DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings via Reinforcement Learning

Guanyu Gao, Jie Li, and Yonggang Wen, *Fellow, IEEE*

***Abstract*—Heating, ventilation, and air conditioning (HVAC) are extremely energy consuming, accounting for 40% of total building energy consumption. It is crucial to design some energy-efficient building thermal comfort control strategy which can reduce the energy consumption of the HVAC while maintaining the comfort of the occupants. However, implementing such a strategy is challenging, because the changes of the thermal states in a building environment are influenced by various factors. The relationships among these influencing factors are hard to model and are always different in different building environments. To address this challenge, we propose a deep-reinforcement-learning-based framework, DeepComfort, for thermal comfort control in buildings. We formulate the thermal comfort control as a cost-minimization problem by jointly considering the energy consumption of the HVAC and the occupants' thermal comfort. We first design a deep feedforward neural network (FNN)-based approach for predicting the occupants' thermal comfort and then propose a deep deterministic policy gradients (DDPGs)-based approach for learning the optimal thermal comfort control policy. We implement a building thermal comfort control simulation environment and evaluate the performance under various settings. The experimental results show that our approaches can improve the performance of thermal comfort prediction by 14.5% and reduce the energy consumption of HVAC by 4.31% while improving the occupants' thermal comfort by 13.6%.**

***Index Terms*—Deep-reinforcement learning (DRL), heating, ventilation and air conditioning (HVAC), smart building, thermal comfort control, thermal comfort prediction.**

I. Introduction

THERMAL comfort control in buildings is important for providing high-quality working and living environments, because feeling comfortable can directly impact the occupants' mood and productivity. However, the ambient thermal condition may change dramatically, leading to the fluctuation of the indoor thermal condition, which may cause the discomfort of the occupants. Therefore, thermal comfort control is necessary for maintaining the satisfactory indoor thermal condition. Heating, ventilation, and air conditioning (HVAC) systems are the main ways to control indoor thermal conditions.

One main concern of the HVAC system is the high energy consumption, which makes building energy consumption account for 20%–40% of the total energy consumption [1]. On the other hand, the occupants may feel too cold or too hot if the setpoints of the HVAC system are inappropriate, although more energy may be consumed. Thus, it is necessary to study how to reduce the energy consumption of the HVAC system while keeping occupants comfort, especially with the rising of the electricity price and the increasing of electricity consumption and environmental pollution.

We categorize the factors which may influence the thermal comfort control in buildings into three parts, namely: 1) the HVAC-related factors; 2) the building thermal environment-related factors; and 3) the human subject-related factors. The HVAC can control the building thermal condition by adjusting the setpoints of air temperature and humidity, which will also consequently change the energy consumption. The building thermal environment is determined by the physical structures of the building, the indoor and outdoor thermal conditions (e.g., weather), and the heat sources (e.g., bulbs and computers). The above two families of factors determine the indoor thermal condition, and the human subjects can be seen as a subjective thermal comfort evaluator of the indoor thermal condition [2]. The HVAC system takes appropriate control actions to meet the satisfaction of the human subjects. These factors are correlated, which requires to consider them as a whole to design a proper thermal comfort control policy.

Many approaches have been proposed for building thermal comfort control and energy optimization [3]. The model-based approaches aimed to model the thermal comfort control in buildings with simplified mathematical models, such as proportional–integral–derivative (PID) [4], [5], model predictive control (MPC) [6], [7], fuzzy control [6], [7], and linear-quadratic regulator [8]. However, the complexity of the building thermal environments and the relationships among various influencing factors is hard to be precisely modeled. Moreover, different building environments may require different modelings, it is hard to implement a generalized approach which can be directly applied in all building environments.

Some other works adopted the learning-based approaches to learn the optimal control policy, such as the reinforcement learning approach. The reinforcement learning approach can lean the optimal control policy by interactions

Manuscript received January 17, 2020; revised April 16, 2020; accepted April 22, 2020. Date of publication May 4, 2020; date of current version September 15, 2020. This work was supported in part by the Project Fund from DSAIR@NTU, and a BSEWWT Project Fund from National Research Foundation Singapore, administrated through the BSEWWT Program Office under Grant BSEWWT2017_2_06, and in part by the Green Buildings Innovation Cluster under Grant NRF2015ENC-GBICRD001-012, administered by Building and Construction Authority Singapore. *(Corresponding author: Guanyu Gao.)*

Guanyu Gao is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: ggao001@ntu.edu.sg).

Jie Li and Yonggang Wen are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: lijie@ntu.edu.sg; ygwen@ntu.edu.sg).

Digital Object Identifier 10.1109/JIOT.2020.2992117

with the thermal environment. However, the plain reinforcement learning approaches (e.g., [9]–[14]) cannot achieve high performance if the state–action space is large. Deep-reinforcement learning (DRL) approaches (e.g., [15] and [16]) have been recently adopted to address this limitation by using deep learning to learn the representation of the large state–action space. However, the methods adopted in [15] and [16] require discretization of the state–action space, which will also limit the performance of the thermal control policy.

In this article, we propose a deep-learning-based framework, DeepComfort, for the thermal comfort control in buildings. We first design a deep feedforward neural network (FNN) with Bayesian regularization for predicting the occupants' thermal comfort. The thermal comfort prediction will be used as a feedback for thermal comfort control. Then, we propose a deep deterministic policy gradients (DDPGs)-based approach [17] for learning the thermal comfort control policy. The control variables in thermal control (e.g., temperature and humidity) are all continuous, and DDPG is a natural solution for the continuous control problem, because it can avoid the discretization of the control variables and significantly improve the performances. We implement a building thermal comfort control simulation environment to evaluate the performances of our method. The main contributions of this article are as follows.

1) We propose a DDPG-based approach for learning the thermal comfort control policy in buildings. It can reduce the energy consumption by 4.31% and improve the thermal comfort by 13.6%. The proposed method converges faster and requires less training data.
2) We design a deep FNN with Bayesian regularization for predicting thermal comfort. Our method can improve the prediction accuracy by 14.5% in terms of mean-square error (MSE).
3) We implement a building thermal comfort control simulation environment and conduct extensive experiments to evaluate the performances under different settings.

The remainder of this article is organized as follows. Section II presents the preliminary and related works. Section III introduces the design and workflow of the thermal comfort control system. Section IV presents the system model and problem formulation. Section V presents the algorithms for predicting thermal comfort and learning the thermal comfort control policy. Section VI evaluates the performance of the proposed method. Section VII concludes this article.

II. Related Work

In this section, we first introduce the preliminary and then review the existing approaches for thermal comfort control.

A. Preliminary

1) HVAC: The main functionalities of HVAC include heating, ventilation, and air conditioning [18]. Specifically, heating is to generate heat to raise the air temperature in the building. Ventilation is to exchange air with the outside and circulate the air within the building. Air conditioning provides cooling and humidity control. In this article, we mainly study the setpoints and the energy consumption of HVAC. We do not consider the inside mechanisms of HVAC (i.e., how the setpoints are achieved by HVAC via its inside functioning). Different HVACs may have different inner mechanisms, and we design the algorithms independent of the inner mechanisms so that they can be applied in different types of HVACs.

2) Thermal Comfort: Thermal comfort reflects the occupants' satisfaction with the thermal condition [19]. To quantitatively evaluate thermal comfort, different thermal comfort models are introduced for predicting the occupants' satisfaction under different thermal conditions. Because the occupant's feeling about the thermal condition is subjective, thermal comfort is usually assessed through subjective evaluation. Many subjects will be invited to evaluate their degrees of satisfaction under different thermal conditions, such as cold (-2), cool (-1), neutral (0), warm (1), and hot (2). Then, different methods can be adopted to fit the data. Many approaches have been developed for evaluating the thermal comfort of the occupants under different thermal conditions [20], for instance, predicted mean vote (PMV), actual mean vote (AMV), predicted percentage dissatisfied (PPD), etc.

3) Reinforcement Learning: Reinforcement learning has been applied in many areas for intelligent control, e.g., autonomous vehicle, video streaming [21], [22], resource provisioning [23], etc. It is concerned with how the agent should take action in a dynamic environment to maximize the overall rewards [24]. The agent first observes the current state of the environment and selects an action to take, followed by obtaining the rewards for the action. These steps will be iterated during the learning stage and the control policy will be updated until it is converged. The optimal policy can be learned during the trails without directly modeling the system dynamics, therefore, reinforcement learning is a suitable solution if the policy can be learned by the interactions with the environment and the dynamics of the system is hard to model precisely. DRL [25] can significantly improve the performance compared with reinforcement learning if the dimensions of state and action are large, because deep learning has a higher capacity for learning the representation of large space and tremendous data.

B. Thermal Comfort Control in Buildings

The existing approaches for building thermal comfort control can be generally classified into the following two categories: 1) the model-based approaches derive the control policy by modeling the dynamics of the environment and 2) the learning-based approaches derive the control policy by learning from the interactions with the environment.

1) Model-Based Approaches: Levermore [4] and Dounis *et al.* [5] used the PID method for building energy management and indoor air quality control. Shepherd and Batty [6] and Calvino *et al.* [7] proposed a fuzzy control method for managing building thermal conditions and energy cost. Kummert *et al.* [26] and Wang and Jin [27] proposed the optimal control method for controlling the HVAC system. Ma *et al.* [28] introduced an MPC-based approach for controlling building cooling systems by considering thermal energy storage. Wei *et al.* [29] adopted the MPC-based

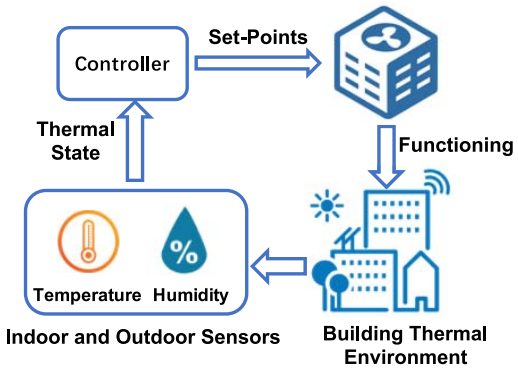


Fig. 1. Reference design of the building thermal comfort control system. The controller obtains the thermal state information from the sensors and makes control actions by adjusting the setpoints of the HVAC.

approach for jointly scheduling HVAC, electric vehicle, and battery usage for reducing building energy consumption while keeping the temperature within the comfort zone. Maasoumy *et al.* [8] proposed a tracking linear-quadratic regulator for balancing human comfort and energy consumption in buildings. Oldewurtel *et al.* [30] proposed a bilinear model under stochastic uncertainty for building climate control while considering weather predictions.

The model-based approaches strive to model the dynamics of the building thermal environment with some simplified mathematical models. However, the thermal environment is affected by various factors and complicated in nature, and the changing of the thermal condition is hard to be modeled precisely. Moreover, the model-based approaches are always designed for a specified building environment, and it is hard to derive a generalized model-based approach that is applicable in various building environments.

2) Learning-Based Approaches: Barrett and Linder [9], Li and Xia [10], and Nikovski *et al.* [11] adopted Q-learning-based approaches for the HVAC control. Zenger *et al.* [12] adopted state–action–reward–state–action (SARSA) for achieving the desired temperature while reducing energy consumption. Fazenda *et al.* [13] proposed a neural fitted reinforcement learning approach for learning how to schedule thermostat temperature setpoints. Yu *et al.* [31] proposed a DRL-based approach for energy management in smart homes. Dalamagkidis *et al.* [14] designed the linear reinforcement learning controller (LRLC) using linear function approximation of the state–action-value function to achieve thermal comfort with minimal energy consumption. Anderson *et al.* [32] proposed a robust control framework for combined proportional–integral (PI) control and reinforcement learning control for HVAC of buildings. Wei *et al.* [15] adopted a neural-network-based deep Q learning method for the HVAC control. Wang *et al.* [16] adopted a long short-term memory (LSTM) recurrent neural-network-based reinforcement learning controller for controlling the air conditioning system.

The tabular Q learning approach, SARSA, and other plain reinforcement learning approaches adopted in [9]–[14] and [32] are not suitable for problems with large state–action spaces, partly due to that plain reinforcement learning approaches fail to achieve satisfying generalization of the value function and policy function in large spaces.

Deep Q learning [15] and LSTM-based reinforcement learning [16] can improve the performances with neural networks, which have better generalization capacity. However, the proposed approaches in these works require discretization of the state–action space, which will decrease the control precision and performance. To fill the performance gap, we propose a DDPG-based approach for thermal comfort control in buildings. Our method also provides the capacity for customizing the thermal comfort settings. One can customize the thermal comfort threshold according to the occupants' thermal requirements to reduce energy consumption.

III. Thermal Comfort Control System Overview

In this section, we present the system design and the control flow for thermal comfort control in buildings.

A. Reference System Design

The reference design of the thermal comfort control system [33], [34] is illustrated in Fig. 1. The system mainly consists of the following components.

Sensors: The sensors periodically measure the thermal conditions of the indoor and outdoor building environments [35], including temperature, humidity, etc. The sensors are connected with the controller via Internet-of-Things (IoT) networks [36], [37], and the sensors will send the collected information to the controller for making thermal control decisions.

Controller: The controller collects the building thermal state information from the sensors and collects the energy consumption information from the HVAC [35]. Based on this information, the controller will make control actions by updating the setpoints of the HVAC periodically according to the thermal comfort control policy.

HVAC: The HVAC will function according to the setpoints updated by the controller. For instance, if the setpoint of the temperature is lower than the current indoor temperature, the HVAC will start cooling until the indoor temperature matches the setpoint temperature. If the setpoint temperature is higher than the current indoor temperature, the HVAC will start heating until achieving the specified indoor temperature.

B. Control Flow

We illustrate the control flow of the system in Fig. 2. We adopt deep FNN for predicting the occupants' thermal comfort given the current indoor thermal state, and we use DRL for making thermal comfort control decisions. The neural-network-based thermal comfort predictor can be trained offline using the existing thermal prediction data set. After training, the indoor building thermal state information will be input into the thermal prediction model for predicting thermal comfort. The thermal comfort prediction and the energy consumption information of the HVAC will be used for calculating the reward during each time slot. The DRL-based controller can learn the control policy by observing the rewards for taking actions on different states. We train the DRL-based controller

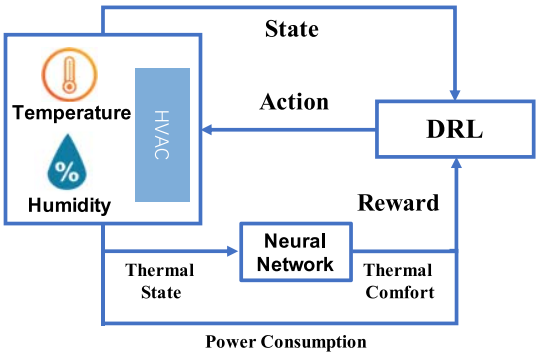


Fig. 2. Control flow of the building thermal comfort control system. The deep FNN is adopted for thermal comfort prediction and DRL is adopted for thermal comfort control.

using the building thermal comfort control simulation system (detailed in Section VI-A).

IV. System Model and Problem Formulation

In this section, we present the system models and problem formulation. We adopt a discrete-time model, where the time is denoted as $t = 0, 1, 2, \ldots$ The duration of each time slot is from several minutes to 1 h. The main notations used in this article are summarized in Table I.

TABLE I
Key Notation and Definition

t	the discrete time slot, $t = 0, 1, 2, ...$
T_t^{in}	the indoor air temperature at time slot t
H_t^{in}	the indoor air humidity at time slot t
T_t^{out}	the outdoor air temperature at time slot t
H_t^{out}	the outdoor air humidity at time slot t
T_t^{set}	the set-point air temperature at time slot t
H_t^{set}	the set-point air humidity at time slot t
M_t	predicted thermal comfort value at time slot t
Φ	thermal comfort prediction model
P_t	the energy consumption of the HVAC system at time slot t
S_t	the state at time slot t
A_t	the action taken at time slot t
R_t	the received reward at time slot t
π	the thermal control policy
γ	the discount factor for the reward
D_1, D_2	the threshold for thermal comfort
$g(\cdot), h(\cdot)$	the penalty function for thermal comfort
β	the weight of the penalty for energy consumption
α_1, α_2	Bayesian hyper-parameters
n	the number of training samples
m	the number of weights in the neural network
w_j	the j-th weight in the neural network
θ^Q, θ^μ	the parameters of the critic network and the actor network
$N(t)$	the exploration noise for training
τ	the discount factor for model update

A. Building Thermal State

The energy consumption of the HVAC is affected by both of the indoor thermal environment and the outdoor thermal environment. For the thermal state of the indoor and outdoor thermal environments, we consider the air temperature and humidity, which have the greatest influences on the energy consumption of the HVAC and the occupants' comfort. We denote the indoor air temperature and humidity at time slot t as T_t^{in} and H_t^{in}, and the outdoor air temperature and humidity at time slot t as T_t^{out} and H_t^{out}. The air temperature and humidity can be obtained by the sensors in a smart building.

B. Setpoints of HVAC

The controller can change the setpoints (e.g., air temperature and humidity) of the HVAC for adjusting the indoor thermal state. We denote the setpoint air temperature of the HVAC at time slot t as T_t^{set} and denote the setpoint air humidity as time slot t as H_t^{set}. At the beginning of each time slot, the controller updates the setpoint air temperature and humidity of the HVAC according to the indoor and outdoor thermal state to control thermal comfort and energy consumption.

C. Thermal Comfort Prediction

Thermal comfort prediction is to quantitatively evaluate the occupants' subjective satisfaction toward the thermal state. There are six primary factors which directly affect the occupants' thermal comfort, namely, metabolic rate, clothing insulation, air temperature, mean radiance, air speed, and humidity [38]. Air temperature and humidity can be easily measured in real time. For the other influencing factors, there are large variations from person to person even in the same environment. Specifically, metabolic rate and clothing insulation are personal factors and are determined by the characteristics of each individual occupant. The air speed and mean radiance for an occupant are determined by the structures of the building and the occupant's distance to the air outlet. For these factors, we can only estimate them from an average occupant's perspective, because it is impossible to measure these factors for each occupant in real time due to the privacy concern or the complexity of the implementation. Therefore, the indoor air temperature and humidity are considered as the time-varying variables which are directly influenced by the control actions of the HVAC [19]. The values of the other factors can be obtained by evaluating them from an average occupant's perspective and are considered as fixed for some durations (e.g., one season). Thus, we predict the occupants' thermal comfort value at time slot t as

$$M_t = \Phi\left(T_t^{\text{in}}, H_t^{\text{in}}\right) \tag{1}$$

where M_t is the predicted thermal comfort and Φ is the thermal comfort prediction model. We adopt the deep FNN-based method for predicting the occupants' thermal comfort (detailed in Section V-A).

D. Energy Consumption of HVAC

The HVAC will consume energy for heating, cooling, and dehumidification. The energy consumption of the HVAC during each time slot can be obtained from the smart meter. We denote the units of energy consumed by the HVAC measured in kW·h during time slot t as P_t, which can be obtained from the smart meter at the end of time slot t. We only consider the overall energy consumption of the HVAC during each time slot. The detailed energy consumptions for heating, cooling, and dehumidification are considered as unknown.

E. Problem Formulation

We formulate the energy optimization and thermal comfort control in buildings as a reinforcement learning problem which aims to maximize the overall rewards over time.

State: The state is the current indoor and outdoor thermal states at the beginning of each time slot. The state greatly influences the comfort of the occupants and the energy consumption of the HVAC. We denote the state as

$$S_t = \left(T_t^{\text{in}}, H_t^{\text{in}}, T_t^{\text{out}}, H_t^{\text{out}}\right) \tag{2}$$

where S_t is the state at time slot t.

Action: The control actions are the setpoints of the air temperature and humidity of the HVAC. The control action influences the indoor thermal state, the occupants' thermal comfort, and the energy consumption of the HVAC. We denote the control action as

$$A_t = \left(T_t^{\text{set}}, H_t^{\text{set}}\right) \tag{3}$$

where A_t is the action taken at time slot t. The control action is determined by the control policy and the current state. This relationship can be denoted as

$$A_t = \pi(S_t) \tag{4}$$

where π is the control policy for thermal comfort control.

Reward: The reward of DRL is a quantitative evaluation of the performances of the thermal comfort control policy. The performance metrics of thermal comfort control mainly include two components, namely: 1) the energy cost incurred by the HVAC and 2) the comfort level of the occupants. The goal of thermal comfort control is to improve the thermal comfort level of the occupants while reducing the energy cost. Therefore, thermal comfort and energy cost should be both considered when designing the reward function.

There are two main challenges for designing an appropriate reward function for thermal comfort control. First, the thermal comfort of the occupants and the energy cost incurred by the HVAC are of different types. However, the reward of DRL is 1-D. Thus, it needs to find an appropriate way to map the 2-D values into 1-D. Second, in different building environments, one may have different preferences on which component is more important. For instance, in some building environments, one may prefer a higher level of thermal comfort regardless of the energy cost. On the contrary, one may prefer to save more energy cost by sacrificing the comfort of the occupants.

One commonly adopted approach for addressing the above challenges is to define the reward as a weighted sum of the two parts. The weight represents the relative importance of the two components. It allows the reward to be configured according to the different preferences on each contributing component. In our case, the predicted thermal comfort value (M_t) ranges from -3 to 3, where -3 is too cold and $+3$ is too hot and 0 is neutral. The occupants can feel comfort when the predicted thermal comfort value is within an acceptable range. We denote the range as $[D_1, D_2]$, where D_1 and D_2 are the thresholds for thermal comfort and $D_1 < D_2$.

As illustrated in Fig. 3, if the thermal comfort value lies within $[D_1, D_2]$, it will not incur a penalty on thermal comfort, because the occupants feel comfortable. Otherwise, it will incur a penalty for the occupants' dissatisfaction with the building thermal state. Specifically, the occupants will feel too cold if the thermal comfort value is less than D_1 or too hot if the thermal comfort value is larger than D_2.

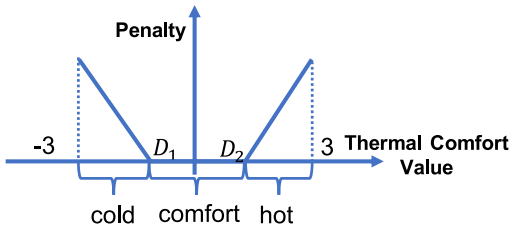


Fig. 3. Penalty on thermal comfort. If the thermal comfort value is less than D_1, the occupants will feel too cold. If it is larger than D_2, the occupants will feel too hot. $[D_1, D_2]$ is the comfort zone.

By jointly considering the occupants' thermal comfort M_t and the energy consumption of the HVAC P_t (defined in Section IV-D), we calculate the reward during time slot t as the weighted sum of the two components

$$R_t(S_t, A_t) = -\beta P_t - \begin{cases} 0 & D_1 < M_t < D_2 \\ M_t - D_2 & M_t > D_2 \\ D_1 - M_t & M_t < D_1 \end{cases} \tag{5}$$

where R_t is the reward for time slot t, and β is the weight of the energy consumption of the HVAC. The weigh β reflects the relative importance of energy consumption compared to the occupants' thermal comfort. The physical meaning of the equation is that the penalty for $(1/\beta)$ units (kW·h) of energy consumption is equal to the penalty for the deviation of the thermal comfort value from the thermal comfort thresholds by one. If the occupants' thermal comfort is more important, β should be set as a smaller value. Otherwise, β should be set as a larger value for achieving energy efficiency.

Optimization Objective: The objective of DRL is to maximize the overall discount rewards from the current time slot by deriving the optimal thermal comfort control policy. The objective function can be denoted as follows:

$$\max_{\pi} \sum_{t'=0}^{\infty} \gamma^{t'} R_{t+t'}\left(S_{t+t'}, A_{t+t'}\right) \tag{6}$$

where γ is the discount factor. If the precise information of the system dynamics is known, the optimal policy can be derived using value iteration or policy iteration through model-based approaches. However, it is impossible to obtain the transition probabilities in such a complex system. It motivates us to adopt the learning-based approach to learn the optimal control policy.

V. Algorithms for Thermal Comfort Control

In this section, we first introduce the neural network method for thermal comfort prediction. Then, we introduce the DDPG method for the learning thermal control policy.

A. Deep Neural Network for Thermal Comfort Prediction

We adopt the deep feedforward neural network for predicting thermal comfort. The structure of the neural network for predicting thermal comfort is illustrated in Fig. 4.

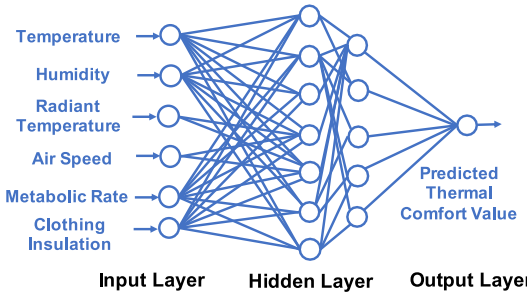


Fig. 4. Structure of the deep FNN for predicting thermal comfort. The inputs of the neural network include temperature, humidity, radiant temperature, air speed, metabolic rate, and clothing insulation. The output of the neural network is the predicted thermal comfort value.

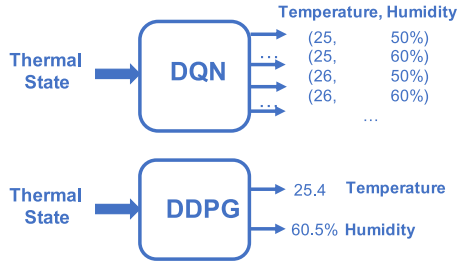


Fig. 5. Comparison of DQN and DDPG for thermal control. Temperature and humidity are continuous values, DDPG can be directly applied for thermal control, and DQN needs the discretization of the action space.

The inputs of the neural network include air temperature, humidity, mean radiant temperature, air speed, metabolic rate, and clothing insulation. All of these values are numerical. The hidden layer of the neural network has two layers, and the output layer has one neuron. The output of the neural network is the predicted thermal comfort value. The activation function of the hidden layer is a sigmoid function, and the activation function of the output layer is a linear function.

For training the neural network, the thermal comfort prediction data sets should be adopted as the training data. These data sets are labeled by the subjects for evaluating their thermal comfort levels under different thermal states. The labeled data can be noisy, due to the variations of the psychological and physiological characteristics of the occupants. To interpolate the noisy data, we adopt Bayesian regularization [39] to avoid overfitting. The cost function for training the neural network with Bayesian regularization is to minimize the training error using the minimal weights of the neural network

$$\phi = \alpha_1 \sum_{i=1}^{n} \left(Y_i - Y_i'\right)^2 + \alpha_2 \sum_{j=1}^{m} w_j^2 \tag{7}$$

where ϕ is the cost function, α_1 and α_2 are Bayesian hyperparameters for specifying the direction of the learning process to seek (i.e., minimize error or weights), n is the number of training samples, Y_i is the ith labeled value by the subject, Y_i' is the predicted value by the neural network, m is the number of weights in the neural network, and w_j is the jth weight.

The training algorithm is Levenberg–Marquardt backpropagation [40]. The convergence of the training algorithm will be validated in Section VI-C. The training process is performed offline. After training, the thermal comfort prediction model will be applied for thermal comfort prediction.

B. DDPG-Based Thermal Comfort Control

1) Why DDPG?: The setpoints of the HVAC, such as air temperature and humidity, are continuous. Deep Q network (DQN) is designed for handling problems with discrete and low-dimensional action spaces [17], therefore, it cannot be directly applied to the continuous control problem. Specifically, the optimal action for a given state in DQN must be obtained by enumerating the action-value of each action

$$a^* = \arg\max_{a \in A} Q(s, a) \tag{8}$$

where $Q(s, a)$ is the action-value of action a given state s, and A is the set of all actions. The optimal action a^* is the action which has the largest action-value given the current state s. Since there is an infinite number of actions for continuous actions, therefore, DQN cannot be directly applied.

Compared with DQN, DDPG is designed for the continuous control problem, and we can directly obtain the setpoints of the HVAC from the outputs of DDPG. We illustrate the comparisons of DQN and DDPG in Fig. 5. With DDPG, the network only has two outputs, namely, the setpoints of air temperature and humidity. To apply DQN in thermal comfort control, we need to discretize the action space, creating a finite number of control actions. However, if one wants to achieve finer grained dicretization, it may lead to an explosion of the number of actions. For instance, the range of humidity is from 0 to 100, if we discretize with the granularity of one, it will turn into 101 possible actions. Similarly, suppose that the range of the air temperature is from 15 to 35, and there will be 200 possible actions if the granularity is 0.1. The setpoints of the HVAC require one temperature value and one humidity value, and there are 20 200 possible combinations, resulting in 20 200 possible setpoints of the HVAC.

As illustrated in Fig. 5, in the implementation of DQN, each action is one possible combination of the temperature value and the humidity value. Each output of DQN represents one action, and the value of the output is the action-value of the corresponding action. The action which has the largest action-value will be chosen as the optimal control action. The temperature value and the humidity value of the optimal control action will be the setpoints of the HVAC. Therefore, the DQN network will have 20 200 outputs in the above case. It will require more training data to train the DQN network, and the dicretization of the action space will also decrease the performance. The detailed performance comparisons of DQN and DDPG will be presented in Section VI.

2) Training Methodology: DDPG adopts an actor–critic framework based on the deterministic policy gradient (DPG) [41]. We illustrate the network architecture of DDPG in Fig. 6. The actor network is denoted as $A_t = \mu(S_t|\theta^\mu)$, where S_t is the thermal state, θ^μ represents the weights of the actor network, and A_t represents the control action. The actor network maps the thermal state to a specific control action (i.e., the setpoints of the HVAC). The critic network is denoted as $Q(S_t, A_t|\theta^Q)$, where A_t is the specified control action by the actor network and θ^Q represents the weights of the critic

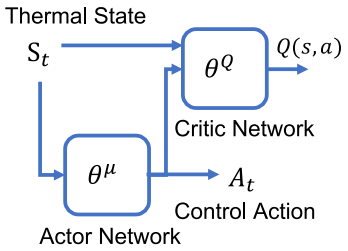


Fig. 6. Network architecture of DDPG. The actor network specifies a control action given the current thermal state and the critic network outputs an evaluation of the action generated by the actor network.

network. The action-value function $Q(S_t, A_t|\theta^Q)$ describes the expected reward by taking action A_t at state S_t by following the policy. The control action specified by the actor network will be used for selecting actions during the training. After training, only the actor network is required for making thermal comfort control actions.

The training for the DDPG network is performed by interacting with the building thermal environment. The training procedure is illustrated in Algorithm 1. At the beginning of each time slot t, we first obtain the current indoor and outdoor thermal state S_t, and input the thermal state into the policy network, which will output the control action. During the training, we need to explore the state space so that the policy will not converge to local optimal solutions. Therefore, we add a random noise to the obtained control action for exploration

$$A_t = \mu\left(S_t|\theta^\mu\right) + N(t) \tag{9}$$

where $N(t)$ is the exploration noise and A_t is the control action added with the exploration noise. In this article, we use an Ornstein–Uhlenbeck process [42] for generating the noise $N(t)$ for exploration. Then, the control action A_t will be applied to the HVAC. At the end of time slot t, we will obtain the new thermal state S_{t+1} and calculate the overall reward R_t during the time slot. The transition (S_t, A_t, R_t, S_{t+1}) will be stored in the replay buffer B for training the policy network and the actor network.

In each training episode, we will randomly sample N transitions from the replay buffer for training the network. For each transition $(S_i, A_i, R_i, S_{i+1}) \in N$, the estimated reward is calculated as follows:

$$R_i' = R_i + \gamma Q'\left(S_{i+1}, \mu'\left(S_{i+1}|\theta^{\mu'}\right)|\theta^{Q'}\right). \tag{10}$$

The critic network will be updated by minimizing the MSE between the estimated reward [R_i' calculated from (10)] and the reward predicted by the critic network [$Q(S_i, A_i)$] over the sampled minibatch, and the actor network will be updated using the sampled policy gradient [17]. Then, the target networks will be updated using the following equation:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \tag{11}$$

where τ is the discount factor for model update.

3) Complexity Analysis: The training for the DDPG model is performed offline. After training, only the actor network is required for making control action, and the computational complexity for making control actions is only determined by the actor network. The complexity of the actor network is determined by the number of hidden layers and the number of neurons in each hidden layer. In our case, the actor network has two hidden layers. Suppose that each hidden layer has n neurons, the complexity for making a control action is $O(n^2)$.

Algorithm 1 Training Thermal Comfort Control Policy

1: Initialize critic network $Q(S_t, A_t|\theta^Q)$ and actor network $\mu(S_t|\theta^\mu)$ with random weights θ^Q and θ^μ
2: Initialize target network $Q'(S_t, A_t|\theta^{Q'})$ and actor network $\mu'(S_t|\theta^{\mu'})$ with $\theta^{Q'} \leftarrow \theta^Q$ and $\theta^{\mu'} \leftarrow \theta^\mu$
3: Initialize replay buffer B
4: **for** episode = 0,1,...,M **do**
5: Obtain the initial thermal state S_0
6: **for** t = 0,1,...,T **do**
7: Obtain control action A_t according to (9)
8: Update the setpoints of the HVAC according to control action A_t
9: Obtain new thermal state S_{t+1} and calculate reward R_t according to Eq. (5) at the end of time slot t
10: Store (S_t, A_t, R_t, S_{t+1}) into replay buffer B
11: Randomly sample N transitions from replay buffer B
12: Calculate the estimated reward for each sampled transition using Eq. (10)
13: Update the critic network by minimizing the MSE over the sampled minibatch and update the actor network using the sampled policy gradient
14: Update target network Q' and μ' using Eq. (11)
15: **end for**
16: **end for**

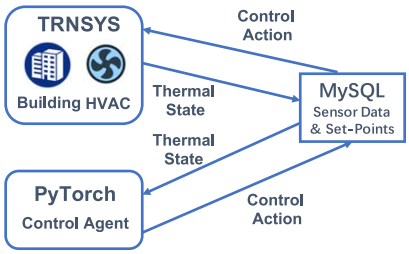


Fig. 7. Implementation of the thermal comfort control simulation system. TRNSYS is adopted for simulating the building thermal environment and the HVAC, and MySQL is adopted for storing sensor data and setpoints.

VI. Performance Evaluation

In this section, we first present the implementation of the thermal comfort control simulation system and then evaluate the performances of our proposed methods.

A. Simulation Environment

The thermal comfort control simulation environment is implemented to simulate the changing of the thermal state in a building and the energy consumption of the HVAC. The main components of the thermal comfort control simulation are illustrated in Fig. 7. We use TRNSYS [43] to simulate the HVAC and the thermal state in a building. The thermal comfort control algorithm and the thermal comfort prediction

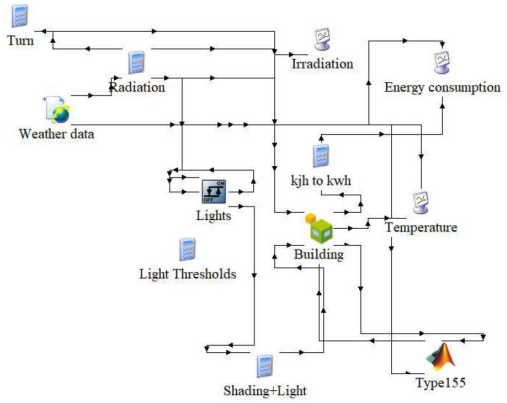


Fig. 8. Control diagram of building thermal comfort control simulation in TRNSYS. The thermal comfort control simulation environment simulates the energy consumption of the HVAC and building thermal states.

algorithm are implemented with PyTorch [44], which is an open-source machine learning library. The control diagram of the thermal simulation in TRNSYS is illustrated in Fig. 8. The thermal comfort control simulation environment in TRNSYS can only be programmatically controlled and accessed with MATLAB (Type 155 Module [45]), therefore, the control agent which is implemented with PyTorch cannot directly interact with TRNSYS. We use MySQL [46] as the pipeline for the interactions between the control agent and TRNSYS.

At the beginning of each time slot, the thermal state in the simulation environment will be read from TRNSYS and written into MySQL via the MATLAB interface. The control agent will read the thermal state information from MySQL and make a control action, which is the new setpoints of the HVAC. The setpoints will be written into MySQL, and TRNSYS will read the setpoints from MySQL using MATLAB interface and update the setpoints of the HVAC. TRNSYS will use the new setpoints of the HVAC and environment information (e.g., outside temperature and solar irradiation) to calculate the energy consumption of the HVAC and the indoor temperature and humidity. The control agent can improve the control policy during the iterations in the training process.

B. Experiment Setting

We simulate the building environment of a laboratory which is 307 m^2. It has 30 occupants and 40 computers with monitors. The air change rate is 0.67/h. The weather data set we use for simulation is SG-Singapore-Airp-486980, which is collected from Singapore. We use 10 000 h of the simulation data in TRNSYS for training the models and use 5000 h of data for testing the performances.

In our implementation, the actor network and the critic network of DDPG have two hidden layers, and each layer has 128 neurons. We use the tanh activation function and batch normalization in each layer. We adopt Adam for gradient-based optimization and the learning rate is 0.001. The discount factor τ for model update is 0.001 and the batch size is 128. The duration of each time slot is 30 min and each episode consists of 48 time slots. The default weight of energy cost is 0.05. The initial exploration noise scale is 0.7 and the final noise scale is 0.1, and the exploration noise decreases linearly over 300 episodes to 0.1. The default discount factor is 0.5.

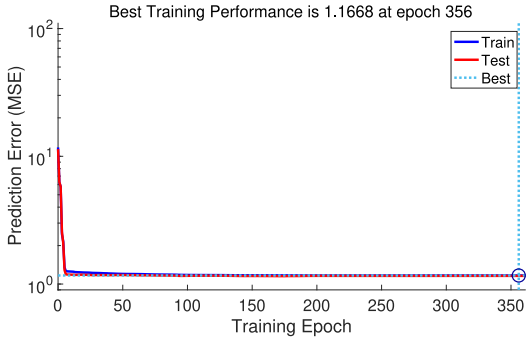


Fig. 9. Convergence of the deep neural network method for thermal comfort prediction. The prediction error can converge after several epochs of training, and the best training performance is 1.1668.

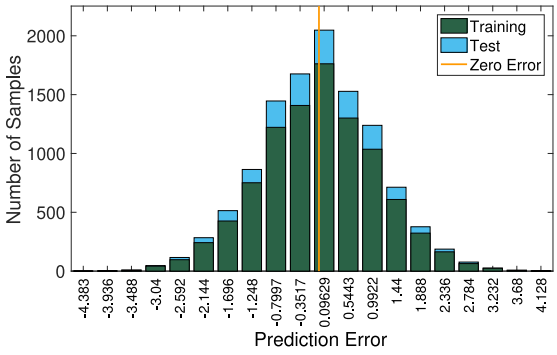


Fig. 10. Distribution of the prediction errors of the training and test samples. The prediction errors of most of the training samples and test samples are within the range $[-1, 1]$.

C. Performance of Thermal Comfort Prediction

We adopt the ASHRAE RP-884 thermal comfort data set [47] for training our model for thermal comfort prediction. We use 11 164 samples from the data set for training and testing, and each sample is a subject's evaluation of his/her thermal comfort level under a certain thermal state. The samples are randomly divided, 80% of the samples are used for training and 20% of the samples are used for testing. We compare the performance of our method with the following baselines, namely, linear regression (LR), support vector machine (SVM), Gaussian process regression (GPR), and ensemble regression. We evaluate the prediction errors of different methods using MSE.

We first illustrate the prediction error of our method during different training epochs in Fig. 9. The training error can converge after several epochs, and the best training performance is 1.1668 at epoch 356. The prediction error distribution of our method is illustrated in Fig. 10. From the test, we can observe that for most of the training samples and test samples, the prediction error is within $[-1, 1]$. The performance comparison of different methods is illustrated in Fig. 11. The MSE of our method is 1.1583, and the MSEs of LR, SVM, GPR, and ensemble regression are 1.3555, 1.4026, 2.1486, 1.4374, and 1.8145, respectively. The thermal comfort

TABLE II
THERMAL COMFORT PREDICTION PERFORMANCE IMPROVEMENT OF DNN COMPARED WITH BASELINES

	LR	SVM	RT	GPR	Ensemble
DNN	14.5%	17.4%	46.1%	19.4%	36.2%

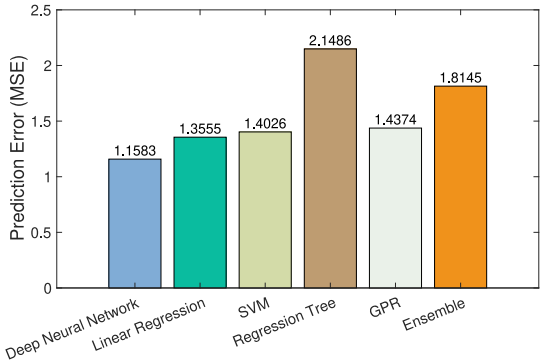


Fig. 11. Comparison of prediction errors of different methods. Our method can achieve smaller prediction error compared with the baselines.

prediction performance improvement of DNN compared with the other baseline methods is illustrated in Table II. DNN can improve the prediction performance by 14.5%, 17.4%, 46.1%, 19.4%, and 36.2% in terms of MSE compared with LR, SVM, GPR, and ensemble regression, respectively. The results verify that our method can predict the occupants' thermal comfort more precisely compared with the baselines.

D. Convergence Rates Under Different Settings

In this section, we analyze the convergence rates of the thermal comfort control algorithm under different settings. The comparisons among different settings are conducted using the same data without exploration noise.

Convergence Rates Under Different Learning Rates: We evaluate the convergence rates of the thermal comfort control algorithm under different learning rates, and the results are illustrated in Fig. 12. The algorithm can converge to the best performance when the learning rate is set as 0.001 and 0.0001. When the learning rate is too large (e.g., 0.01) and too small (e.g., 0.00001), the algorithm cannot converge to the optimal performances. The algorithm converges slightly faster when the learning rate is 0.0001 than 0.001.

Convergence Rates Under Different Hidden Sizes: We illustrate the convergence rates of the thermal comfort control algorithm under different hidden sizes in Fig. 13. In our implementation, the actor network and the critic network both have two hidden layers. We adjust the number of neurons in the hidden layers to observe the convergence rates of the algorithm. We can observe from Fig. 13 that a larger size of the hidden layers can increase the learning speed and the convergence rate of the algorithm. Specifically, the convergence rates with 512 neurons and 1024 neurons are faster than the convergence rates with 256 neurons and 128 neurons. This is because more neurons in the hidden layers can improve the learning capacities. However, when the number of neurons in the hidden layers exceeds 512, the convergence rates are no differences.

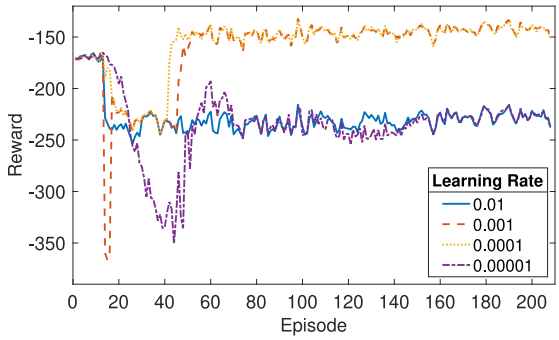


Fig. 12. Convergence rates of the thermal comfort control algorithm under different learning rates. The algorithm can achieve the best performances when the learning rate is set as 0.001 and 0.0001.

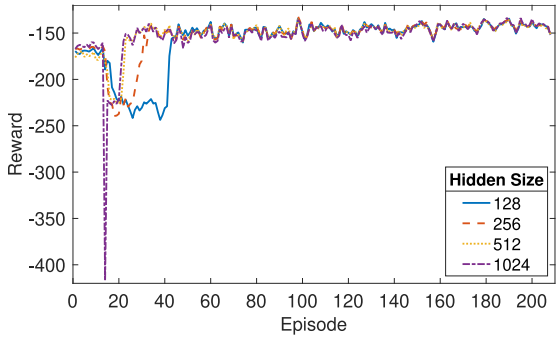


Fig. 13. Convergence rates under different hidden sizes. A larger number of neurons in the hidden layers can make the learning and convergence faster, but there will be no differences if the hidden size is larger than 512.

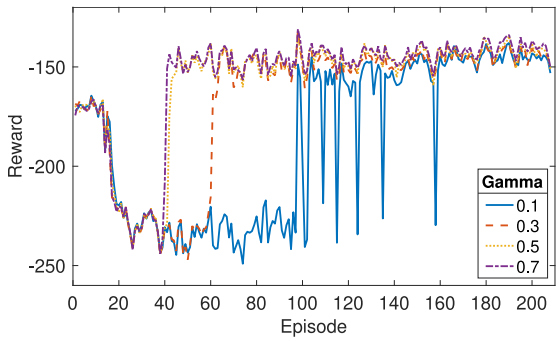


Fig. 14. Convergence rates under different discount factors. The algorithm converges faster with larger discount factors.

Convergence Rates Under Different Discount Factors: We illustrate the convergence rates of the thermal comfort control algorithm under different discount factors in Fig. 14. The discount factor γ represents the weight of the future rewards. In Fig. 14, we can observe that the algorithm converges faster with larger discount factors. Meanwhile, a larger discount factor can also slightly increase the reward. When the discount factor is larger than 0.5, the differences are not significant.

Convergence Rates Under Different Decay Rates: We illustrate the convergence rates of the thermal comfort control algorithm under different decay rates τ in Fig. 15. With larger decay rates, the algorithm will converge faster. However,

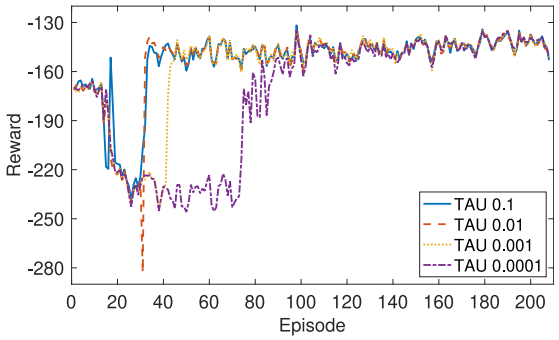


Fig. 15. Convergence rates under different decay rates. Larger decay rates can make the algorithm converge faster.

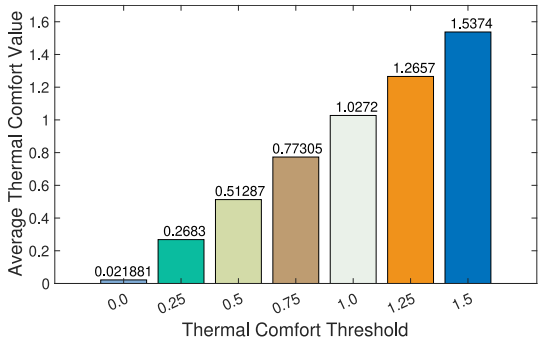


Fig. 16. Average thermal comfort values under different thermal comfort thresholds. The average thermal comfort values are close to the thresholds.

when the decay rate is larger than 0.01, the convergence rates under different decay rates are not significant. The difference between the rewards under different decay factors is not significant.

E. Performance Under Different Experiment Settings

In this section, we evaluate the performances of the DDPG-based thermal control method under different settings.

Performance Under Different Thermal Comfort Thresholds: With our method, one can set different thermal comfort thresholds according to the occupants' thermal comfort requirements. We evaluate the energy consumption of the HVAC and the distribution of the thermal comfort values under different thermal comfort thresholds. In Fig. 16, we illustrate the average thermal comfort values under different prescribed thermal comfort thresholds, and it can be observed that the average actual thermal comfort value measured from the building is close to the prescribed thresholds. In Fig. 17, we illustrate the distribution of the thermal comfort value of each time slot under different prescribed thermal comfort thresholds. It can be verified that the thermal comfort values in the indoor environment are closely centered around the prescribed thresholds. Therefore, our method can control the thermal comfort of the occupants precisely.

Fig. 18 illustrates the average cooling load of the HVAC under different thermal comfort thresholds. It will lead to less energy consumption of the HVAC if the thermal comfort threshold is set to a larger value. Therefore, if the occupants do not have stringent thermal comfort requirement, the thermal

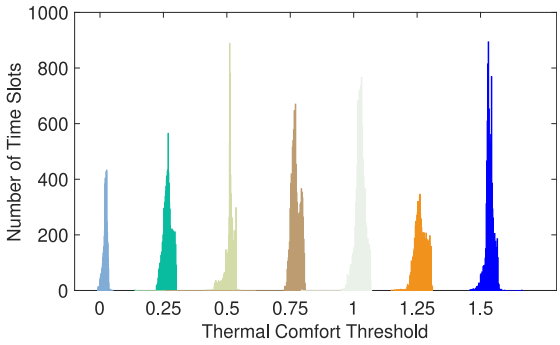


Fig. 17. Distribution of the thermal comfort values under different thermal comfort thresholds. The thermal comfort value of each time slot is closely centered around the prescribed thermal comfort threshold.

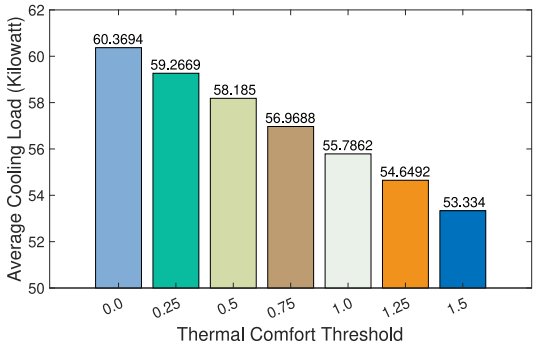


Fig. 18. Average cooling loads under different thermal comfort thresholds. Increasing the thermal comfort threshold can reduce energy consumption.

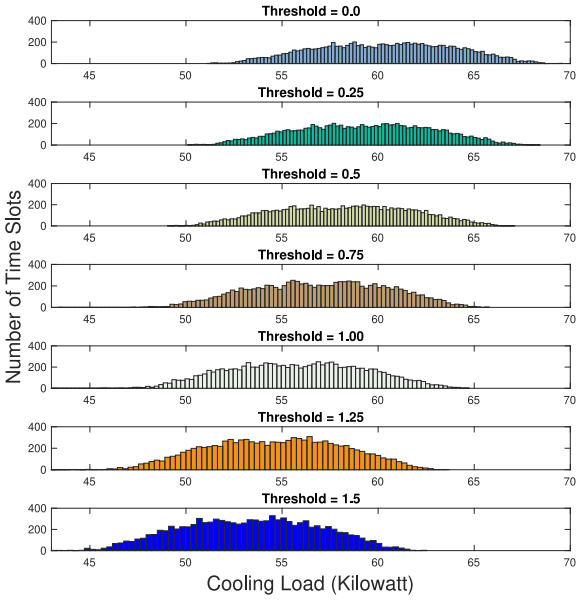


Fig. 19. Distribution of the cooling load of the HVAC under different thermal comfort thresholds. The distribution of cooling load moves to smaller values under a larger thermal comfort threshold.

comfort threshold can be set to a larger value for energy efficiency. We illustrate the distribution of the cooling load of the HVAC under different thermal comfort thresholds in Fig. 19. We can observe that the distribution of the cooling load moves toward small values when the thermal comfort threshold is

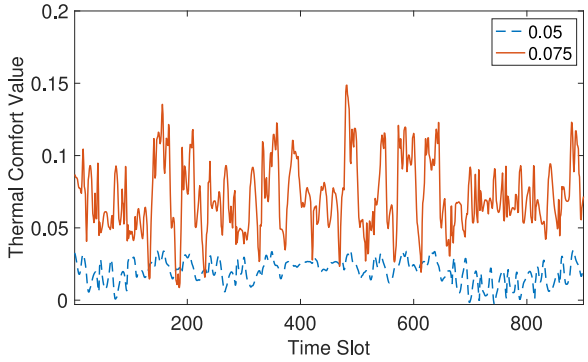


Fig. 20. Thermal comfort values under different weights of energy cost over time. If the weight of the energy cost is small, the thermal comfort value will be close to the threshold, and the changes of the value will be smaller.

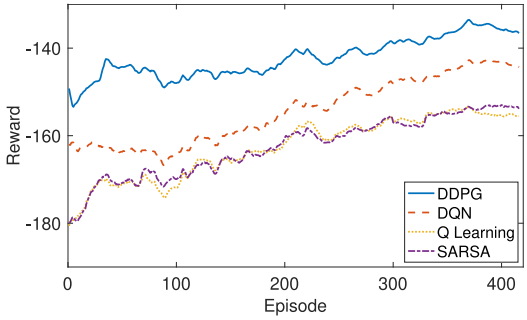


Fig. 21. Convergence of different algorithms. DDPG can converge faster and achieve a higher reward than the other baseline methods.

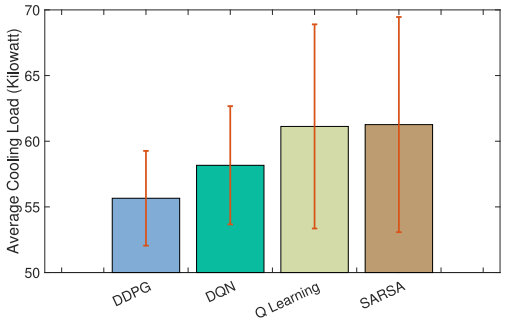


Fig. 22. Average cooling load of different methods. DDPG can achieve less energy consumption compared with the baseline methods. It can reduce the energy consumption by 4.31%, 8.95%, and 9.15% compared with DQN, Q learning, and SARSA-based approaches, respectively.

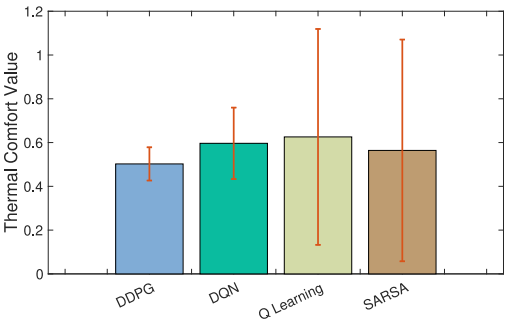


Fig. 23. Average thermal comfort values of different methods. DDPG can achieve higher thermal comfort compared with baseline methods. It can improve the occupants' thermal comfort by 13.6%, 17.6%, and 8.6% compared with DQN, Q learning, and SARSA-based approaches, respectively.

larger. Therefore, adjusting the thermal comfort threshold can control the energy consumption of the HVAC.

Performance Under Different Weights of Energy Cost: We can set different weights for the cost of the energy consumption in the reward function (5). If one puts more interest on the occupants' thermal comfort, the weight can be set smaller. On the contrary, if one puts more interest on energy cost, the weight can be set larger. Fig. 20 illustrates the changes of the thermal comfort value over time under different weights of energy cost. The thermal comfort threshold in the experiment is 0.0. We can observe that if the weight is small, the thermal comfort value will be close to the threshold, and the changes of the thermal comfort value will be small. This is because the penalty for violating the thermal comfort threshold is larger than the energy cost, due to the smaller weight of the energy cost. Therefore, the thermal comfort value will be kept close to the threshold for reducing the cost incurred by violating the thermal comfort threshold. On the contrary, if the weight of the energy cost is large, the thermal comfort value may fluctuate largely for reducing the energy consumption.

F. Performance Comparison With Different Methods

We compare the performance of our thermal control method (DDPG) with the following baseline methods, namely, Q learning, SARSA, and DQN. In the baselines, the temperature is discretized with the granularity of one centi-degree and the humidity is discretized with the granularity of 5%. We illustrate the convergences of different methods in Fig. 21. It can be observed that DDPG can achieve a faster convergence rate compared with the other baselines. This is because DDPG does not need the discretization of the action space and has fewer number of outputs of the network. Therefore, DDPG can learn the thermal comfort control policy more efficiently and require less training data. Moreover, DDPG can also achieve a higher overall reward compared with the baselines. This verifies that our method can achieve higher performances compared with the baselines. This is because Q learning and SARSA adopt the tabular methods for storing and updating the state–action values without generalization, and DQN needs the discretization of the action space.

We illustrate the average cooling load of different methods in Fig. 22. It can be observed that the average cooling load of our method is lower than the other baseline methods. Compared with DQN, Q learning, and SARSA-based approaches, our proposed DDPG-based approach can reduce the energy consumption by 4.31%, 8.95%, and 9.15%, respectively. Therefore, our method can achieve higher energy efficiency. We illustrate the average thermal comfort values of different methods in Fig. 23. The average thermal comfort value of DDPG is more close to the preset threshold 0.5 and lower than the baseline methods. Compared with DQN, Q learning, and SARSA-based approaches, our proposed DDPG-based approach can improve the occupants' thermal comfort by 13.6%, 17.6%, and 8.6%, respectively. Therefore, our method can achieve a higher degree of thermal comfort for the occupants compared with the other methods while consuming

less energy. The standard deviation (SD) of the thermal comfort value of DDPG is 0.07, compared to 0.16 of DQN, 0.49 of Q learning, and 0.50 of SARSA. This means that our method has a much smaller variation of the thermal comfort value and the occupants can feel more comfortable with our thermal comfort control method.

VII. Conclusion

In this article, we proposed a learning-based optimization framework for optimizing the occupants' thermal comfort and the energy consumption of the HVAC in buildings. We first designed a deep neural-network-based method with Bayesian regularization for thermal comfort prediction and then we adopted DDPG for controlling the HVAC for optimizing the energy consumption whiling meeting the occupants' thermal comfort requirements. We implemented a building thermal comfort control simulation system using TRNSYS and evaluated the performances under different settings. The results showed that our method can achieve better prediction performances for thermal comfort prediction, and it can achieve higher thermal comfort and energy efficiency compared with the baseline methods. In future works, we may consider using transfer learning to improve learning efficiency via the transfer of knowledge from different HVACs, or using model-based reinforcement learning [48] to improve the learning efficiency of thermal comfort control by modeling system dynamics.

References

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy Build.*, vol. 40, no. 3, pp. 394–398, 2008.

[2] S. P. Corgnati, M. Filippi, and S. Viazzo, "Perception of the thermal environment in high school and university classrooms: Subjective preferences and thermal comfort," *Build. Environ.*, vol. 42, no. 2, pp. 951–959, 2007.

[3] A. I. Dounis and C. Caraiscos, "Advanced control systems engineering for energy and comfort management in a building environment—A review," *Renew. Sustain. Energy Rev.*, vol. 13, nos. 6–7, pp. 1246–1261, 2009.

[4] G. J. Levermore, *Building Energy Management Systems: An Application to Heating and Control*. London, U.K.: E&FN Spon, 1992.

[5] A. I. Dounis, M. Bruant, M. Santamouris, G. Guarracino, and P. Michel, "Comparison of conventional and fuzzy control of indoor air quality in buildings," *J. Intell. Fuzzy Syst.*, vol. 4, no. 2, pp. 131–140, 1996.

[6] A. Shepherd and W. Batty, "Fuzzy control strategies to provide cost and energy efficient high quality indoor environments in buildings with high occupant densities," *Build. Services Eng. Res. Technol.*, vol. 24, no. 1, pp. 35–45, 2003.

[7] F. Calvino, M. La Gennusa, G. Rizzo, and G. Scaccianoce, "The control of indoor thermal comfort conditions: Introducing a fuzzy adaptive controller," *Energy Build.*, vol. 36, no. 2, pp. 97–102, 2004.

[8] M. Maasoumy, A. Pinto, and A. Sangiovanni-Vincentelli, "Model-based hierarchical optimal control design for HVAC systems," in *Proc. ASME Dyn. Syst. Control Conf. Bath/ASME Symp. Fluid Power Motion Control*, 2011, pp. 271–278.

[9] E. Barrett and S. Linder, "Autonomous HVAC control, a reinforcement learning approach," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Disc. Databases*, 2015, pp. 3–19.

[10] B. Li and L. Xia, "A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, 2015, pp. 444–449.

[11] D. Nikovski, J. Xu, and M. Nonaka, "A method for computing optimal set-point schedules for HVAC systems," in *Proc. REHVA World Congr. CLIMA*, 2013, pp. 1–12.

[12] A. Zenger, J. Schmidt, and M. Krödel, "Towards the intelligent home: Using reinforcement-learning for optimal heating control," in *Proc. Annu. Conf. Artif. Intell.*, 2013, pp. 304–307.

[13] P. Fazenda, K. Veeramachaneni, P. Lima, and U.-M. O'Reilly, "Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems," *J. Ambient Intell. Smart Environ.*, vol. 6, no. 6, pp. 675–690, 2014.

[14] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, and G. S. Stavrakakis, "Reinforcement learning for energy conservation and comfort in buildings," *Build. Environ.*, vol. 42, no. 7, pp. 2686–2698, 2007.

[15] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building HVAC control," in *Proc. 54th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, 2017, pp. 1–6.

[16] Y. Wang, K. Velswamy, and B. Huang, "A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems," *Processes*, vol. 5, no. 3, p. 46, 2017.

[17] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015. [Online]. Available: arXiv:1509.02971.

[18] F. McQuiston and J. Parker, *Heating, Ventilating, and Air Conditioning: Analysis and Design*. New Delhi, India: Wiley, 2018.

[19] *ASHRAE STANDARD: An American Standard: Thermal Environmental Conditions for Human Occupancy*, Amer. Soc. Heating refrigeration Air Conditioning Eng., Atlanta, GA, USA, 1992.

[20] Y. Cheng, J. Niu, and N. Gao, "Thermal comfort models: A review and numerical investigation," *Build. Environ.*, vol. 47, pp. 13–22, Jan. 2012.

[21] G. Gao, L. Dong, H. Zhang, Y. Wen, and W. Zeng, "Content-aware personalised rate adaptation for adaptive streaming via deep video analysis," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–8.

[22] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan, "DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction," *IEEE Trans. Multimedia*, early access, Feb. 14, 2020, doi: 10.1109/TMM.2020.2973828.

[23] G. Gao, H. Hu, Y. Wen, and C. Westphal, "Resource provisioning and profit maximization for transcoding in clouds: A two-timescale approach," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 836–848, Apr. 2017.

[24] M. van Otterlo and M. Wiering, "Reinforcement learning and Markov decision processes," in *Reinforcement Learning*. Heidelberg, Germany: Springer, 2012, pp. 3–42.

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

[26] M. Kummert, P. André, and J. Nicolas, "Optimal heating control in a passive solar commercial building," *Solar Energy*, vol. 69, no. 6, pp. 103–116, 2001.

[27] S. Wang and X. Jin, "Model-based optimal control of VAV air-conditioning system using genetic algorithm," *Build. Environ.*, vol. 35, no. 6, pp. 471–487, 2000.

[28] Y. Ma, F. Borrelli, B. Hencey, B. Coffey, S. Bengea, and P. Haves, "Model predictive control for the operation of building cooling systems," *IEEE Trans. Control Syst. Technol.*, vol. 20, no. 3, pp. 796–803, May 2012.

[29] T. Wei, Q. Zhu, and M. Maasoumy, "Co-scheduling of HVAC control, EV charging and battery usage for building energy efficiency," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design*, 2014, pp. 191–196.

[30] F. Oldewurtel *et al.*, "Energy efficient building climate control using stochastic model predictive control and weather predictions," in *Proc. Amer. Control Conf.*, 2010, pp. 5100–5105.

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

[32] C. W. Anderson *et al.*, "Robust reinforcement learning for heating, ventilation, and air conditioning control of buildings," in *Handbook of Learning and Approximate Dynamic Programming*. Hoboken, NJ, USA: IEEE Press, 2004, pp. 517–534.

[33] M. Feldmeier and J. A. Paradiso, "Personalized HVAC control system," in *Proc. Internet Things (IoT)*, 2010, pp. 1–8.

[34] A. Lifson and M. F. Taras, "Control of conditioned environment by remote sensor," U.S. Patent App. 12/746,432, Sep. 30, 2010.

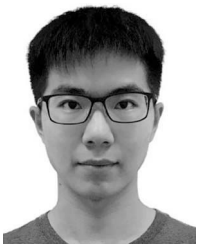
[35] J. Pan, R. Jain, S. Paul, T. Vu, A. Saifullah, and M. Sha, "An Internet of Things framework for smart energy in buildings: Designs, prototype, and experiments," *IEEE Internet Things J.*, vol. 2, no. 6, pp. 527–537, Dec. 2015.

[36] 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 Things J.*, vol. 4, no. 1, pp. 269–283, Feb. 2017.

[37] 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 Things J.*, vol. 5, no. 5, pp. 4164–4177, Oct. 2018.
[38] P. O. Fanger *et al.*, *Thermal Comfort. Analysis and Applications in Environmental Engineering*. Malabar, FL, USA: Krieger, 1970.
[39] F. D. Foresee and M. T. Hagan, "Gauss–Newton approximation to Bayesian learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 3, 1997, pp. 1930–1935.
[40] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 989–993, Nov. 1994.
[41] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. ICML*, 2014, pp. 1–9.
[42] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," *Phys. Rev.*, vol. 36, no. 5, p. 823, 1930.
[43] *TRNSYS: Transient System Simulation Tool*. Accessed: Oct. 20118. [Online]. Available: http://www.trnsys.com
[44] *Pytorch: An Open Source Deep Learning Platform*. Accessed: Oct. 2018. [Online]. Available: https://pytorch.org
[45] *Type 155: TRNSYS-MATLAB Link*. Accessed: Jun. 2019. [Online]. Available: http://sel.me.wisc.edu/trnsys/trnlib/trnsys-matlab/type155-manual.html
[46] *MySQL: The World's Most Popular Open Source Database*. Accessed: Oct. 2018. [Online]. Available: https://www.mysql.com
[47] R. J. De Dear, "A global database of thermal comfort field experiments," *ASHRAE Trans.*, vol. 104, p. 1141, Oct. 1998.
[48] L. Dong, G. Gao, Y. Li, and Y. Wen, "Baconian: A unified opensource framework for model-based reinforcement learning," 2019. [Online]. Available: arXiv:1904.10762.

Guanyu Gao received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2012, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2017.

He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include multimedia networking, edge/cloud computing, resource scheduling, and deep-reinforcement learning.



Jie Li is currently pursuing the M.Eng. degree with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

His research interests include smart building, energy trading on blockchain, and deep learning.

Yonggang Wen (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science (minor in Western Literature) from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008.

He is a Professor of computer science and engineering with Nanyang Technological University (NTU), Singapore, where he has been serving as an Associate Dean (Research) with the College of Engineering since 2018. He was the Acting Director with Nanyang Technopreneurship Centre, NTU, from 2017 to 2019, and an Assistant Chair (Innovation) with the School of Computer Science and Engineering from 2016 to 2018. He led product development in content delivery network with Cisco, San Jose, CA, USA, which had a revenue impact of $3 billion globally. He has worked extensively in learning-based system prototyping and performance optimization for large-scale networked computer systems. His work in Multiscreen Cloud Social TV has been featured by global media (more than 1600 news articles from over 29 countries). His research interests include cloud computing, green data center, distributed machine learning, blockchain, big data analytics, multimedia network, and mobile computing.

Prof. Wen is a co-recipient of the Best Transaction Paper Awards for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2019 and *IEEE Multimedia Magazine* in 2015, the Best Conference Paper Awards at IEEE Globecom in 2016, IEEE Infocom MuSIC Workshop in 2016, EAI/ICST Chinacom in 2015, IEEE WCSP in 2014, and IEEE Globecom in 2013. He received the IEEE ComSoc MMTC Distinguished Leadership Award in 2016 and the ASEAN ICT Awards (Gold Medal) in 2013. His recent work on Cloud3DView, as the only academia entry, has won ASEAN ICT Awards (Gold Medal) in 2016 and the Datacentre Dynamics Awards—APAC ("Oscar" Award of data centre industry) in 2015. He is the sole winner of Nanyang Awards in Innovation and Entrepreneurship at NTU in 2016. He serves on editorial boards for multiple transactions and journals, including the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *IEEE Wireless Communication Magazine*, IEEE COMMUNICATIONS SURVEY & TUTORIALS, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, IEEE ACCESS, and *Ad Hoc Networks* (Elsevier). He was elected as the Chair for the IEEE ComSoc Multimedia Communication Technical Committee from 2014 to 2016.