Toward intelligent multizone thermal control with multiagent deep reinforcement learning

Li, Jie; Zhang, Wei; Gao, Guanyu; Wen, Yonggang; Jin, Guangyu; Christopoulos, Georgios

2021

Li, J., Zhang, W., Gao, G., Wen, Y., Jin, G. & Christopoulos, G. (2021). Toward intelligent multizone thermal control with multiagent deep reinforcement learning. IEEE Internet of Things Journal, 8(14), 11150-11162. <https://dx.doi.org/10.1109/JIOT.2021.3051400>

<https://hdl.handle.net/10356/152738>

<https://doi.org/10.1109/JIOT.2021.3051400>

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: <https://doi.org/10.1109/JIOT.2021.3051400>.

Downloaded on 26 Apr 2023 20:19:23 SGT

Towards Intelligent Multi-Zone Thermal Control with Multi-Agent Deep Reinforcement Learning

Jie Li, Wei Zhang, *Member, IEEE*, Guanyu Gao, *Member, IEEE*, Yonggang Wen, *Fellow, IEEE*,
Guangyu Jin, and Georgios Christopoulos

Abstract—Energy usage and thermal comfort are the pillars of smart buildings. Many research works have been proposed to save energy while maintaining a comfortable thermal condition. However, most of them either make the over-simplified assumption on thermal comfort with unsatisfied comfort performance or deal with the single-zone thermal control only with limited practical impact. A few preliminary pieces of research on multi-zone control are available, but they fail to keep pace with the latest advancements in the deep learning-based control techniques. In this paper, we investigate the multi-zone thermal control with optimized energy usage and canonical thermal comfort modeling. We adopt the emerging multi-agent deep reinforcement learning techniques and propose to model each zone as an agent. A multi-agent framework is established to support the information exchange among the agents and enable intelligent thermal control in the heterogeneous zones. Accordingly, we mathematically formulate a problem to optimize both energy and comfort. A multi-zone thermal control algorithm (MOCA) is proposed to solve the problem by deriving optimal control policies. We validate the performance of MOCA through simulation in professional TRNSYS, configured based on our real-world laboratory. The results are promising with up to 15.4% energy-saving as well as satisfied thermal comfort in different zones.

Index Terms—Multi-agent deep reinforcement learning, neural network, energy efficiency, thermal comfort, smart building

I. INTRODUCTION

The heating, ventilation, and air conditioning (HVAC) system is probably one of the most influential innovations in human history. It regulates the thermal condition and air quality of an indoor environment to keep occupants thermally

comfortable, and eventually changes human behavior in different aspects. Thanks to the HVAC system, people nowadays spend a significant 90% of time in buildings for various activities from working to entertaining. However, thermal comfort comes at a cost. Buildings are currently the dominant energy consumers in most power grids worldwide, and the HVAC system is the major factor contributing to 30 ~ 50% of the building energy usage [1].

We are facing a dilemma. Towards the energy saving, we may reduce the HVAC regulation intensity, i.e., not cooling too much in the tropics and not heating excessively in winter. However, the adjustments if not well investigated and controlled may worsen the comfort¹ condition and in turn affect the long-term health and working productivity of occupants. Moving towards comfort improvement, we are likely to see a surged electricity bill and over-stress the power grid. Thereby, it is critical to strike the tradeoff between energy and comfort through intelligent HVAC control.

Most existing works are energy-optimized and thermal comfort is considered in different granularity. Early research and the majority of the industry practices incorporate comfort as a threshold and assume temperature is the only determinant variable of thermal comfort. For example, the widely used proportional-integral-derivative (PID) algorithm [2] follows a rule-based heuristic control method. Comfort is modeled as a setpoint temperature like 23°C, and the method aims to optimize energy efficiency as long as the indoor temperature is not drifted far from the setpoint. Similar works can also be found in [3]–[6]. However, such empirical assumption on comfort has proven to be inaccurate in many cases, and as a result, the energy saving is often achieved with the significant sacrifice of occupant comfort.

Recently, researches have investigated the adequacy of the consideration of comfort in the HVAC systems [7]–[9]. Instead of making the over-simplified comfort assumptions, they aim to bring in more factors to capture both environmental conditions and occupant's vital states [10]–[12]. For example, advanced comfort models, i.e., predicted mean vote (PMV) [13], are used to model the comfort level of occupants, based on both indoor conditions like temperature and relative humidity (RH) and vital state like metabolic rate. According to the evaluated comfort level, the HVAC system is controlled dynamically to optimize the energy-comfort tradeoff. However, most of those works assume a single-zone setting,

Manuscript received January 1, 2000; revised January 1, 2000; accepted January 1, 2000. This research is funded by National Research Foundation (NRF) via the Green Buildings Innovation Cluster (Grant NO.: NRF2015ENC_GBICRD001-012), administered by Building and Construction Authority (BCA) Singapore. In addition, this research is sponsored by National Research Foundation (NRF) via the Behavioural Studies in Energy, Water, Waste and Transportation Sectors (Grant NO.: BSEWWT2017_2_06), administered by National University of Singapore (NUS). Moreover, this research is funded by Nanyang Technological University (NTU) via the Data Science & Artificial Intelligence Research Centre @ NTU (Grant NO.: DSAIR@NTU). (Corresponding author: W. Zhang.)

J. Li and Y. Wen are with the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore 639798 (e-mail: jie006@e.ntu.edu.sg and ygwen@ntu.edu.sg.)

W. Zhang is with the Infocomm Technology Cluster, Singapore Institute of Technology (SIT), Singapore 567739 (e-mail: wei.zhang@singaporetech.edu.sg.)

G. Gao is with the school of computer science and engineering, Nanjing University of Science and Technology, China (e-mail: ygao@njust.edu.cn)

G. Jin is with the Building and Construction Authority (BCA), Singapore 579700 (e-mail: jin_guang_yu@bca.gov.sg.)

G. Christopoulos is with the Nanyang Business School, Nanyang Technological University (NTU), Singapore 639798 (e-mail: cgeorgios@ntu.edu.sg.)

¹In this paper, we use comfort and thermal comfort interchangeably for simplicity, and use humidity and RH interchangeably as well.

which is untrue for many buildings, where space is divided into zones to serve different functions, i.e., office and lobby. Some latest works [14]–[16] have been proposed to deal with the multi-zone energy-comfort optimization. But the control mechanisms are still preliminary and the latest progress on control algorithms is yet well covered.

In this paper, we propose to optimize both energy and comfort for multi-zone and advance the HVAC control with the latest deep reinforcement learning (DRL) techniques. Specifically, we adopt a multi-agent concept to model the zones, where each zone is associated with one agent. A framework is established to coordinate the information exchange among the agents and control units and support the intelligent control of the HVAC system for the heterogeneous zones. Accordingly, we formulate an optimization problem mathematically for our multi-zone HVAC control system. Our formulated algorithm well aligns with the cutting-edge multi-agent DRL technique. Thus, we design a multi-zone thermal control algorithm (MOCA) based on multi-agent DRL to control the thermal condition of each zone with optimized energy and comfort performance. Finally, we validate our proposed solution by building up a simulation environment according to a real-world laboratory setting with professional TRNSYS software. The results show that our solution achieves considerable energy saving by up to 15.4%. Meanwhile, the solution can also maintain a comfortable indoor environment for each zone.

Our research findings in this work provide fundamental insights and guidelines for applying the DRL techniques to the intelligent multi-zone HVAC control optimization. Our solution can also be used to perform the building-side demand response via thermal load control to stabilize the power grid. Overall, this work owns the potential to advance future researches in both smart buildings and smart grid.

The rest of this paper is organized as follows. We present the related works in Section II. In Section III, we introduce our system design and formulate the optimization problem. Section IV presents our intelligent multi-zone thermal control algorithm. Section V evaluates the performance of our proposed solution. Finally, Section VI concludes this paper.

II. RELATED WORK

Many research works have been conducted to optimize energy usage and thermal comfort in the past decades. Those works can be categorized into two major groups, including the single-objective energy-driven approaches and multi-objective energy-comfort tradeoff optimization approaches.

The first group targets to optimize the energy solely without much focus on the thermal comfort, i.e., not being considered or as long as the temperature is within a certain range. In the 1970s, conventional mechanisms (e.g., basic ON-OFF method and PID [2]) were developed, and those pioneer control schemes often suffer from the overshoots in the setpoint temperature. Even more, control operations might be inappropriate if the settings are not configured properly to well capture the real-world settings, potentially increasing energy consumption as a result.

Various more advanced control and optimization methods were proposed in the following years, to deal with the possible

inappropriate control of the pioneer approaches. Meanwhile, shallow learning-based approaches also drew attention from the research community for intelligent HVAC control (e.g., genetic algorithm [17], neural network [18], and reinforcement learning [19]). However, those new ideas were seldom applied in practice due to industry maturity at that time. In the new millennium, model-based predictive control (MPC) approaches [20]–[22] were also proposed to optimize the HVAC control. With the fast advancement of deep learning in recent years, the earlier learning-based approaches have also been upgraded to improve the control performance (e.g., deep neural network-based [3] and deep reinforcement learning (DRL) based [4], [23]). However, comfort is still not the key concern of those works. Overall, thermal comfort is overlooked by the existing single-objective energy-driven approaches. For some putting certain consideration on comfort, only a temperature setpoint or range is specified, and this is far from accurately modeling the dynamic occupant comfort affected by many factors.

Indeed, some researchers noticed the issues of single-objective control and have worked on optimizing the tradeoff between energy and comfort, i.e., achieving the maximized comfort with minimized energy expenditure. The models to evaluate occupant comfort have been incorporated into the control system. For example, PMV [13], possibly the most popular thermal comfort model, and the adaptive thermal comfort model [24] are used for evaluating and predicting the thermal comfort. In terms of the control methodologies, conventional mathematical formulations are still prevalent in some related works [25], [26]. Deep learning has also been investigated [7]–[9] for intelligent HVAC control. Most of the approaches in this group assume a single-zone environment, which, however, is not true for many buildings where space is divided into multiple zones, i.e., different offices and stores.

Recently, a few works have been proposed for both multi-objective energy-comfort optimization and multi-zone thermal control. A natural way is to use the multi-agent systems [14]–[16], where each agent can model one of the multiple zones. Existing approaches mainly adopt the conventional optimization algorithms (e.g., particle swarm optimization [14], MPC [16], Nash-bargaining [15], and alternating direction multiplier method [16]). Those conventional approaches are often sensitive to the input variations and parameter settings. Thus, we would like to bring in the latest deep learning advancements on the multi-agent systems, to improve the control performance.

In this paper, we propose to adopt the latest multi-agent DRL techniques, for optimizing the multi-zone thermal control. With the rapid development of the DRL techniques [27], [28], various multi-agent DRL systems are developed to address the complex control problems. For example, AlphaStar [29] is trained by multi-agent DRL to play the game StarCraft II and can beat human professional players. Multi-agent DRL shows its outstanding learning ability to learn the optimal control policy for handling a large volume of state information and complicated control actions. Overall, our proposed approach not only considers the energy-comfort tradeoff in each zone but also allows the agents to learn from each other to speed up the learning process of the optimal control strategy for the

Fig. 1: The overall workflow of intelligent multi-zone HVAC control. According to the updated zone information (e.g., indoor sensing data and energy expenditure of the HVAC systems) and outdoor sensing data (e.g., temperature and humidity), the learning-based algorithm keeps optimizing the thermal control policy to send out the control actions (e.g., change the setpoint temperature and/or the setpoint humidity) to the HVAC controller for zone-based thermal condition adjustment.

multiple zones.

III. SYSTEM DESIGN & PROBLEM FORMULATION

In this section, we first describe the overall workflow of the system design and then formulate the optimization problem of the multi-zone HVAC control.

A. System Design & Workflow

We illustrate the system design and the corresponding multi-zone control workflow in Fig. 1, which consists of the following four major modules.

1) *Building Zones*: Based on different functionalities, buildings can be divided into various zones (e.g., workspace and meeting rooms). The HVAC systems are responsible for maintaining the thermal conditions for these zones.

2) *Indoor & Outdoor Sensors*: They are used to monitor the weather and zone conditions. The measured environmental variables (e.g., temperature and humidity) will be inputted into the learning-based algorithm for control decision making.

3) *Learning-based Algorithm*: With the updated indoor and outdoor sensing data, the learning-based algorithms learn the control policies and send the optimal control actions (e.g., change the setpoint temperature) to the HVAC controller.

4) *HVAC Controller*: After receiving the optimal control actions, the HVAC controller executes accordingly to adjust the thermal conditions of each zone.

Worth to mention that the overall workflow forms a self-learning loop for optimizing the control policies. First, the algorithm learns to take control actions based on the latest indoor and outdoor sensing information. Then, the thermal environment of each zone is affected by the executed control actions and the environment change can be sensed by indoor sensors. Meanwhile, the weather changes are monitored by the outdoor sensors. Next, the learning-based algorithm embarks

another iteration of the self-learning loop with the newly obtained sensing data.

B. State, Action, and Reward

In this paper, we model the multi-zone HVAC control as an Markov decision process (MDP) problem. We utilize a time-slotted model with the discrete time slots $t = 0, 1, \dots$ and we assume m zones in the environment, as shown in Fig. 2. Zone i is managed by agent i for $0 \leq i \leq m - 1$. More details are as follows.

1) *State*: The state of an environment is a group of factors that can reflect its current condition. Specifically, indoor environmental factors include indoor temperature and humidity and outdoor factors are outdoor temperature and humidity. We use s_t^i to denote the state of zone i at time slot t . Combining the states of the m zones, we have $s_t = (s_t^0, s_t^1, \dots, s_t^{m-1})$. According to the actions taken by the m agents at time slot t , the environment moves into the next state s_{t+1} at time slot $t + 1$.

2) *Action*: The action refers to various control operations taken by the agents. Each agent controls the HVAC system to adjust the thermal conditions. Here, we assume a common HVAC system with controllable setpoint temperature T_t^i and setpoint humidity H_t^i for zone i at time t . Given the current state s_t^i , agent i evaluates the current state and takes an action $a_t^i \in \mathcal{A}$ to adjust the thermal condition of zone i , where $\mathcal{A}$ is the set of possible actions. The behavior of each agent i is determined by a deterministic control policy μ^i as,

$$a_t^i = \mu_t^i(s_t^i). \quad (1)$$

Combining the actions and policies of m agents at time slot t , we have actions $a_t = (a_t^0, a_t^1, \dots, a_t^{m-1})$ and policies $\mu_t = (\mu_t^0, \mu_t^1, \dots, \mu_t^{m-1})$. Here, with one exclusive policy for each zone, the policy can be specialized based on the zone's

Fig. 2: The schematic diagram of the multi-agent MDP. For zone i , the related agent i observes the state s_t , takes the action a_t^i , and receives the reward r_t^i at time slot t for zone i .

unique setup. Each policy is guided by the corresponding reward function, as defined in the following part.

3) *Rewards*: Reward function plays a vital role in the policy learning process of DRL with a significant impact on the convergence of the model training. In this paper, we design the reward function based on the following two aspects.

- *Energy Consumption*: The energy expenditure of the HVAC systems mainly consists of the energy consumed by cooling, heating, ventilation, humidification, and dehumidification. In this work, we focus on the cooling energy and dehumidification energy based on the following reasons. First, our testbed locates in the tropics, and the year-round hot and humid climate makes the local HVAC systems rarely equipped with the heating and humidification modules. Second, ventilation is essential to ensure the indoor air quality, which significantly affects occupants' health. So it is risky to reduce the ventilation intensity for saving energy with a potentially negative impact on the air quality and occupant health. Also, ventilation energy expenditure is relatively insignificant compared to the cooling and dehumidification energy. Overall, we mainly consider the energy usage of cooling and dehumidification, and calculate the total energy usage ε_t^i in zone i at time slot t via the following equation,

$$\varepsilon_t^i = c_t^i + d_t^i, \quad (2)$$

where c_t^i and d_t^i are the cooling and dehumidification energy, respectively.

In practice, both energy terms can be tracked with the power meters for cooling and dehumidification units. Some energy systems directly monitor the combined energy usage ε_t^i without differentiating the cooling and dehumidification components. Our formulation is compatible with both aggregated and separate monitoring. As long as we have the information about the total energy usage, we can quantify the corresponding impact based on our reward function as described below. Given that a real test site sometimes is unavailable, a simulator is also a venue. For example, the energy trend is accessible in TRNSYS given the specified building settings. More details are available in our experimental study.

- *Thermal Comfort*: Thermal comfort is a critical factor to evaluate occupants' satisfaction levels of their surrounding environment. In this paper, we adopt the popular PMV thermal comfort model and predict the occupants' average thermal comfort index p_t^i as follows,

$$p_t^i = \mathcal{PMV}(T_t^i, H_t^i, V_t^i, M_t^i, I_t^i, B_t^i), \quad (3)$$

where the model inputs are six PMV parameters, namely temperature T_t^i , RH H_t^i , air velocity V_t^i , mean radiant temperature (MRT) M_t^i , clothing insulation I_t^i , and metabolic rate B_t^i . The value of the output p_t^i is from -3 to $+3$, where the seven integers within the range indicate cold, cool, slightly cool, comfortable, slightly warm, warm, and hot. The comfort range is often set to $[-0.5, 0.5]$ or $[-1.0, 1.0]$ to maintain a comfortable thermal condition. The computation detail is available [13]. Note that last four parameters are hardly accessible in reality, i.e., we cannot monitor metabolic rate due to user privacy concern and MRT sensors are rarely deployed. In this research, we follow the common practice to assume some fixed values based on existing empirical study, i.e., 1.0MET metabolic rate and MRT equals to indoor temperature for a typical office setting.

Based on Eq. (2) and Eq. (3), we define the reward function $\mathcal{R}$ to jointly consider the energy expenditure ε_t^i and the occupants' average thermal comfort index p_t^i as follows,

$$r_t^i = \mathcal{R}(\varepsilon_t^i, p_t^i) = -\delta_i \varepsilon_t^i + \begin{cases} 0, & \text{if } L_p \leq p_t^i < 0, \\ p_t^i, & \text{if } 0 \leq p_t^i \leq U_p, \\ -|p_t^i|, & \text{otherwise,} \end{cases} \quad (4)$$

where $\delta = (\delta_0, \delta_1, \dots, \delta_{m-1})$ and δ_i is the control parameter to balance the impact of energy consumption and thermal comfort in the reward function and r_t^i is the reward for zone i at time t . Note that the parameters do not have to be the same and we can specify different weight for each zone to fulfill our goal of zone-based energy-comfort optimization. Moreover, we set the expected thermal comfort upper bound $U_p \geq 0$ and lower bound $L_p \leq 0$ to guide the algorithm to maintain the acceptable thermal comfort within the desired range, i.e., $L_p = -0.5$ and $U_p = 0.5$ based on the ASHRAE standard.

The rationale of our energy-comfort formulation lies in the following aspects. First, we would like to minimize energy usage, which is a positive value. We impose a negative coefficient ahead of energy so that the reward is minimized with large energy usage and vice versa. Second, we also want to minimize the discomfort due to a poorly regulated environment. But the comfort's impact on energy is relatively not straightforward. Here, we specify an expected acceptable range of comfort index. If the comfort index is within the range and negative (towards the cool sensation), we maintain a neutral position and assign a zero reward. Within the range and positive (towards the warm sensation), we encourage such a state by increasing the reward accordingly. Our motivation is that in a typical cooling scenario like for Singapore, a certain level of tolerance about a slightly warm environment helps reduce the cooling load and thus save energy. The index can also be beyond the specified range and breach the

expected comfort. In this case, we shall penalize to avoid such a situation from happening, and we add a negative term in the reward function to reflect the penalization. Overall, we aim to maximize the reward with good energy efficiency and satisfying comfort. Despite the maximization problem, the reward can be negative based on our formulation due to the negative terms for energy and discomfort, as will be shown in the experimental study.

Note that our formulation can be easily extended to capture further aspects like light comfort by incorporating new terms if necessary. Altogether, we have the rewards of the m zones as $r_t = (r_t^0, r_t^1, \dots, r_t^{m-1})$.

C. Problem Formulation

We formulate the multi-zone thermal control optimization problem to strike the tradeoff between energy and comfort. Following the DRL framework, at every time slot t we aim to optimize the summation of discounted future reward R_t^i of each agent i as below,

$$R_t^i = \sum_{j=1}^{\mathcal{T}} \lambda_i^{j-1} r_{t+j}^i, \quad (5)$$

where $\mathcal{T}$ is the optimization horizon and r_{t+j}^i is the reward at time slot $t+j$ and $j = 1, 2, \dots, \mathcal{T}$. Moreover, $\lambda_i \in [0, 1]$ denotes the discounted factor attached to agent i . With m agents, let $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_{m-1})$ and the expected reward of all agents be $\sum_{i=0}^{m-1} \mathbb{E}[R_t^i]$. Here, we assume the temperature and humidity setpoints of the HVAC system are controllable and mathematically formulate the optimization problem based on Eqs. (2)-(5) as follows,

$$\max_{T_t^i, H_t^i} \sum_{i=0}^{m-1} \sum_{j=1}^{\mathcal{T}} \lambda_i^{j-1} \mathcal{R}(\varepsilon_{t+j}^i, p_{t+j}^i), \quad (6)$$

$$\text{s.t.} \quad L_{T_t^i} \leq T_t^i \leq U_{T_t^i}, \quad (7)$$

$$L_{H_t^i} \leq H_t^i \leq U_{H_t^i}, \quad (8)$$

where $\mathcal{R}$ is the reward function in Eq. (4), ε_{t+j}^i is the total energy expenditure of the HVAC systems of zone i at time slot $t+j$ by Eq. (2), and p_{t+j}^i denotes the average comfort index of zone i at time slot $t+j$ by Eq. (3). Moreover, T_t^i is the zone temperature controlled by agent i , which is constrained by Eq. (7), and H_t^i is the zone humidity controlled by agent i following the constraint in Eq. (8). The corresponding lower bounds, $L_{T_t^i}$ and $L_{H_t^i}$, and upper bounds, $U_{T_t^i}$ and $U_{H_t^i}$, can be configured into the HVAC control system.

IV. PROBLEM SOLVING & ALGORITHM DESIGN

In this section, we solve the formulated optimization problem in Eq. (6). The optimized control policies for multiple zones is learned by one of the latest DRL algorithms, called multi-agent deep deterministic policy gradients (DDPG) [30], a general-purpose DRL algorithm that extends DDPG [31] to the multi-agent applications. Multi-agent DDPG is built upon the classic actor-critic structure that each agent is designed with the architecture of a centralized critic and decentralized actor, where both actor and critic are approximated via the

Fig. 3: The architecture of multi-agent DRL for multi-zone thermal control. Each agent contains a centralized critic to observe the global information, i.e., zone states and actions, and a decentralized actor to take action based on the state.

deep neural networks, as shown in Fig. 3. Multi-agent DDPG is designed to support multiple agents to collaborate to solve a complex problem together. Agents learn from each other via the shared global information (e.g., the states and actions of the other agents) to optimize their local action-taking policies. It not only improves the policy learning efficiency but also avoids over-fitting in a single zone with degraded overall HVAC control performance.

A. MOCA

Following the multi-agent DDPG framework, we present MOCA in this part.

1) *Critic Network*: Critic network, generally named as Q -function, is the approximation of action-value function to measure the quality of the action taken by the actor network. Given the set of observed states $s_t = (s_t^0, s_t^1, \dots, s_t^{m-1})$. For agent i , the critic network with weight ω_t^i aims to learn a centralized action-value function with respect to the policies $\mu_t = (\mu_t^0, \mu_t^1, \dots, \mu_t^{m-1})$ of all agents as follows,

$$Q_i^{\mu_t}(s_t, a_t | \omega_t^i), \quad (9)$$

where $\omega_t^i \in \omega_t$ and $\omega_t = (\omega_t^0, \omega_t^1, \dots, \omega_t^{m-1})$ is the set of weights of all critic networks at time t .

The weight of each agent's critic network is updated by minimizing the following loss function,

$$\ell(\omega_t^i) = [Q_i^{\mu_t}(s_t, a_t | \omega_t^i) - y_i]^2, \quad (10)$$

where $\ell(\omega_t^i)$ is the squared temporal-difference error and y_i is the target action-value function of agent i as follows,

$$y_i = r_t^i + \lambda_i Q_i^{\mu_{t+1}}(s_{t+1}, a_{t+1} | \omega_{t+1}^i), \quad (11)$$

where μ_{t+1} is the set of target policies at time $t+1$.

Finally, the critic network of agent i is updated with the following updating rule,

$$\omega_{t+1}^i = \tau_c \omega_t^i + (1 - \tau_c) \omega_{t+1}^i, \quad (12)$$

where τ_c is used for target update of the critic networks.

Algorithm 1 MOCA

```

1: initialize the replay memory  $\mathcal{D}$ , number of episodes  $n_e$ , and the number of time slots  $n_t$ .
2: for agent  $i = 0, 1, \dots, m-1$  do
3:   initialize the critic network  $Q_t^\mu(s, a|\omega^i)$  with random weights  $\omega^i$ , where  $s = (s^0, \dots, s^{m-1})$  and  $a = (a^0, \dots, a^{m-1})$ 
4:   initialize the actor network  $a^i = \mu^i(s^i|\theta^i)$  with random weights  $\theta^i$ 
5: end for
6: for episode  $= 1, 2, \dots, n_e$  do
7:   initialize a random process  $\eta$  for action exploration
8:   all agents observe the initial state of the multi-zone environment
9:   for  $t = 0, 1, \dots, n_t$  do
10:    for agent  $i = 0, 1, \dots, m-1$  do
11:      observe the current state  $s_t$  of the environment
12:      select and execute action  $a_t^i$  via Eq. (13) based on the current local zone state  $s_t^i$ 
13:      predict occupants' average thermal comfort  $p_t^i$  via Eq. (3)
14:      calculate the reward  $r_t^i$  via Eq. (4)
15:      observe the next state  $s_{t+1}$  of the environment and obtain the local transition  $(s_t, a_t^i, r_t^i, s_{t+1})$ 
16:    end for
17:    combine the local transitions of  $m$  agents and buffer the global transition  $(s_t, a_t, r_t, s_{t+1})$  in the replay memory  $\mathcal{D}$ 
18:    if exploration is completed then
19:      for agent  $i = 0, 1, \dots, m-1$  do
20:        sample a random mini-batch of  $K$  transitions from the replay memory  $\mathcal{D}$ 
21:        update the critic network via minimizing the loss function Eq. (10) and the network weight  $\omega^i$  via Eq. (12)
22:        update the actor network via the sampled policy gradients Eq. (15) and the network weight  $\theta^i$  via Eq. (16)
23:      end for
24:    else
25:      continue the exploration to sample sufficient transitions and store them in the replay memory  $\mathcal{D}$ 
26:    end if
27:  end for
28: end for

```

2) *Actor Network*: Actor network is the approximation of the policy function which controls agent's action guided by the critic network. For agent i , the actor network parameterized by θ_t^i aims to learn a decentralized policy function based on the local zone state s_t^i . Here, we add random noise to our data for better generalization. Based on the Eq. (1), we get the following exploration policy function at time slot t ,

$$a_t^i = \mu_t^i(s_t^i|\theta_t^i) + \eta_t, \quad (13)$$

where η_t is the exploration noise generated by the Gaussian random process and $\theta_t = (\theta_t^0, \theta_t^1, \dots, \theta_t^{m-1})$ is the set of weights of all the agent actor networks.

During the exploration, first, based on the current zone observation s_t^i , agent i executes the action a_t^i , receives the reward r_t^i , and observes the next zone observation s_{t+1}^i . All agents together, the general transition is expressed as,

$$(s_t, a_t, r_t, s_{t+1}), \quad (14)$$

where given the current state s_t of multi-zone environment, the m agents take m actions $a_t = (a_t^0, a_t^1, \dots, a_t^{m-1})$ and receive m rewards $r_t = (r_t^0, r_t^1, \dots, r_t^{m-1})$, leading to the new state s_{t+1} . After that, these transitions will be buffered in the replay memory $\mathcal{D}$ for the mini-batch sampling to train the actor-critic networks of all agents.

Then, based on the Q -function in Eq. (9) and the policy function in Eq. (13), with one sampled transition

$(\hat{s}_t, \hat{a}_t, \hat{r}_t, \hat{s}_{t+1})$ from $\mathcal{D}$, the actor network of agent i is updated with the following sampled policy gradient $\nabla_{\theta_t^i} J(\theta_t^i)$,

$$\nabla_{\theta_t^i} J(\theta_t^i) \approx \nabla_{\theta_t^i} \mu_t^i(\hat{s}_t^i|\theta_t^i) \nabla_{\hat{a}_t^i} Q_t^{\mu_t}(\hat{s}_t, \hat{a}_t|\omega_t^i)|_{\hat{a}_t^i = \mu_t^i(\hat{s}_t^i|\theta_t^i)}. \quad (15)$$

Besides, the actor network of agent i is updated as,

$$\theta_{t+1}^i \leftarrow \tau_a \theta_t^i + (1 - \tau_a) \theta_{t+1}^i, \quad (16)$$

where τ_a is used for target update of the actor networks.

Here, we present the training procedures of the MOCA in Algorithm 1 and time overhead discussion as follows.

B. Time Overhead

Time overhead is a key concern of system deployment. The model training is normally performed offline infrequently, so its time overhead is negligible. Once the training is done, we deploy the actor-network in real HVAC systems. Then, in each time slot, the network observes the latest state and takes it as input, and derives the action to be performed. Such observe-derive-perform operations can be implemented repeatedly to realize automatic system control. Here, the time complexity of the actor-network is determined by the number of hidden layers and the number of neurons in each hidden layer. Given m hidden layers and n neurons per layer, the time complexity is $O(mn^2)$. In our experimental study as shown in the following section, we configure the actor-network with 2 hidden layers, each with 128 neurons, and the configuration

Fig. 4: The design diagram of the MOCA simulation environment in TRNSYS. It consists of various modules, i.e., lighting configuration, weather data loader, radiation settings, multi-zone building model, and MATLAB communication channel.

TABLE I: Heat Gain Settings of Each Zone in TRNSYS.

Zone	Function	# of People	# of Computers
Zone 0	Project Demo Area	3	12
Zone 1	Student Self-study Area	10	12
Zone 2	Meeting Room Area	4	6
Zone 3	Staff Working Area	16	20

works well. Given such a minimal configuration, the actions can be derived instantly, i.e., in merely a few milliseconds.

V. PERFORMANCE EVALUATION

In this section, we first introduce the implemented simulation environment. Then, we illustrate the experimental settings. Finally, we analyze the experimental results and evaluate the performance of our proposed MOCA.

A. Simulation Environment

To verify the effectiveness of our approach, we create a multi-zone simulation environment with the HVAC systems by using professional TRNSYS software. As shown in Fig. 4, it consists of the following two major components.

1) *Building Model*: We follow the multi-zone building modeling tutorials of TRNSYS to create a building model with four zones. Each zone is equipped with an individual air-node for thermal control, and we only enable the cooling, ventilation, and humidification functions of each air-node. Other settings (e.g., construction of walls, floors, ceilings, roofs and windows, lighting scheduling control, and radiation) follows the default TRNSYS settings. Moreover, we configure the built-in monitor components to track the energy consumption and thermal conditions of each zone in real-time.

2) *Control Interface*: In this paper, we implement MOCA using PyTorch in Python. One problem is that TRNSYS cannot communicate with Python directly. As such, we use MATLAB to create a proxy communication channel for parameter

exchange between Python and TRNSYS. Moreover, we also record the operation information (e.g., each zone's thermal condition time-series and control operations) in a MySQL database.

B. Experimental Settings

In this part, we describe the experiment settings of the simulation environment and our control algorithm.

1) *Simulation Environment*: We create a multi-zone simulation environment referring to our real-world research laboratory settings. The total laboratory area is 301m² and the area is divided into four equal-size zones. We configure the heat gains for each zone by considering two heat sources for people and computers. Detailed heat gain settings are shown in Table I to reflect the different functions of different zones. Besides, we follow a prevalent setting of 67% air change rate per hour for all zones. The working hours of the laboratory are from 9am to 6pm. Weather data plays a vital role in our simulation, and we chose a local weather dataset, SG-Singapore-Airp-486980, available in TRNSYS to simulate the variation of the outdoor thermal environment.

2) *Control Algorithm*: Our implemented MOCA involves several parameters, detailed as follows. We configure two hidden layers, each with 128 neurons, for both actor and critic networks for each agent. Both networks adopt the adam optimizer. Note that the key job of the critic network is to evaluate the actions derived by the actor-network, so the former shall have a faster pace to be able to guide the training of the latter. Thus, we configure the learning rates to 0.001 and 0.0001 for the critic- and actor-network, respectively, to enforce such a pace difference. We also set the discounted factor $\lambda = 0.95$ and the training batch size as 128, normally the power of 2 to well utilize the computing resource, for all agents. Besides, let $\tau_c = \tau_a = 0.01$ for the target updates and we set the replay memory size as 10^6 . Each transition in the replay memory is a 4-tuple, including the current state with 4 dimensions, 2-dimensional action, 1-dimensional reward, and the next state also with 4 dimensions, resulting from the current state and action. In total, we have 11 dimensions for each transition. For our experiments, we consider 4 zones and we create one replay memory for each zone.

For the model training, we sample one transition per hour (i.e., one hour for each time slot) and we configure 100 time slots per episode. With 100 transitions in every episode, we store them in the replay memory and update the actor-critic network weights accordingly. In total, we simulate for 150 episodes, where 50 for action exploration and transition buffering, and 100 for control policy learning. Afterward, we execute 50 episodes on modeling testing and the analysis of the result will be presented in the following parts.

On the building side, we set the indoor temperature range to [20.0, 30.0]°C and RH to [50, 70]%. Besides, we assume the same energy-comfort weight for each zone in our test case, but the weights are allowed to be zone-specified if necessary.

C. Training Convergence Performance

In this part, we analyze the convergence performance of MOCA. We refer to the prevalent ASHRAE standards and set

Fig. 5: The training function reward varies with episode. The policy learning of each zone fast converges with 55 episodes.

the expected comfort bounds to $L_p = -0.5$ and $U_p = 0.5$. We show the returned rewards of our algorithm in different training episodes in Fig. 5. For the total reward of all the zones as shown in Fig. 5(a), we can see that our algorithm starts to sample the stored transitions for training at episode 50. The convergence speed is outstanding and the algorithm has converged only after 5 more episodes. After that, the reward fluctuates and remains relatively stable in the rest of the training episodes. Note that the reward can be a negative value despite our objective of maximizing the reward. The reason is that we penalize the high energy usage, which is a positive value, by multiplying a negative coefficient in our reward function in Eq. 4. As a result, the higher the energy usage is, the lower (possibly negative) the reward will be. Besides, comfort can also introduce a negative term into the reward, when the comfort index is beyond the expected comfort bound. In short, the negative reward values align with our objective of reward maximization and accordingly saving energy with satisfying comfort.

Here, we zoom in to investigate the rewards for each zone and show the results in Fig. 5(b). We can see that the fast convergence is also true for each zone. Note that the four zones in our test are heterogeneous in terms of the function and heat gain as shown in Table I. As a result, the cooling and dehumidification demand are different for the zones. Such a difference is captured by our algorithm and we can observe the different reward trends for the zones. In summary, MOCA demonstrates its outstanding learning ability and convergence performance in our multi-zone setting.

D. Thermal Control Performance

In this part, we present the thermal control performance of MOCA. First, we investigate the hourly indoor and outdoor thermal conditions as well as energy consumption. Second, we discuss the performance impact of different comfort demands.

1) *Hourly Thermal Control*: We analyze the thermal control performance with 5,000-hour testing data. Note that although our control system starts operating at 9am, it takes some time for the indoor condition to stabilize based on the

Fig. 6: The hourly outdoor temperature and humidity are tightly correlated, i.e., humidity decreases with the increment of the temperature, and vice versa.

default control settings. So, our control actions will be tuned and optimized since 10am.

- *Outdoor Thermal Condition*: First, the hourly statistical outdoor thermal condition in the tropics are shown in Fig. 6. We can see that the average outdoor temperature increases from 26.6°C at 10am to the maximum 29.8°C at 3pm, and then decreases to 28.7°C at 6pm. Meanwhile, the average outdoor humidity decreases from the maximum 86% at 10am to the minimum 72% at 3pm and then increases to 77% at 6pm. Overall, the outdoor humidity decreases with the increment of temperature.
- *Indoor Temperature*: Under the simulated outdoor thermal condition, we demonstrate the hourly indoor temperature, humidity, comfort index, and cooling-dehumidification energy consumption for each zone in Fig. 7. First, Fig. 7(a) shows that the average temperature of each zone fluctuates around 26.6°C. Overall the fluctuation of indoor temperature is insignificant within around half a degree.
- *Indoor humidity*: Fig. 7(b) shows the hourly average RH of each zone and the fluctuation is tiny. Zone 1 shows the largest RH difference over a day, and even though the maximum difference is only 0.5%.

Fig. 7: The hourly indoor temperature, humidity, thermal comfort index, and cooling-dehumidification energy consumption of each zone. Temperature is the major factor for comfort and the energy usage is high for the zones with large heat gain.

Fig. 8: The average temperature, humidity, thermal comfort index, and energy consumption of each zone under different thermal comfort bounds. The change pattern of thermal comfort matches with that of temperature, and humidity does not show much variation. Energy consumption of each zone decreases with the less strict thermal comfort requirement.

- *Thermal Comfort*: With the stable indoor condition, the average thermal comfort index of each zone is also well-controlled as shown in Fig. 7(c). Intuitively, high indoor temperature helps save cooling energy. Thus, we observe that our control system maintains the comfort index around 0.5 towards slightly warm. Note that we do not impose a hard limit on the comfort index in our reward function. Instead, we jointly optimize the tradeoff between energy and comfort, controlled by a weightage coefficient in Eq. 4. Given the excessive energy usage, we also explore the potential of going beyond the specified expected comfort bound to hold back the energy's impact on reward reduction. Indeed, occupants normally cannot tell the difference of a few tenths of the index change. So we allow our system to push the comfort slightly over 0.5 sometimes to save energy. Despite such green light, the comfort condition is well managed by MOCA and the resulting comfort index is at most 0.15 higher, i.e., for Zone 3 at 3pm.
- *Energy Consumption*: Finally, Fig. 7(d) shows that the hourly cooling-dehumidification energy consumption of each zone. The energy usage is mainly affected by the thermal condition. We observe that the zone with high heat gain consumes much energy, i.e., Zone 3 for the working area with the most occupants and computers.

In summary, MOCA can well maintain a comfortable indoor environment under different outdoor thermal conditions. The

indoor temperature is the key factor that affects occupant thermal comfort. Also, the zones with large heat gain draw much energy for cooling and dehumidification.

2) *Performance Impact of Comfort Bounds*: In this part, we analyze the thermal control performance of MOCA with different thermal comfort demands. MOCA adjusts the energy-comfort weight to guide the algorithm to fulfill different expected thermal comfort bounds from $[-0.1, 0.1]$ to $[-1.0, 1.0]$. For each bound, we report the statistical results of the 5,000-hour testing, and the results are shown in Fig. 8.

As seen from the figures, when we relax the thermal comfort constraint, the temperature and comfort index both increase and the energy usage decreases. Specifically, Fig. 8(a) and Fig. 8(c) show that the change in temperature and comfort index is approximately linear. With the most strict expected comfort bound $[-0.1, 0.1]$, the temperature is 25.3°C and the average comfort is 0.11. For the relaxed bound like $[-1.0, 1.0]$, temperature increases to 27.9°C and comfort index is 0.99.

Humidity remains stable with different comfort bounds as shown in Fig. 8(b). The only less stable humidity is 64.4%, observed in Zone 0 with bound $[-0.1, 0.1]$. Even though, it is only around 1.5% different from the stable level. The results indicate that humidity does not own a strong correlation with thermal comfort. It is true that a recent research [32] shows that humidity, even changing significantly from 50% to 60%, only affects comfort by the same amount as changing the temperature by merely 0.4 Celsius degree. This can be a reason

TABLE II: Energy-Comfort Performance of the Comparison Algorithms with Comfort Bound $[-0.5, 0.5]$.

Thermal Comfort Index					
Approach	Zone 0	Zone 1	Zone 2	Zone 3	Average
Baseline	-0.4	-0.4	-0.4	-0.4	-0.4
DDPG	0.5	0.5	1.0	1.0	0.8
MOCA	0.5	0.5	0.5	0.5	0.5
Energy Consumption (kWh)					
Baseline	9.4	10.3	8.4	12.6	40.7
DDPG	8.6	9.5	6.9	11.0	35.9
MOCA	8.6	9.5	7.6	11.7	37.3

why RH is not adjusted too much in our system.

The energy consumption for cooling and dehumidification in Fig. 8(d) also follows an approximately linear relationship with comfort bound. For the most strict bound, the energy is relatively high at 39.6kWh. When the thermal demand is relaxed, we see a reduction of the energy of 34.6kWh.

Note that our experiments are based on TYNYSYS. Despite professional software, inevitably, there is a gap between the simulation and real-world dynamics. We would like to clarify that our results and some discussions are based on the simulation study. In the real world, observations and interpretations could be different.

In summary, the simulation results demonstrate the effectiveness of our approach on multi-zone energy-comfort management. Under different thermal comfort demands, our approach can learn a suitable control policy automatically to balance the energy-comfort of each zone.

E. Comparison Study on Thermal Control

In this part, we conduct a comparison study to investigate the thermal control performance of MOCA and several comparison algorithms. First, we provide a brief description of the comparison algorithms as follows.

- **Baseline:** Our baseline algorithm is a typical rule-based control method. It aims to maintain the room temperature based on a common pre-configured setpoint, i.e., 23.0°C , via the direct digital control, an improved PID control [2]. The implementation of such control is easy and the performance is relatively acceptable, so it is widely used in modern HVAC control systems.
- **DDPG:** As introduced above, DDPG is a model-free off-policy actor-critic algorithm based on deterministic policy gradient and deep Q-learning. We let DDPG control the zones with one agent, as a benchmark to evaluate our performance on multi-agent control.

We present the comparison results as follows.

1) **Energy Saving Comparison:** Here, we consider the default comfort bounds $[-0.5, 0.5]$ first for the comparative analysis and we report the results in Table II. The baseline follows a constant configuration with the same level of comfort all the time. We can see that such a constant setting is neither comfort optimal nor energy optimal. Comfort-wise, occupants feel slightly cool with index -0.4 in the indoor environment

Fig. 9: The energy saving of each zone under different thermal comfort bounds. Relaxed bound helps achieve energy-saving for all zones. The zones with few occupants and computers show good energy-saving potential.

instead of comfortable. This is unreasonable in our tropical region as people are normally used to a warm atmosphere and the cooling energy can be significant. Indeed, this can be reflected in the energy results, where the baseline has the highest energy usage of 40.7kWh. Note that the energy usage for each zone is different for the baseline algorithm, and it is because of the heterogeneous heat gains of the zones.

DDPG adopts a more advanced control algorithm compared to the baseline. The energy performance of DDPG is outstanding, i.e., 35.9kWh. However, such energy performance is at the cost of poor thermal comfort. Same as the baseline, DDPG does not capture the difference of the zones and derives control actions based on the whole space. As a result, some zones with low heat gain are still manageable within the comfort bounds, and the others are not fortunate enough. For example, Zone 3 with the highest heat gain fails to receive enough cooling, and the comfort index is 2x of the expected upper bound. Even worse, Zone 3 has the most occupants as a staff working area, so such comfort violation shall be avoided.

Our MOCA well balances the tradeoff between energy and comfort through zone-based control. As seen from the table, the comfort bounds can be met for all the zones, i.e., 0.5. Note that high comfort index helps save energy in the tropics, so pushing to the maximum of the expected comfort bounds is energy-friendly. With such comfort performance, the resulting energy usage is 37.3kWh, much better compared to the baseline.

2) **Energy Saving of Each Zone:** In this part, we consider the comparison results with different comfort bounds. Due to the significant comfort violation of DDPG, we specifically compare MOCA and baseline here and the results are shown in Fig. 9.

Seen from the figure, the energy saving is significant with a relaxed comfort bound, and the saving is relatively minimal with tight bounds. For example, given the bound $[-0.1, 0.1]$, MOCA consumes 2.8% less energy compared to the baseline. When a large bound say $[-1.0, 1.0]$ is configured, the saving improves to 15.4%, 5.5x more than the case with tight bound

$[-0.1, 0.1]$. Consider our previous default comfort bound of $[-0.5, 0.5]$, MOCA achieves 8.5% energy saving on average. Different zones own different energy-saving potentials. Among them, Zone 1 shows the largest saving of almost 20% with bound $[-1.0, 1.0]$.

Moreover, we adopt the back-of-envelope calculation to estimate the annual electricity cost-saving under different thermal comfort bounds. As shown in Table II, the total energy consumption of baseline method is 40.7kWh. We refer to the energy-saving 8.5% with the default bound $[-0.5, 0.5]$. Let us assume 9 working hours a day, 22 working days a month, and 12 months a year. We have the annual energy saving 8,221kWh. According to the latest electricity tariff of 26 cents per kWh in Singapore, the money-saving can be a significant \$2,138 per year. Note that for some less critical areas, the comfort bound can be relaxed and the energy saving can be further improved, i.e., \$3,873 saving per year with bound $[-1.0, 1.0]$. Besides, our calculation is based on a small laboratory of only 301m². Given a large space or even the whole building, the energy-saving can be remarkable. Meanwhile, energy-saving also helps meet the national targets of carbon emission reduction.

In summary, compared to the baseline, our approach can learn the optimal control policy based on the heterogeneous thermal conditions of each zone and the comfort requirements. Accordingly, we can optimize energy-comfort and achieve significant energy expense reduction.

VI. CONCLUSION

In this paper, we investigate the thermal control for multiple zones based on the multi-agent DRL. We model each zone as an agent and design a framework to coordinate different agents and perform the HVAC controls. An optimization problem is formulated mathematically to minimize energy usage under a specified thermal demand. Then, we present MOCA, a multi-agent DRL-based algorithm, to optimize the energy-comfort tradeoff for multiple zones. We use professional TRNSYS software to build the multi-zone thermal control simulation environment to evaluate the performance of our solution. The simulation results demonstrate that our solution can achieve considerable electricity saving of the HVAC systems by up to 15.4% while meeting the thermal comfort requirement in multiple zones of a building.

In the future, we would like to extend this work to support intelligent demand response in smart buildings. We also want to look for opportunities to deploy and validate our solution in real buildings. Last but not least, the impact of the new technologies in machine learning and comfort modeling domains would be interesting to be investigated for the multi-zone thermal control.

REFERENCES

- [1] J. D. Chiara Delmastro and T. Abergel, "Cooling: Tracking clean energy progress," <http://www.iea.org/tcep/buildings/cooling/>, 2019, [Accessed on 14-April-2020].
- [2] J. Clifford and W. Stephenson, "An introduction to mechanical engineering," *Writing Culture: The Poetics and Politics of Ethnography*, 1975.
- [3] J. W. Moon, S.-H. Yoon, and S. Kim, "Development of an artificial neural network model based thermal control logic for double skin envelopes in winter," *Building and Environment*, vol. 61, pp. 149–159, 2013.
- [4] Z. Zhang, A. Chong, Y. Pan, C. Zhang, S. Lu, and K. P. Lam, "A deep reinforcement learning approach to using whole building energy model for hvac optimal control," in *2018 Building Performance Analysis Conference and SimBuild*, 2018.
- [5] 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*, 2020.
- [6] S. Nagarathinam, V. Menon, A. Vasan, and A. Sivasubramaniam, "Marco-multi-agent reinforcement learning based control of building hvac systems," in *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, 2020, pp. 57–67.
- [7] M. Manic, K. Amarasinghe, J. J. Rodriguez-Andina, and C. Rieger, "Intelligent buildings of the future: Cyberaware, deep learning powered, and human interacting," *IEEE Industrial Electronics Magazine*, vol. 10, no. 4, pp. 32–49, 2016.
- [8] A. Javed, H. Larijani, A. Ahmadiania, and D. Gibson, "Smart random neural network controller for hvac using cloud computing technology," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 1, pp. 351–360, 2016.
- [9] A. Afram, F. Janabi-Sharifi, A. S. Fung, and K. Raaheimifar, "Artificial neural network (ann) based model predictive control (mpc) and optimization of hvac systems: A state of the art review and case study of a residential hvac system," *Energy and Buildings*, vol. 141, pp. 96–113, 2017.
- [10] W. Zhang, W. Hu, and Y. Wen, "Thermal comfort modeling for smart buildings: A fine-grained deep learning approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2540–2549, 2018.
- [11] W. Hu, Y. Wen, K. Guan, G. Jin, and K. J. Tseng, "itcm: Toward learning-based thermal comfort modeling via pervasive sensing for smart buildings," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 4164–4177, 2018.
- [12] G. Gao, J. Li, and Y. Wen, "Deepcomfort: Energy-efficient thermal comfort control in buildings via reinforcement learning," *IEEE Internet of Things Journal*, 2020.
- [13] P. O. Fanger *et al.*, "Thermal comfort. analysis and applications in environmental engineering," *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.
- [14] R. Yang and L. Wang, "Multi-zone building energy management using intelligent control and optimization," *Sustainable cities and society*, vol. 6, pp. 16–21, 2013.
- [15] H. Hao, J. Lian, K. Kalsi, and J. Stoustrup, "Distributed flexibility characterization and resource allocation for multi-zone commercial buildings in the smart grid," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 3161–3168.
- [16] J. Cai, J. E. Braun, D. Kim, and J. Hu, "A multi-agent control based demand response strategy for multi-zone buildings," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 2365–2372.
- [17] W. Huang and H. Lam, "Using genetic algorithms to optimize controller parameters for hvac systems," *Energy and Buildings*, vol. 26, no. 3, pp. 277–282, 1997.
- [18] P. S. Curtiss, J. Kreider, and G. Shavit, "Neural networks applied to buildings—a tutorial and case studies in prediction and adaptive control," *American Society of Heating, Refrigerating and Air-Conditioning Engineers . . .*, Tech. Rep., 1996.
- [19] J. Schmidhuber, "Curious model-building control systems," in *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*. IEEE, 1991, pp. 1458–1463.
- [20] S. Wang and X. Jin, "Model-based optimal control of vav air-conditioning system using genetic algorithm," *Building and Environment*, vol. 35, no. 6, pp. 471–487, 2000.
- [21] T. Chen, "Real-time predictive supervisory operation of building thermal systems with thermal mass," *Energy and buildings*, vol. 33, no. 2, pp. 141–150, 2001.
- [22] Y. Ma, F. Borrelli, B. Hencsey, A. Packard, and S. Bortoff, "Model predictive control of thermal energy storage in building cooling systems," in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009, pp. 392–397.
- [23] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building hvac control," in *Proceedings of the 54th Annual Design Automation Conference 2017*. ACM, 2017, p. 22.
- [24] R. De Dear and G. S. Brager, "Developing an adaptive model of thermal comfort and preference," *ASHRAE Transactions*, vol. 104, pp. 145–167, 1998.

- [25] E. Donaisky, G. H. Oliveira, R. Z. Freire, and N. Mendes, "Pmv-based predictive algorithms for controlling thermal comfort in building plants," in *2007 IEEE International Conference on Control Applications*. IEEE, 2007, pp. 182–187.
- [26] R. Z. Freire, G. H. Oliveira, and N. Mendes, "Predictive controllers for thermal comfort optimization and energy savings," *Energy and buildings*, vol. 40, no. 7, pp. 1353–1365, 2008.
- [27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [28] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.
- [29] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser *et al.*, "Starcraft ii: A new challenge for reinforcement learning," *arXiv preprint arXiv:1708.04782*, 2017.
- [30] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, 2017, pp. 6379–6390.
- [31] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
- [32] W. Zhang, Y. Wen, K. J. Tseng, and G. Jin, "Demystifying thermal comfort in smart buildings: an interpretable machine learning approach," *IEEE Internet of Things Journal*, pp. 1–1, 2020.