This challenge primarily arises from the inherent volatility and intermittency associated with power generation from these systems. Consequently, enhancing the penetration of RESs within various energy systems has emerged as a crucial area of research focus. The advent of the home energy management system (HEMS) presents a potential answer to tackling the challenges mentioned above by offering an integrated energy system that incorporates advanced communication, sensing, measurement, and control technologies [9,10]. These HEMSs can store excess renewable energy through energy storage systems (ESSs) during off-peak periods, utilizing this stored energy as a dependable power source during periods of high demand [11]. As a result, the issue of power mismatch in grids with high-penetration intermittent renewable energy can be effectively resolved.

Indeed, the implementation of HEMSs encounters a multitude of uncertainties, posing significant obstacles to efficient energy scheduling within the system. A key challenge originates from the inherent nature of renewable energy generation, which is strongly influenced by environmental factors, resulting in pronounced intermittency and uncertainty [12]. Secondly, to promote collaboration between distributed energy suppliers, users, and the public grid, more and more countries and regions have embraced price-driven demand response control strategies, including floating feed-in tariff (FiT) subsidies and real-time electricity price (RTP) [13]. Thirdly, for residential customers, variations in seasons and disparate living habits contribute to uncertainties in electricity demand [14,15]. Consequently, accurately modeling demand response (DR) for multiple household devices, effectively managing uncertainties, and advancing comprehensive decision-making methods in high-dimensional settings pose significant challenges in HEMS research. Currently, classical control methods are commonly employed in HEMSs, but they often lack the necessary precision due to limited domain-specific knowledge and historical data utilization. Model predictive control (MPC) represents a viable approach to address this limitation by formulating uncertainties as constrained optimization problems [16,17]. Nevertheless, the effectiveness of MPC models depends significantly on the precision of predictive models and the alignment of constraints, both demanding ample data and tailored adjustments. Thus, developing standardized MPC models that can cater to the diverse needs of residential customers remains a significant challenge [18].

1.2. Related Work

As machine learning (ML) methods gain popularity, data-driven reinforcement learning (RL) emerges as an effective solution for optimizing the operations of HEMSs [19,20]. RL algorithms can learn optimal decision strategies through interaction with the environment without relying on large amounts of labeled training data. This characteristic enables RL to achieve real-time learning and decision-making capabilities. As a result, RL technology is well suited for adapting to dynamically changing requirements and effectively handling real-time flexible load control tasks. Q-learning was one of the initial RL algorithms proposed and applied in the operation optimization of HEMS [21–23]. However, when confronted with high-dimensional action space problems, Q-learning methods always encounter the "curse of dimensionality" challenge, making it computationally difficult to store and update Q-values. Therefore, there may be limitations when applying the Q-learning method to complex HEMS problems.

Deep reinforcement learning (DRL) aims to effectively address the issue of the curse of dimensionality. The DRL algorithm could cope with the complexities of high-dimensional states or action spaces by the function approximation method based on the deep neural network, enabling the DRL methods to make accurate and flexible decisions [24]. Xiao et al. [25] developed an energy scheduling algorithm that combines Deep Q-Networks (DQNs) with the Long Short-Term Memory (LSTM) network and attention mechanism. Experimental results show that this method demonstrated a 4.11% reduction in economic consumption and a 24.4% increase in energy storage utilization compared to the baseline models. In [26], the authors proposed an energy scheduling framework for hydrogen production systems utilizing the Deep Deterministic Policy Gradient (DDPG). Compared to traditional methods, the DDPG algorithm delivers improved economic benefits and enhances the utilization of renewable energy. Wang et al. [27] assessed the effectiveness of different DRL algorithms in managing HVAC systems within buildings, and their findings revealed that DDPG achieved a significant 10.06% energy-saving effect relative to the original control approach. Ren et al. [28] illustrated a HEMS optimization framework based on data-driven DRL, using a neural network with bidirectional gated recurrent units for PV generation and RTP prediction, as well as a soft actor–critic (SAC) algorithm for optimal decision-making. The results exhibited a 17.7% decline in household electricity costs and an 8.4% decrease in total costs. The work [26] addressed the operational challenges of enhancing energy cost optimization and off-grid operation duration in HEMSs. The DDPG and the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithms underwent training and verification. The findings illustrate that the TD3 algorithm achieves effective optimization by reducing the average hourly discrepancy in grid power purchases to below 2 kWh and maintaining battery safety for up to 7.72 h.

Since the equipment often includes cooling, heating, and power, in practical application scenarios, the HEMS always requires the coordinated control of multiple devices to achieve optimal operation. Increasing the dimensionality of the action space is one potential solution to address this issue. Some researchers have successfully applied the DRL method with high-dimensional action spaces to control and regulate complex energy systems, yielding promising outcomes. For example, Ruan et al. [29] proposed a novel optimization framework utilizing DRL to independently control PV generation and ESS to minimize the operating costs of CCHP systems. Through a comparative analysis with traditional methods, the results demonstrate superior performance achieved by the DRL approaches. In [30], the authors introduced a model-free dynamic optimization management strategy using DLR for HEMSs, which uses the DQN algorithm to control and manage diverse devices dynamically. A case study was undertaken to validate the strategy's efficacy, and the results showed a 36.7% decrease in carbon trading costs and a 50.2% reduction in penalties, both associated with user satisfaction. Langer et al. [31] investigated applying a DRL method to operate a smart house with various RESs. By comparing the results obtained using the DDPG algorithm with the MPC and rule-based benchmarks, the DRL approach achieved a self-reliance level of 75% while maintaining acceptable comfort breaches.

The literature review above demonstrates the extensive application of RL technology in the optimization of building energy systems scheduling. Nonetheless, three notable limitations persist in the current research. First, most studies primarily examine the performance of DRL methods in scenarios with a fixed coefficient of performance (COP) of the heat pump, which often overlooks the significant influence of the ambient temperature on the energy usage of heat pumps. Secondly, many studies have not considered price-driven demand–response control strategies, as they often utilize fixed or stepped electricity prices as experimental conditions. The absence of these strategies limits the exploration of more dynamic and real-time price variations in the research. Furthermore, when applying the DRL method to optimize HEMS, many studies prioritize cost reduction and enhancing human comfort as single-objective optimizations, while overlooking the crucial objective of increasing the incorporation of RES into the public grid. Although this strategy may enhance the system's economic performance, it is crucial to acknowledge that it simultaneously reduces the consumption rate of renewable energy. Therefore, adopting a multi-objective optimization approach in the operation of HEMSs is crucial for enhancing energy flexibility, enabling flexible loads, and facilitating the integration of RESs.

1.3. Contributions

Based on the literature reviewed above, this study's contributions can be summarized as follows:

- Using measured data of a zero energy house, we proposed a novel multi-objective optimization algorithm for residential hybrid energy system operations based on

MDPs with high-dimensional action spaces and evaluated optimization performances of various DRL algorithms, including TD3, DDPG, SAC, and PPO.

- Regarding system constraints, we developed a new multi-objective optimization reward function that guarantees optimizing goals, specifically reducing system energy costs and increasing the ratio of PV self-consumption. Furthermore, we have optimized the environment model and reward function by incorporating expert experience, which enhanced data utilization and improved the model's adaptability to small-sample data.
- All cases in this study used simulated dynamic COP and RTP as experimental conditions. In addition, we tested the effect of the DRL models on a floating FiT scenario. These findings provide valuable insights into the practical application of DRL, offering practical implications for integrating dynamic pricing mechanisms and renewable energy incentives into real energy systems.

2. Model Formulation

This section introduces the energy management of an intelligent HEMS. Subsequently, the system's operational optimization problem is translated into a mathematical model. Finally, all mathematical models are transformed into an MDP, applying DRL methods to determine the optimal solution.

2.1. Energy Management Optimization Model

This study focuses on an intelligent HEMS implemented in an existing zero energy house (ZEH), whose structure is shown in Figure 1. The building comprises a rooftop PV panel, an air source heat pump, a battery, and other household appliances. Electricity demand is satisfied through a combination of PV generation and grid supply, while the battery facilitates the charging or discharging of electric power for optimized scheduling. The air source heat pump caters to the heating demand, supplemented by a domestic hot-water tank for thermal buffering. Notably, the interplay between the battery storage and the grid is disregarded in this study to prevent the arbitrage effect that may arise when RTP is considered.

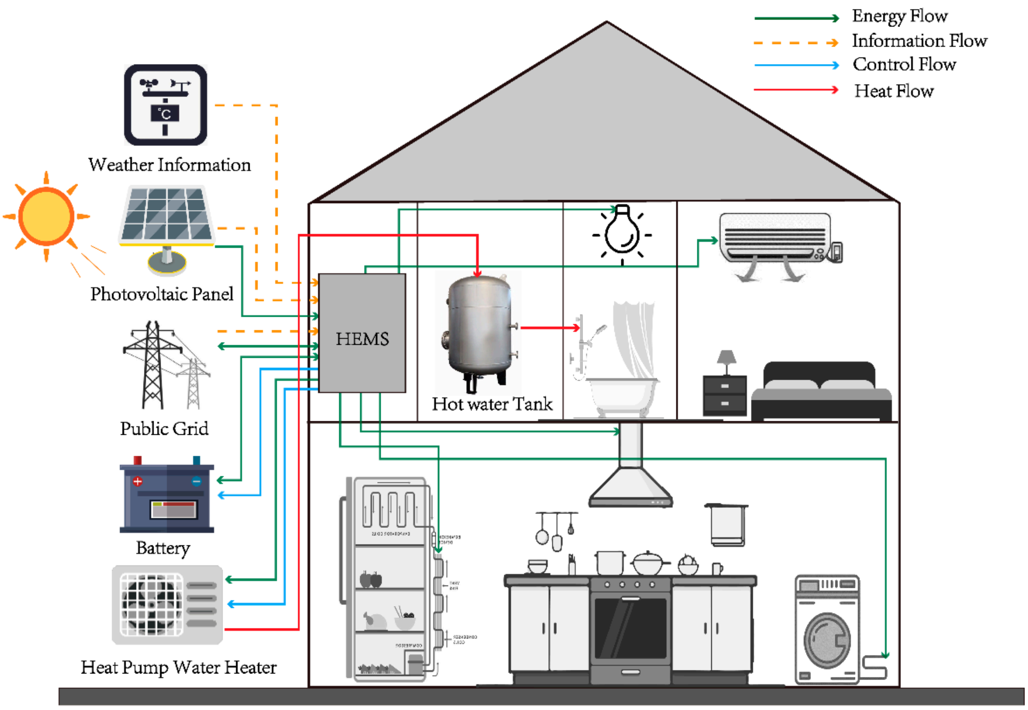

Figure 1. Structure of the examined building energy system.

2.1.1. Objective Function

This study aims to minimize the operational cost and maximize the PV self-utilization percentage of the HEMS by optimizing and regulating the power of the battery and air source heat pump. Hereby, the objective functions encompass the average operating cost and the average PV self-utilization ratio during system operation, defined as Equations (1) and (2).

$$\underset{t\in T}{min} C_{per} = \frac{1}{T}\sum_{t\in T}\Big(E_{gird}(t) * P_{gird}(t) - E_{sell}(t) * P_{sell}(t)\Big) \tag{1}$$

$$\underset{t\in T}{max} r_{pv} = \frac{\sum_{t\in T}(E_{pv}(t) - E_{sell}(t))}{\sum_{t\in T}(E_{pv}(t))} \times 100\% \tag{2}$$

where C_{per} denotes the operational cost, r_{pv} denotes the PV self-consumption ratio, and T is the entire duration of time increments. $E_{gird}(t)$ and $E_{sell}(t)$ are the purchased and sold electricity at time slot t, respectively. $P_{gird}(t)$ is the tariff for purchasing electricity from the public grid at time slot t, considering dynamic pricing in this study. On the other hand, $P_{sell}(t)$ is the fixed tariff at time slot t.

2.1.2. Energy Balance Constraints

This study assumes that the option to buy electricity from and sell electricity to the grid is always available. The system's energy balance constraints are expressed in Equations (3) and (4).

$$E_t^{demand} = E_t^{PV} + E_t^{battery} + E_t^{gird} \tag{3}$$

$$E_t^{PV} = E_t^{PV\to demand} + E_t^{PV\to battery} + E_t^{PV\to gird} + E_t^{PV\to hp} \tag{4}$$

Equation (3) guarantees that the electricity demand (E_t^{demand}) is met at time slot t, utilizing the available sources, such as the PV system (E_t^{PV}), battery ($E_t^{battery}$), and grid (E_t^{gird}). In addition, Equation (4) guarantees that the total flow from the PV system to fulfill the demand ($E_t^{PV\to demand}$), battery ($E_t^{PV\to battery}$), grid ($E_t^{PV\to gird}$), and heat pump ($E_t^{PV\to hp}$) is equal to the total electricity generated (E_t^{PV}).

2.1.3. Battery Constraints

The battery plays a crucial role in balancing power and transferring loads within the HEMS. The following equations represent the battery model:

$$SOC_{t+1} = \Big(1 - \varepsilon_{battery}\Big)SOC_t + \eta_{battery}^{ch}E_t^{ch} - E_t^{dch}/\eta_{battery}^{dch} \tag{5}$$

$$SOC_{min} \le SOC_t \le SOC_{max} \tag{6}$$

$$0 \le E_t^{ch} \le E_{max}^{ch} \tag{7}$$

$$0 \le E_t^{dch} \le E_{max}^{dch} \tag{8}$$

where SOC_t is the state of charge (SOC) at time slot t, $\varepsilon_{battery}$ represents the self-discharging rate of the battery due to energy dissipation. $\eta_{battery}^{ch}$ denotes the power charging efficiency and $\eta_{battery}^{dch}$ denotes the power discharging efficiency; E_t^{ch} denotes the charging power flows, and E_t^{dch} denotes the discharging power flows. E_{max}^{ch} is the highest charging threshold of the battery, and E_{max}^{dch} is the battery's maximum discharging capacity.

2.1.4. Heat Pump Constraints

The correlation between the electricity consumption (E_t^{hp}) and thermal production ($P_t^{thermal}$) at the time interval t is described by Equation (8):

$$P_t^{thermal} = COP_t^{hp} \times E_t^{hp} \tag{9}$$

As previously discussed, the ambient temperature stands out as the primary factor influencing the COP of air-source heat pumps. By performing regression calculations on the data provided by the manufacturer, the COP_t^{hp} can be expressed as a function dependent on ambient temperature ($Temp$), as shown in Equation (10) [21]:

$$COP_t^{hp} = 3.41/\left(1 - 0.014 \times Temp_t - 0.0000035 \times Temp_t^2\right) \tag{10}$$

The incorporation of a thermal storage tank enhances the flexibility of the heat pump operation, and its operational constraints can be described as follows:

$$P_{min}^{thermal} \leq P_t^{thermal} \leq P_{max}^{thermal} \tag{11}$$

$$P_{t+1}^{tank} = (1 - \varepsilon_{thermal})P_t^{tank} + \eta_{tank}^{ch} P_t^{ch} - P_t^{dch}/\eta_{tank}^{dch} \tag{12}$$

$$0 \leq E_t^{ch} \leq E_{max}^{ch} \tag{13}$$

The thermal energy balance is described as follows:

$$P_t^{thermal} - P_t^{ch} + P_t^{dch} = d_t^{thermal} \tag{14}$$

where P_t^{tank} is the stored energy of the hot water tank at the time slot t; $\varepsilon_{thermal}$ represents the loss of energy during thermal energy storage. The variables P_t^{ch} and P_t^{dch} describe the input and output power of thermal energy, respectively. Additionally, $d_t^{thermal}$ denotes the demand for thermal energy.

2.2. Markov Decision Process

Modeling HEMS as an MDP is an essential step in optimizing its behavior and decision-making processes using DRL algorithms. Formally, an MDP is characterized by a quintuple (S, A, R, γ, P), where S signifies the state space, A signifies the action space, R denotes the reward function, γ represents the discount factor, and P stands for the state-transition probability.

2.2.1. State Space

The state space S holds the information or data acquired by the agent upon observing the current environment, which reflects the current state and serves as the foundation for the agent's decision-making process. It is important to highlight that during pre-processing, all observation values must be normalized, which entails scaling each variable's values to the range [0, 1]. The state space in this research primarily comprises four parts:

1. Energy features: Through latent pattern analysis of the data (see Section 4.1), we identified that residential users' PV generation, electricity demand, heat demand, and electricity prices exhibit periodicity in their time series. To facilitate the agent in learning the underlying rules by capturing these patterns, we designed a sliding time window of 24 steps (12 h).
2. Time series features: the time of day (X_t^{hour}), the day of the month (X_t^{month}).
3. Environmental features: outdoor temperature (X_t^{temp}), illumination (X_t^{lux}).
4. Episode step: the present time slot's location within the optimization window (T).

In summary, the S at time slot t is defined in Equation (15):

$$S_t = [T,\ s^{pv}_{t-23}, \ldots s^{pv}_t,\ s^{ele\_demand}_{t-23}, \ldots s^{ele\_demand}_t, s^{ther\_demand}_{t-23}, \ldots s^{ther\_demand}_t, \\ s^{price}_{t-23}, \ldots s^{price}_t X^{hour}_t,\ X^{month}_t,\ X^{temp}_t, X^{lux}_t\Big] \tag{15}$$

2.2.2. Action Space

In this study, the optimization of HEMS is achieved by the continuous action space of battery charging and discharging power and the heat pump's power. As a result, the action space is defined by the battery control factor and the heat pump control factor. These factors represent the control parameters for the battery and heat pump, which are defined as follows:

$$a_t = [a^b_t, a^{hp}_t] \tag{16}$$

where a^b_t is the battery control factor, and a^{hp}_t is the heat pump control factor. The range of a^b_t is from -1 to 1, with the negative value indicating battery discharging and the positive value indicating battery charging. The actual battery power is computed by multiplying the highest charge and discharge rate per hour by a^b_t. The a^{hp}_t is the heat pump control factor, which varies between 0 and 1, and the actual power of the heat pump is calculated by multiplying the maximum power per hour by a^{hp}_t.

2.2.3. Reward Function

This section aims to convert the objective function outlined in Section 2.1.1 into a multi-objective optimization problem, where the reward function usually comprises multiple components and constraints. Currently, two recognized methodologies exist in DRL for formulating reward functions: discrete and continuous. Discrete reward functions are straightforward to implement and converge, but they may lack detailed information, limiting the algorithm's adaptability to environmental changes. On the other hand, continuous reward functions provide richer information, but they can lead to challenges in training due to sparse rewards and slower convergence [32]. Based on previous engineering experience, it has been observed that training often becomes more challenging to converge when both optimization objectives are defined using continuous reward functions. Therefore, the reward function (R) is divided into two components: the energy cost reward (R_{eco}), functioning as a continuous reward, and the PV usage reward (R_{pv}), operating as a discrete reward, as illustrated in Equation (17):

$$R = a \times R_{eco} + \beta \times R_{pv} \tag{17}$$

$$R_{eco} = -\left(\frac{1}{T}\sum\nolimits_{t \in T}\Big(E_{gird}(t) * P_{gird}(t) - E_{sell}(t) * P_{sell}(t)\Big)\right) \tag{18}$$

$$R_{pv} = \begin{cases} 1\ if\ r^{DRL}_{pv} > r^{Baseline}_{pv} \\ -10\ if\ r^{DRL}_{pv} \leq\ r^{Baseline}_{pv} \end{cases} \tag{19}$$

where a and β are reward factors used to regulate the magnitude and significance of R_{eco} and R_{pv}. The minus sign in Equation (18) indicates that a larger value of R_{eco} corresponds to a lower average energy cost. In Equation (19), r^{DRL}_{pv} and $r^{Baseline}_{pv}$ signify the PV self-consumption rates of the DRL and baseline models, respectively. It is also evident from Equation (19) that optimizing PV utilization in this study is accomplished by penalizing the DRL model's PV self-consumption rate when it falls below that of the baseline model. The specifics of the baseline model will be described in the subsequent chapter.

3. Algorithm

3.1. The Selection of the Algorithms

Table 1 outlines the key DRL algorithms, typically classified into value-based and policy-based methodologies. Value-based approaches estimate rewards for selecting actions, while the policy-based ones prioritize actions with higher expected returns by training probability distributions [33]. Value-based methods suit discrete action spaces, while policy-based ones excel in continuous control. The actor–critic method, combining aspects of both, is widely adopted. Here, the actor network selects actions based on policy distributions, and the critic network evaluates action values, which is particularly useful for continuous control tasks. Hence, this study will focus on exploring the actor–critic approaches.

Table 1. Shared characteristics of prevalent RL algorithms.

Algorithm	DQN	DDQN	SAC	A3C	DDPG	PPO	TD3
Category	Value-based	Value-based	Actor-critic	Actor-critic	Actor-critic	Actor-critic	Actor-critic
Data Utilization	Off-policy	Off-policy	Off-policy	On-policy	Off-policy	On-policy	Off-policy
Action Space	Discrete	Discrete	Continuous	Discrete/Continuous	Continuous	Discrete/Continuous	Continuous

DRL algorithms can also be categorized based on adopting either an off-policy or on-policy approach, which depends on the agent's method of interplaying with the environment. Specifically, off-policy methods enable agents to learn from accumulated experience or direct interaction with the environment [34]. Conversely, on-policy methods restrict learning to direct interaction only. Obviously, off-policy methods exhibit a higher utilization rate for data samples. As this study examines measurement data from an actual HMES with limited and slow data collection, the off-policy method is preferred for its sample efficiency. On the other hand, on-policy methods are better suited for scenarios with data generated using simulators.

In summary, this study opts for three prevalent off-policy actor-critic algorithms (SAC, DDPG, TD3) to tackle optimization issues in continuous action spaces. Their performance is assessed to determine the optimal solution for this scenario. Furthermore, an on-policy actor-critic algorithm (PPO) is selected for comparison to ascertain the superiority of the off-policy approach. The model development process is depicted in Figure 2.

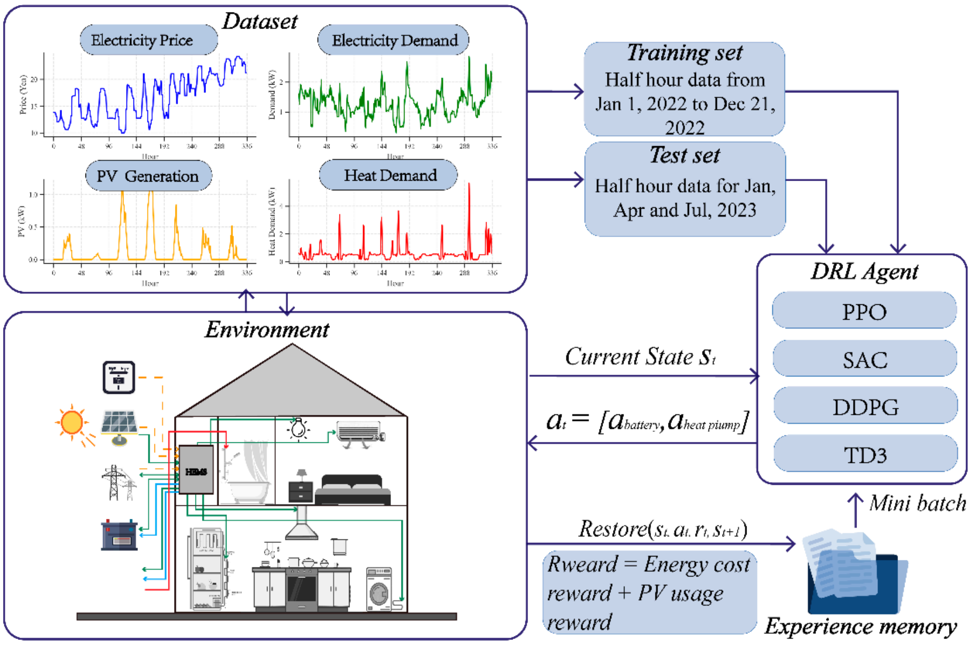

Figure 2. The comprehensive diagram of the HEMS optimization approach.

3.2. Deep Deterministic Policy Gradient (DDPG)

The DDPG stands as a classic actor-critic algorithm, stemming from the Deterministic Policy Gradient (DPG) and integrating key elements of DQN. It inherits two pivotal advantages from DQN: an empirical replay mechanism and an independent target network, and its learning procedure is shown in Figure 3.

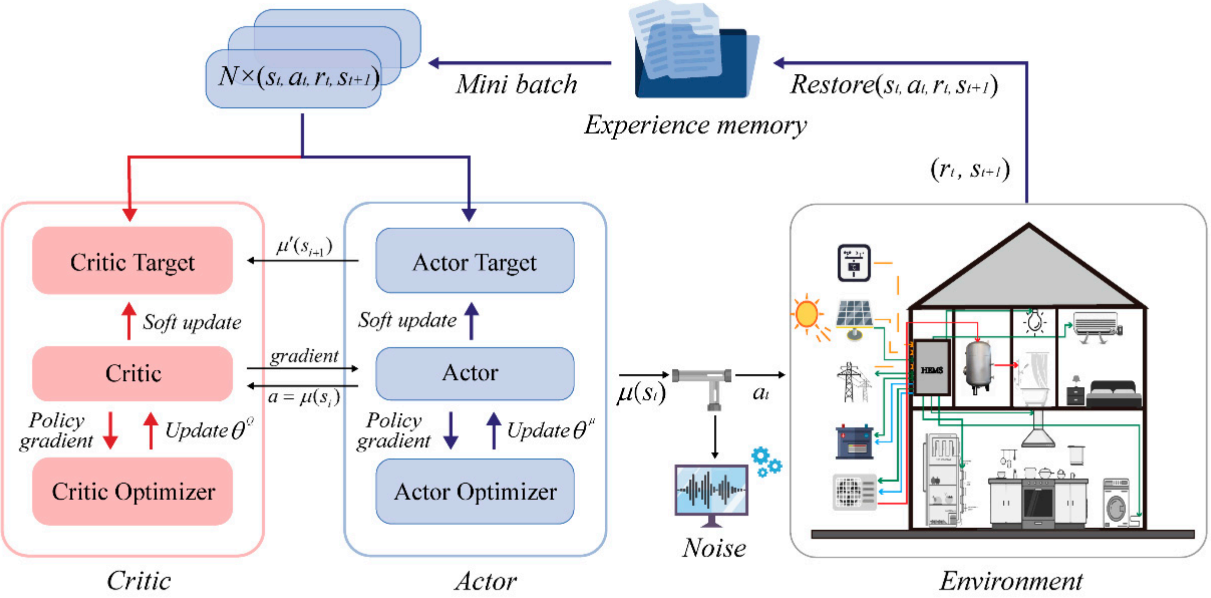

Figure 3. The learning procedure of the DDPG's interaction with the environment.

During the exploration phase, the action-selection process involves feeding the current state, denoted as s_t, along with random noise, represented as x_t, through the actor network, as depicted in Equation (20). Subsequently, upon the environment executing a_t, the (r_t, s_{t+1}) generated at the present slot is computed and stored in the experience replay buffer alongside (a_t, s_t) that was passed into the environment. In the subsequent training process, the agent selects a subset of these transition tuples from the replay buffer and feeds them into the actor network for learning through small-batch sampling. The critic network and its target counterpart in the actor network evaluate the target value y_i for the tuples (a_t, s_t, r_t, s_{t+1}), followed by updating the network by minimizing the loss function L, as described in Equations (21) and (22):

$$a_t = \mu(s_t|\theta^\mu) + x_t \tag{20}$$

$$y_i = r_i + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1}|\theta^{\mu'}\right)|\theta^{Q'}\right) \tag{21}$$

$$L = \frac{1}{N}\Sigma_i\left(y_i - Q\left(s_i, a_i|\theta^Q\right)\right)^2 \tag{22}$$

Equation (23) illustrates the agent's process of computing the gradient of the actor-network policy. Subsequently, upon obtaining the policy gradient, the agent updates the entire network's parameters by ascending the gradient, as described in Equations (24) and (25) [26].

$$\nabla_{\theta^\mu}\mu|_{s_t} \approx \frac{1}{N}\sum_i \nabla_a Q\left(s, a|\theta^Q\right)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i} \tag{23}$$

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \tag{24}$$

$$\theta^\mu \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \tag{25}$$

3.3. Twin Delayed Deep Deterministic Policy Gradient (TD3)

Section 3.2 indicates that the DDPG is a modification of the DQN. Consequently, the DDPG inherits numerous advantages from the DQN; however, it also inherits certain limitations, notably the issue of overestimation. For example, when updating the value function in a traditional DQN, the agent always chooses the a_t with the calculated maximum Q-values, leading to an overestimating issue.

To tackle this overestimation challenge, Hasselt initially introduced the double Q-learning method and implemented it within the framework of the DQN, known as the DDQN [35]. The DDQN utilizes a pair of value functions to determine the most advantageous action for the next interaction, thereby efficiently mitigating the overestimation challenge linked to Q-values. Aiming to tackle the overestimation challenge within the DDPG, the TD3 algorithm follows a comparable strategy by integrating a second critic and target critic pair [36]. The TD3 algorithm effectively mitigates the q-value overestimation by employing two critic networks and calculating the target value y_i as the minimum output from these networks, as depicted by Equation (26) [27]:

$$y_i = r_t + \gamma \underset{j=1,2}{min} Q'_j(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'}) \tag{26}$$

Additionally, the TD3 algorithm enhances training stability by decreasing the frequency of updates to the actor network. Observations made by the creators of the TD3 algorithm indicate that stabilizing the Q value before learning the policy results in fewer erroneous updates in the actor network, thereby contributing to training stabilization [37]. As the TD3 algorithmic flow closely resembles that of the DDPG, it will not be elaborated upon in this section.

3.4. Soft Actor-Critic (SAC) Method

The SAC algorithm also falls under the actor-critic methodology that integrates elements of maximum entropy, initially proposed by Tuomas Haarnoja in 2018 [38]. Its key distinguishing characteristic lies in entropy regularization. The SAC algorithm discourages excessively deterministic strategies by promoting broader exploration through entropy regularization, thereby preventing the agent from becoming trapped in local optima [39]. The loss function in the SAC algorithm consists of three components: the critic's loss, the actor's loss, and the entropy regularization term. While the update process in the SAC algorithm shares similarities with the DDPG, the SAC model considers both the Bellman error and policy entropy during the minimization of the loss function [28], as explained below:

$$y_i = r_i + \gamma \underset{i=1,2}{min} Q'(s_{i+1}, a_{i+1}) - \alpha log_{\pi_\theta}(a_{i+1}|s_{i+1}) \tag{27}$$

where the policy π_θ is typically parameterized using a Gaussian distribution, and the parameters γ and α are hyperparameters that manage the balance between the actor's loss and the entropy regularization term. Specifically, the SAC algorithm also adopts the strategy of simultaneous learning of dual Q functions to reduce the overestimation bias associated with a single Q function [40]. By incorporating these three components into the loss function, the SAC algorithm simultaneously considers expected returns and policy entropy during the optimization of both the policy and the value function, which enables it to learn robust and diverse strategies for continuous control problems.

4. Case Study

4.1. Data Source

This research will validate all the simulation models using energy generation and consumption data sourced from an operational zero energy house (ZEH) in Japan's Kitakyushu region [21]. The HEMS collected data automatically every 30 min and spanned approximately 20 months, encompassing half-hourly data from 1 January 2022, to 30 August 2023.

Table 2 presents the basic variables and their associated value configurations for the HEMS utilized in the proposed RL environment.

Table 2. Parameters used in the simulation models.

Parameter	Descriptions	Value
$\varepsilon_{battery}$	Power dissipation rate	0.01
$\eta^{ch}_{battery}$	Power charging efficiency	0.95
$\eta^{dch}_{battery}$	Power discharging efficiency	0.95
$Cap_{battery}$	Energy storage capacity of the battery	6.0 kWh
P^{ch}_{max}	Peak charging power rate	0.75 kWh
P^{dch}_{max}	Peak discharging power rate	0.75 kWh
SOC_{min}	Minimum levels for battery SoC	0.20
SOC_{max}	Maximum levels for battery SoC	0.90
$\varepsilon_{thermal}$	Thermal energy dissipation rate	0.20
$\eta^{ch}_{thermal}$	Charging efficiency of thermal energy	0.90
$\eta^{dch}_{thermal}$	Discharging efficiency of thermal energy	0.90
$Cap_{thermal}$	Thermal energy storage capacity	20.0 kWh
$P^{thermal}_{max}$	Lowest thermal energy generation	0
$P^{thermal}_{min}$	Highest thermal energy generation	3.0 kW
u_{max}	Maximum thermal energy input	3.0 kW
u_{min}	Minimum thermal energy input	0

The final dataset is derived from raw data after correlation analysis, comprising nine features: PV generation (kWh), electricity demand (kWh), thermal demand (kWh), electricity prices (JPY), the COP of the heat pump, month, hour, outdoor temperature, and illumination. Measured data comprises all values, except the RTP and dynamic COP generated through simulation. Figure 4 offers an overview of the dataset, illustrating the relationship between energy and time features. It is evident from Figure 4 that the PV generation, electricity demand, thermal demand, and the RTP exhibit strong seasonal characteristics. Hence, it is not comprehensive to assess the optimization effectiveness of a model with only a small dataset, such as a few weeks or individual months. To address this, we first categorized the remaining eight months of data into cooling, heating, and transition seasons, based on energy consumption characteristics. Subsequently, we selected a representative working month from each of these three seasons to serve as the test set.

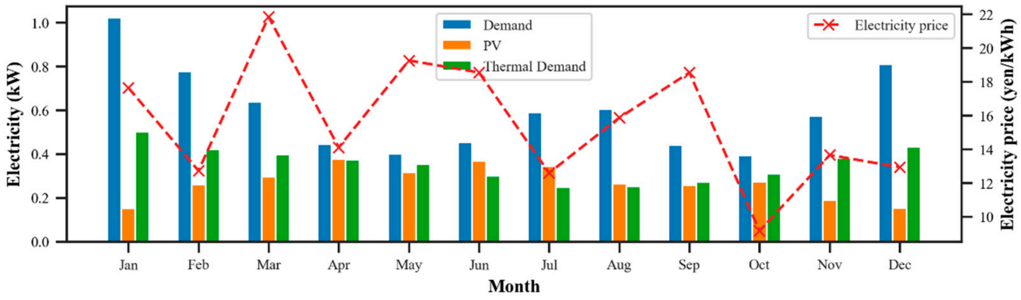

Figure 4. The average monthly distribution of electricity demand, thermal demand, PV generation, and electricity prices.

4.2. Environment Setup

When the agent explores the RL environment, it experiments with various actions to discover the optimal strategy. However, excessive exploration can complicate DRL model training. This study aims to integrate the expert experience into the environment design to mitigate unnecessary or inefficient exploration, thereby enhancing model performance on small-sample datasets. Therefore, we aim to develop constraints based on expert experience and apply these constraints to limit the agent's exploration during interactions with the environment. The scheduling rules based on these constraints are as follows:

- For the battery, the agent controls the charging and discharging power. When PV generation exceeds the electricity and thermal demand, and the battery has available charging capacity, the system prioritizes storing the excess PV in the battery, and the extra remainder is then supplied to the public grid for profit. Conversely, when PV generation falls short of meeting the electricity and thermal demands, and the battery has available discharge capacity, and the system prioritizes discharging the battery. Any remaining shortfall in power is then purchased from the public grid. If neither of the above conditions is met, the battery will remain inactive, refraining from any action.
- For the heat pump, the agent controls the power. When the heat pump power exceeds the thermal demand, if the capacity of the hot water tank is less than the maximum thermal holding capacity, the excess thermal is flushed into the hot water tank. Otherwise, the excess hot water is discarded, and the agent is penalized for exceeding the limit. When the heat pump power is less than the thermal demand, the tank releases hot water if the excess hot water can satisfy the remaining thermal demand. If the excess hot water is insufficient, the electric water heater is activated to supplement the thermal, while the agent incurs a penalty for exceeding the limit.

Agents can significantly mitigate unnecessary exploration behaviors by implementing the predefined constraints within the interactive environment. This approach also prevents instances of battery or hot water tank overcharging or discharging and the exploitation of electricity prices for arbitrage. Furthermore, minimizing the number of agent experiments and errors can enhance the utilization rate of the training samples, thereby achieving the desired experimental outcomes, even with limited data.

4.3. Model Setup

To validate the optimization effects of various DRL algorithms in this environment, we constructed five models, comprising a baseline model and four DRL models, selected as outlined in Section 3.1. They are as follows:

- M.0: It serves as the baseline, accurately reflecting the real-world usage state of the user. The HEMS currently employed in the target house operates on a rule-based control approach. Specifically, the battery and the heat pump operate at full power without real-time power control, following the constraints outlined in Section 4.2. The heat pump's operation schedule is determined based on the user's actual usage, with fixed full-power operation scheduled daily from 4 a.m. to 7 a.m. to refill the hot water tank. The heat pump provides thermal demand through real-time heating during the remaining time.
- M.1: It utilizes the PPO as the optimization approach, representing an on-policy DRL method for comparison.
- M.2: It adopts the SAC as the optimization approach.
- M.3: It adopts the DDPG as the optimization approach.
- M.4: It adopts the TD3 as the optimization approach. Notably, M3 and M4 utilize identical hyperparameters for comparative purposes.

The experimental environment utilized in this study was constructed using the Python language, leveraging the OpenAI gym framework [41]. Meanwhile, the DRL algorithms employed in the experimentation were implemented by the Stable Baselines framework [42]. Table 3 displays the essential hyperparameters for various algorithm configurations. Given the orientation of this research toward practical applications rather than optimizing hyperparameters for specific scenarios, we aim to utilize the default hyperparameters of the Stable Baselines framework while ensuring algorithm performance. This approach could ensure the universality of the proposed models.

Table 3. Hyperparameters for various algorithm configurations.

Category	M.1 (PPO)	M.2 (SAC)	M.3 (DDPG)	M.4 (TD3)
Activation function	Tanh	Relu	Relu	Relu
Learning rate	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
Batch size	256	256	256	256
Replay memory capacity	10^6	10^6	10^6	10^6
Discount factor	0.99	0.99	0.99	0.99
Hidden Layer Dimensions	64	256	256	256
Trace-decay parameter	0.95	None	None	None
Polyak averaging	None	5×10^{-3}	None	None
Delay steps in TD3	None	None	None	2

4.4. Experimental Setup

In this study, we utilized 17,520 data points collected at half-hour intervals over the timeframe from 1 January 2022, to 31 December 2022, as the training set. The test set comprised data from 1 January 2023 to 31 August 2022. Each scheduling interval was set at half an hour. Furthermore, the test set was partitioned into the cooling season, the heating season, and the transition season for separate evaluation of the models. This study encompassed two steps:

Step.1: All real data are utilized to train and evaluate the proposed DRL models, determining the best optimization algorithm in this scenario.

Step.2: In this step, the optimal optimization algorithm identified in step 1 is selected as the research focus. The training and test sets remain unchanged, but the feed-in tariff (FiT) gradually decreases by JPY 2 in each training iteration. This decrease aligns with the observed trend of FiT in Japan over the past 14 years, as illustrated in Figure 5. We will simulate this change by sequentially reducing the FiT by JPY 2 in both the training and test environments. The FiT in the training environment will consistently remain JPY 2 higher than that in the test environment. It is worth noting that both the test and training datasets will remain unchanged throughout this simulation.

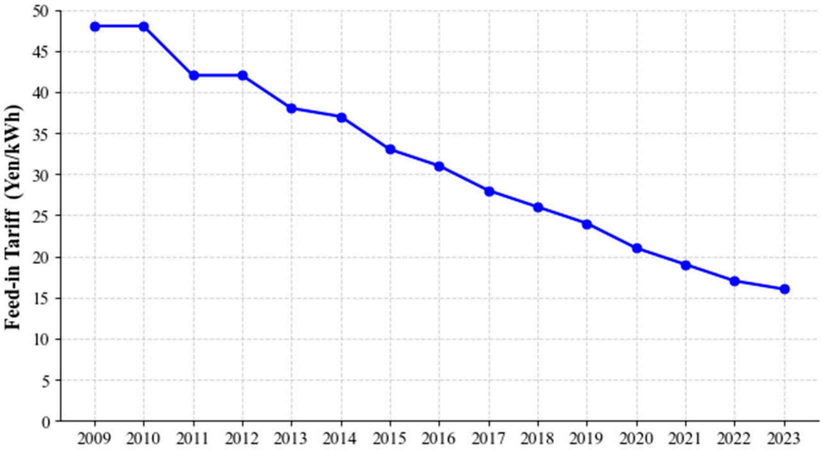

Figure 5. Changes in the FiT in Japan.

5. Results and Discussion

5.1. Training Process Analysis

As outlined in the previous section, we utilized a year's worth of data as a training set, simulating 17,520 steps of operational optimization per set. Each model was saved using the callback function provided by the stable baseline for optimal training results [42]. Reducing the impact of randomness on experimental results is crucial to provide a more objective assessment of the algorithm's performance [43]. Therefore, we introduce three different random seeds, utilized across the training loops to control different random streams in each model execution, ensuring comparability among diverse experiments and repeatability within the same experiment [31].

Figure 6 illustrates the changes in training episodes for the different methods. It is evident that all models experience a sharp increase in training gains within the first 20 rounds of training, followed by a decline in growth rates and a leveling off within 60 to 100 rounds. It should be noted that the reward of the PPO algorithm in the first 20 rounds is significantly lower than that of other algorithms. Although its performance stabilizes as the number of training epochs grows, its average reward is still lower than that of other algorithms, which indicates that the limitations of the on-policy method led to its weaker training performance on small data samples compared to other off-policy algorithms. Among the three off-policy methods, the training curves of the TD3 and DDPG algorithms are more stable compared to that of the SAC algorithm. Additionally, the average reward of the TD3 is the highest among all off-policy methods, suggesting that the TD3 algorithm achieved the best training effect on the training set.

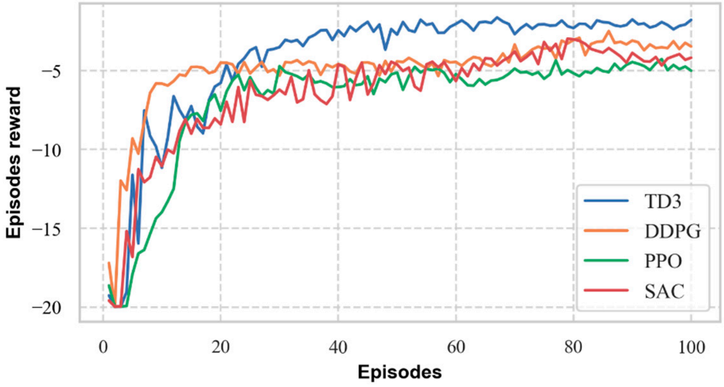

Figure 6. Average episode rewards during the training process.

5.2. Performance Evaluation

5.2.1. Energy Cost Optimization

Given the significant fluctuations in the characteristics of the test dataset across different months, we adopted a comprehensive approach to evaluate the online decision-making prowess of the DRL agents. Specifically, we selected three months characterized by distinctly varied data distributions to form our test set: January, the heating season; July, the cooling season; and April, the transition season. The input dataset, encompassing customer load demand, PV generation, and RTP, among others, comprises measured user data from 2022, with a FiT set at JPY 17. Subsequently, we computed the energy cost savings achieved by each of the five models within these three test months. The outcomes of these analyses are briefly summarized in Table 4.

Table 4. Energy cost optimization results.

		Baseline	**PPO**	**SAC**	**DDPG**	**TD3**
January	Cost (JPY)	23,003.61	21,253.36	20,775.69	20,737.13	20,591.87
	VS Baseline		7.61%	9.69%	9.85%	10.48%
April	Cost (JPY)	14,565.58	12,474.97	12,670.92	12,164.27	12,370.42
	VS Baseline		14.35%	13.01%	16.49%	15.07%
July	Cost (JPY)	10,201.38	9074.35	8709.357	8531.925	8219.691
	VS Baseline		11.05%	14.63%	16.37%	19.43%
Total	Cost (JPY)	47,770.57	42,802.68	42,155.97	41,433.32	41,181.98
	VS Baseline		10.40%	11.75%	13.27%	13.79%

All DRL models successfully attained the energy cost optimization objective across each test period, with overall energy cost reductions of 10.40%, 11.75%, 13.27%, and 13.79% compared to the baseline model. Notably, TD3 exhibited the most notable overall optimization effect, demonstrating superior performance in January and July, with improvements of 10.48% and 19.43% over the baseline model, respectively. Following closely, the DDPG achieved a comparable optimization level to the TD3 model, with its performance in April surpassing that of the TD3 model, with a 16.49% cost reduction compared to the baseline model, suggesting the efficacy of the actor–critic framework algorithm in this context. Conversely, the PPO algorithm, operating as an on-policy comparison group, demonstrated the lowest total optimization, highlighting the preference for off-policy algorithms, with improved sample efficiency in practical applications within this context. Moreover, it is notable that while the optimization effects of the four algorithms appear similar in January and April, they exhibit significant disparities in July. Upon analyzing the distribution of the test data, we observed a weaker periodicity in the cooling season data compared to other periods. Specifically, this was evident in the pronounced fluctuations in energy demand and electricity prices, posing higher demands on the learning capabilities of agents. Overall, the TD3 and DDPG algorithms demonstrated optimal and suboptimal optimization effects during the cooling season, indicating their enhanced capacity to discern patterns in feature changes.

Moreover, it is notable that while the optimization effects of the four algorithms appear similar in January and April, they exhibit significant disparities in July. Upon analyzing the distribution of the test data, we observed a weaker periodicity in the cooling season data compared to other periods. Specifically, this was evident in the pronounced fluctuations in energy demand and electricity prices, posing higher demands on the learning capabilities of THE agents. Overall, the TD3 and DDPG algorithms demonstrated optimal and suboptimal optimization effects during the cooling season, indicating their enhanced capacity to discern patterns in feature changes. The variances in optimization strategies among different algorithms will be thoroughly discussed in Section 5.2.3.

5.2.2. PV Self-Consumption Ratio Optimization

Figure 7 illustrates the statistics of the PV consumption ratio across the three test datasets. The scheduling outcomes indicate that, compared to the baseline model, the DRL model notably enhanced PV consumption and energy self-sufficiency ratios in April and July. However, its performance was less commendable than the January baseline model. This discrepancy can be attributed to January being characterized by the highest energy demand and the lowest PV generation, resulting in saturation of PV consumption by the baseline model. Consequently, any decision by the DRL model to sell additional PV to the public grid would inevitably impact the PV consumption ratio in this scenario.

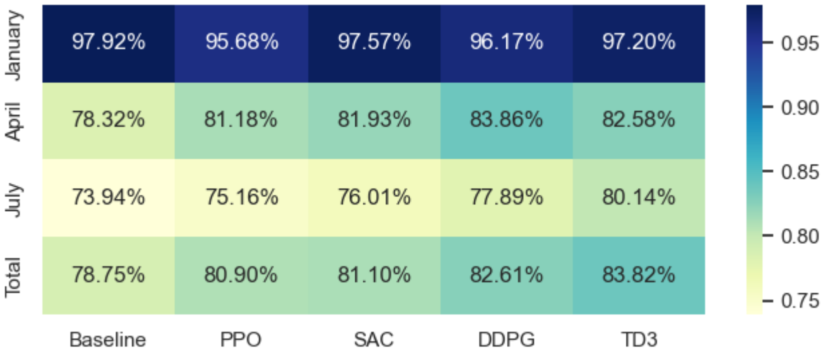

Figure 7. Thermal visualization of PV self-consumption rate.

It should be noted that this study employs a reward function with discrete constraints to optimize the PV self-consumption, as outlined in Section 2.2.3. Consequently, the effect of DRL models exhibits certain limitations compared to the baseline model. Hence, the augmentation of PV self-consumption is markedly influenced by cost optimization, demonstrating a notable negative correlation, as evidenced by the performance in July. As depicted in Figure 7, the optimization effectiveness of DDPG and TD3 surpasses that of the PPO and SAC algorithms. Specifically, the PV self-consumption rate increased by 5.54% for the DDPG and 4.26% for TD3 in April. Similarly, these figures stood at 3.95% and 4.98% in July. These findings suggest that the DDPG and TD3 algorithms are more adept at optimizing the utilization of PV generation in this scenario, and the TD3 model achieves the best optimization effect.

5.2.3. Comparison of Operation Strategy

In this section, we will delve into the optimization strategies of each algorithm across three test months, elucidating specific optimization strategies concerning the SOC of the battery and heat pump power. To facilitate this analysis, we selected one week of data from each of these three months as the test set, adhering to the experimental process outlined in Section 4.4. Figure 8 illustrates the optimal scheduling outcomes for each experimental week, where the left ordinate denotes the SOC of the battery, the right ordinate shows the electricity price, while the horizontal axis denotes the hour of the day.

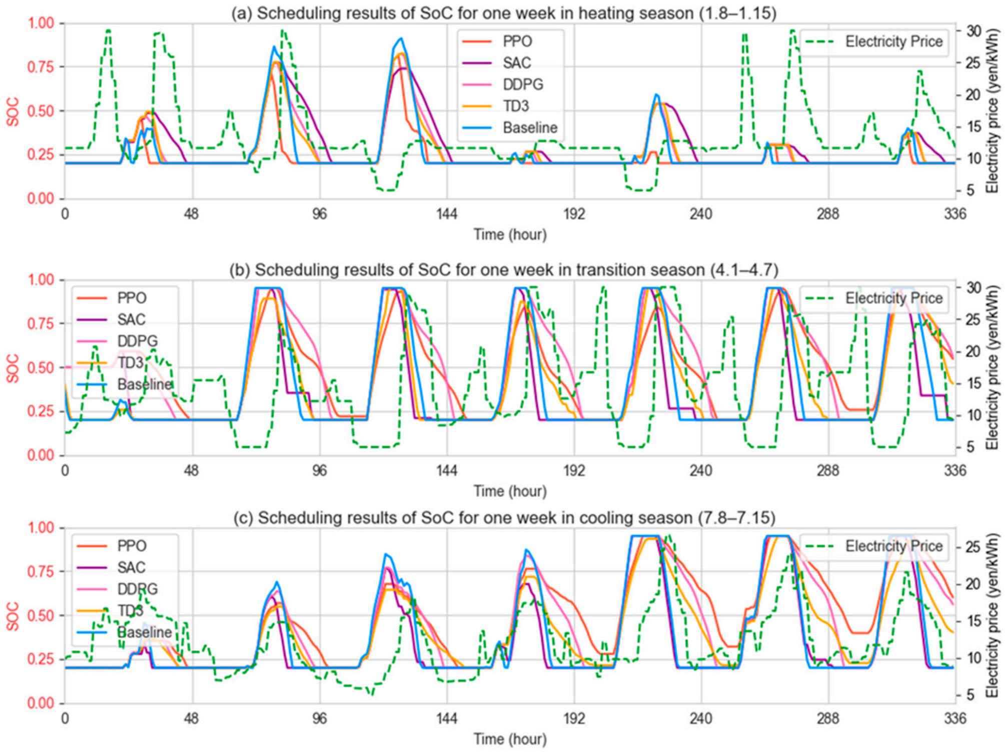

Figure 8. The optimal battery operation strategy of different DRL models under a typical working week.

By comparing Figure 8, we can discern the differences in operational approaches between the four DRL methods and the baseline model, particularly in the selection of charging and discharging time points and the control of battery power. In the operational strategy derived from the DRL methods, the battery power during the charging period decreases compared to the baseline model, which fosters a higher transfer of PV generation to the heat pump or the public grid. While discharging, the DRL methods tend to regulate the discharge power according to real-time electricity prices, prioritizing electricity retention for peak price periods, particularly noticeable in Figure 8a,b. It is noteworthy that the operational strategy of the battery is also influenced by the operational strategy of the heat pump, which will be elaborated on in the discussion of the heat pump strategy.

Upon examination of Figure 8, it becomes apparent that while the operational strategies of the four DRL algorithms are similar, their execution effects vary considerably, leading to differences in optimization outcomes. As depicted in Figure 8a, the RTP is notably lower than the FiT during periods of PV generation. Consequently, all DRL methods opt to charge at full power in the initial half of the charging period, reduce the charging power in the latter half, and opt to vend surplus electricity to the public grid to achieve higher cost optimization. However, the PPO algorithm over-implements this strategy on days 2, 4, and 5, resulting in insufficient electricity available for release during evening peak pricing, thereby diminishing the cost optimization effect. During the discharge period, the PPO model exhibits similar inadequacies, conserving power for peak tariff periods only on days 3 and 5, while the three actor–critic framework DRL approaches consistently execute the strategy effectively. Notably, the discharge strategies of the DDPG and TD3 algorithms are similar. Still, the SAC algorithm demonstrates a phenomenon of over-conservation, wherein its power selection during the discharge stage is too conservative, resulting in the battery withholding excess electricity during peak pricing, thereby diminishing the cost optimization effect.

From the observations of Figure 8b,c, it is evident that the strategies employed by the four DRL models during the transition and cooling seasons closely resemble those in the heating season. The primary distinction lies in the fact that, due to sufficient PV generation during these quarters, the issue of the battery not reaching full capacity due to reduced charging power has been alleviated. However, differences between the models primarily manifest in the control of battery power during the discharge period. Across the two test weeks, all models, except the SAC algorithm, prefer storing electricity and releasing it during the evening peak in both load and electricity prices. Notably, TD3's choice of charging and discharging times shows a more pronounced response to load peaks and electricity prices in both the transition and cooling seasons, which aligns with the cost optimization results presented in Table 4. From the results shown in Table 4 and the dynamic dispatch, it is evident that TD3 responds better, leveraging RTP differences to achieve cost savings. Conversely, the PPO algorithm continues to opt for reducing battery charging and discharging power, while neglecting the response to real-time electricity prices. It is important to highlight that the SAC's optimization effect in this study is inferior to that of the DDPG and TD3 algorithms, which can be attributed to the SAC's suitability for tasks requiring exploration and diversity, whereas TD3 excels in accuracy and stability-focused tasks. Thereby, the TD3 algorithm emerges as the optimal choice for battery regulation in this scenario.

Figures 9–11 illustrate the heat pump optimization strategies across different DRL methods during typical work weeks, across three seasons. In the heating season, as depicted in Figure 9, DDPG and TD3 prioritize battery charging during the ascending phase of PV generation. Once the battery reaches full capacity, any surplus photovoltaic and battery electricity is utilized to heat the hot water tank. In contrast, the SAC algorithm operates the heat pumps at moderate power levels during low electricity prices. While PPO's strategy aligns somewhat with the DDPG and TD3 models, there are discrepancies in the timing and power selection, resulting in an overall less effective operation compared to the DDPG and TD3 models. Figure 10 shows that the strategies employed by the DDPG and TD3 models mirror those observed in the heating season, with the SAC algorithm also adopting a similar approach. Given the surplus PV generation during the transition season and the reduced heating load demand, it proves more economically viable to use renewable energy to fill the heat storage tank with hot water rather than relying on battery power for heating. Among the algorithms, the TD3 algorithm more frequently chooses to activate the heat pump during periods when both low electricity prices and PV generation coexist, thereby achieving optimal cost and photovoltaic self-consumption optimization, as evidenced by Table 4 and Figure 7. Notably, PPO's strategy of running the heat pump in the morning contributes to the insufficient charging power observed in PPO's charging strategy in Figure 8, resulting in inadequate storage power—a flawed strategy. Figure 11 illustrates that the heat pump optimization strategies employed by different DRL methods vary significantly from those observed in other seasons. This disparity arises from the cooling season's minimal heat demand, juxtaposed with abundant photovoltaic power, making it challenging for the DRL models to accurately forecast future heat demand and select appropriate heat pump operation times. In terms of strategic choices, the PPO algorithm opts to utilize battery power for heat storage, the SAC algorithm operates the heat pump at low power levels during periods of low electricity prices, the DDPG runs the heat pump during the low electricity price period in the morning, while TD3 schedules heat pump operation both during the low electricity price period in the morning and in the afternoon, when there is a surplus of PV generation. This dual scheduling approach maximizes the use of renewable energy while also taking advantage of lower electricity prices. By aligning heat pump operation with the periods of high PV output, TD3 effectively reduces reliance on the grid and further lowers operating costs. As shown by the cost optimization results in Table 4, TD3's strategy is the most effective in this scenario, integrating low-cost grid power and excess PV generation to achieve cost-efficient optimization.

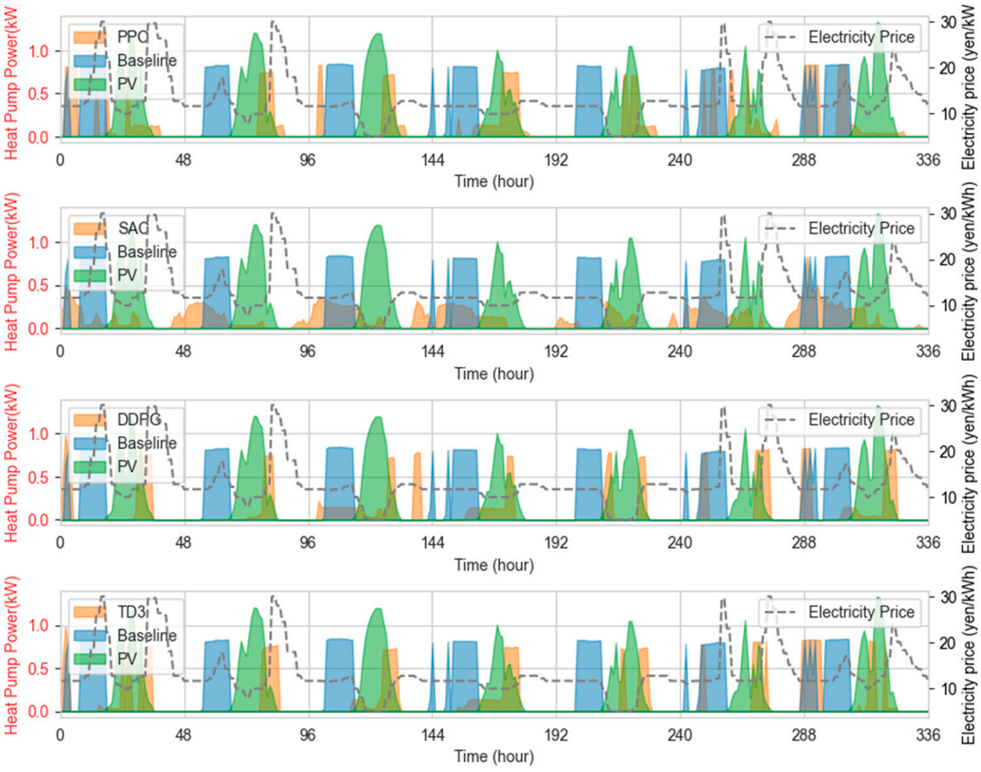

Figure 9. The optimal heat pump operation strategy of different DRL models under a typical working week in the heating season.

Through the series of experiments conducted, a notable phenomenon emerged: the SAC algorithm's performance was consistently weaker compared to TD3 in both battery and heat pump regulation. This discrepancy can be attributed to two main factors within the experimental framework. Firstly, pre-tuned models within the Stable Baselines framework are trained with default hyperparameters or specific selections based on prior knowledge, which compensates for the difficulty of tuning hyperparameters with the TD3 model. Secondly, rule-based approaches are integrated into environmental models to plan or optimize actions. This supplementary framework may impose constraints or limitations on the entropy-maximizing exploration in the SAC algorithm. In summary, the comprehensive nature of TD3's optimization effectiveness is reflected in its ability to balance multiple key factors—minimizing operating costs, maximizing the use of local PV generation, and strategically scheduling heat pump operation to take advantage of low electricity prices and high renewable energy availability. Therefore, the TD3 algorithm emerges as the most well-rounded and effective optimization algorithm in this scenario.

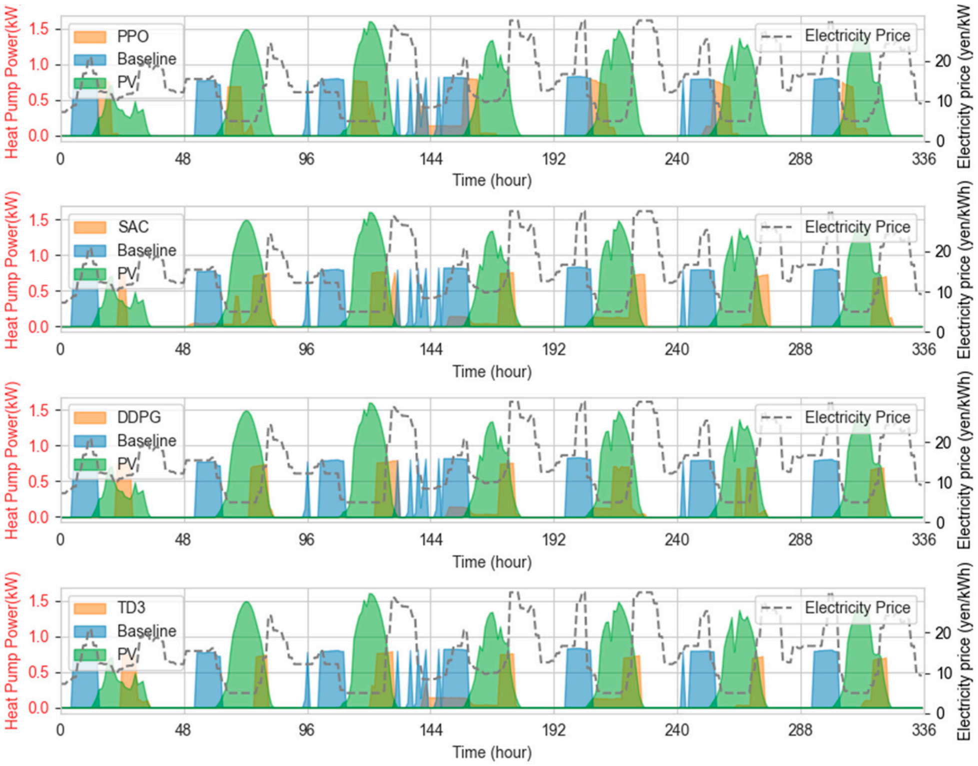

Figure 10. The optimal heat pump operation strategy of different DRL models under a typical working week in the transition season.

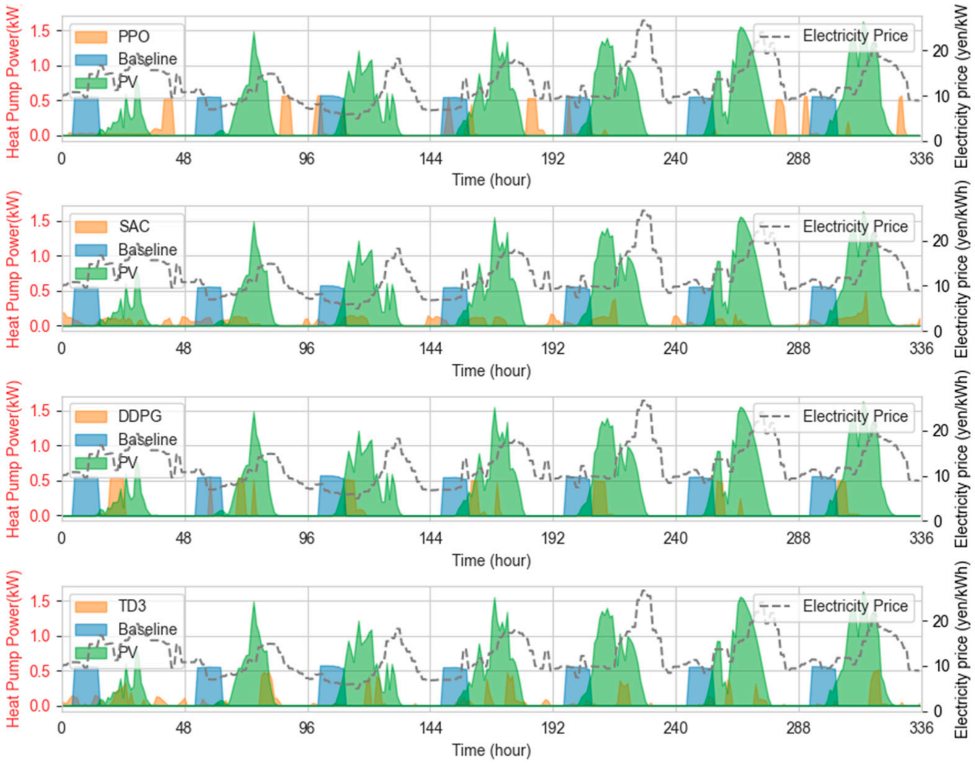

Figure 11. The optimal heat pump operation strategy of different DRL models under a typical working week in the cooling season.

5.2.4. Effects of FiT on Optimization

To explore the optimization potential of the proposed model amid a yearly decline in FiT following its practical application, we conducted a series of tests using the TD3 model. This choice was based on its superior performance observed in the 2022 data. The settings for the training and test sets remained the same, while the FiT was reduced by JPY 2 each year in the training and test environments. Importantly, the FiT in the training environment was always JPY 2 higher compared to the test environment, while the training and test sets remained unchanged. The experimental results are presented in Figure 12, where the horizontal axis represents the FiT, and the y-axis displays the cost optimization impact of the TD3 model relative to the baseline model. The findings suggest that as the FiT decreases, the outcome of cost optimization for the TD3 model first improves, reaching its peak when the FiT drops to JPY 15 and JPY 13 and then gradually declines. When the FiT is reduced to JPY 7, the optimization effect is slightly less compared to 2022, but the cost savings remain significant. This phenomenon can be attributed to two main factors: (1) the decrease in FiT reduces the baseline model's income from selling photovoltaic power to the grid, allowing the DRL model to achieve a higher income by increasing the PV self-consumption rate; and (2) while the lower FiT also reduces the DRL model's revenue from selling PV to the grid, the DRL model compensates by effectively managing battery and heat pump operations to avoid RTP peaks.

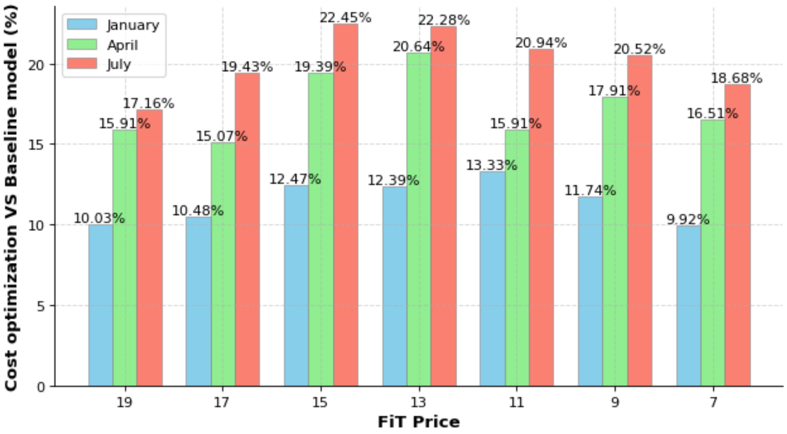

Figure 12. Cost optimization effect of TD3 algorithm with different FiT.

6. Conclusions

DRL-based energy control methods have demonstrated significant potential in both reducing building energy costs and enhancing renewable energy penetration by leveraging their ability to acquire strategies from energy system data, offering a promising approach for optimizing energy management systems. This paper introduces a data-driven DRL control method for optimizing operations of distributed energy systems to lower energy costs, while guaranteeing the PV self-absorption ratio. This study introduces a novel interactive environment and a multi-objective optimization reward function, with their efficacy being substantiated by experimental validation.

In this case study, we meticulously examined the disparities in battery and heat pump optimization strategies among the PPO, SAC, DDPG, and TD3 algorithms in scenarios considering dynamic COP and RTP, systematically evaluating their optimization efficacy across various periods. TD3 not only demonstrated the best average episode rewards on the training set but also achieved the lowest overall operating costs on the test set, with a reduction of 13.79% compared to the baseline model and a 5.07% increase in PV self-consumption. TD3 achieved the best performance in January and July, and only slightly lagged behind the DDPG algorithm in April. Additionally, the study simulated the impact of the FiT on TD3's performance, revealing that the TD3 model maintains high optimization performance even with decreasing the FiT. These findings illustrate the efficacy of the proposed DRL-based approach in enhancing HEMS operation and underscore the potential of advanced DRL algorithms such as TD3 in achieving significant cost savings and improved renewable energy utilization, facilitating informed decisions regarding the selection of DRL algorithms for specialized scenarios.

Future research endeavors will initially concentrate on integrating multi-agent-based DRL technology into this scenario [44,45], alongside developing associated interaction environments and reward function designs. Subsequently, efforts will be directed toward refining the algorithms outlined in this study to enhance their generalization capabilities, facilitating their seamless deployment across other buildings within the same community. Moreover, attention will be devoted to exploring model-based reinforcement learning methods, incorporating additional expert knowledge into model design to bolster learning efficiency [46].

Author Contributions: Conceptualization, Y.X. and Y.L.; methodology, Y.X.; software, Y.X.; validation, Y.L. and W.G.; formal analysis, Y.X.; investigation, W.G.; resources, W.G.; data curation, Y.X.; writing—original draft preparation, Y.X.; writing—review and editing, Y.L.; visualization, Y.X.; supervision, W.G.; project administration, W.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shandong Natural Science Foundation, 'Research on Flexible Control Strategies for Building Energy Systems Based on Deep Reinforcement Learning' (grant number: ZR2023ME160), and the Shandong Natural Science Foundation 'Research on Flexible District Integrated Energy System under High Penetration Level of Renewable Energy' (grant number: ZR2021QE084).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: This study received support from National Natural Science Foundation of China (grant number: 52308098), the Shandong Provincial Natural Science Foundation (grant number: ZR2021QE084), and the Shandong Provincial Natural Science Foundation (grant number: ZR2023ME160).

Conflicts of Interest: We declare that we have financial and personal relationships with other people or organizations that can inappropriately influence our work, and there is professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "Comparative analysis of reinforcement learning approaches for multi-objective optimization in residential hybrid energy systems".

References

1. Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; Meger, D. Deep Reinforcement Learning That Matters. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
2. Ahmed, A.; Ge, T.; Peng, J. Assessment of the renewable energy generation towards net-zero energy buildings: A review. *Energy Build.* **2022**, *256*, 111755. [CrossRef]
3. Balasubramanian, C.; Lal Raja Singh, R. IOT based energy management in smart grid under price based demand response based on hybrid FHO-RERNN approach. *Appl. Energy* **2024**, *361*, 122851. [CrossRef]
4. Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zaremba, W. OpenAI Gym[A]. *arXiv* **2016**, arXiv:1606.01540.
5. Cai, Q.; Qing, J.; Xu, Q. Techno-economic impact of electricity price mechanism and demand response on residential rooftop photovoltaic integration. *Renew. Sustain. Energy Rev.* **2024**, *189*, 113964. [CrossRef]
6. Chowdhury, M.A.; Al-Wahaibi, S.S.; Lu, Q. Entropy-maximizing TD3-based reinforcement learning for adaptive PID control of dynamical systems. *Comput. Chem. Eng.* **2023**, *178*, 108393. [CrossRef]
7. Ding, B.; Li, Z.; Li, Z. A CCP-based distributed cooperative operation strategy for multi-agent energy systems integrated with wind, solar, and buildings. *Appl. Energy* **2024**, *365*, 123275. [CrossRef]
8. Duryea, E.; Ganger, M.; Wei, H. Deep Reinforcement Learning with Double Q-learning. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
9. Fujimoto, S.; Van Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the 35th International Conference on Machine Learning, Stockholm Sweden, 10–15 July 2018; Volume 80, pp. 1587–1596.

10. Gao, Y.; Hu, Z.; Shi, S. Adversarial discriminative domain adaptation for solar radiation prediction: A cross-regional study for zero-label transfer learning in Japan. *Appl. Energy* **2024**, *359*, 122685. [CrossRef]
11. Ghaderi, R.; Kandidayeni, M.; Boulon, L. Q-learning based energy management strategy for a hybrid multi-stack fuel cell system considering degradation. *Energy Convers. Manag.* **2023**, *293*, 117524. [CrossRef]
12. Hou, H.; Ge, X.; Chen, Y. Model-free dynamic management strategy for low-carbon home energy based on deep reinforcement learning accommodating stochastic environments. *Energy Build.* **2023**, *278*, 112594. [CrossRef]
13. Huang, R.; He, H.; Zhao, X. Battery health-aware and naturalistic data-driven energy management for hybrid electric bus based on TD3 deep reinforcement learning algorithm. *Appl. Energy* **2022**, *321*, 119353. [CrossRef]
14. Jia, C.; Li, K.; He, H. Health-aware energy management strategy for fuel cell hybrid bus considering air-conditioning control based on TD3 algorithm. *Energy* **2023**, *283*, 128462. [CrossRef]
15. Jiang, K.; Wang, K.; Wu, C. Trajectory simulation and optimization for interactive electricity-carbon system evolution. *Appl. Energy* **2024**, *360*, 122808. [CrossRef]
16. Kim, D.; Wang, Z.; Brugger, J. Site demonstration and performance evaluation of MPC for a large chiller plant with TES for renewable energy integration and grid decarbonization. *Appl. Energy* **2022**, *321*, 119343. [CrossRef]
17. Kontokosta, C.E.; Spiegel-Feld, D.; Papadopoulos, S. The impact of mandatory energy audits on building energy use. *Nat. Energy* **2020**, *5*, 309–316. [CrossRef]
18. Langer, L.; Volling, T. A reinforcement learning approach to home energy management for modulating heat pumps and photovoltaic systems. *Appl. Energy* **2022**, *327*, 120020. [CrossRef]
19. Li, Q.; Zhang, M.; Shen, Y. A hierarchical deep reinforcement learning model with expert prior knowledge for intelligent penetration testing. *Comput. Secur.* **2023**, *132*, 103358. [CrossRef]
20. Li, Y.; Ding, Y.; He, S. Artificial intelligence-based methods for renewable power system operation. *Nat. Rev. Electr. Eng.* **2024**, *1*, 163–179. [CrossRef]
21. Li, Y.; Zhang, X.; Xiao, F. Modeling and management performances of distributed energy resource for demand flexibility in Japanese zero energy house. *Build. Simul.* **2023**, *16*, 2177–2192. [CrossRef]
22. Liang, T.; Chai, L.; Cao, X. Real-time optimization of large-scale hydrogen production systems using off-grid renewable energy: Scheduling strategy based on deep reinforcement learning. *Renew. Energy* **2024**, *224*, 120177. [CrossRef]
23. Liu, X.; Liu, J.; Ren, K. An integrated fuzzy multi-energy transaction evaluation approach for energy internet markets considering judgement credibility and variable rough precision. *Energy* **2022**, *261*, 125327. [CrossRef]
24. Lyu, J.; Wan, L.; Li, X. Off-policy RL algorithms can be sample-efficient for continuous control via sample multiple reuse. *Inf. Sci.* **2024**, *666*, 120371. [CrossRef]
25. Mahmud, K.; Khan, B.; Ravishankar, J. An internet of energy framework with distributed energy resources, prosumers and small-scale virtual power plants: An overview. *Renew. Sustain. Energy Rev.* **2020**, *127*, 109840. [CrossRef]
26. Mnih, V.; Kavukcuoglu, K.; Silver, D. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
27. Pan, C.; Jia, Z.; Wang, J. Optimization of liquid cooling heat dissipation control strategy for electric vehicle power batteries based on linear time-varying model predictive control. *Energy* **2023**, *283*, 129099. [CrossRef]
28. Pang, X.; Wang, Y.; Yu, Y. Optimal scheduling of a cogeneration system via Q-learning-based memetic algorithm considering demand-side response. *Energy* **2024**, *300*, 131513. [CrossRef]
29. Park, K.; Moon, I. Multi-agent deep reinforcement learning approach for EV charging scheduling in a smart grid. *Appl. Energy* **2022**, *328*, 120111. [CrossRef]
30. Patel, I.; Shah, A.; Shen, B. Stochastic optimisation and economic analysis of combined high temperature superconducting magnet and hydrogen energy storage system for smart grid applications. *Appl. Energy* **2023**, *341*, 121070. [CrossRef]
31. Raffin, A.; Hill, A.; Gleave, A. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.
32. Ren, K.; Liu, J.; Wu, Z. A data-driven DRL-based home energy management system optimization framework considering uncertain household parameters. *Appl. Energy* **2024**, *355*, 122258. [CrossRef]
33. Ruan, Y.; Liang, Z.; Qian, F. Operation strategy optimization of combined cooling, heating, and power systems with energy storage and renewable energy based on deep reinforcement learning. *J. Build. Eng.* **2023**, *65*, 105682. [CrossRef]
34. Sharma, S.; Xu, Y.; Verma, A. Time-Coordinated Multienergy Management of Smart Buildings Under Uncertainties. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4788–4798. [CrossRef]
35. Saloux, E.; Runge, J.; Zhang, K. Operation optimization of multi-boiler district heating systems using artificial intelligence-based model predictive control: Field demonstrations. *Energy* **2023**, *285*, 129524. [CrossRef]
36. Sinha, A.; Ghosh, V.; Hussain, N. Green financing of renewable energy generation: Capturing the role of exogenous moderation for ensuring sustainable development. *Energy Econ.* **2023**, *126*, 107021. [CrossRef]
37. Wang, C.; Zhang, J.; Wang, A. Prioritized sum-tree experience replay TD3 DRL-based online energy management of a residential microgrid. *Appl. Energy* **2024**, *368*, 123471. [CrossRef]
38. Wang, M.; Lin, B. MFˆ2: Model-free reinforcement learning for modeling-free building HVAC control with data-driven environment construction in a residential building. *Build. Environ.* **2023**, *244*, 110816. [CrossRef]

39. Wang, Z.; Xiao, F.; Ran, Y. Scalable energy management approach of residential hybrid energy system using multi-agent deep reinforcement learning. *Appl. Energy* **2024**, *367*, 123414. [CrossRef]
40. Wu, J.; He, H.; Peng, J. Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus. *Appl. Energy* **2018**, *222*, 799–811. [CrossRef]
41. Xiao, H.; Fu, L.; Shang, C. Ship energy scheduling with DQN-CE algorithm combining bi-directional LSTM and attention mechanism. *Appl. Energy* **2023**, *347*, 121378. [CrossRef]
42. Zhang, X.; Li, Y.; Xiao, F. Energy efficiency measures towards decarbonizing Japanese residential sector: Techniques, application evidence and future perspectives. *Energy Build.* **2024**, *319*, 114514. [CrossRef]
43. Zhang, X.; Xiao, F.; Li, Y. Flexible coupling and grid-responsive scheduling assessments of distributed energy resources within existing zero energy houses. *J. Build. Eng.* **2024**, *87*, 109047. [CrossRef]
44. Zhang, X.; Xiao, F.; Li, Y. Energy flexibility and resilience analysis of demand-side energy efficiency measures within existing residential houses during cold wave event. *Build. Simul.* **2024**, *17*, 1043–1063. [CrossRef]
45. Zhang, Y.; Zhang, C.; Fan, R. Energy management strategy for fuel cell vehicles via soft actor-critic-based deep reinforcement learning considering powertrain thermal and durability characteristics. *Energy Convers. Manag.* **2023**, *283*, 116921. [CrossRef]
46. Zhang, Z.; Yang, Z.; Yau, D.K.; Tian, Y.; Ma, J. Data security of machine learning applied in low-carbon smart grid: A formal model for the physics-constrained robustness. *Appl. Energy* **2023**, *347*, 121405. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.