Deep Reinforcement Learning for Real-Time Energy Management in Smart Home

Guixi Wei, Ming Chi, Zhi-Wei Liu, *Senior Member, IEEE*, Mingfeng Ge, *Member, IEEE*, Chaojie Li, and Xianggang Liu

***Abstract*—Energy management in the smart home can help reduce residential energy costs by scheduling various energy consumption activities. However, accurately modeling factors, such as user behavior, renewable power generation, weather conditions, and real-time electricity prices can be challenging, making the design of an efficient energy management strategy difficult. This article proposes a real-time energy management algorithm based on deep reinforcement learning (DRL) for smart homes equipped with rooftop photovoltaics, energy storage systems, and smart appliances. The algorithm aims to minimize the energy cost while ensuring user comfort. A policy network that can output both discrete and continuous actions is designed to generate actions for different types of devices in a smart home. The proposed DRL-agent is trained using a proximal policy optimization approach with historical data and is used for real-time scheduling. Finally, simulations based on real-world data demonstrate the effectiveness and robustness of the proposed algorithm.**

***Index Terms*—Deep reinforcement learning (DRL), energy storage system (ESS), home energy management, proximal policy optimization (PPO).**

NOMENCLATURE

Indices

t	Index of time slot.
s_i	Index of shiftable appliance.
c_i	Index of controllable appliance.
n_i	Index of nonshiftable appliance.

Variables

$P_t^{s_i}$	Power consumption of s_i at time slot t (kW).
$I_t^{s_i}$	Control variable of s_i at time slot t.
$\rho_t^{s_i}$	Task progress of appliance s_i at time slot t.
$t_{\text{ini}}^{s_i}$	Initial time of appliance s_i's scheduling window.
$t_{\text{end}}^{s_i}$	End time of appliance s_i's scheduling window.
$P_t^{c_i}$	Power consumption of c_i at time slot t (kW).
T_t^{in}	Indoor air temperature at time slot t (°C).
T_t^{out}	Outdoor air temperature at time slot t (°C).
P_t^{HVAC}	Power consumption of HVAC at time slot t (kW).
T_t^{wa}	Water temperature in the EWH at time slot t (°C).
P_t^{EWH}	Power consumption of the EWH at time slot t (kW).
W_t	Hot water flow during time slot t (L).
SoC_t^{EV}	State of Charge of the EV at time slot t.
P_t^{EV}	Charging power of the EV at time slot t (kW).
$t_{\text{ini}}^{\text{EV}}$	Time EV arrives home.
$t_{\text{end}}^{\text{EV}}$	Time EV leaves home.
$\text{SoC}_t^{\text{ESS}}$	State of charge of the ESS at time slot t.
P_t^{ESS}	Charging power of the ESS at time slot t (kW).
C_t^{SoC}	SoC-related degradation cost of the ESS at time slot t (\$).
ΔL_t^{DoD}	DoD of a particular discharging process at time slot t.
C_t^{DoD}	DoD-related degradation cost of the ESS at time slot t (\$).
C_t^{ESS}	Degradation cost of ESS at time slot t.
$P_t^{n_i}$	Power consumption of appliance n_i at time slot t.
$t_{\text{ini}}^{n_i}$	Task starting time of appliance n_i.
$t_{\text{end}}^{n_i}$	Task deadline of appliance n_i.
P_t^{PV}	Output power of PV at time slot t.
i_t	Solar irradiation at time slot t (kW/m^2).
P_t^g	Exchange power between smart home and utility grid (kW).
s_t	State of the MDP at time step t.
a_t	Action of the MDP at time step t.
r_t	Reward of the MDP at time step t.
θ	Weights of the policy network.
ω	Weights of the value network.

Constants

$P_{\text{rated}}^{s_i}$	Rated power of appliance s_i (kW).
K^{s_i}	Required time for appliance s_i to complete its task.
$P_{\text{min}}^{c_i}$	Minimum power of appliance c_i (kW).
$P_{\text{max}}^{c_i}$	Maximum power of appliance c_i (kW).
η^{HVAC}	Thermal conversion efficiency of the HVAC.
A	Mean thermal conductivity of house (kW/°C).
ε	Inertia factor of the HVAC.

Manuscript received 10 February 2022; revised 29 May 2022 and 20 August 2022; accepted 17 February 2023. Date of publication 10 March 2023; date of current version 8 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62222205, Grant 61973133, Grant 61972170, and Grant 62073301, and in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFA052 and Grant 2021CFB343, and in part by the Australian Research Council under Grant DE210100274. *(Corresponding author: Ming Chi.)*

Guixi Wei, Ming Chi, Zhi-Wei Liu, and Xianggang Liu are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: gx_wei@hust.edu.cn; chiming@hust.edu.cn; zwliu@hust.edu.cn; 13230909571@163.com).

Mingfeng Ge is with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China (e-mail: fmgabc@163.com).

Chaojie Li is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: cjlee.cqu@163.com).

Digital Object Identifier 10.1109/JSYST.2023.3247592

$P_{\max}^{\text{HVAC}}$	Maximum power of the HVAC.
$T_{\text{set}}^{\text{in}}$	Set value of the indoor temperature (°C).
$\Delta T_{\text{set}}^{\text{in}}$	Threshold of the indoor temperature deviation (°C).
α_1	Thermal comfort conversion factor of HVAC ($).
V	Water tank capacity of the EWH (L).
ρ	Density of water (kg/m^3).
c_p	Specific heat of water (kJ/(°C · kg)).
G	Thermal conductivity of the water tank (W/°C).
T_{co}	Temperature of cold water (°C).
$P_{\max}^{\text{EWH}}$	Maximum power of the EWH (kW).
$T_{\text{set}}^{\text{wa}}$	Set value of the water in the EWH (°C).
$\Delta T_{\text{set}}^{\text{wa}}$	Threshold of the water temperature deviation (°C).
α_2	Thermal comfort conversion factor of EWH ($).
$\eta_{\text{ch}}^{\text{EV}}$	Charging efficiency of the EV.
$E_{\max}^{\text{EV}}$	Battery capacity of the EV (kWh).
$P_{\max}^{\text{EV}}$	Maximum charging power of the EV (kW).
$\text{SoC}_{\max}^{\text{EV}}$	Maximum SoC of the EV.
α_3	Range anxiety conversion factor ($/kWh^2).
$\text{SoC}_{\max}^{\text{EV}}$	Maximum SoC of the EV.
$\eta_{\text{ch}}^{\text{ESS}}$	Charging efficiency of the ESS.
$\eta_{\text{dis}}^{\text{ESS}}$	Discharging efficiency of the ESS.
$E_{\max}^{\text{ESS}}$	Battery capacity of the ESS (kWh).
$P_{\max}^{\text{ESS}}$	Maximum charging power of the ESS (kW).
$\text{SoC}_{\max}^{\text{ESS}}$	Maximum SoC of the ESS.
$\text{SoC}_{\min}^{\text{ESS}}$	Minimum SoC of the ESS.
C_0	Installation cost of battery in ESS ($).
$P_{\text{rated}}^{n_i}$	Rated power of appliance n_i.
ρ_{pv}	Conversion efficiency of the PV panel.
A_{pv}	Surface area of roof PV (m^2).
$P_{\max}^{g}$	Maximum total power demand limitation (kW).

I. Introduction

SMART homes are usually equipped with an advanced automation system called a home energy management system (HEMS), which can interact with the utility grid, monitor, and schedule the smart appliances in the home. By applying HEMS, smart homes can participate in demand response programs, which increase the flexibility of the power grid. Moreover, HEMS is enabled to reduce the household energy cost while maintaining the satisfaction of the user's comfort.

In early work, Paterakis et al. [1] developed a mixed-integer linear optimization method to reduce the household energy cost by scheduling smart home appliances a day ahead, but the comfort level of residents was not considered. Therefore, Althaher et al. [2] proposed a mixed-integer nonlinear programming approach considering the comfort level of residents. However, the randomness of energy consumption and electricity price was not considered in [1] and [2]. Then, a stochastic model was adopted in [3] and [4] to describe the uncertainties of electric vehicles (EV) availability and an HEMS was proposed to solve the energy management problem. In [5], a real-time energy management method considering the uncertainty of photovoltaic (PV) output and electricity price was proposed to minimize the energy cost while considering the user's comfort. In [6], the model predictive control (MPC) strategy was adopted to minimize the energy cost and reduce the EV battery degradation. It is worth noting that the methods proposed in [1], [2], [3], [4], [5], and [6] require a great deal of knowledge of systems and environments, which are often difficult to acquire and, therefore, may be difficult to apply in some cases.

On the other hand, machine learning is an effective method that can model complex systems based on data without a lot of prior knowledge. It has attracted many scholars to simplify the modeling of HEMS by applying machine learning. Zhang et al. [7] adopted an artificial neural network (ANN) to learn the dynamics of heating, ventilating, and air conditioning (HVAC) and proposed an optimal schedule algorithm for appliances in smart home. But, it is also an offline optimization method, which is hard to make a real-time decision. The real-time energy management algorithm for the PV-storage system based on approximate dynamic programming (ADP) was proposed in [8], in which, an ANN-based policy function was trained with historical data. However, the state transition probability of the environment is still essential in the ADP methods.

Without the state transition probability, model-free reinforcement learning (RL) can train an agent to complete a certain task by interacting with the environment. In each interaction, the agent observes the state of the environment and generates an action, which can achieve real-time scheduling for the appliances in the smart home. Q-learning is a classical model-free RL algorithm, which was adopted in different cases [9], [10], [11], [12], [13] to reduce the energy cost of the smart home. But, the Q-learning algorithm can only deal with the problem that both the state and action are discrete and low-dimensional. Therefore, it may be hard to deal with the complex environment in smart home.

The deep RL (DRL) method has been proposed to solve the problem that both state and action are continuous and high-dimensional by integrating the perception of deep learning and the decision-making ability of RL [14], [15]. For instance, Mathew et al. [16] proposed a multiobjective residential energy management algorithm based on Deep Q Network (DQN) [17] to minimize the load profile deviation and energy cost. In [18], an online optimal scheduling method based on DQN and deterministic policy gradient (DPG) was proposed to reduce the residential energy cost, and it was found that DPG is more effective than DQN. Furthermore, a real-time scheduling strategy in a multienergy smart home based on deep DPG (DDPG) was proposed in [19] but the action was discrete, and the continuous version was proposed in [20]. The works in [16], [18], [19], and [20] can only solve the problem that with either discrete or continuous actions. However, some appliances in a smart home require discrete control actions while the others require continuous control actions, which need additional studies.

Motivated by the above discussion, this article proposes a real-time energy management approach based on proximal policy optimization (PPO) [21] algorithm to minimize the energy cost while maintaining satisfaction of the user's comfort in a smart home. A policy network, which can generate both discrete and continuous action, is adopted to schedule different types of appliances. The proposed algorithm solves the energy management problem in a smart home equipped with PV, energy storage

system (ESS), and different types of smart appliances while considering the uncertainty of user's behavior and electricity price. Compared with [22], the scheduling of ESS and PV is further investigated. The charging/discharging of ESS enhances the coupling of actions at each time slot and the uncertainty of PV output results in a new disturbance, which increased the difficulty of residential energy management. Besides, a surrogate objective and entropy of the policy are considered, which speeds up the training process and improves the exploration ability of the method. Finally, simulations based on real-world data (e.g., solar irradiation, outdoor temperature, electricity price, etc.) are adopted to estimate the effectiveness of the proposed algorithm.

The main contributions of this article are as follows.

1) A model of the smart home incorporating PV, ESS, and another three kinds of smart appliances is established, of which the control variables included both discrete and continuous types, in contrast to [16], [18], [19], and [20], which considered either discrete or continuous only. Also, the degradation cost of ESS is considered.
2) A Markov decision process (MDP) form of the energy management problem in the smart home is formulated, in which the user's comfort is considered even if the future information associated with the outdoor temperature, irradiation, real-time electricity prices (RTP), and users' behaviors is unknown.
3) A real-time energy management approach based on the PPO [21] is proposed, which can be trained with historical data and executed for the real-time scheduling. To achieve the control of different types of appliances, the proposed method incorporates a mixed output policy (MOP), which can output both discrete and continuous actions compared to the work in [19] and [20].
4) Simulations based on the real-world data are carried out. The results show that the proposed method performs better in reducing the energy cost of the smart home and is more robust to handle the outdoor temperature deviation than RL-based method DDPG [20] and trust region policy optimization (TRPO) [22].

The remainder of this article is organized as follows. Section II gives the smart home model for simulations. Section III describes the MDP form of the HEMS problem. Section IV explains the approach of HEMS based on the PPO algorithm. Section V shows simulation results, and the conclusion is given in Section VI.

II. Model of Smart Home

A smart home considered in this article is shown in Fig. 1, including rooftop PV, ESS, and other appliances. There is a bidirectional connection between the smart home and the utility grid, which can exchange power in both directions. RTP with inclining block rates (RTP-IBR) scheme [23] is adopted in the utility grid. Appliances in a smart home are mainly divided into shiftable, nonshiftable, and controllable appliances. HEMS schedules in a set of time slots $T = \{1, 2, 3, \ldots, T_{\text{end}}\}$ and the interval of each time slot is ΔT. The rest of the models associated with the smart home are given as follows.

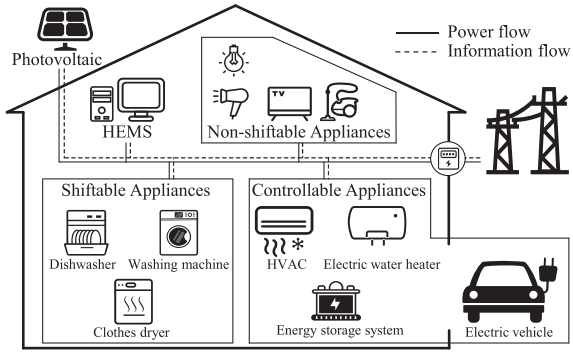

Fig. 1. Architecture of a smart home.

A. Shiftable Appliances

Let $S = \{s_1, s_2, \ldots, s_j\}$ denote the set of the shiftable appliances in the smart home. For the appliance $s_i \in S$, $P_t^{s_i}$ denotes the power consumption in the time slot t described as

$$P_t^{s_i} = I_t^{s_i} \cdot P_{\text{rated}}^{s_i} \quad \forall t \in T \tag{1}$$

where $P_{\text{rated}}^{s_i}$ and $I_t^{s_i} \in \{0, 1\}$ are rated power and control variable of s_i, respectively; when $I_t^{s_i} = 1$, s_i is consuming power at time slot t, otherwise s_i is dormant.

Assume that the shiftable appliances can be operated by the HEMS within a scheduling window $[t_{\text{ini}}^{s_i}, t_{\text{end}}^{s_i}]$ and cannot be interrupted once they start working [24], $I_t^{s_i}$ satisfies

$$\begin{cases} I_t^{s_i} = 1, & I_{t-1}^{s_i} = 1, \rho_t^{s_i} < K^{s_i} \quad \forall t \in T & \text{(2a)} \\ 1, & \rho_t^{s_i} = 0, t_{\text{end}}^{s_i} - t = K^{s_i} & \text{(2b)} \\ 0, & t \notin [t_{\text{ini}}^{s_i}, t_{\text{end}}^{s_i}] & \text{(2c)} \end{cases}$$

where

$$\rho_t^{s_i} = \sum_{i=t_{\text{ini}}^{s_i}}^{t-1} I_i^{s_i}, t \in [t_{\text{ini}}^{s_i}, t_{\text{end}}^{s_i}] \tag{3}$$

denotes task progress of s_i; $t_{\text{ini}}^{s_i}$ and $t_{\text{end}}^{s_i}$ are the initial time and the end time of the scheduling window; $K^{s_i} \leq t_{\text{end}}^{s_i} - t_{\text{ini}}^{s_i}$ denotes the time required by s_i to complete the task. Equation (2a) forces s_i to work continually. Equation (2b) forces s_i to finish its task within the scheduling window. Equation (2c) ensures s_i is turned OFF outside the scheduling window.

B. Controllable Appliances

Let $C = \{c_1, c_2, \ldots, c_j\}$ denotes the set of controllable appliances in smart home. For the controllable appliance $c_i \in C$, the power $P_t^{c_i}$ can be adjusted continuously, satisfying

$$P_{\min}^{c_i} \leq P_t^{c_i} \leq P_{\max}^{c_i} \quad \forall t \in T \tag{4}$$

where $P_{\min}^{c_i}$ and $P_{\max}^{c_i}$ are the minimum and maximum values of an adjustable range of power, respectively. We mainly focus on HVAC, electrical water heater (EWH), EV, and ESS.

1) HVAC: In the heating mode of HVAC, the thermodynamic model [22] of the indoor temperature is

$$T_{t+1}^{\text{in}} = \varepsilon T_t^{\text{in}} + (1-\varepsilon)\left(T_t^{\text{out}} + \eta^{\text{HVAC}} P_t^{\text{HVAC}} \Delta T / A\right) \quad (5a)$$

$$0 \le P_t^{\text{HVAC}} \le P_{\max}^{\text{HVAC}} \quad (5b)$$

where T_t^{in} and T_t^{out} are the indoor and outdoor temperatures, respectively; η^{HVAC} is the thermal conversion efficiency; A is the mean thermal conductivity of house; ε is a factor of inertia; P_t^{HVAC} is the power of HVAC; $P_{\max}^{\text{HVAC}}$ is the maximum power of HVAC.

To maintain T_t^{in} within a comfortable range, consider the following form of the thermal comfort cost:

$$C_t^{\text{HVAC}} = \alpha_1 \cdot \left(|T_t^{\text{in}} - T_{\text{set}}^{\text{in}}| / \Delta T_{\text{set}}^{\text{in}}\right)^2 \quad (6)$$

where $\alpha_1 > 0$ is the conversion coefficient of indoor temperature comfort, in unit \$; $T_{\text{set}}^{\text{in}}$ and $\Delta T_{\text{set}}^{\text{in}}$ are the preset value and maximum allowable deviation of the indoor temperature set by the users, respectively.

2) EWH: According to the work in [25], the thermodynamic of the water temperature in a tank of EWH is

$$T_{t+1}^{\text{wa}} = T_t^{\text{wa}} \cdot e^{-\Delta T/\tau} + Q \cdot \left(1 - e^{-\Delta T/\tau}\right) \quad (7a)$$

$$\tau = \rho c_p V / (G + \rho c_p W_t) \quad (7b)$$

$$Q = \left(G T_t^{\text{in}} + \rho c_p W_t T_{\text{co}} + P_t^{\text{EWH}}\right) / (G + \rho c_p W_t) \quad (7c)$$

$$0 \le P_t^{\text{EWH}} \le P_{\max}^{\text{EWH}} \quad (7d)$$

where T_t^{wa} is the water temperature in a tank; V is the capacity of the water tank and the unit is L; ρ and c_p are density and specific heat capacity of water, respectively; G is the thermal conductivity of the water tank; W_t is the hot water flow; T_{co} is the temperature of the replenishing cold water; P_t^{EWH} and $P_{\max}^{\text{EWH}}$ are heating power and maximum heating power of EWH, respectively.

Similar to HVAC, the thermal comfort cost is designed as

$$C_t^{\text{EWH}} = \alpha_2 \cdot \left(|T_t^{\text{wa}} - T_{\text{set}}^{\text{wa}}| / \Delta T_{\text{set}}^{\text{wa}}\right)^2 \quad (8)$$

where $\alpha_2 > 0$ is a conversion coefficient of water temperature comfort, in unit \$; $T_{\text{set}}^{\text{wa}}$ and $\Delta T_{\text{set}}^{\text{wa}}$ are the preset temperature and maximum allowable deviation of the water temperature set by the users.

3) EV: EV arrives home at time $t_{\text{ini}}^{\text{EV}}$ and leaves home at time $t_{\text{end}}^{\text{EV}}$ every day. EV can charge during it is at home and should be fully charged when leaving. The charging model of EV is expressed as

$$\text{SoC}_{t+1}^{\text{EV}} = \text{SoC}_t^{\text{EV}} + \eta_{\text{ch}}^{\text{EV}} \cdot P_t^{\text{EV}} \Delta T / E_{\max}^{\text{EV}} \quad (9a)$$

$$0 \le P_t^{\text{EV}} \le \min\left(P_{\max}^{\text{EV}}, \frac{\left(\text{SoC}_{\max}^{\text{EV}} - \text{SoC}_t\right) \cdot E_{\max}^{\text{EV}}}{\Delta T \cdot \eta_{\text{ch}}^{\text{EV}}}\right) \quad (9b)$$

where SoC_t^{EV} is the state of charge (SoC) of EV; $\eta_{\text{ch}}^{\text{EV}}$ denotes the charging efficiency; $E_{\max}^{\text{EV}}$ is a battery capacity of EV; P_t^{EV} and $P_{\max}^{\text{EV}}$ are charging power and maximum charging power of EV, respectively; $\text{SoC}_{\max}^{\text{EV}}$ is maximum SoC of EV. Overcharge of EV is avoided by (9b).

The user's range anxiety is denoted by C_t^{EV} and occurs when EV is not well charged at the departure time

$$C_t^{\text{EV}} = \begin{cases} \alpha_3 \left(\left(\text{SoC}_t^{\text{EV}} - \text{SoC}_{\max}^{\text{EV}}\right) E_{\max}^{\text{EV}}\right)^2, & t = t_{\text{end}}^{\text{EV}} \\ 0, & t \ne t_{\text{end}}^{\text{EV}} \end{cases} \quad (10)$$

where α_3 is a range anxiety conversion factor in unit \$/kWh^2.

4) ESS: ESS is controlled by HEMS to charge or discharge. Let $P_t^{\text{ch}} > 0$ and $P_t^{\text{dis}} < 0$ denote the charging and discharging powers of ESS, respectively. The SoC model of ESS can be expressed as

$$\text{SoC}_{t+1}^{\text{ESS}} = \text{SoC}_t^{\text{ESS}} + \eta_{\text{ch}}^{\text{ESS}} \cdot P_t^{\text{ch}} \Delta T / E_{\max}^{\text{ESS}} + 1/\eta_{\text{dis}}^{\text{ESS}} \cdot P_t^{\text{dis}} \Delta T / E_{\max}^{\text{ESS}} \quad (11)$$

where $\text{SoC}_t^{\text{ESS}}$ is the SoC of ESS; $E_{\max}^{\text{ESS}}$ is the battery capacity of ESS; $\eta_{\text{ch}}^{\text{ESS}}$ and $\eta_{\text{dis}}^{\text{ESS}}$ denote the charging and discharging efficiencies of ESS, respectively. Similar to EV, the charging and discharging powers should satisfy

$$0 \le P_t^{\text{ch}} \le \min\left(P_{\max}^{\text{ESS}}, \frac{\left(\text{SoC}_{\max}^{\text{ESS}} - \text{SoC}_t^{\text{ESS}}\right) E_{\max}^{\text{ESS}}}{\Delta T \cdot \eta_{\text{ch}}^{\text{ESS}}}\right) \quad (12a)$$

$$\max\left(-P_{\max}^{\text{ESS}}, \frac{\left(\text{SoC}_{\min}^{\text{ESS}} - \text{SoC}_t^{\text{ESS}}\right) E_{\max}^{\text{ESS}} \eta_{\text{dis}}^{\text{ESS}}}{\Delta T}\right) \le P_t^{\text{dis}} \le 0 \quad (12b)$$

where $\text{SoC}_{\max}^{\text{ESS}}$ and $\text{SoC}_{\min}^{\text{ESS}}$ are maximum and minimum SoCs of ESS, respectively; $P_{\max}^{\text{ESS}}$ is the maximum power of ESS.

Let P_t^{ESS} denotes control variable of ESS. Then, charging and discharging powers can be expressed as

$$\left[P_t^{\text{ch}}, P_t^{\text{dis}}\right] = \begin{cases} \left[P_t^{\text{ESS}}, 0\right], P_{\max}^{\text{ESS}} \ge P_t^{\text{ESS}} \ge 0 \\ \left[0, P_t^{\text{ESS}}\right], -P_{\max}^{\text{ESS}} \le P_t^{\text{ESS}} < 0. \end{cases} \quad (13)$$

Note that the ESS is already installed in the smart home and the economy of its investment is already guaranteed while improper short-term scheduling operation may cause the rapid degradation of the battery units in the ESS, resulting in the battery units needing to be replaced before they reach the design service life. To avoid that, the SoC-related degradation cost and the depth-of-discharge (DoD)-related degradation cost for one discharging cycle are considered. According to the work in [26], the SoC-related degradation cost during a time slot is

$$C_t^{\text{SoC}} = C_0 \cdot \frac{\kappa \cdot \text{SoC}_t^{\text{ESS}} - \psi}{F_{\max} \cdot 15 \cdot 365 \cdot 24} \cdot \frac{\Delta T}{60} \quad (14)$$

where C_0 is the cost of the battery units, in unit \$, which is predicted to dropping below \$100/kWh by 2030 [27], [28]; κ and ψ are linear regression factors in battery test data; $F_{\max}$ is the maximum capacity fade constant, usually is 20%. The DoD-related degradation cost for one discharging cycle is

$$C_t^{\text{DoD}} = C_0 \cdot \frac{\Delta L_t^{\text{DoD}}}{f\left(\Delta L_t^{\text{DoD}}\right)}$$

$$\Delta L_t^{\text{DoD}} = -P_t^{\text{ESS}} \cdot \Delta T / \left(\eta_{\text{dis}}^{\text{ESS}} \cdot E_{\max}^{\text{ESS}}\right), P_t^{\text{ESS}} < 0 \quad (15)$$

where ΔL_t^{DoD} is the difference of DoD before and after a particular discharging process. $f\left(\Delta L_t^{\text{DoD}}\right)$ is the fitting function obtained from the battery discharge experiment [26]:

$$f\left(\Delta L_t^{\text{DoD}}\right) = \begin{pmatrix} 1.06 \cdot \left(\Delta L_t^{\text{DoD}}\right)^4 - 2.80 \cdot \left(\Delta L_t^{\text{DoD}}\right)^3 + 2.66 \\ \cdot \left(\Delta L_t^{\text{DoD}}\right)^2 - 1.07 \cdot \left(\Delta L_t^{\text{DoD}}\right) + 0.17 \end{pmatrix} \cdot 10^5. \tag{16}$$

Therefore, the degradation cost of ESS can be expressed as

$$C_t^{\text{ESS}} = C_t^{\text{SoC}} + C_t^{\text{DoD}}. \tag{17}$$

C. Nonshiftable Appliances

Nonshiftable appliances cannot be scheduled by HEMS. Let $N = \{n_1, n_2, \ldots, n_j\}$ denotes a set of nonshiftable appliances in the smart home. For $n_i \in N$, it starts working at $t_{\text{ini}}^{n_i}$ and finishes its task at $t_{\text{end}}^{n_i}$, which is associated with users' behaviors. The power of n_i is

$$P_t^{n_i} = \begin{cases} P_{\text{rated}}^{n_i}, & t \in [t_{\text{ini}}^{n_i}, t_{\text{end}}^{n_i}] \\ 0, & t \notin [t_{\text{ini}}^{n_i}, t_{\text{end}}^{n_i}] \end{cases} \tag{18}$$

where $P_{\text{rated}}^{n_i}$ and $P_t^{n_i}$ denote the working power and the rated power, respectively, of n_i.

D. Photovoltaic

Let P_t^{PV} denotes the output power of PV. P_t^{PV} is correlated with the environmental condition (e.g., solar irradiation), which is uncontrollable. According to the work in [29], the estimation of P_t^{PV} is

$$P_t^{\text{PV}} = \rho_{\text{pv}} \cdot A_{\text{pv}} \cdot i_t \tag{19}$$

where ρ_{pv} is the conversion efficiency of PV panel per unit area; A_{pv} is the surface area of roof PV (in m^2); i_t is the solar irradiation (in kW/m^2).

E. Power Balance Constraint

In each time slot, the power balance between smart home and utility grid can be expressed as

$$P_t^g + P_t^{\text{PV}} = \sum_{s_i \in S} P_t^{s_i} + \sum_{c_i \in C} P_t^{c_i} + \sum_{n_i \in N} P_t^{n_i} \tag{20}$$

where P_t^g denotes the exchange power between smart home and utility grid; $P_t^g > 0$ indicates that HEMS purchases electricity from the utility grid, on the contrary, sells electricity to utility grid. The price of electricity sold by HEMS to utility grid is 0.9 times of the RTP $\mathcal{P}_t$. The RTP-IBR price in time slot t is given

$$\text{price}_t = \begin{cases} \mathcal{P}_t, & 0 \le P_t^g \le P_{\max}^g \\ \xi \cdot \mathcal{P}_t, & P_{\max}^g < P_t^g \\ 0.9 \cdot \mathcal{P}_t, & P_t^g < 0 \end{cases} \tag{21}$$

where ξ is the IBR factor [23]; $P_{\max}^g$ is the maximum power limitation. Therefore, the energy cost of smart home in time slot $t \in T$ is

$$C_t^{\text{E}} = P_t^g \cdot \text{price}_t \cdot \Delta T. \tag{22}$$

F. Objective Function of Energy Management

HEMS aims at minimizing the energy cost while ensuring user comfort within a time period via regulating $\{I_t^{s_i}, P_t^{c_i} | s_i \in S, c_i \in C\}$. Specifically, the objective function can be formulated as

$$\min \sum_{t=0}^{T_{\text{end}}} \left(C_t^{\text{E}} + C_t^{\text{HVAC}} + C_t^{\text{EWH}} + C_t^{\text{EV}} + C_t^{\text{ESS}}\right). \tag{23}$$

III. MDP Formulation

This section converts the energy management problem to an MDP(S, A, P, r), where, $S, A, P,$ and r are the state space, action space, state transition probability, and reward function of the environment, respectively. In time slot t, an agent observes a state s_t from S_t and selects an action a_t from A. After executing a_t in the environment, the state s_t is transformed to s_{t+1} based on $P(s_t, a_t)$, and the agent receives a reward $r_t = r(s_t, a_t)$. In this article, the state transition probability P remains unknown because of the uncertainties of the electricity prices, outdoor temperature, irradiations, and user's behaviors in the future. Therefore, we mainly focus on the formulations of the state s_t, action a_t, and reward function r_t.

A. State

The state of the environment, which is measurable, should reflect the characteristic of the environment at time slot t. Therefore, the state of the smart home is defined as follows.

1) Shiftable Appliances: Defining the state of the shiftable appliance $s_i \in S$ at time slot t as

$$s_t^{s_i} = \begin{cases} [\rho_t^{s_i}, t_{\text{ini}}^{s_i} - t], & t \in [t_{\text{ini}}^{s_i}, t_{\text{end}}^{s_i}] \\ [0, 0], & t \notin [t_{\text{ini}}^{s_i}, t_{\text{end}}^{s_i}] \end{cases}$$

then, $s_t^S = \{s_t^{s_1}, \ldots, s_t^{s_j}\}$ denotes the state of all shiftable appliances in smart home.

2) Controllable Appliances: The states of each controllable appliance are defined as follows.

a) HVAC: $s_t^{\text{HVAC}} = T_t^{\text{in}} - T_{\text{set}}^{\text{in}} \quad \forall t \in T.$

b) EWH: $s_t^{\text{EWH}} = T_t^{\text{wa}} - T_{\text{set}}^{\text{wa}} \quad \forall t \in T.$

c) EV: $s_t^{\text{EV}} = \begin{cases} \text{SoC}_t^{\text{EV}}, & t \in [t_{\text{ini}}^{\text{EV}}, t_{\text{end}}^{\text{EV}}] \\ 0, & t \notin [t_{\text{ini}}^{\text{EV}}, t_{\text{end}}^{\text{EV}}]. \end{cases}$

d) ESS: $s_t^{\text{ESS}} = \text{SoC}_t^{\text{ESS}} \quad \forall t \in T.$

Therefore, the state of all controllable appliances in smart home is $s_t^C = \{s_t^{\text{HVAC}}, s_t^{\text{EWH}}, s_t^{\text{EV}}, s_t^{\text{ESS}}\}$.

3) Nonshiftable Appliances: Define the state of nonshiftable appliance n_i as:

$$s_t^{n_i} = \begin{cases} t - t_{\text{ini}}^{n_i}, & t \in [t_{\text{ini}}^{n_i}, t_{\text{end}}^{n_i}] \\ 0, & t \notin [t_{\text{ini}}^{n_i}, t_{\text{end}}^{n_i}] \end{cases}$$

then, the state of all nonshiftable appliances is described as $s_t^N = \{s_t^{n_1}, \ldots, s_t^{n_j}\}$.

Combining with other information (e.g., outdoor temperature, RTP etc.), the state in time slot $t \in T$ can be expressed as a vector, namely

$$s_t = \left\{s_t^S, s_t^C, s_t^N, P_t^{PV}, \mathcal{P}_{t-T_{\text{end}}+1}, \ldots, \mathcal{P}_t,\right.$$

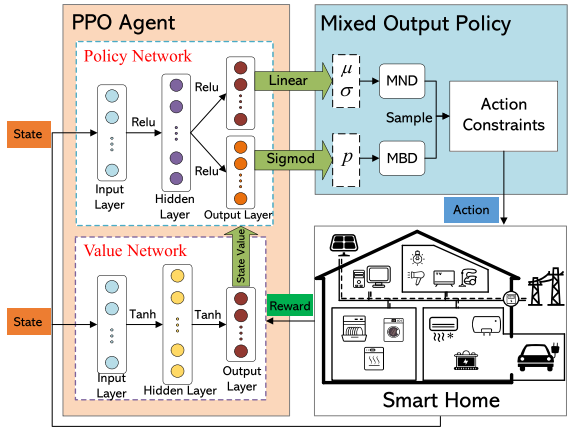

Fig. 2. Architecture of the PPO-based energy management algorithm.

$$i_{t-T_{\text{end}}+1},\ldots,i_t,T^{\text{out}}_{t-T_{\text{end}}+1},\ldots,T^{\text{out}}_t\} \tag{24}$$

where $\mathcal{P}_{t-T_{\text{end}}+1},\ldots,\mathcal{P}_t$, $i_{t-T_{\text{end}}+1},\ldots,i_t$, $T^{\text{out}}_{t-T_{\text{end}}+1},\ldots,T^{\text{out}}_t$ are electricity prices, solar irradiation, and outdoor temperature in past T_{end} time steps, respectively.

B. Action

The action a_t generated by agent after observing s_t at the beginning of time slot t. It can be expressed as

$$a_t=\left\{I^{s_1}_t,\ldots,I^{s_j}_t,P^{c_1}_t,\ldots,P^{c_j}_t\right\} \tag{25}$$

where $I^{s_1}_t,\ldots,I^{s_j}_t$ are the control variables of shiftable appliances, $I^{s_j}_t,P^{c_1}_t,\ldots,P^{c_j}_t$ are the power of the controllable appliances.

C. Reward

The reward function should reflect the objective function (23). Therefore, the reward is modeled as

$$r_t=-C^{\text{E}}_t-C^{\text{HVAC}}_t-C^{\text{EWH}}_t-C^{\text{EV}}_t-C^{\text{ESS}}_t. \tag{26}$$

The agent learns a policy by maximizing the discount cumulative reward calculated by (26), which is equivalent to the objective in Section II-F.

IV. Energy Management Algorithm

A. Algorithm Design

Based on PPO [21], an energy management algorithm is proposed to solve the MDP problem formulated in Section III. PPO is a model-free DRL approach based on the actor–critic framework (AC), which can deal with the continuous actions and states. Policy network (actor) π_θ with weights θ and value network (critic) v_ω with weights ω are adopted in the PPO agent, which are for the approximations to policy function and state value function in DRL, respectively, shown in Fig. 2. The input of π_θ and v_ω is the state of the environment. A MOP is considered to generate actions for different types of appliances. Besides, the outputs of π_θ serve as inputs to the MOP. The output of v_ω is state value, which is for updating π_θ.

1) PPO Agent: According the policy gradient method [30] and gradient boosting algorithm, the weights θ of π_θ is updated by

$$\theta_{\text{new}}=\theta_{\text{old}}+\text{lr}\cdot\nabla_{\theta_{\text{old}}}J(\theta_{\text{old}}) \tag{27}$$

where lr is the learning rate; $J(\theta)$ is the objective function of π_θ; $\nabla_\theta J(\theta)$ is policy gradient. Equation (27) shows that the policy π_θ is updated by maximizing the objective $J(\theta)$. PPO applied a surrogate objective [21]

$$\begin{aligned}J(\theta)&=\mathbb{E}_t\left(L^{\text{CLIP}}(\theta)\right)\\ L^{\text{CLIP}}(\theta)&=\min\left(k_t(\theta)\hat{A}_t,\text{clip}(k_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t\right)\\ k_t(\theta)&=\pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)\end{aligned} \tag{28}$$

where $\text{clip}(\ldots)$ is the truncation function that limits the deviation of the old policy and new policy, in which, the lower bound is $1-\epsilon$, and the upper bound is $1+\epsilon$; $\hat{A}_t$ is the advantage function, which can reflect the effeteness of the policy. The generalized advantage estimation method [31] is adopted to calculate $\hat{A}_t$

$$\begin{aligned}\hat{A}_t&=\sum_{l=0}^{T_{\text{end}}-t}(\gamma\lambda)^l\delta^V_{t+l}\\ \delta^V_t&=r_t+\gamma V^\pi(s_{t+1})-V^\pi(s_t)\\ V^\pi(s_t)&=\mathbb{E}_\pi\left[\sum_{l=0}^{T_{\text{end}}-t}\gamma^l r_{t+l}\right]\end{aligned} \tag{29}$$

where $\lambda\in[0,1]$ is the estimating factor; δ^V_t is the temporal difference error; $V^\pi(s_t)$ is the state value function, which is approximated by the value network v_ω; γ is the discount factor; r_t is the reward function. Both the policy network and the value network are modeled by multilayer perceptrons (MLP). Therefore, the loss function of v_ω can be defined as

$$L^v(\omega)=\mathbb{E}\left[v_\omega(s_t)-\sum_{l=0}^{T_{\text{end}}-t}\gamma^l r_{t+l}\right]^2. \tag{30}$$

To maximize the surrogate objective function $J(\theta)$, the loss function of the π_θ in the PPO method is defined as

$$\begin{aligned}L^\pi(\theta)&=\mathbb{E}\left[-L^{\text{CLIP}}(\theta)-c_1\cdot H^{\pi_\theta}(s_t)\right]\\ H^{\pi_\theta}(s_t)&=\mathbb{E}_{a_t=\pi_\theta}\left[\pi_\theta(a_t|s_t)\log\pi_\theta(a_t|s_t)\right]\end{aligned} \tag{31}$$

where $c_1\in[0,1]$ is the hyperparameter; $H^{\pi_\theta}(s_t)$ is the entropy of the policy. The exploration ability of the DRL algorithm can be enhanced by maximizing the entropy. To improve the stability of the training process, the parameters of the policy network and the value network will be shared, and the shared parameters are denoted as θ. Then, the overall loss function can be expressed as

$$L^{\text{Total}}(\theta)=L^\pi_t(\theta)+c_2\cdot L^v(\theta) \tag{32}$$

where $c_2\in[0,1]$ is the hyperparameter. Therefore, the PPO agent updates the policy by minimizing the overall loss $L^{\text{Total}}(\theta)$.

2) Mixed Output Policy: The control variables of the shiftable appliances and the controllable appliances in smart home are discrete and continuous, respectively. Therefore, a MOP is proposed, where the discrete actions are sampled from multivariate Bernoulli distribution (MBD) and the continuous actions are sampled from multivariate normal distribution (MND), shown in Fig. 2. The output of policy network includes two parts. The sigmoid function is used to output the p vector $p_\theta = [p_1, p_2, \ldots, p_j]^T$ of MBD. The linear function is used to output mean vector $\mu_\theta = [\mu_1, \mu_2, \ldots, \mu_j]^T$ and variance vector $\sigma_\theta^2 = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_j^2]^T$ of MND. Furthermore, sampling discrete actions of the shiftable appliances from

$$\pi_\theta\left(I_t^S|s_t\right)_{p_\theta} = \prod_{i=1}^{j} p_i^{I_t^{s_i}} (1-p_i)^{1-I_t^{s_i}} \tag{33}$$

where $I_t^S = \{I_t^{s_1}, \ldots, I_t^{s_j}\}$ is a control variable set of shiftable appliances. The continuous actions of the controllable appliances are sampled from

$$\pi_\theta\left(P_t^C|s_t\right)_{\mu_\theta,\sigma_\theta} = \frac{1}{(2\pi)^{j/2}\prod_{i=1}^{j}\sigma_i} \exp\left(-\sum_{i=1}^{j}\frac{\left(P_t^{c_i}-\mu_i\right)^2}{2\sigma_i^2}\right) \tag{34}$$

where $P_t^C = \left\{P_t^{c_1}, \ldots, P_t^{c_j}\right\}$ is a control variable set of controllable appliances. The MOP is expressed as

$$\pi_\theta(a_t|s_t) = \pi_\theta\left(I_t^S|s_t\right)_p \cdot \pi_\theta\left(P_t^C|s_t\right)_{\mu,\sigma}. \tag{35}$$

The action sampled from (35) are truncated according to the action constraints of the different appliances, and finally executed in the smart home. Note that the proposed MOP can generate both discrete and continuous actions, which provides more accurate control for different kinds of appliances.

B. Multiprocess Data Sampling

In the training process, the data used for training are acquired by interacting with the environment, the pseudocode of data sampling is shown in Algorithm 1. Specifically, in each time slot t, the agent observes the state s_t of the environment and selects an action a_t, which is executed in the environment. Then, the state of the environment is transformed by a_t and a reward r_t is returned from the environment. PPO is an on-policy RL algorithm that uses the same policy for sampling and updating in the training process, which limits training efficiency. In this article, the multiprocess data sampling method is adopted in the training process to accelerate data sampling speed, where the pseudocode is shown in Algorithm 2. In each training episode, the multiprocess pool technic is adopted to sample the data used for training. Then, the agent was trained with the minibatch sampling technic. After training, the trained agent can be adopted to real-time scheduling.

Algorithm 1: Data Sampling Method.

```
Create an environment ENV of smart home
Initialize data buffer m_d
Initialize ENV with d to get state S_0
for t = 1, 2, ..., T_end do
    Sample action a_t and log π_θ(a_t|s_t) based on (35)
    Constrain a_t with (2), (5b), (7d), (9b), (12a)
    Execute a_t and observe S_{t+1}
    Calculate r_{t+1} based on (26)
    Store (s_t, a_t, s_{t+1}, r_{t+1}, log π_θ(a_t|s_t)) to m_d
end for
The data in m_d are for training
```

Algorithm 2: Training Process of PPO Agent.

```
Initialize policy network π_θ and state value network v_θ
Initialize the maximum episodes M_ep, the trajectory size D
for episode = 1, 2, ..., M_ep do
    Initialize multiprocess pool P and trajectory buffer M
    P generates subprocesses with tasks d = 1, 2, ..., D
    Execute Algorithm 1 in each subprocess
    Push data buffer of subprocess to M
    for i = 1, 2, ..., N do
        Sample minibatch data from M
        Calculate Â_t based on (29)
        Calculate L_t^Total(θ_t) based on (32)
        Update networks by minimizing loss L_t^Total(θ_t)
    end for
end for
Use π_θ for real-time energy management in smart home
```

V. Simulations

A. Simulation Setup

Appliances adopted in simulations are shown in Table I. A day (started at 8 A.M. and ended at 8 A.M. of next day) is divided into 96 time slots, that is, $T_{\text{end}} = 96$, $\Delta T = 0.25$ h. At the beginning of each time slot, HEMS will monitor the state of smart home and then execute scheduling action, ultimately minimizing the energy cost of the day. The IBR factor is set to $\xi = 1.4423$ [23] and the maximum power limitation $P_{\max}^g = 6$ kW. We adopted truncated normal distribution TN (μ, σ, min, max) [32] to simulate the behaviors of users, which can generate t_{ini} and t_{end} of each appliances in smart home randomly. The parameters related to the PV are set as $\rho_{\text{pv}} = 0.2$; $A_{\text{pv}} = 15$. In addition, the parameters of HVAC, EWH, EV, and ESS are set as Table II.

Both the policy network and the value network have 3 hidden layers with 128 neurons in each layer. The activation functions *Relu* and *Tanh* are adopted in hidden layers of the policy network and the value network, respectively.

Hourly RTP data from CAISO [33], outdoor temperature data in Los Angeles from NOAA/NCDC [34], solar irradiation data in Golden city of USA from NREL/MIDC [35] are adopted in simulations. The scheduling windows of appliances and W_t of EWH are generated by TN. Also, the data from December 1, 2019 to January 31, 2020, is for training, the test after training using data from December 1, 2020 to January 31, 2021.

TABLE I
PARAMETERS OF APPLIANCES USED IN SIMULATIONS

Shiftable appliances			
Name	$P_{\text{rated}}/\text{kW}$	t_{ini}	t_{end}
dish washer	1.8	TN(40,44,42,2)	TN(58,62,60,2)
wash machine	0.4	TN(6,10,8,4)	TN(36,44,40,4)
clothes dryer	1.2	WM done	WM done+8
Controllable appliances			
Name	P_{max}/kW	t_{ini}	t_{end}
AC	2.5	0	96
EWH	4.5	0	96
EV	6	TN(36,44,40,4)	TN(92,96,94,2)
ESS	2.4	0	96
Nonshiftable appliances			
Name	$P_{\text{rated}}/\text{kW}$	t_{ini}	Duration
TV	0.1	TN(40,44,42,2)	TN(14,18,16,2)
refrigerator	0.2	0	96
light	0.2	TN(32,40,36,4)	TN(18,22,20,2)
vacuum cleaner	1.2	TN(24,32,28,4)	TN(1,5,3,2)
hairdryer	1	TN(48,52,50,2)	1

TABLE II
PARAMETERS OF CONTROLLABLE APPLIANCES

Appliances	parameters	value
HVAC	ε	0.968
	η^{HVAC}	1.0
	A	$7.27 \cdot 10^{-3}\text{kW}/^{\circ}\text{C}$
	α_1	0.01
	$T_{\text{set}}^{\text{in}}$	22°C
	$\Delta T_{\text{set}}^{\text{in}}$	2°C
EWH	ρ	$1000\text{kg}/\text{m}^3$
	c_p	$4.2 \cdot 10^3\text{J}/(\text{kg} \cdot {}^{\circ}\text{C})$
	G	$2.60\text{W}/^{\circ}\text{C}$
	T_{co}	15°C
	α_2	0.01
	$T_{\text{set}}^{\text{wa}}$	60°C
	$\Delta T_{\text{set}}^{\text{wa}}$	3°C
EV	$\eta_{\text{ch}}^{\text{EV}}$	0.95
	$E_{\text{max}}^{\text{EV}}$	24kWh
	α_3	0.05
ESS	$\eta_{\text{ch}}^{\text{ESS}}$	0.95
	$\eta_{\text{dis}}^{\text{ESS}}$	0.95
	$E_{\text{max}}^{\text{ESS}}$	6.4kWh
	C_0	$E_{\text{max}}^{\text{ESS}} \cdot 120\$/\text{kWh}$ [28]

Parameters of PPO agent are shown in Table III. The DRL agent is implemented by Pytorch-1.0, Python-3.8, and trained on Windows 10 platform with I5-10400 CPU(2.9 GHz), 16 GB of RAM. The training process took around 2 h but the trained agent makes decision only in milliseconds at each time slot, which can be used for real-time scheduling.

B. Performance Analysis

1) Learning Performance: Like other learning-based methods, it is necessary to analyze the training result before testing. It can be observed from Fig. 3 that the mean reward agent got increases rapidly after the training process begins and gradually converges, ultimately, reaching -1.6 at around 800 episodes, and remains stable, which means the agent has successfully learned an effective policy. The training process took around 2 h.

2) Test Day Scheduling Results: Applying the trained agent in the testing process, the scheduling results of a certain day are shown in Figs. 4–7. We can observe from Fig. 4 that the shiftable appliances are scheduled for the periods where the electricity price is low. As shown in Fig. 5, the indoor temperature and the water temperature are kept near the set temperature,

TABLE III
PARAMETERS OF PPO AGENT WHILE TRAINING

Parameters	Value	Parameters	Value
D	60	c_2	0.5
M_{ep}	2500	ϵ	0.2
N	20	γ	0.995
c_1	0.01	λ	0.97

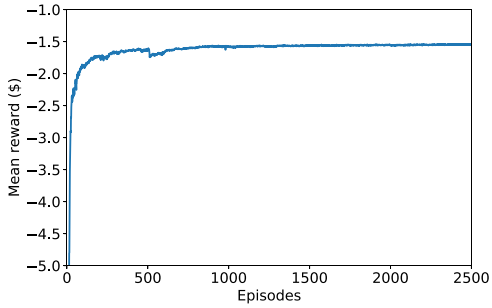

Fig. 3. Mean reward in each episode while in training.

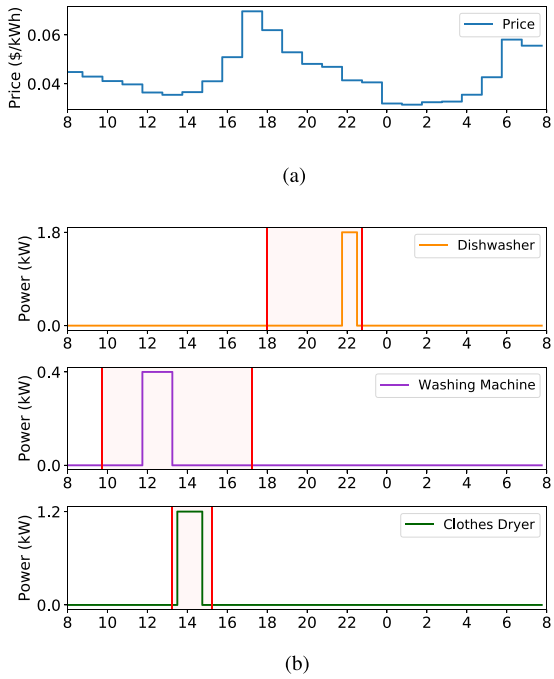

Fig. 4. Performance of shiftable appliances on a test day. (a) Real-time price [33]. (b) Shiftable appliances. The appliance can be operated by HEMS within the region between the two red lines.

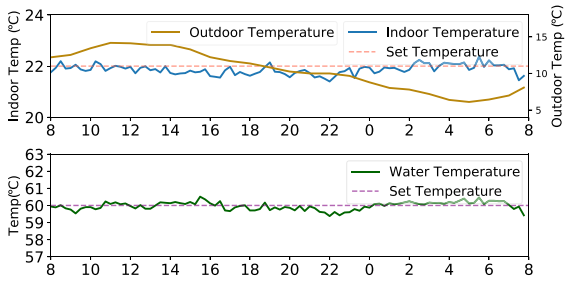

Fig. 5. Performance of thermal loads on a test day.

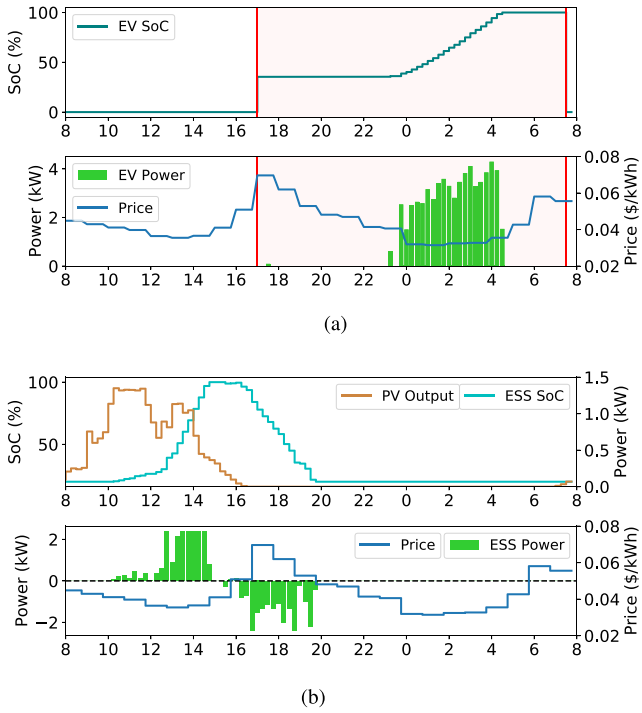

Fig. 6. Performance of EV and ESS on a test day. (a) EV. The region between red lines denotes the time when EV is at home. (b) ESS.

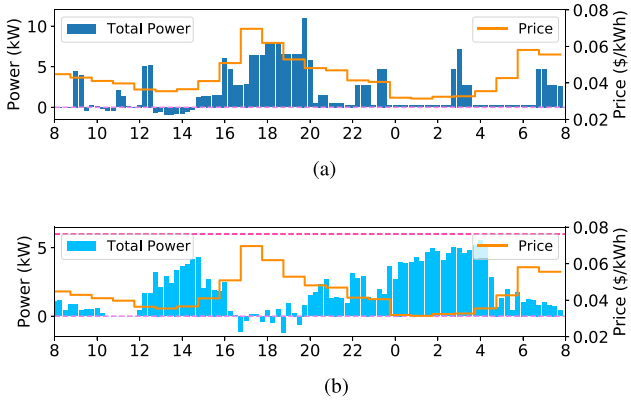

Fig. 7. Total power during the test day. (a) Without scheduling. (b) After scheduling.

indicating that thermal comfort is guaranteed. Meanwhile, the charging task of EV is scheduled for the low-price period instead of charging immediately after getting home and is completed before leaving the home, shown in Fig. 6(a). Although the ESS operates throughout the day, it performs one charge–discharge cycle to avoid battery degradation caused by overuse, which is shown in Fig. 6(b). Note that the ESS charges from 12:00 to 15:00 when the PV output is surplus and the electricity prices are low instead of at midnight when the electricity prices are cheapest, which makes full use of the PV output and reduces the energy cost of charging. The proposed method avoids purchasing too much electricity from the grid when the electricity prices are high while also keeping the total power demand below 6 kW to avoid the IBR price, shown in Fig. 7.

C. Performance Comparison

1) Comparison Setup: The real-world data from December 1, 2020 to January 31, 2021, are used in our comparison experiment. The proposed algorithm is compared with DDPG and TRPO, which are state-of-the-art DRL algorithms based on the actor–critic frame. DDPG and TRPO adopt the same policy network and value network as the proposed algorithm. Besides, another three baselines are considered, which are shown as follows.

1) *Baseline 1 (B1):* Without HEMS and ESS. HVAC and EWH adopt ON/OFF control, stop working when the temperature is higher than T_{set}, otherwise work at maximum power. EV charges at a maximum power immediately after getting home. Other appliances, which are assigned a task will work without delay.
2) *Baseline 2 (B2):* Assuming that all future information (RTP, outdoor temperature, etc.) is known. The energy management problem can be formulated into a constrained optimization problem, which can be solved by a solver (e.g., Gurobi [36]). Note that *B2* represents the best scheduling performance but cannot be truly achieved since the future information is unpredictable.
3) *Baseline 3 (B3):* Adopting MPC to schedule the appliances in the smart home [6]. The MPC controller forecasts the future information (RTP, outdoor temperature, etc.) in each time slot t for a receding horizon (t, T). Then, the MPC solves an optimization problem and executes the first step of the optimized result. We assume that the prediction of the future information is equal to the real data plus a bias sampling from a normal distribution $N(0, {\sigma_\tau}^2)$ truncated by $[-0.15\sigma_\tau, 0.15\sigma_\tau]$, where ${\sigma_\tau}^2$ is the variance of the real data of the horizon (t, T).

2) Comparison Results: It can be seen from Fig. 8 that the proposed algorithm converges faster and reaches a higher reward, which means the proposed algorithm has learned a better policy to deal with the energy management problem in the smart home. Note that *B2* represents the lower bound of the energy cost but cannot be truly achieved. As shown in Fig. 9, compared with *B1*, the cumulative energy cost of the proposed algorithm has been reduced by 42.9%, which is close to that of *B2* (50.9%).

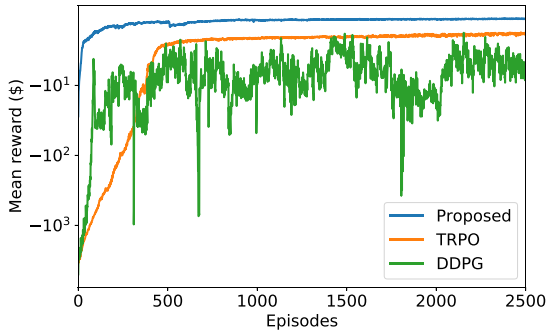

Fig. 8. Comparison of TRPO, DDPG, and the proposed algorithm while training.

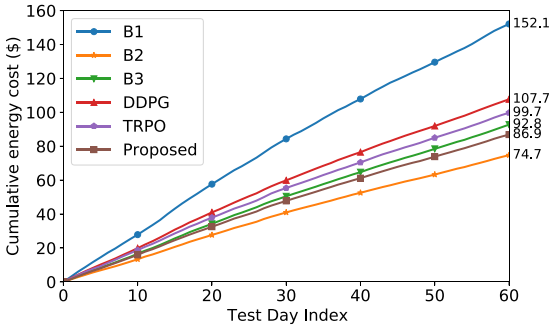

Fig. 9. Cumulative energy cost on the test days.

TABLE IV
DISTRIBUTION OF MEAN COST IN TEST DAYS ($)

	B1	B2	B3	DDPG[20]	TRPO[22]	Proposed
Energy cost	2.53	1.25	1.55	1.79	1.66	1.45
ESS cost	0	0	0.12	0.01	0.08	0.05
Thermal comfort	2.27	0.09	0.03	5.64	0.16	0.06
Range anxiety	0	0	0.01	0	0.02	0.05

TABLE V
CALCULATION TIME OF DIFFERENT METHODS WHILE MAKING DECISION AT EACH TIME SLOT

	B1	B2	B3	DDPG[20]	TRPO[22]	Proposed
calculation time	1 ms	-	30 s	1 ms	1 ms	1 ms

However, B3, DDPG, and TRPO just reduce the energy cost by 39.0%, 29.2%, and 34.5%, respectively. Besides, the thermal comfort cost of DDPG and TRPO is higher than that of the proposed method, shown in Table IV. Compared with *B3*, the proposed method reached a lower ESS cost, which means less degradation of the ESS battery. The calculation times of each method while making decision in the testing process is shown in Table V. The learning-based methods (i.e., TRPO, DDPG, and the proposed method) cost much time in training process but cost milliseconds in the testing process, which can be used in real-time scheduling. However, the MPC controller (i.e., *B3*) has to solve a mixed-integer nonconvex optimization problem, which costs much more time.

TABLE VI
CUMULATIVE ELECTRICITY COST ON THE TEST DAYS ($)

Disturbance	B1	B2	B3	DDPG[20]	TRPO[22]	Proposed
1°C	151.4	74.1	96.7	106.7	99.5	88.2
2°C	153.0	76.3	92.4	106.3	100.8	85.8
3°C	154.9	77.4	99.9	105.4	103.5	86.6

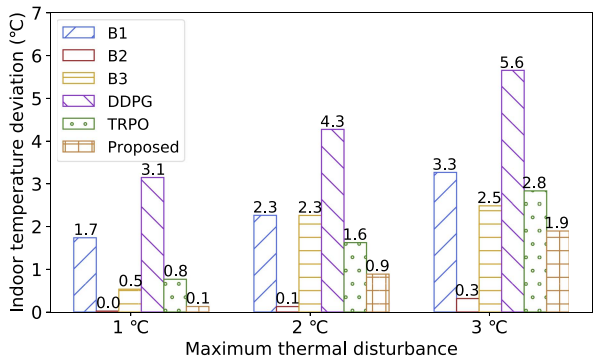

Fig. 10. Average of indoor temperature deviation on the test days.

D. Algorithmic Robustness

The model of HVAC in (5a) is a simplification of that in reality. In the real world, the indoor temperature is affected by many external factors (e.g., weather, user's behaviors, etc.), which cannot be accurately described by (5a). Therefore, a thermodynamic model with disturbance is considered to verify the robustness of the proposed algorithm, specifically, $T_{t+1}^{\text{in}} = \varepsilon T_t^{\text{in}} + (1-\varepsilon)(T_t^{\text{out}} - \eta^{\text{HVAC}} P_t^{\text{HVAC}} \Delta T/A) + u_t$, where u_t is a random variable that obeys a uniform distribution $U(a,b)$. In this section, three cases are considered, which is, $-a = b = 1, 2, 3$. Compared with baselines *B1, B3, DDPG, and TRPO*, the proposed algorithm achieves a lower energy cost and is close to the *B2* in all three cases, shown in Table VI. Meanwhile, it can be observed from Fig. 10 that the proposed algorithm can achieve a lower indoor temperature deviation in all three cases. The indoor temperature is maintained in a comfortable range. In general, the robustness of the proposed algorithm shows its potential practicability.

VI. CONCLUSION

In this article, a PPO-based home energy management algorithm had been proposed to minimize the household energy cost. A policy network with discrete and continuous outputs was adopted to generate actions for different types of devices in smart home. Also, the user's thermal comfort requirements as well as the uncertainties of user' behaviors, RTP, outdoor temperature, and PV power generation had been taken into consideration. Besides, the degradation cost model of the battery in ESS was considered. Simulations based on the real-world data had shown that the proposed algorithm could effectively solve the energy management problem in the smart home. The results show that the proposed method performed better in reducing the household energy cost while maintaining the comfort of users and minimizing the ESS degradation cost. The robustness

test has shown that the algorithm had potential practicability. In future work, we will further investigate the energy management problem of aggregated smart homes, e.g., apartments and commercial buildings. In this situation, the interaction between smart homes is considered to reduce the energy cost of aggregated smart homes. The privacy of users while interacting should be guaranteed as well.

REFERENCES

[1] N. G. Paterakis, O. Erdinc, A. G. Bakirtzis, and J. P. Catalao, "Optimal household appliances scheduling under day-ahead pricing and load-shaping demand response strategies," *IEEE Trans. Ind. Informat.*, vol. 11, no. 6, pp. 1509–1519, Dec. 2015.

[2] S. Althaher, P. Mancarella, and J. Mutale, "Automated demand response from home energy management system under dynamic pricing and power and comfort constraints," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1874–1883, Jul. 2015.

[3] M. Shafie-Khah and P. Siano, "A stochastic home energy management system considering satisfaction cost and response fatigue," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 629–638, Feb. 2018.

[4] X. Wu, X. Hu, X. Yin, and S. J. Moura, "Stochastic optimal energy management of smart home with PEV energy storage," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2065–2075, May 2018.

[5] S. Paul and N. P. Padhy, "Real-time energy management for smart homes," *IEEE Syst. J.*, vol. 15, no. 3, pp. 4177–4188, Sep. 2021.

[6] M. Yousefi, A. Hajizadeh, M. N. Soltani, and B. Hredzak, "Predictive home energy management system with photovoltaic array, heat pump, and plug-in electric vehicle," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 430–440, Jan. 2021.

[7] D. Zhang, S. Li, M. Sun, and Z. ONeill, "An optimal and learning-based demand response and home energy management system," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1790–1801, Jul. 2016.

[8] C. Keerthisinghe, A. C. Chapman, and G. Verbič, "Energy management of PV-storage systems: Policy approximations using machine learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 1, pp. 257–265, Jan. 2019.

[9] F. Ruelens, B. J. Claessens, S. Quaiyum, B. De Schutter, R. Babuška, and R. Belmans, "Reinforcement learning applied to an electric water heater: From theory to practice," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3792–3800, Jul. 2018.

[10] T. Remani, E. Jasmin, and T. I. Ahamed, "Residential load scheduling with renewable generation in the smart grid: A reinforcement learning approach," *IEEE Syst. J.*, vol. 13, no. 3, pp. 3283–3294, Sep. 2019.

[11] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6629–6639, Nov. 2019.

[12] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3201–3211, Jul. 2020.

[13] M. Ahrarinouri, M. Rastegar, and A. R. Seifi, "Multiagent reinforcement learning for energy management in residential buildings," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 659–666, Jan. 2021.

[14] C. Lei, "Deep reinforcement learning," in *Deep Learning and Practice With MindSpore*. Berlin, Germany: Springer, 2021, pp. 217–243.

[15] Y. Du et al., "Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning," *Appl. Energy*, vol. 281, 2021, Art. no. 116117.

[16] A. Mathew, A. Roy, and J. Mathew, "Intelligent residential energy management system using deep reinforcement learning," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5362–5372, Dec. 2020.

[17] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[18] E. Mocanu et al., "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.

[19] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3068–3082, Jul. 2020.

[20] L. Yu et al., "Deep reinforcement learning for smart home energy management," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 2751–2762, Apr. 2020.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[22] H. Li, Z. Wan, and H. He, "Real-time residential demand response," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4144–4154, Sep. 2020.

[23] Y. F. Du, L. Jiang, Y. Li, and Q. Wu, "A robust optimization approach for demand side scheduling considering uncertainty of manually operated appliances," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 743–755, Mar. 2018.

[24] H. Li, Z. Wan, and H. He, "A deep reinforcement learning based approach for home energy management system," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf.*, 2020, pp. 1–5.

[25] M. Shad, A. Momeni, R. Errouissi, C. P. Diduch, M. E. Kaye, and L. Chang, "Identification and estimation for electric water heaters in direct load control programs," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 947–955, Mar. 2017.

[26] Y. Sun, H. Yue, J. Zhang, and C. Booth, "Minimization of residential energy cost considering energy storage system and EV with driving usage probabilities," *IEEE Trans. Sustain. Energy*, vol. 10, no. 4, pp. 1752–1763, Oct. 2019.

[27] N. Kittner, F. Lill, and D. M. Kammen, "Energy storage deployment and innovation for the clean energy transition," *Nature Energy*, vol. 2, no. 9, pp. 1–6, 2017.

[28] W. Cole, A. W. Frazier, and C. Augustine, "Cost projections for utility-scale battery storage: 2021 update," Nat. Renewable Energy Lab. (NREL), Golden, CO, USA, Tech. Rep. NREL/TP-6A20-79236, 2021.

[29] L. Yu, T. Jiang, and Y. Zou, "Online energy management for a sustainable smart home with an HVAC load and random occupancy," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1646–1659, Mar. 2019.

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[31] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2015, *arXiv:1506.02438*.

[32] J. Burkardt, "The truncated normal distribution," Dept. Scientific Computing Website, Florida State Univ., pp. 1–35, 2014. [Online]. Available: https://people.sc.fsu.edu/jburkardt/presentations/truncated_normal.pdf

[33] CAISO, "Open access same-time information system." [Online]. Available: http://oasis.caiso.com/

[34] NOAA, "National climatic data center," 2021. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/

[35] NREL, "The measurement and instrumentation data center." 2021. [Online]. Available: https://midcdmz.nrel.gov/

[36] Gurobi, "Gurobi optimization." 2021. [Online]. Available: https://www.gurobi.com//