Deep Reinforcement Learning for Smart Building Energy Management: A Survey

Liang Yu, *Member, IEEE*, Shuqi Qin, Meng Zhang, Chao Shen, *Senior Member, IEEE*,
Tao Jiang, *Fellow, IEEE*, Xiaohong Guan, *Fellow, IEEE*

Abstract

Global buildings consumed 30% of total energy and generated 28% of total carbon emission in 2018, which leads to economic and environmental concerns. Therefore, it is of great significance to reduce energy consumption, energy cost and carbon emission of buildings while maintaining user comfort. To this end, several challenges have to be addressed. Firstly, it is very challenging to develop a building thermal dynamics model that is both accurate and efficient enough for building control. Secondly, there are many kinds of uncertainties. Thirdly, there are many spatially and temporally operational constraints. Fourthly, building energy optimization problems may have extremely large solution spaces, which can not be solved in real-time by traditional methods. Fifthly, traditional building energy management methods have respective applicable premises, which means that they have low versatility when confronted with varying building environments. As a general artificial intelligence technology, deep reinforcement learning (DRL) has the potential of addressing the above challenges. Thus, this paper presents a comprehensive literature review on DRL for smart building energy management (SBEM). To be specific, we first introduce the fundamentals of DRL and provide the classification of DRL methods used in existing works related to SBEM. Then, we review the applications of DRL in a single building energy subsystem, multiple energy subsystems of buildings, and building microgrids, respectively. Furthermore, we identify the unsolved issues and point out the possible research directions of applying DRL. Finally, we summarize the lessons learned from this survey.

Index Terms

Deep reinforcement learning (DRL), artificial intelligence, smart buildings, energy management, smart home, building microgrids, uncertainty, energy cost, carbon emission

I. Introduction

Buildings account for a large portion of total energy consumption and total carbon emission in the world [1] [2]. For example, global buildings consumed 30% of total energy and generated 28% of total carbon emission in 2018 [3]. Moreover, the energy demand of buildings is expected to increase by 50% in the next 30 years [4]. Under the above background, smart buildings have received more and more attention in recent years, which can provide sustainable, economical and comfortable operation environments for occupants using many advanced technologies, e.g., Internet of Things (IoT), cloud computing, machine learning and big data analytics [5]–[7]. To support the above features, smart building energy management (SBEM) is of great importance [8]. To be specific, by intelligently scheduling building energy systems, the optimal tradeoff between energy consumption, carbon emission, energy cost, and user comfort can be achieved [9]–[14].

Although SBEM has many advantages, the following challenges have to be addressed. Firstly, it is often intractable to develop a building thermal dynamics model that is accurate and efficient enough for building control [15]. Secondly, there are many sources of uncertainties related to SBEM [16], e.g., renewable generation output, electricity price, indoor temperature, outdoor temperature, CO_2 concentration, number of occupants, and power demand of appliances. Thirdly, there are many temporally and spatially coupled operational constraints related to energy subsystems [17] [18], e.g., heating, ventilation, and air conditioning (HVAC) systems, energy storage systems (ESSs), electric vehicles (EVs), which means that the current system decision will affect the future decisions and the decisions among different subsystems should be coordinated. Fourthly, it is difficult to solve large-scale building energy optimization problems in real-time when traditional optimization methods are adopted [19]. Finally, it is hard to develop a generalized building energy management method that can be applied in all building environments [14]. In existing SBEM methods, most of them have strong applicable premises, which may not be satisfied by some building environments [20]. For example, stochastic programming and model predictive control need the prior or forecasting information of uncertain parameters [21] [22], and Lyapunov optimization techniques require some strict usage conditions [12] [23].

L. Yu is with the College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, and also with Systems Engineering Institute, Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China. (email: liang.yu@njupt.edu.cn)

S. Qin is with the College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. (email: shuqi_qin@163.com)

M. Zhang, C. Shen and X. Guan are with Systems Engineering Institute, Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China. (email: mengzhang2009@xjtu.edu.cn, {cshen,xhguan}@sei.xjtu.edu.cn)

T. Jiang is with Wuhan National Laboratory for Optoelectronics, School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China. (email: Tao.Jiang@ieee.org)

TABLE I
THE COMPARISON BETWEEN OUR WORK AND RELATED SURVEYS

Literature	System type(s)	Main focus	Involved Methods/Algorithms	DRL methods for SBEM classified	Future directions in DRL-based SBEM provided
Han *et al.* [42]	HVAC	Occupant comfort control	RL	No	No
Leitão *et al.* [43]	Smart home	Energy optimization	LP, NLP, CP, DP, GA, PSO, MPC, RL	No	No
Mason *et al.* [44]	HVAC, EWH, home management systems, smart home	Building energy optimization	RL	No	No
Wang *et al.* [45]	HVAC, batteries, home appliances, EWH, windows, lighting	Building control	RL	No	No
Rajasekhar *et al.* [46]	HVAC	Building energy optimization	ANN, RNN, WNN, RT, SVM, PSO, MPC, SMPC, FL, RL, DQN	No	No
Zhang *et al.* [47]	Smart grid	Load forecasting, microgrid, demand response, cyber security	RL, DQN, DDPG, NAF, A3C	No	No
Yang *et al.* [48]	Microgrid, ESS, HVAC, home appliances, PV	Operation optimization, system control, energy markets, cyber security	RL, DQN, DDPG, A3C, DDQN, TRPO	No	No
Our work	HVAC, multiple energy subsystems, building microgrids	Building energy optimization	DQN, DDQN, BDQ, DDPG, PDDPG, MADDPG, FH-DDPG, A2C, A3C, TRPO, PPO, MAPPO, MuZero, MAAC, EB-C-A2C, EB-C-DQN	Yes	Yes

To address the above challenges, deep reinforcement learning (DRL) [24] can be used, which is a general artificial intelligence technology and has shown great success in many application fields [25]–[32]. Different from other types of machine learning such as supervised learning and unsupervised learning, DRL or RL can be used in sequential decision-making problems under uncertainty for optimal control [25] [33]. In 2017, the first work that adopts DRL algorithm for SBEM has been done [15]. To be specific, a deep Q-network (DQN) algorithm has been adopted for the control of building HVAC systems and simulation results have showed the effectiveness of the designed control algorithm in reducing energy cost and maintaining thermal comfort of occupants. Since then, many DRL-based methods for SBEM have been proposed [14] [17] [34]. In general, DRL-based methods have following advantages in dealing with above-mentioned challenges.

- Model-free DRL methods can learn an optimal control policy merely based on the interactive information with building environments. In other words, they can operate without knowing explicit building thermal dynamics models [15].
- DRL methods can operate in an online way without knowing any forecast information or statistics information of building environments, which can effectively overcome the challenges brought by system uncertainties and temporally-coupled constraints related to HVAC systems, ESSs, EVs and so on [16] [35].
- Multi-agent DRL methods support the flexible coordination among different building energy subsystems, which can deal with spatially-coupled operational constraints very well [17].
- DRL methods can support "end-to-end" control for large-scale building energy optimization problems. To be specific, the DRL agent can determine the optimal control actions instantly (e.g., few milliseconds) given the high-dimensional raw state data [19] [36] [37].
- DRL methods have wide applicable premises in building energy optimization. For example, the training of agents in model-free DRL methods is conducted by the trial-and-error process, which means that rigorous mathematical models and premise conditions are not required. Thus, the trained DRL agent can still work or even be improved persistently by online learning when confronted with varying building environments [19] [38].

There are many surveys related to DRL in the literature. However, they do not focus on SBEM. For example, the applications of DRL in power systems, communications and networking, autonomous IoT, cyber security, and multi-agent systems can be found in [20], [29], [39]–[41]. In addition, there are several surveys on building energy systems, but the involved methods are RL [42]–[45] or other artificial intelligence methods (e.g., model predictive control (MPC), fuzzy logic (FL)) [46]. Although some DRL algorithms are mentioned in [47] and [48], they mainly focus on different applications (ranging from load forecasting to cyber security) of RL/DRL in sustainable energy and electric systems. To the best of our knowledge, there is no survey that completely focuses on DRL for SBEM. Based on the above observation, we are motivated to conduct a comprehensive survey on DRL for SBEM and identify the unsolved issues as well as the possible research directions in this field. For convenience, we provide the comparison between our work and related surveys in Table I. It can be observed that our work completely focuses on DRL for SBEM from the perspective of system complexities (i.e., a single building energy subsystem, multiple energy subsystems in buildings, and building microgrids), while works in [42]–[46] mainly focus on RL and other artificial intelligence methods for occupant comfort control and building energy optimization. Compared with [48], we provide a deeper analysis of DRL-based building energy optimization. For example, we summarize DRL advantages for building energy optimization comprehensively. Moreover, more DRL methods for SBEM are classified and reviewed. Furthermore, we point out challenges and the future research directions of DRL-based building energy management.

The rest of this paper is organized as follows. In Section II, we give an overview of DRL. In Section III, we introduce the background of SBEM, the procedure of solving SBEM problems using DRL, and the classification of DRL methods for SBEM. In Section IV, we discuss DRL applications in a single building energy subsystem. In Section V, we discuss DRL applications in multiple energy subsystems of buildings. In Section VI, we discuss DRL applications in building microgrids. In Section VII, we identify some unsolved issues and point out the future research directions. Finally, conclusions and lessons learned are provided in Section VIII. For easy understanding, the list of abbreviations commonly appeared in this paper is given in Table II.

II. An Overview of Deep Reinforcement Learning

DRL can be regarded as the combination of deep learning [49] and reinforcement learning (RL) [50]. To be specific, deep neural networks are adopted to approximate the optimal value functions or optimal policies in RL. Therefore, DRL has a powerful representation ability through deep learning and a strong decision-making ability under uncertainty through reinforcement learning. Since DRL algorithms are mainly based on Markov decision process (MDP) framework or its variants (e.g., Partially observable MDP [29], Markov game [17]), we first give the background of MDP. Then, we introduce the basic principle of RL. Finally, we introduce DQN, which is the first DRL algorithm used in the field of SBEM.

A. MDP

Typically, an MDP is defined by a five-tuple ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$, $\mathcal{R}$, γ), where $\mathcal{S}$ and $\mathcal{A}$ denote the sets of state and action, respectively. $\mathcal{P}: \ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\rightarrow[0,1]$ is the state transition probability function, which models the uncertainty in the evolution of system states based on the action taken by the agent. $\mathcal{R}:\mathcal{S}\times\mathcal{A}\rightarrow\mathbb{R}$ is the reward function and $\gamma\in[0,\ 1]$ is a discount factor. Note that MDP provides a mathematical framework for sequential optimal decision-making problems under uncertainty. In other words, the decision maker (i.e., the agent) observes a state S_t and takes an action A_t at each time slot t. Next, the state of the system (i.e., the environment) evolves into another one. Then, the agent finds itself in a new state S_{t+1} and receives a reward R_{t+1}. In addition, the aim of the agent at time slot t is to maximize the expected return it receives over the future [33], where is given by $\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$.

B. RL

As shown in the literature, RL has been widely used in solving MDPs [33], [42], [51]–[56]. In a RL process, the agent learns its optimal policy π by interacting with the environment, where a policy π is a mapping from states to the probabilities of selecting every possible action [33]. In particular, the agent observes a state and takes an action at slot t. Then, it receives a reward and a new state, which are used to update the policy. The above process repeats until the policy converges. Since Q-learning is the most effective algorithm of learning an optimal policy in RL, we will introduce its basic principle.

Let the value of taking action a in state s under a policy π be $Q_\pi(s,a)$, which is defined as follows,

$$Q_\pi(s,a)\doteq\mathbb{E}_\pi[\sum\nolimits_{k=0}^{\infty}\gamma^k R_{t+k+1}(S_t=s,A_t=a)], \tag{1}$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expected value of a random variable given that the agent follows policy π. Then, the optimal action-value function $Q^*(s,a)$ is $\max_\pi Q_\pi(s,a)$ and can be calculated by the following Bellman optimality equation in a recursive manner [16], i.e.,

$$\begin{aligned} Q^*(s,a)&=\mathbb{E}[R_{t+1}+\gamma\text{max}_{a'}Q^*(S_{t+1},a')|S_t=s,A_t=a]\\ &=\sum\nolimits_{s',r}P(s',r|s,a)[r+\gamma\text{max}_{a'}Q^*(s',a')], \end{aligned}$$

where $s'\in\mathcal{S}$, $\boldsymbol{r}\in\mathcal{R}$, $a'\in\mathcal{A}$, and $P(s',r|s,a)$ denotes a conditional probability function. To obtain the value of $Q^*(s,a)$, the information of $P(s',r|s,a)$ must be known, which may be unavailable in practice. To address this challenge, Q-learning algorithm is proposed to approximate $Q^*(s,a)$ using the following way,

$$Q(S_t,A_t)=Q(S_t,A_t)+\Delta_t, \tag{2}$$

where $\Delta_t=\alpha\Big[R_{t+1}+\gamma\max_a Q(S_{t+1},a)-Q(S_t,A_t)\Big]$ and α is the step size. It is obvious that $Q(S_t,A_t)=R_{t+1}+\gamma\max_a Q(S_{t+1},a)$ when $\Delta_t=0$. At this time, $Q(S_t,A_t)$ will not be updated and the learned action-value function Q directly

TABLE II
THE LIST OF ABBREVIATIONS

Abbreviation	Description
IoT	Internet of Things
SBEM	Smart Building Energy Management
HVAC	Heating, Ventilation and Air Conditioning
DRL/RL	Deep Reinforcement Learning/Reinforcement Learning
MPC	Model Predictive Control
FL	Fuzzy Logic
LP	Linear Programming
NLP	Non-Linear Programming
CP	Convex Programming
DP	Dynamic Programming
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
ANN/RNN/WNN	Artificial/Recurrent/Wavelet Neural Network
RT	Regression Tree
SVM	Support Vector Machine
SMPC	Stochastic Model Predictive Control
DQN	Deep Q-Network
A3C	Asynchronous Advantage Actor-Critic
D-DNFQI	Double Deep Neural Fitted Q Iteration
DDQN	Double Deep Q-Network
BDQ	Branching Dueling Q-Network
DDPG	Deep Deterministic Policy Gradient
NAF	Normalized Advantage Functions
PDDPG	Prioritized Deep Deterministic Policy Gradient
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
FH-DDPG	Finite-Horizon Deep Deterministic Policy Gradient
A2C	Advantage Actor-Critic
TRPO	Trust Region Policy Optimization
PPO	Proximal Policy Optimization
MAPPO	Multi-Agent Proximal Policy Optimization
MAAC	Multi-Actor Attention-Critic
EB-C-A2C	Entropy-Based Collective Advantage Actor-Critic
EB-C-DQN	Entropy-Based Collective Deep Q-Network
MDP	Markov Decision Process
LSTM	Long Short-Term Memory
AHU	Air Handling Unit
MES	Multiple Energy Subsystem
BEM	Building Energy Model
BAS	Building Automation System
ESS	Energy Storage System
PV	Photovoltaic Panels
WT	Wind Turbine
DG	Diesel Generator
EV	Electric Vehicle
EWH	Electric Water Heater
DW	Dishwasher
WM	Washing Machine
TES	Thermal Energy Storage
EHP	Electric Heat Pump
GB	Gas Boiler
CD	Clothes Dryer
IAQ	Indoor Air Quality
MCTS	Monte-Carlo Tree Search
FH-RDPG	Finite-Horizon Recurrent Deterministic Policy Gradient
VAV	Variable Air Volume

approximates the optimal action-value function $Q^*(s,a)$. Note that Q-learning algorithm is effective when state and action spaces are small. However, the spaces of some parameters may be large and continuous in practice, e.g., temperature and CO_2 concentration. To deal with large state and action spaces, a nonlinear function approximator such as a neural network can be used to represent the action-value function. Under this situation, RL is known to be unstable or even divergent.

C. DQN

To overcome the drawback of Q-learning, Mnih *et al.* proposed a novel method named DQN in 2015, which adopts several techniques of stabilizing learning process, e.g., experience replay and target network [24]. To be specific, experience replay mechanism intends to store the experience transitions $(S_t, A_t, S_{t+1}, R_{t+1})$ in a replay memory and draw samples of them uniformly at random for training, which brings greater data efficiency when compared with standard online Q-learning algorithm.

Moreover, randomizing the samples contributes to decreasing their correlations and reducing the variance of weight updates. In addition, target network is adopted to improve the stability of training process by copying a separate network with longer update period for the computation of target value (i.e., $R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$).

III. DRL-based Smart Building Energy Management

In this section, we briefly introduce the main research problems in the field of SBEM and more details will be explained in the next three sections. Then, we classify DRL methods in the field of SBEM by pointing out their respective advantages, disadvantages, representative algorithms, and application scenarios, which contributes to the selection of appropriate DRL methods for SBEM.

A. SBEM

In smart buildings, there are several types of energy systems, e.g., photovoltaic panels (PVs), wind turbines (WTs), diesel generators (DGs), electric energy storage systems, thermal energy storage systems, HVAC systems, lighting systems, blind systems, window systems, electric water heaters (EWHs), EVs, washing machines (WMs), gas boilers (GBs), and clothes dryers (CDs). Since the operations of such systems will affect building performances, it is very necessary to schedule them intelligently. Typical building performance metrics consist of energy cost [15], energy consumption [57], thermal comfort [14], indoor air quality (IAQ) [17], non-uniformity of radiant temperature [58], peak demand [19], consumers' satisfaction degree [59], lighting comfort [60] [61], productivity [62], operating cost [63], the overall levelized energy cost [64], and carbon emission [65]–[67]. Generally speaking, such metrics can be classified into three types, i.e., economy-related (e.g., energy cost or operation cost), environment-related (e.g., carbon emission), society-related (e.g., comfort or satisfaction degree).

Since HVAC systems account for a large portion of energy consumption in buildings, we take them for example to illustrate a typical research problem in SBEM. Similar to our previous work in [17], we intend to minimize the long-term energy cost of an HVAC system in a N-zone commercial building with the consideration of random zone occupancy, thermal comfort, and indoor air quality comfort. Then, the formulated HVAC energy cost minimization problem can be given by

$$
\begin{aligned}
(\mathbf{P1}) \ \min_{m_{i,t},\sigma_t} \ & \sum\nolimits_{t=1}^{L} \mathbb{E}\{C_t(m_{i,t},\sigma_t)\} && \text{(3a)}\\
s.t.\ & T_i^{\min} \le T_{i,t} \le T_i^{\max},\ \forall\ i,t,\ K_{i,t} > 0, && \text{(3b)}\\
& T_{i,t+1} = \mathcal{F}(T_{i,t},\ T_{z,t}|_{\forall z\in\mathcal{N}_i},\ T_t^{\text{out}},\ m_{i,t},\varsigma_{i,t}), && \text{(3c)}\\
& O_{i,t} \le O_i^{\max},\ \forall\ i,t,\ K_{i,t} > 0, && \text{(3d)}\\
& O_{i,t+1} = \mathcal{G}(O_{j,t}|_{\forall j\in\mathcal{N}}, K_{i,t}, m_{i,t}, \sigma_t) && \text{(3e)}\\
& m_{i,t} \in \mathcal{M},\ \forall\ i,t, && \text{(3f)}\\
& \sigma_t \in \Omega,\ \forall\ t. && \text{(3g)}
\end{aligned}
$$

where $\mathbb{E}$ denotes the expectation operator, which acts on random system parameters, e.g., outdoor temperature T_t^{out}, number of occupants $K_{i,t}$. Decision variables of **P1** are air supply rate in each zone $m_{i,t}$ and damper position in air handling unit (AHU) σ_t, L is the considered total number of time slots. $C_t(m_{i,t},\sigma_t)$ is the energy cost at slot t, $T_{i,t}$ and $O_{i,t}$ are indoor air temperature and indoor CO_2 concentration at slot t, respectively. It is obvious that they should be controlled within comfortable ranges, which can be captured by (3b) and (3d), respectively. The dynamics of $T_{i,t}$ and $O_{i,t}$ are represented by (3c) and (3e), respectively. Note that $\varsigma_{i,t}$, $\mathcal{N}_i$, and $\mathcal{N}$ are thermal disturbance, the set of neighbors related to zone i, and the set of zones, respectively. The discrete solution spaces of $m_{i,t}$ and σ_t are shown in (3f) and (3g), respectively. As mentioned in [17], there are several challenges in solving the sequential decision-making problem **P1**. Firstly, it is difficult to obtain a building thermal dynamics model $\mathcal{F}$ that is accurate and efficient enough for HVAC control. Secondly, there are temporally and spatially coupled constraints about indoor temperature and CO_2 concentration. Thirdly, the solution space in each time slot is extremely large. Fourthly, the objective function is often non-convex and non-separable. Lastly, there are some uncertain parameters, e.g., outdoor temperature and number of occupants. When facing with such challenges, existing building energy optimization approaches are not applicable.

Note that the above example is just related to the management of a single building energy subsystem. With the increase of system complexities, more and more challenges have to be addressed, which will be discussed in next four sections.

B. Procedure of Solving SBEM Problems using DRL

In this subsection, we introduce the procedure of solving SBEM problems using DRL. Firstly, we should reformulate the original problem (e.g., **P1**) as a MDP problem or Markov game (i.e., a multi-agent extension of MDP) according to the number of agents. Take **P1** for example, $N+1$ agents are adopted since there are $N+1$ discrete decision variables and the solution space grows rapidly with the increase of zone number. Therefore, we should reformulate the original problem as a Markov game and design its components, e.g., state, action, reward function. Secondly, an appropriate DRL algorithm should be proposed to solve the reformulated problem. For instance, in order to solve the Markov game associated with **P1**, we propose an HVAC control algorithm based on multi-agent DRL with attention mechanism, which can overcome all challenges mentioned in above subsection [17]. Above all, the architecture of the proposed algorithm is scalable to the number of agents. Thirdly, the computational complexities of the designed DRL-based energy management algorithms (including training algorithm and execution algorithm if offline learning is considered) should be analyzed to show algorithm performance. Typical factors that affect the computational complexity of DRL algorithms are summarized as follows [68] [16], e.g., the network architecture, the number of training episodes, the number of weight update, batch size, and the number of testing periods. Finally, evaluations should be conducted to show the performance of the designed algorithm in the aspect of convergence, effectiveness, scalability, or robustness and so on.

C. *Classification of DRL Methods for SBEM*

DRL methods for SBEM can be generally classified into two types as shown in Tables III and V, i.e., model-free DRL methods and model-based DRL methods. In model-free methods, agents intend to learn policies by directly interacting with the unknown environment, which means that no explicit building environment model is needed. However, they need to collect sufficient experience transitions for training, which may result in high exploration cost (e.g., time or financial loss). In contrast, model-based methods intend to construct a model to simulate the environment and use it to generate future episodes for training. Therefore, model-based methods outperform model-free methods in terms of sample complexity. However, for model-based methods, it is often challenging to obtain an accurate and useful model.

Model-free DRL methods can be further divided into value-based methods and policy-based methods. To be specific, the former learns an approximation of optimal value function (i.e., learn a deterministic policy indirectly), while the latter learns an approximation of optimal policy directly. Typically, value-based methods sometimes update value function in an "off-policy" (i.e., the learned policy is different from the behavior policy used for selecting actions [69]) manner, which means that the previous collected experience transitions in the same environment can be used for training. In contrast, "on-policy" means that all of the updates are made using the data from the trajectory distribution induced by the current policy [70]. Therefore, "on-policy" methods are more stable but less data-efficient compared with "off-policy" methods. In existing DRL methods for SBEM, eight representative algorithms are provided, their features and SBEM application scenarios can be identified in Tables III and V.

TABLE III
MODEL-FREE DRL METHODS FOR SBEM

Description	Agents learn optimal policies based on the interaction information with unknown environment			
Advantages	No need to know building environment model			
Disadvantages	Require a large number of samples, high exploration cost (e.g., time or financial loss)			
Categories	Value-based, off-policy	Policy-based, off-policy	Policy-based, on-policy	
Representative Algorithms	DQN	DDPG	PPO	A2C/A3C
Features	Support only for discrete action space	Support only for continuous action space	Stable and robust to hyper-parameter and network architecture	Support stable and fast training
SBEM Application Scenarios	HVAC control [15] [73] [75] [76], MES optimization [19] [62], Microgrid energy management [64] [85] [86] [87]	HVAC control [14], MES optimization [16] [81], Microgrid energy management [89] [90]	MES optimization [18], Microgrid energy management [91]	HVAC control [58], Microgrid energy management [37]

TABLE IV
MODEL-BASED DRL METHODS FOR SBEM

Description	Agents construct a stimulated environment model and use it to generate future episodes for training			
Advantages	High sample efficiency			
Disadvantages	Generating an accurate and useful model is often challenging			
Categories	Model-based, off-policy		Model-based, on-policy	
Representative Algorithms	MuZero	LSTM-DDPG	Differentiable MPC-PPO	BEM-A3C
Features	Learn a network model with accurate planning performance	Use LSTM and historical data to learn transition function and reward function	Pre-train a differentiable MPC policy based on imitation learning, which is both sample-efficient and interpretable	Use EnergyPlus and measured data to simulate building environment
SBEM Application Scenarios	Microgrid energy management [63]	HVAC control [57]	HVAC control [77]	HVAC control [78]

In the next three sections, we will introduce DRL applications in SBEM considering different building system complexities as shown in Fig. 1, i.e., a single building energy subsystem, multiple energy subsystems (MES) in buildings, and building energy systems in microgrid environment.

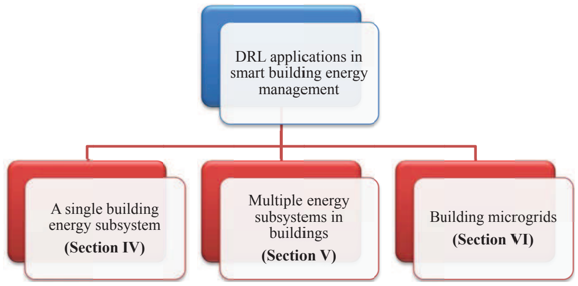

Fig. 1. Taxonomy of DRL Applications in SBEM

IV. Applications of DRL in a Single Building Energy Subsystem

In existing works, DRL techniques have been adopted to optimize the operation cost or energy consumption of a single building energy subsystem, e.g., a hot water heater [71] [72], an HVAC system [15]. Due to the large power consumption of HVAC systems, we mainly focus on them in this section. To be specific, model-free DRL methods and model-based DRL methods for HVAC control are introduced in section IV-A and section IV-B, respectively.

A. Model-free DRL Methods for HVAC Control

Since the operations of HVAC systems place an economic burden on building operators, it is very necessary to minimize energy cost/consumption while maintaining thermal comfort for occupants. In existing works, many DRL-based HVAC control methods have been proposed. For example, Wei *et al.* [15] proposed a DQN-based HVAC control method to save energy cost in office buildings while maintaining the room temperature requirements. In [38], Nagy *et al.* proposed a model-free DRL-based HVAC control method in a residential building to save energy cost and reduce the loss of occupant comfort based on D-DNFQI. In [14], Gao *et al.* presented a DDPG-based HVAC control method to optimize energy consumption and thermal comfort in a laboratory. In [73], Yoon *et al.* developed a DQN-based HVAC control method to optimize energy consumption and thermal comfort in an office building. In [74], Valladares *et al.* proposed a DDQN-based control algorithm to optimize the energy consumption from air-conditioning units and ventilation fans while maintaining thermal comfort and indoor air quality comfort for occupants in a classroom and a laboratory. In [75], Gupta *et al.* proposed a DQN-based heating controller to improve thermal comfort and minimize energy cost in smart buildings. In [76], Sakuma *et al.* proposed a DQN-based airflow direction control algorithm to achieve a uniform comfort within an indoor housing environment. In [58], Morinibu *et al.* proposed a A2C-based HVAC control method to decrease the non-uniformity of radiation temperature in the room. In [36], Nagarathinam *et al.* proposed a multi-agent DRL based algorithm to minimize HVAC energy consumption without sacrificing user comfort by adjusting both the building and chiller set-points. To be specific, each DDQN-based agent coordinate with each other to learn an optimal HVAC control policy. Note that the coordination is achieved by allocating the same reward for each agent. Since a large building may have few hundreds of AHUs and few tens of chillers, it is time-consuming to train all agents centrally. To speed up the training process, transfer learning is adopted, i.e., training a multi-agent on a sub-set of HVAC systems (including one AHU and one chiller) and the learned network weights are used to initialize the multiple agents related to other HVAC subsystems, which can be depicted by Fig. 2.

B. Model-based DRL Methods for HVAC Control

Although the above-mentioned works are effective, there are two major drawbacks in training a DRL agent. Firstly, it is impractical to let the DRL agent to explore the state space fully in a real building environment since unacceptably high cost may be incurred [44] [57] [77]. Secondly, it may take a long time for the DRL agent to learn an optimal policy if trained in a real-world environment [57] [77]. To reduce the dependency on a real building environment, many model-based DRL control methods have been developed [57] [78]. For example, Zhang *et al.* [78] proposed and implemented a building energy model (BEM)-based DRL control framework for a novel radiant heating system in an existing office building. The proposed control framework consists of four steps as shown in Fig. 3, i.e., building energy modeling, model calibration, DRL training and real deployment. To be specific, EnergyPlus is used to develop a building energy model for the office building. Next, based on the observed data, the building energy model can be calibrated. Then, the calibrated building energy model is used as the simulator of environment to train the DRL agent off-line based on A3C algorithm. Finally, the learned optimal control policy will be deployed in the building automation system (BAS) for generating HVAC control signals in real-time. Experimental

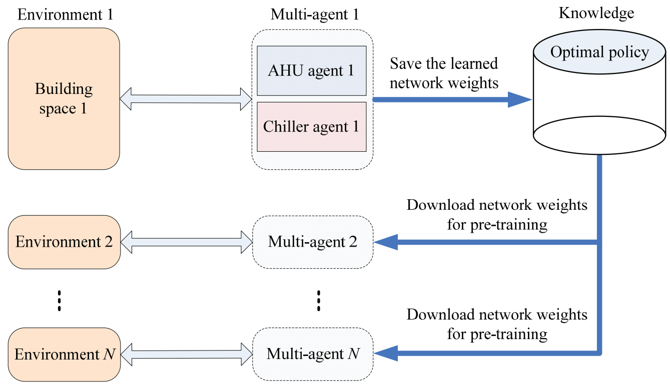

Fig. 2. The proposed transfer learning framework for multi-agent training

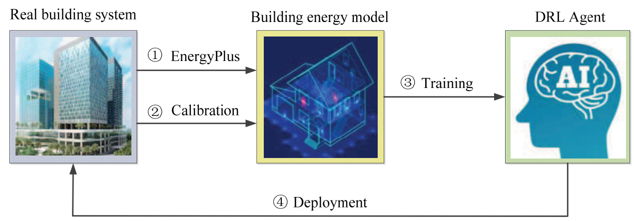

Fig. 3. BEM-based DRL control framework

results show that the obtained DRL-based control strategy can reduce 16.7% heating demand with more than 95% probability compared with the rule-based control strategy.

Since building energy models are calibrated based on the observed data in a short period of time (e.g., three months), the DRL agent's exposure to real-world HVAC operational data is limited [78]. To overcome this drawback, Zou *et al.* [57] proposed a DRL-based HVAC control framework to minimize energy consumption while maintaining thermal comfort levels for occupants based on operational data within two years. The proposed control framework is composed of two parts as shown in Fig. 4, i.e, creating DRL training environment and training DRL agent with the created environment. To be specific, LSTM models are built based on BAS historical data, which can approximate HVAC operations. Note that the inputs of LSTM models are current state and action, while their outputs are next state and reward. After LSTM networks are trained, they can be used to create DRL training environment. Next, DRL agent interacts with the training environment until it converges to an optimal HVAC control policy, which can be deployed for controlling AHUs in real-time. Moreover, DRL agent contains an actor network and a critic network, which are trained using DDPG algorithm. Algorithmic testing results show that DRL agents can save energy by 27% to 30% while maintaining the predicted percentage of discomfort at 10%.

TABLE V
SUMMARY OF EXISTING WORKS ON DRL FOR OPTIMAL HVAC CONTROL

Research work	Object(s)	Primary objective	Secondary objective(s)	DRL algorithm(s)	Performance improvement	Practical implementation
Gao *et al.* [14]	Laboratory	Energy cost	Thermal comfort	DDPG	4.31%∼9.15%	No
Wei *et al.* [15]	Office	Energy cost	Thermal comfort	DQN	19.1%∼ 71.2%	No
Nagarathinam *et al.* [36]	A campus building	Energy consumption	Thermal comfort	DDQN	17%	No
Nagy *et al.* [38]	Residential buildings	Energy cost	Thermal comfort	D-DNFQI	5.5%∼10%	No
Zou *et al.* [57]	Office	Energy consumption	Thermal comfort	LSTM-DDPG	27%∼31.27%	No
Yoon *et al.* [73]	Office	Energy consumption	Thermal comfort	DQN	12.4%∼32.2%	No
Valladares *et al.* [74]	Laboratory and Classroom	Energy cost	Thermal comfort, Air quality	DDQN	4%∼5%	No
Gupta *et al.* [75]	Residential buildings	Energy cost	Thermal comfort	DQN	5%∼12%	No
Sakuma *et al.* [76]	Residential buildings	Energy consumption	Thermal comfort	DQN	34.5%	No
Morinibu *et al.* [58]	Smart home	Non-uniformity of radiant temperature	Thermal comfort	A2C	—	No
Chen *et al.* [77]	Office	Energy cost	Thermal comfort	Differentiable MPC-PPO	16.7%	Yes
Zhang *et al.* [78]	Office	Energy cost	Thermal comfort	BEM-A3C	7.06%∼16.7%	Yes

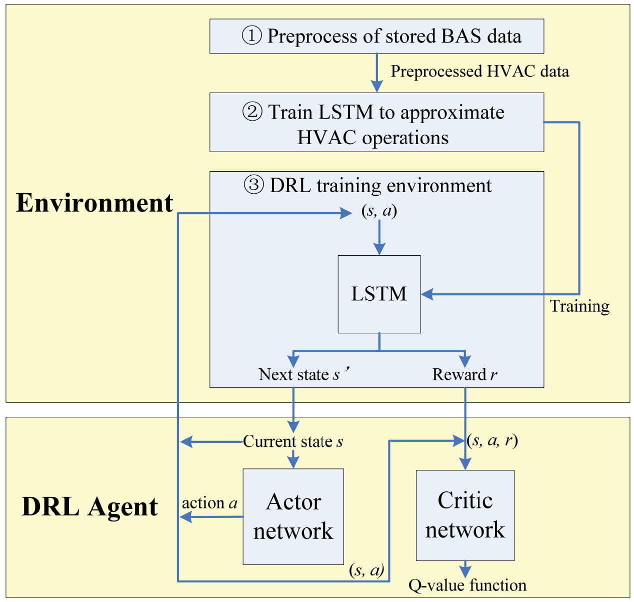

Fig. 4. LSTM-based DRL control framework

In [77], Chen *et al.* proposed a novel approach to enable the practical deployment of DRL for HVAC control and the framework of the approach is shown in Fig. 5. Specifically, historical data from existing HVAC controllers are used to pre-train a differentiable MPC policy based on imitation learning. Note that the pre-trained policy can encode domain knowledge into planning and system dynamics, making it both sample-efficient and interpretable. Next, the pre-trained control policy is improved continually in the process of interacting with the real building environment using online learning algorithm. Since PPO is robust to hyperparameters and network architectures, it is adopted to improve the pre-trained policy. Practical experimental results show that the proposed approach can save 16.7% of cooling demand compared with the existing controller and track temperature set-point better.

In this section, we review existing works on DRL for optimal HVAC control. For easy reading, the specific details including objectives, DRL algorithms, and implementation methods are summarized in Table V. It can be found that most of objectives are related to energy cost and thermal comfort. In addition, a few of DRL-based methods are evaluated by practical implementation. In next section, we will introduce DRL applications in multiple energy subsystems of buildings.

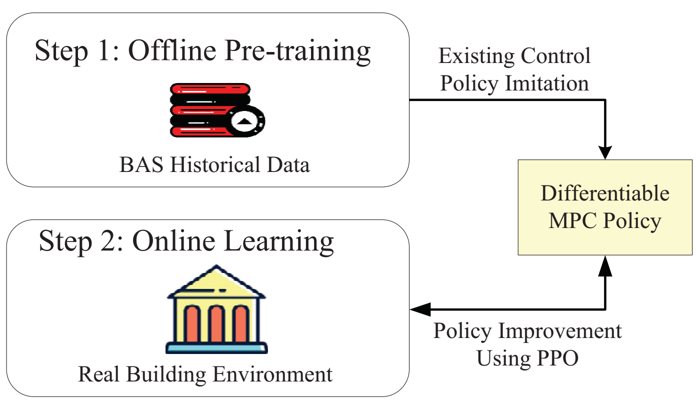

Fig. 5. Differentiable MPC policy based HVAC control framework

V. Applications of DRL in Multiple Energy Subsystems of Buildings

In this section, we will introduce the applications of DRL in multiple energy subsystems of residential buildings and commercial buildings, respectively. To be specific, section V-A focuses on the coordination of home energy management system, HVAC systems, ESSs, EVs, WMs, PVs, and EWHs in residential buildings, while section V-B focuses on the coordination of HVAC systems, lighting systems, blind systems, window systems, and personal electric devices in commercial buildings.

A. Multiple Energy Subsystems in Residential Buildings

There are many existing works on residential building energy optimization with the consideration of multiple energy subsystems. For example, Mocanu *et al.* proposed a DQN-based algorithm to minimize energy cost and peak load of residential buildings with HVAC systems, EVs, and dishwashers (DWs) [19]. Since ESS can be used to save energy cost by exploiting temporal diversity of dynamic prices [79] [80], Yu *et al.* proposed a DDPG-based home energy management algorithm to minimize energy cost for the joint scheduling of HVAC systems and ESSs in [16]. Similarly, Liu *et al.* proposed a DDQN-based home energy management algorithm to minimize energy cost with the consideration of PV systems, ESSs, HVAC systems, EVs, EWHs, and DWs in [59]. To improve the training performance, Ye *et al.* [81] proposed an autonomous control method for a residential multi-energy system based on DDPG with prioritized experience replay (i.e., PDDPG) to reduce energy cost.

Although some advances have been made in above works, their methods can only deal with discrete or continuous action spaces. However, both discrete and continuous actions may appear in practical residential energy management. To support discrete and continuous actions simultaneously, Li *et al.* proposed a TRPO-based approach to jointly optimize the schedules of different types of appliances in a smart home, e.g., HVAC systems, EWHs, EVs, DWs, WMs, CDs, a refrigerator, and a hairdryer [82]. With the increase of home number, the scheduling of all energy subsystems would be more difficult since more coupling constraints and control decisions should be considered. To deal with this challenge, Zhang *et al.* proposed a multi-household energy management method for residential units connected to the same transformer with the consideration of PVs, ESSs and controllable loads based on cooperative multi-agent DRL [18], which can reduce total energy cost and violate the transformer capacity at a low probability.

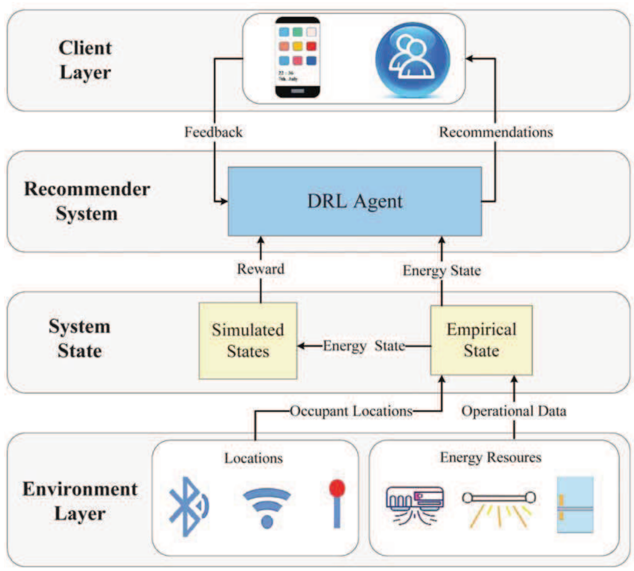

Fig. 6. The architecture of the proposed control framework

B. Multiple Energy Subsystems in Commercial Buildings

In existing works, some DRL-based approaches have been proposed to reduce commercial building energy consumption [57] [74] [78]. Although some advances have been made, these works only consider a single subsystem in buildings (e.g., an HVAC system) without noticing that other subsystems can also affect energy consumption and user comfort in terms of thermal, air quality and illumination conditions. In fact, some research results show that jointly controlling HVAC systems and other building energy subsystems (e.g., blind systems, lighting systems, and window systems) has great potential of saving energy [83] [84]. For example, HVAC energy consumption can be reduced by 17%-47% if window-based natural ventilation is adopted [84]. Based on the above observation, Ding *et al.* proposed a DRL-based framework in Fig. 6 for efficiently controlling four building energy subsystems (including HVAC systems, lighting systems, blind systems, and window systems [60]) so that the total energy consumed by all subsystems can be minimized while still maintaining user comfort. To solve the high-dimensional action problem, a Branching Dueling Q-Network (BDQ) algorithm is used. Moreover, a calibrated EnergyPlus simulation model is adopted to generate enough data for the training of the DRL agent. Simulation results show that the proposed framework can save energy by 14.26% compared with the rule-based method while maintaining human comfort within a desired range.

The above-mentioned works mainly focus on building energy system itself and treat occupants as immovable objects, which may decrease the potential of reducing energy consumption. Therefore, it is very necessary to investigate the potential of saving energy by shaping occupant behavior. To this end, Wei *et al.* [62] designed a DRL-based recommender system in commercial buildings, which can learn actions with high energy saving potential and distribute recommendations to occupants. Based on

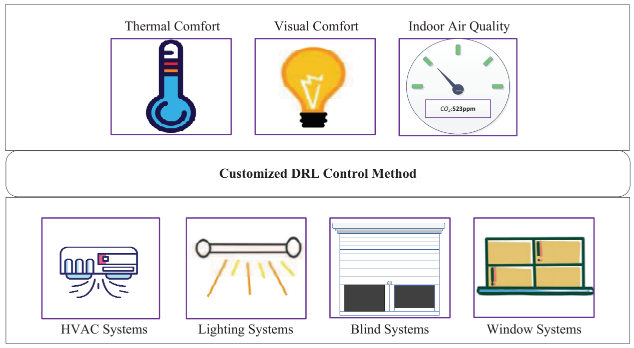

Fig. 7. The system architecture of the designed recommender

the feedback from occupants, better recommendations can be learned. The system architecture of the designed recommender is shown in Fig. 7, which consists of four layers, i.e., *environment layer*, *system state layer*, *recommender system*, and *client layer*. To be specific, environment layer measures building environment (e.g., occupant locations, energy consumption information) and sends such information to system state layer. System state layer contains two components, i.e., an empirical state, which maintains the current building state, and simulated states, which are used to represent the next state after the potential energy saving actions are taken. The recommender system layer learns the potential of different recommendation actions (including *move recommendation*, *schedule change*, *reduce personal resources*, and *reduce service in spaces*). The client layer receives recommendations and allows clients to provide feedback (e.g., accept or reject the recommendation). A four week user study shows that the designed recommender system can reduce building energy consumption by 19% to 26% when compared with a passive-only strategy.

TABLE VI
SUMMARY OF EXISTING WORKS ON DRL FOR MULTI-ENERGY SUBSYSTEMS IN BUILDINGS

Research work	Object(s)	Energy subsystems	Primary objective	Secondary objective(s)	DRL algorithm(s)	Performance improvement	Practical implementation
Yu *et al.* [16]	Smart home	PV, ESS, HVAC	Energy cost	Thermal comfort	DDPG	8.10%∼15.21%	No
Zhang *et al.* [18]	Residential buildings	PV, ESS, EV	Energy cost	Transformer capacity violation	PPO	59.77%	No
Mocanu *et al.* [19]	Residential buildings	PV, HVAC, EV, DW	Energy cost	Peak demand, load operational time or condition	DQN	14.1%∼27.4%	No
Liu *et al.* [59]	Smart home	PV, ESS, HVAC, EV, Heater, DW	Energy cost	Consumers' satisfaction degree	DDQN	41.8%∼59%	No
Ding *et al.* [60]	Commercial buildings	HVAC, lighting, blind and window	Energy consumption	Thermal comfort, IAQ, lighting comfort	BDQ	14.26%	No
Li *et al.* [82]	Smart home	HVAC, EV, EWH, DW, WM	Energy cost	Thermal comfort and range anxiety	TRPO	31.6%	No
Ye *et al.* [81]	Residential buildings	PV, ESS, TES, EHP, GB	Energy cost	Excess energy sale revenue	PDDPG	6.28%∼10.21%	No
Wei *et al.* [62]	Commercial buildings	HVAC, lighting, plug load	Energy consumption	Safety, comfort, productivity	DQN	19%∼26%	Yes

In this section, we review existing works on DRL applications in multiple energy subsystems of buildings. For easy understanding, the research objects, considered energy subsystems, research objectives, DRL algorithms, performance improvement, and implementation methods in existing works are summarized in Table VI. It can be observed that there is a great potential in reducing energy cost of buildings by scheduling multiple energy subsystems coordinately, e.g., relative energy cost reduction is up to 59% while maintaining satisfaction degree of occupants. Compared with the optimal HVAC control in Table V, more advanced DRL algorithms are adopted to deal with more complex problems, e.g., PDDPG, BDQ, and TRPO. In addition, most of DRL methods are evaluated by simulations. In next section, we will introduce the DRL applications in building microgrids.

VI. APPLICATIONS OF DRL IN BUILDING MICROGRIDS

In this section, we review the existing works on DRL-based energy optimization for building microgrids. To be specific, section V-A introduces DRL-based energy management algorithms for microgrids with uncontrollable building loads, while section VI-B introduces DRL-based microgrid optimization algorithms considering the flexibility of building loads.

A. Microgrid Optimization without Considering Controllable Building Loads

In existing works, many DRL-based methods have been proposed for residential microgrids [63], [64], [85]–[87], where a microgrid is a low voltage distribution network comprising various distributed generation, storage devices, and responsive loads [88]. For example, Francois-Lavet *et al.* proposed a DQN-based energy management algorithm for a residential microgrid with the consideration of battery and hydrogen storage device to minimize the levelized energy cost, which is an economic assessment of the cost that covers all the expenses over the lifetime of the microgrid [64]. In [85], Dominguez-Barbero *et al.* proposed a DQN-based energy management algorithm for an isolated residential microgrid to minimize the operating cost, which is the sum of DG generation cost and the penalty of non-served power demand. In [86], Chen *et al.* proposed a DQN-based energy trading strategy for a microgrid to maximize the utility function, which is related to trading profit, retail profit, battery wear cost, demand penalty, and virtual penalty. In [87], Ji *et al.* proposed a DQN-based energy management algorithm for a microgrid to minimize daily operating cost.

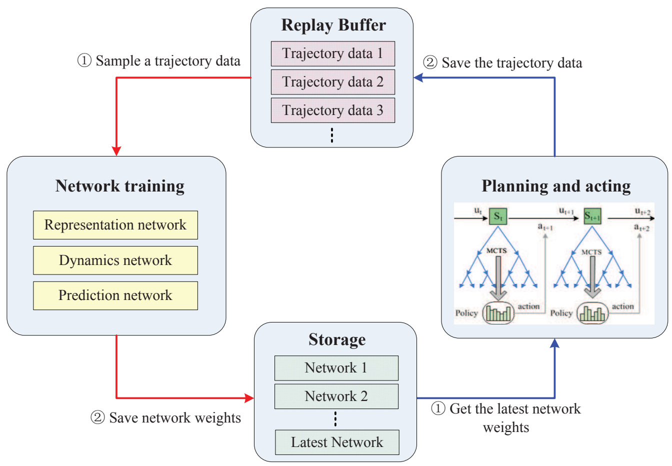

Fig. 8. The training process of the network model

In the above research efforts, the proposed DQN-based methods can not deal with DRL problems with continuous actions, e.g., the generation output of diesel generators [89]. To support continuous actions, DDPG-based methods could be adopted. For example, Lei *et al.* proposed a FH-DDPG based energy management algorithm for an isolated microgrid to minimize the sum of power generation cost and the power unbalance penalty [89]. Since model-free based DRL algorithms in existing works have low sample efficiency, Shuai *et al.* proposed a model-based DRL algorithm (i.e., *MuZero*) for the online scheduling of a residential microgrid under uncertainties [63] based on Monte-Carlo tree search (MCTS) strategy with a learned network model. Note that the off-line learning process of the network model can be depicted by Fig. 8, where four components can be identified, i.e., *network training*, *replay buffer*, *storage*, *planning and acting*. First, the latest network weights are obtained from a storage and used for planning implemented by MCTS. Next, an action is sampled from the search policy, which is proportional to the visit count for each action from the root node. Then, the environment returns a new state and reward. At the end of the episode, the trajectory data is stored into a replay buffer. When conducting network training, a trajectory data will be randomly sampled from the replay buffer and the updated network weights will be saved in a storage device. It is obvious that network training and trajectory data generation are independent process, which can be implemented in parallel. Once the training process of the network model (including three components, i.e., representation, dynamics, and prediction) is completed, MCTS with the learned network model can be used to obtain the optimal online decision at each time step. Note that the proposed algorithm can operate without relying on any forecast information and statistic distribution information of the system.

B. Microgrid Optimization with Controllable Building Loads

Although some advances have been made in above-mentioned works, they take building loads as uncontrollable resources. In fact, the energy cost of a microgrid could be reduced by scheduling loads flexibly. For example, Yang *et al.* proposed a DDPG-based scheduling algorithm for a data center microgrid with renewable sources to reduce energy cost by choosing the execution time and the quantity of served workloads flexibly [90]. In [37], Yang *et al.* investigated a large-scale smart home energy management problem in a residential microgrid and proposed an entropy-based collective multiagent reinforcement learning framework to learn home control strategies for scheduling EVs and ESSs. Simulation results based on real-world traces show the effectiveness of the proposed framework in reducing the operating cost and the peak load. In [91], Lee *et al.* proposed a MAPPO-based algorithm to solve the demand response problem in a microgrid of residential district. The proposed algorithm intends to train multiple household agents centrally. Once an optimal policy is learned by each household agent,

TABLE VII
SUMMARY OF EXISTING WORKS ON DRL FOR MICROGRIDS

Research work	Microgrid type	Energy systems	Controllable building load considered	Optimization objective(s)	DRL algorithm(s)	Cost reduction	Practical implementation
Lei *et al.* [89]	An isolated microgrid	PV, DG, ESS	No	Power generation cost, power unbalance	FH-DDPG, FH-RDPG	80%	No
Shuai *et al.* [63]	A residential microgrid	PV, WT, ESS	No	Operating cost	MuZero	9.28%∼28.93%	No
Francois-Lavet *et al.* [64]	A residential microgrid	PV, Battery, hydrogen storage device	No	The overall levelized energy cost	DQN	5%∼12%	No
Dominguez-Barbero *et al.* [85]	An isolated residential microgrid	PV, DG, Battery, hydrogen storage device	No	Operating cost	DQN	58.5%∼67.20%	No
Chen *et al.* [86]	A residential microgrid	PV, ESS	No	Profit minus cost	DQN	>30%	No
Ji *et al.* [87]	A residential microgrid	PV, WT, DG, ESS	No	Daily operating cost	DQN	20.75%	No
Yang *et al.* [37]	A residential microgrid	PV, ESS, EV	Yes, EV	Energy cost and peak load	EB-C-A2C, EB-C-DQN	24.69%	No
Lee *et al.* [91]	A residential microgrid	WM, CD, WH, DW and refrigerator.	Yes, household appliances	Energy cost and peak load	MAPPO	—	No
Yang *et al.* [90]	A data center microgrid	PV, ESS, servers	Yes, servers	Energy cost	DDPG	6.24%	No

it can schedule household appliances under real-time pricing environment for reducing energy cost without knowing specific information about other households.

In this section, we review existing works on DRL applications in building microgrids. For easy understanding, we summarize the details of existing works in Table VII. It can be observed that the proposed DRL-based methods can bring economic benefits for microgrid operators. However, most of them neglect the control of building loads and all of them are not implemented in practice.

VII. FUTURE RESEARCH DIRECTIONS

In this section, we describe some unsolved issues related to DRL for SBEM and present the possible directions for future research.

A. Multi-time scale building energy optimization

Most existing DRL-based methods focus on single-time scale building energy optimization problems. In fact, there are many multi-time scale decision-making problems in the field of building energy optimization. For example, supply air temperature and the ratio of re-use air in a commercial building HVAC system can be adjusted once every hour since the frequent adjustment can cause damage to HVAC components [92]–[94]. In contrast, supply air rate in each zone can be changed every 10-15 minutes [94]. When confronted with multi-time scale decision-making problems, existing DRL-based methods are not applicable. A possible way is to design energy optimization algorithms based on the framework of hierarchical DRL [95] [96], which can support multi-time scale DRL problems with delayed rewards. In hierarchical DRL, actions can be divided into two types with different time scales. To be specific, actions with slow time scale are first taken in the upper level based on system state. Then, actions with fast time scale are taken in the lower level based on system state and the chosen actions in the upper level. By coordinating the actions of upper level and lower level, hierarchical DRL-based methods can explore the environments efficiently.

B. Multi-objective building energy optimization

As shown in Section IV, multiple objectives are pursued by SBEM, e.g., energy cost/consumption minimization, carbon emission minimization, and comfort maximization. Moreover, such objectives are often conflicting with each other [17]. A typical way of dealing with conflicting objectives in existing DRL-based methods is to design a synthetic reward function as a weighted sum of different objectives [16] [17]. Since the weight parameters related to different objectives typically have different units and/or scales, it is very challenging to decide their proper values beforehand. Moreover, learned policies based on the above-mentioned way can not support flexible operation of building energy systems, e.g., switching flexibly between low energy cost mode and high comfort mode. To avoid deciding weighted parameters for multiple objectives and support flexible operations, a possible way is to design building energy optimization algorithms based on the framework of multi-objective DRL [97] [98] or multi-objective meta-DRL [99].

C. Multi-zone building energy optimization

In existing works on building HVAC systems, the proposed DRL-based control methods mainly focus on a single-zone building. In [15], Wei *et al.* proposed a heuristic algorithm for variable air volume (VAV) HVAC control in a multi-zone office building and DRL agent for each zone is trained separately. Although the proposed algorithm is effective when 5 zones are considered, it is not scalable due to the lack of multi-zone coordination. In other words, the proposed algorithm may diverge or show degraded performance when the number of zones is very large. In [34], Hu *et al.* proposed a MADDPG-based method to decide temperature and humidity setpoints in a four-zone building. Since the input of each critic in MADDPG is the concatenation of state and action information from all agents, the scalability of the MADDPG-based method is not very high. To overcome the drawback of MADDPG, Iqbal *et al.* proposed a MAAC method in [100] by adopting attention mechanism. To be specific, MAAC can learn the critic for each agent by selectively paying attention to the information from other agents. Therefore, MAAC is more scalable than MADDPG. In [17], Yu *et al.* proposed a MAAC-based VAV HVAC control method for a multi-zone commercial building with the consideration of thermal comfort, indoor air quality comfort and random occupancy. Extensive simulation results show that the proposed method is still effective when 30 zones are considered. As mentioned in [17], a larger capacity of a memory replay buffer is required with the increase of zone number. Therefore, it is very challenging to effectively learn an optimal policy given a large buffer capacity. A possible way is to adopt prioritized experience replay [101], which can learn experience more efficiently by replaying important transitions more frequently. In addition, more scalable DRL-based algorithms should be designed since the number of zones in a practical commercial building may exceed one hundred or even one thousand.

D. Efficient training of DRL agents in multi-building energy optimization

As introduced in Section III, model-based DRL methods for building energy optimization are sample-efficient. For example, authors in [57] and [77] developed state transition prediction model and reward prediction model based on historical data within several years, which can provide enough training data for the DRL agent. However, a large amount of historical data may be unavailable for some buildings (e.g., brand-new buildings). At this time, a building energy model has to be developed by collecting and exploiting a limited amount of actual operational data [78], which will affect the training performance of the DRL agent. To improve this situation, a possible way is to combine DRL with transfer learning. To be specific, the key idea of transfer learning is to apply the knowledge from one task to a related but different task, which contributes to the reduction of training time [102]. In [103], Zhang *et al.* proposed a transfer learning based scheme for thermal dynamics modeling in smart buildings, which can solve the problem of generalizing an established model from one building with a large amount of historical data to another building with a limited amount of data. In addition, control strategies can be also transferred from source buildings to target buildings with different zones, materials, layouts, and weather conditions. For instance, Xu *et al.* proposed a DQN-based transfer learning approach for multiple buildings, which can effectively reduce the training time, energy cost, and temperature violations [104]. In summary, combining DRL with transfer learning is beneficial to the fast training of DRL agents in multiple buildings.

E. DRL-based energy optimization for building microgrids

Most existing works focus on the case that building loads are uncontrollable, which means that the advantage of demand side management can not be fully utilized, e.g., reducing peak load or energy cost [105]. In building loads, HVAC systems have high and flexible power consumption. Under some operational constraints (e.g., comfortable indoor temperature range and comfortable indoor air quality), HVAC systems can be scheduled flexibly to save energy cost as a response to dynamic prices [106], [107], which can also offer many benefits for microgrids. For example, incorporating building HVAC control in microgrid scheduling [88] and planning [108] are beneficial to reduce operation cost and total annualized cost (including investment cost and operation cost), respectively. Therefore, it is worthwhile to design DRL-based energy management algorithms for building microgrids with the consideration of load flexibility. Since there are both discrete and continuous variables in energy optimization problem of building microgrids with HVAC loads, the designed DRL-based algorithms should support different kinds of actions as in [82]. In addition, the designed DRL-based algorithms should be scalable since the number of HVAC systems in residential building microgrids or the number of zones served by an HVAC system in commercial building microgrids is large. Last but not least, the designed algorithms should be capable of solving multi-time scale optimization problem (e.g., hydrogen storage devices and battery operate in microgrids have different time scales [4]).

VIII. Conclusions and Lessons Learned

This paper reviewed the applications of deep reinforcement learning in smart building energy management with the consideration of different system complexities comprehensively. Firstly, we provided an overview of deep reinforcement learning. Then, we introduced the background knowledge of SBEM and classified the DRL methods for SBEM. Next, we introduced the existing works in the aspect of a single energy subsystem, multiple energy subsystems in buildings, and building microgrids. Finally, we identified unsolved issues and pointed out the future research directions.

Few major lessons that we learned from this review are summarized as follows. Firstly, most of DRL-based building energy optimization methods are still not implemented in practice. The main reason is that model-free DRL approaches for building energy optimization require a large number of interactions between DRL agents and environments, which is time-consuming and costly. Secondly, model-based DRL approaches for building energy optimization are more practical than model-free DRL approaches since the former can generate enough training data for DRL agents and reduce the dependency on real environment. When the amount of historical data is not enough in the current environment, transfer learning can be used to pre-train building thermal dynamics model or policy model based on the large amount of historical data in a related, but different building environment. Thirdly, compared with some traditional methods, DRL-based energy management methods have the potential of improving some building performance metrics (e.g., energy cost, peak load, and occupant dissatisfaction degree) simultaneously. However, there are still many challenges caused by multiple time scales, multiple optimization objectives, multiple zones, multiple buildings, and multiple building microgrids. Therefore, research on DRL for smart building energy management is still in its infancy and remains to develop.

REFERENCES

[1] S. Hu, C. Hore, P. Raftery, and J. O'Donnell, Environmental and energy performance assessment of buildings using scenario modelling and fuzzy analytic network process, *Applied Energy*, vol. 255, pp. 113788-113799, 2019.

[2] S. Hu, E. Corry, M. Horrigan, C. Hore, M. Dos Reis, and J. O'Donnell, Building performance evaluation using OpenMath and Linked Data, *Energy and Buildings*, vol. 174, pp. 484-494, 2018.

[3] The Global Alliance for Buildings and Construction (GABC), The Global Status Report 2019, https://www.unenvironment.org/resources/publication/2019-global-status-report-buildings-and -construction-sector

[4] S. Sharma, Y. Xu, A. Verma, and B. Panigrahi, "Time-coordinated multi-energy management of smart buildings under uncertainties," *IEEE Trans. Industrial Informatics*, vol. 15, no. 8, pp. 4788-4798, 2019.

[5] B. Qolomany, A. Al-Fuqaha, A. Gupta, D. Benhaddou, S. Alwajidi, J. Qadir, and A. Fong, "Leveraging machine learning and big data for smart buildings: a comprehensive survey," *IEEE Access*, vol. 7, pp. 90316-90356, 2019.

[6] D. Minoli, K. Sohraby, and B. Occhiogrosso, "IoT considerations, requirements, and architectures for smart buildings-energy optimization and next-generation building management systems," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 269-283, Feb. 2017.

[7] X. Zhang, M. Pipattanasomporn, T. Chen, and S. Rahman, "An IoT-based thermal model learning framework for smart buildings," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 269-283, Feb. 2017.

[8] W. Feng, Z. Wei, G. Sun, Y. Zhou, H. Zang, and S. Chen, "A conditional value-at-risk-based dispatch approach for the energy management of smart buildings with HVAC systems," *Electric Power Systems Research*, vol. 188, pp. 196535-106534, 2020.

[9] B. Yang, X. Li, Y. Hou, A. Meier, X. Cheng, J. Choi, F. Wang, H. Wang, A. Wagner, D. Yan, A. Li, T. Olofsson, and H. Li, "Non-invasive (non-contact) measurements of human thermal physiology signals and thermal comfort/discomfort poses-A review," *Energy and Buildings*, vol. 224, pp. 110261-110270, 2020.

[10] F. Wang, L. Zhou, H. Ren, X. Liu, S. Talari, and M. Shafie-khah, "Multi-objective optimization model of source-load-storage synergetic dispatch for a building energy management system based on TOU price demand response," *IEEE Transactions on Industry Applications*, vol. 54, no. 2, pp. 1017-1028, 2018.

[11] A. Pallante, L. Adacher, M. Botticelli, S. Pizzuti, G. Comodi, and A. Monteriu, "Decision support methodologies and day-ahead optimization for smart building energy management in a dynamic pricing scenario," *Energy and Buildings*, vol. 216, pp. 109963-109973, 2020.

[12] A. Ahmad and J. Khan, "Real-time load scheduling, energy storage control and comfort management for grid-connected solar integrated smart buildings," *Applied Energy*, vol. 259, pp. 114208-114226, 2020.

[13] R. Zhang, T. Jiang, G. Li, X. Li, and H. Chen, "Stochastic optimal energy management and pricing for load serving entity with aggregated TCLs of smart buildings: A stackelberg game approach," *IEEE Transactions on industrial informatics*, DOI: 10.1109/TII.2020.2993112, 2020.

[14] G. Gao, J. Li, and Y. Wen, "DeepComfort: Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning," *IEEE Internet of Things Journal*, DOI: 10.1109/JIOT.2020.2992117, 2020.

[15] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building HVAC control," *Proc. of DAC*, 2017.

[16] L. Yu, W. Xie, D. Xie, Y. Zou, D. Zhang, Z. Sun, L. Zhang, Y. Zhang, and T. Jiang, "Deep reinforcement learning for smart home energy management," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2751-2762, 2020.

[17] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, and X. Guan, "Multi-agent deep reinforcement learning for HVAC control in commercial buildings," *IEEE Transactions on Smart Grid*, DOI: 10.1109/TSG.2020.3011739, 2020.

[18] C. Zhang, S. Kuppannagari, C. Xiong, R. Kannan, and V. Prasanna, "A cooperative multi-agent deep reinforcement learning framework for real-time residential load scheduling," *Internet of Things Design and Implementation*, 2019.

[19] E. Mocanu, D. Mocanu, P. Nguyen, A. Liotta, M. Webber, M. Gibescu, and J. Slootweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698-3708, 2019.

[20] Z. Zhang, D. Zhang, and R. Qiu, "Deep reinforcement learning for power system: an overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213-225, 2020.

[21] Y. Ma, J. Matuško, and F. Borrelli, "Stochastic model predictive control for building HVAC systems: complexity and conservatism," *IEEE Trans. Control Systems Technology*, vol. 23, no. 1, pp. 101-116, 2015.

[22] X. Guan, Z. Xu, and Q. Jia, "Energy-efficient buildings facilitated by microgrid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 243-252, Dec. 2010.

[23] L. Yu, D. Xie, T. Jiang, Y. Zou, and K. Wang, "Distributed real-time HVAC control for cost-efficient commercial buildings under smart grid environment," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 44-55, 2018.

[24] V. Mnih, et. al. "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-541, 2015.

[25] Y. Li, "Deep reinforcement learning: an overview," http://arXiv:1701.07274v5, 2017.

[26] R. Lu, Y. Li, Y. Li, J. Jiang and Y. Ding, "Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management," *Applied Energy*, vol. 276, pp. 115473-115483, 2020.

[27] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, et al., "Mastering Atari, Go, Chess and Shogi by planning with a learned model," http://arXiv:1911.08265v2, 2020.

[28] J. Chen, B. Yuan and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," https://arxiv.org/abs/1904.09503v2, 2019.

[29] L. Lei, Y. Tan, K. Zheng, S. Liu, K. Zhang, X. Shen, "Deep reinforcement learning for autonomous internet of things: model, applications and challenges," *IEEE Communications Surveys & Tutorials*, DOI:10.1109/COMST.2020.2988367, 2019.

[30] S. Lee and D. Choi, "Energy management of smart home with home appliances, energy storage system and electric vehicle: a hierarchical deep reinforcement learning approach," *Sensors*, vol. 20, pp. 2157-2178, 2019.

[31] C. Lork, W. Li, Y. Qin, Y. Zhou, C. Yuen, W. Tushar and T.K. Saha, "An uncertainty-aware deep reinforcement learning framework for residential air conditioning energy management," *Applied Energy*, Vol. 276, pp. 115426-115438, 2020.

[32] X. Zhang, D. Biagioni, M. Cai, P. Graf, and S. Rahman, "An edge-cloud integrated solution for buildings demand response using reinforcement learning," *IEEE Trans. Smart Grid*, DOI: 10.1109/TII.2020.3007167, 2020.

[33] R.S. Sutton and A.G. Barto, "Reinforcement learning: an introduction," The MIT Press, London, England, 2018.

[34] W. Hu, "Transforming thermal comfort model and control in the tropics: a machine-learning approach," Nanyang Technological University, 2020.

[35] Y. Hu, W. Li, K. Xu, T. Zahid, F. Qin, and C. Li, "Energy management strategy for a hybrid electric vehicle based on deep reinforcemant learning," *Applied Sciences*, vol. 8, no. 187, pp. 1-15, 2018.

[36] S. Nagarathinam, V. Menon, A. Vasan, and A. Sivasubramaniam, "MARCO-Multi-agent reinforcement learning based control of building HVAC systems," *The Eleventh ACM International Conference on Future Energy Systems (e-Energy'20)*, 2020.

[37] Y. Yang, J. Hao, Y. Zheng, and C. Yu, "Large-scale home energy management using entropy-based collective multiagent deep reinforcement learning framework," *Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[38] A. Nagy, H. Kazmi, F. Cheaib, and J. Driesen, "Deep reinforcement learning for optimal control of space heating," https://arxiv.org/abs/1805.03777, 2018.

[39] N. Luong, D. Hoang, S. Gong, D. Niyato, P. Wang, Y. Liang, and D. Kim, "Applications of deep reinforcement learning in communications and networking: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133-3174, 2019.

[40] T. Nguyen and V. Reddi, "Deep reinforcement learning for cyber security," http://arXiv:1906.05799v2, 2020.

[41] T. Nguyen, N. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multi-agent systems: a review of challenges, solutions and applications," http://arXiv:1812.11794v2, 2019.

[42] M. Han, R. May, X. Zhang, X. Wang, S. Pan, D. Yan, Y. Jin, and L. Xu, "A review of reinforcement learning methodologies for controlling occupant comfort in buildings" *Sustainable Cities and Society*, vol. 51, pp. 101748-101762, 2019.

[43] J. Leitão, P. Gil, B. Ribeiro, and A. Cardoso, "A survey on home energy management," *IEEE Access*, vol. 8, pp. 5699-5722, 2020.

[44] K. Mason and S. Grijalva, "A review of reinforcement learning for autonomous building energy management," *Computers & Electrical Engineering*, vol. 78, pp. 300-312, 2019.

[45] Z. Wang and T. Hong, "Reinforcement learning for building controls: the opportunities and challenges," *Applied Energy*, vol. 269, pp. 115036-115056, 2020.

[46] B. Rajasekhar, W. Tushar, C. Lork, Y. Zhou, C. Yuen, N. Pindoriya, and K. Wood, "A survey of computational intelligence techniques for air-conditioners energy management," *IEEE Trans. Emerging Topics in Computational Intelligence*, DOI:10.1109/TETCI.2020.2991728, 2020.

[47] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362-370, 2018.

[48] T. Yang, L. Zhao, W. Li, and A. Zomaya, "Reinforcement learning in sustainable energy and electric systems: a survey," *Annual Reviews in Control*, vol. 49, pp. 145-163, 2020.

[49] Z. Wang, T. Hong, and M. Piette, "Data fusion in predicting internal heat gains for office buildings through a deep learning approach," *Applied Energy*, vol. 240, pp. 386-398, 2019.

[50] N. Raman, A. Devraj, P. Barooah, and S. Meyn, "Reinforcement learning for control of building HVAC systems," *Proc. of American Control Conference (ACC)*, 2020.

[51] X. Deng, Y. Jiang, L. Yang, L. Yi, J. Chen, Y. Liu, and X. Li, "Learning automata based confident information coverage barriers for smart ocean internet of things," *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2020.2989696.

[52] J. Vázquez-Canteli, and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied Energy*, vol. 235, pp. 1072-1089, 2019.

[53] S. Kim and H. Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, pp. 1-19, 2018.

[54] M. Ahrarinouri, M. Rastegar, and A. Seif, "Multiagent reinforcement learning for energy management in residential buildings," *IEEE Trans. Industrial Informatics*, vol. 17, no. 1, pp. 659-666, 2021.

[55] R. Lu, S. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Trans. Smart Grid*, DOI: 10.1109/TSG.2019.2909266, 2019.

[56] F. Ruelens, B. Claessens, S. Vandael, B. Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2149-2159, Sept. 2017.

[57] Z. Zou, X. Yu, and S. Ergan, "Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network," *Building and Environment*, https://doi.org/10.1016/j.buildenv.2019.106535, 2020.

[58] T. Morinibu, T. Noda, and S. Tanaka, "Application of deep reinforcement learning in residential preconditioning for radiation temperature," *International Congress on Advanced Applied Informatics*, 2019.

[59] Y. Liu, D. Zhang, and H. Gooi, "Optimization strategy based on deep reinforcement learning for home energy management," CSEE Journal of Power and Energy Systems, DOI:10.17775/CSEEJPES.2019.02890, 2019.

[60] X. Ding, W. Du, and A. Cerpa, "OCTOPUS: deep reinforcement learning for holistic smart building control", *Proc. of BuildSys'19*, 2019.

[61] J. Park, T. Dougherty, H. Fritz, and Z. Nagy, "Multi-agent deep reinforcement learning for zero energy communities," *Building and Environment*, vol. 147, pp. 397-414, 2019.

[62] P. Wei, S. Xia, R. Chen, J. Qian, C. Li, and X. Jiang, "A deep reinforcement learning based recommender system for occupant-driven energy optimization in commercial buildings," *IEEE Internet of Things Journal*, DOI:10.1109/JIOT.2020.2974848, 2020.

[63] H. Shuai, H. He, and J.Wen, "Online scheduling of a residential microgrid via monte-carlo tree search and a learned model," http://arXiv:2005.06161v2, 2020.

[64] V. Francois-Lavet, D. Taralla, D. Ernst, and R. Fonteneau, "Deep reinforcement learning solutions for energy microgrids management," *European Workshop on Reinforcement Learning*, 2016.

[65] X. Dong, Y. Liu, Z. Xu, J. Wu, J. Liu, and X. Guan, "Optimal scheduling of distributed hydrogen-based multi-energy systems for building energy cost and carbon emission reduction," *Proc. of IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 2020.

[66] E. Haghi, Q. Kong, M. Fowler, K. Raahemifar, and M. Qadrdan, "Assessing the potential of surplus clean power in reducing GHG emissions in the building sector using game theory; a case study of Ontario, Canada," *IET Energy Systems Integration*, vol. 1, no. 3, pp. 184-193, 2019.

[67] N. Zhou, N. Khanna, W. Feng, J. Ke, and M. Levine, "Scenarios of energy efficiency and CO_2 emissions reduction potential in the buildings sector in China to year 2050," *Nature Energy*, vol. 3. no. 978-984, 2018.

[68] H. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, "Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids," *IEEE Trans. Industrial Informatics*, DOI: 10.1109/TII.2020.3007167, 2020.

[69] K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao, "A survey of deep reinforcement learning in video games," https://arxiv.org/pdf/1912.10944.pdf, 2019.

[70] S. Gu, T. Lillicrap, Z. Ghahramani, R.E. Turner, B. Schölkopf, and S. Levine, "Interpolated policy gradient: merging on-policy and off-policy gradient estimation for deep reinforcement learning," *Prof. of NIPS*, 2017.

[71] F. Ruelens, B.J. Claessens, P. Vrancx, F. Spiessens, and G. Deconinck, "Direct load control of thermostatically controlled loads based on sparse observations using deep reinforcement learning," *CSEE Journal Of Power and Energy Systems*, vol. 5, no. 4, 423-432, 2019.

[72] H. Kazmi, F. Mehmood, S. Lodeweyckx, and J. Driesen, "Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems," *Energy*, vol. 144, pp. 159-168, 2018.

[73] Y. Yoon and H. Moon, "Performance based thermal comfort control (PTCC) using deep reinforcement learning for space cooling," *Energy and Buildings*, vol. 203, pp. 109420-109430, 2019.

[74] W. Valladares, M. Galindo, J. Gutiérrez, W. Wu, K. Liao, J. Liao, K. Lu, and K. Wang, "Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm," *Building and Environment*, vol. 155, pp. 105-117, 2019.

[75] A. Gupta, Y. Badr, A. Negahban, and R.G. Qiu, "Energy-efficient heating control for smart buildings with deep reinforcement learning," *Journal of Building Engineering*, https://doi.org/10.1016/j.jobe.2020.101739, 2020.

[76] Y. Sakuma and H. Nishi, "Airflow direction control of air conditioners using deep reinforcement learning," *Proc. of 2020 SICE International Symposium on Control Systems*, Tokushima, Japan, pp. 61-68, 2020.

[77] B. Chen, Z. Cai, and M. Berges, "Gnu-RL: A precocial reinforcement learning solution for building HVAC control using a differentiable MPC policy," *Proc. of BuildSys*, 2019.

[78] Z. Zhang, A. Chong, Y. Pan, C. Zhang, and K. Lam, "Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning", *Energy and Buildings*, vol. 199, pp. 472-490, 2019.

[79] Z. Wan, H. Li, and H. He, "Residential energy management with deep reinforcement learning," *International Joint Conference on Neural Networks (IJCNN)*, 2018.

[80] H. Kumar, P. Mammen, and K. Ramamritham, "Explainable AI reinforcement learning agents for residential cost savings," http://arXiv:1910.08719v2, 2019.

[81] 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. on Smart Grid*, DOI:10.1109/TSG.2020.2976771, 2020.

[82] H. Li, Z. Wan, and H. He, "Real-time residential demand response," *IEEE Trans. Smart Grid*, DOI: 10.1109/TSG.2020.2978061, 2020.

[83] Z. Cheng, Q. Zhao, F. Wang, Y. Jiang, L. Xia, and J. Ding, "Satisfaction based Q-learning for integrated lighting and blind control," *Energy and Buildings*, vol. 127, pp. 43-55, 2016.

[84] L. Wang and S. Greenberg, "Window operation and impacts on building energy consumption," *Energy and Buildings*, vol. 92, pp. 313C321, 2015.

[85] D. Dominguez-Barbero, J. Garcia-Gonzalez, M.A. Sanz-Bobi, and E.F. Sanchez-ubeda, "Optimising a microgrid system by deep reinforcement learning techniques," *Energies*, vol. 13, pp. 2830-2847, 2020.

[86] T. Chen, and S. Bu, "Realistic peer-to-peer energy trading model for microgrids using deep reinforcement learning," *IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, 2019.

[87] Y. Ji, J. Wang, J.Xu, X. Feng, and H. Zhang, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, pp. 2291-2311, 2019.

[88] G. Liu, M. Starke, B. Xiao, X. Zhang and K. Tomsovic, "Community microgrid scheduling considering building thermal dynamics," *IEEE Power & Energy Society General Meeting*, 2017.

[89] L. Lei, Y. Tan, G. Dahlenburg, W. Xiang, and K. Zheng, "Dynamic energy dispatch in isolated microgrids based on deep reinforcement learning," http://arXiv:2002.02581v1, 2020.

[90] X. Yang, Y. Wang, H. He, C. Sun, and Y. Zhang, "Deep reinforcement learning for economic energy scheduling in data center microgrids," *IEEE Power & Energy Society General Meeting (PESGM)*, 2019.

[91] J. Lee, W. Wang, and D. Niyato, "Demand-side scheduling based on deep actor-critic learning for smart grids," http://arXiv:2005.01979v1, 2020.

[92] A. Aswani, N. Master, J. Taneja, A. Krioukov, D. Culler, and C. Tomlin, "Energy-efficient building HVAC control using hybrid system LBMPC," *IFAC Proceedings Volumes*, vol. 45, no. 17, pp. 496-501, 2012.

[93] R.K. Kalaimani, M. Jain, S. Keshav, and C. Rosenberg, 'On the interaction between personal comfort systems and centralized HVAC systems in office buildings," *Journal Advances in Building Energy Research*, vol. 14, no. 1, pp. 129-157, 2020.

[94] R.K. Kalaimani, S. Keshav, and C. Rosenberg, "Multiple time-scale model predictive control for thermal comfort in buildings," *Proc. of e-Energy*, 2016.

[95] T. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, "Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation," *Proc. of NIPS*, 2016.

[96] L. Tan and R. Hu, "Mobility-aware edge caching and computing in vehicle networks: a deep reinforcement learning," *IEEE Trans. Vehicular Technology*, vol. 67, no. 11, pp. 10190-10203, 2018.

[97] T. Nguyen, "A multi-objective deep reinforcement learning framework," https://arxiv.org/abs/1803.02965, 2020.

[98] K. Li, T. Zhang, and R. Wang, "Deep reinforcement learning for multiobjective optimization," *IEEE Transactions on Cybernetics*, DOI:10.1109/TCYB.2020.2977661, 2020.

[99] X. Chen, A. Ghadirzadeh, M. Björkman, and P. Jensfelt, "Meta-learning for multi-objective reinforcement learning," https://arxiv.org/abs/1811.03376, 2019.

[100] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," *Proc. of ICML*, 2019.

[101] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *Proc. of ICLR*, 2016.

[102] M. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: a survey," *Journal of Machine Learning Research*, vol. 10, no. 7, pp. 1633-1685, 2009.

[103] Z. Jiang and Y. Lee, "Deep transfer learning for thermal dynamics modeling in smart buildings," *Proc. of IEEE International Conference on Big Data (Big Data)*, 2019.

[104] S. Xu, Y. Wang, Y. Wang, Z. O'Neill and Q. Zhu, "One for many: transfer learning for building HVAC control," http://arXiv:2008.03625v2, 2020.

[105] A. Mohsenian-Rad, V.W. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Trans. on Smart Grid*, vol. 1, no. 3, pp. 320-331, 2010.

[106] Y. Cao, S. Tang, C. Li, P. Zhang, Y. Tan, Z. Zhang, and J. Li, "An optimized EV charging model considering ToU price and SOC curve," *IEEE Trans. on Smart Grid*, vol. 3, no. 1, pp. 388-393, 2012.

[107] K. M. Tsui and S. C. Chan, "Demand response optimization for smart home scheduling under real-time pricing," *IEEE Trans. Smart Grid*, vol. 3, no. 4, pp. 1812-1821, 2012.

[108] X. Zhang, D. Bian, D. Shi, Z. Wang and G. Liu, "Community microgrid planning considering building thermal dynamics," *IEEE Sustainable Power and Energy Conference*, 2019.