Intelligent Thermal Control Algorithm Based on Deep Deterministic Policy Gradient for Spacecraft

Yan Xiong,* Liang Guo,† Hongliang Wang,‡ Yong Huang,§ and Chunlong Liu¶
Chinese Academy of Sciences, 130033 Changchun, People's Republic of China

https://doi.org/10.2514/1.T5951

In practical applications, it is difficult for the control parameters of the proportional integral derivative (PID) thermal controller to be self-tuning online. As the control object or environment changes, the control parameters are required to change accordingly. An intelligent thermal controller based on the deep deterministic policy gradient, called DRLTC, is proposed. Two types of reinforcement learning agents were designed in DRLTC, which can automatically adjust the control parameters of the thermal controllers and self-optimize online after training. Both theoretical and experimental results revealed that, when the control object was the main mirror support, the DRLTC achieved a control precision of 0.01°C. Additionally, the steady-state error was reduced by 40.2, 62.5, and 33.3% in the simulation and by 5.6, 80.6, and 85.7% in the experiment, compared with the reinforcement learning PID, neural network PID, and adaptive PID control based on fuzzy control, respectively. When the control object was changed to the main mirror installation, the DRLTC achieved a control precision of 0.02°C, and the steady-state error was reduced by 87.5, 91.7, and 90.9% in the simulation and by 80.2, 90.6, and 85.7% in the experiment, compared with the above-mentioned thermal control strategies, respectively. Therefore, the DRLTC has better universality, has stronger robustness, and saves more energy.

Nomenclature

a_{MLI}	=	absorption coefficient of multilayer
c_i	=	specific heat capacity of node i, J/(kg · K)
c_{MMI}	=	specific heat capacity of main mirror installation, J/(kg · K)
c_{MMS}	=	specific heat capacity of main mirror support, J/(kg · K)
D_{ji}	=	heat transfer coefficient from node j to node i
$e(t)$	=	system error at time t, °C
G_{ji}	=	radiative heat transfer coefficient from node j to node i
$I(t)$	=	adaptive compensation of proportional integral derivative output current, A
$K(t)$	=	vector of proportional integral derivative parameters selected by intelligent thermal controller based on deep deterministic policy gradient for space telescope
k_{MMI}	=	thermal conductivity of main mirror installation, W/(m · K)
k_{MMS}	=	thermal conductivity of main mirror support, W/(m · K)
M_a	=	memory capacity of actor
M_c	=	memory capacity of critic
m_i	=	quality of node i, kg
m_{MMI}	=	quality of main mirror installation, kg
m_{MMS}	=	quality of main mirror support, kg
N_a	=	batch size of actor
N_c	=	batch size of critic
P_{atm}	=	atmospheric pressure, kg/m^3
P_{Heat}	=	heating power, W
Q_i	=	external heat flux sources of node i, W/m^2
q_i	=	internal heat power for node i, W
$R(t)$	=	cumulative reward from external environment
$r(t)$	=	reward function for training of neural network
S_a	=	update steps for actor
S_c	=	update steps for critic
$S(t)$	=	a set of reinforcement learning agent states at time t
T	=	sampling period, s
T_i	=	thermal dynamic temperature value of node i, °C
T_{inside}	=	temperature inside incubator, °C
T_{outside}	=	temperature outside incubator, °C
V_{MMI}	=	surface volume of main mirror installation, m^3
V_{MMS}	=	physical volume of main mirror support, m^3
$V(t)$	=	estimated value function at time t
$X(t)$	=	temperature at time t, °C
α	=	learning rate
γ	=	discount factor
δ_{TD}	=	temporal difference
ϵ_{MLI}	=	emission coefficient of main mirror installation
ϵ_{MLS}	=	emission coefficient of main mirror support
μ	=	action value
θ	=	weight of neural network
ρ_{MMI}	=	density of main mirror installation material
ρ_{MMS}	=	density of main mirror support material
τ	=	thermal control strategy
ω	=	policy parameters

Received 5 November 2019; revision received 6 February 2020; accepted for publication 16 March 2020; published online 8 May 2020.

*Ph.D. Candidate, Thermal Control Group, Changchun Institute of Optics, Fine Mechanics and Physics; also University of Chinese Academy of Sciences, 100049 Beijing, People's Republic of China; xiongyan16@mails.ucas.ac.cn.

†Associate Professor, Thermal Control Group, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences; also University of Chinese Academy of Sciences, 100049 Beijing, People's Republic of China; guoliang@ciomp.ac.cn (Corresponding Author).

‡Master's Degree Candidate, Thermal Control Group, Changchun Institute of Optics, Fine Mechanics and Physics; also University of Chinese Academy of Sciences, 100049 Beijing, People's Republic of China; wfaqwhl@126.com.

§Associate Professor, Thermal Control Group, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences; also University of Chinese Academy of Sciences, 100049 Beijing, People's Republic of China; huangyong@ciomp.ac.cn.

¶Associate Professor, Thermal Control Group, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences; also University of Chinese Academy of Sciences, 100049 Beijing, People's Republic of China; 6033610432lcl@163.com.

I. Introduction

TEMPERATURE is the main factor affecting the performance of spacecraft, particularly in the case of the high-resolution low-temperature detector. To ensure the long-term stable operation of a space telescope, a correct thermal control strategy is key for controlling the temperature of space telescopes within an acceptable range. With the improvement of space telescope resolution, the thermal control requirements are increasingly becoming stricter [1,2]. Some space telescopes have a temperature control accuracy of one thousandth of a degree Celsius, but higher precision is required.

Presently, there are many methods for achieving high-precision temperature control. The proportional integral derivative (PID) thermal control is widely used in space telescopes owing to its applicability to static and dynamic characteristics, simple implementation, and high robustness. The control parameters of a traditional PID thermal controller are difficult to adjust online and apply in practice. Many scholars have proposed several PID parameter tuning strategies, such as the fuzzy adaptive PID control (FuzzyPID) [3,4], neural network adaptive PID control (BPPID) [5], evolutionary algorithm adaptive PID control [6], among others.

Lemmen et al. [7] proposed a self-tuning active thermal control system to satisfy the ±0.05°C temperature precision for the optical assembly of Envisat. Zhu et al. [8] proposed a precision thermal control algorithm for an infrared detector based on bang–bang and PID thermal control. Choi [9–11] proposed an adaptive PID thermal control algorithm applied to the Swift satellite thermal controller, which is mainly used for the active thermal control of x-ray telescopes and can achieve a control precision higher than 0.1°C. The key parameters of the PID determine the precision of thermal control performance based on the PID thermal control. However, when the system parameters change, because the thermal control system is often disturbed during operation, problems such as time delay and nonlinearity are encountered, whereby the controller output produces jitter or even fails [12–14]. Xiong et al. [15] proposed an intelligent thermal control strategy based on reinforcement learning (RLPID) for space telescopes. This control method has fast-dynamic response and small amplitude when the system parameters of the thermal control object change. However, the algorithm is slow and difficult to converge, and even produces large steady-state error when the control object changes.

To solve the above mentioned problems and realize high-precision temperature control for space telescopes, this paper proposes DRLTC, which is an intelligent thermal control strategy based on the deep deterministic policy gradient combined with PID control for space telescopes. The deep deterministic policy gradient (DDPG) algorithm was proposed by Lillicrap and is an important branch of reinforcement learning [16,17]. DDPG [18] is a data-driven control method that can learn the mathematical model of the system and realize the optimal control of the system according to the system's input and output data, and control the error of the thermal control system based on the construction of the reward function. Therefore, the control parameters of a space telescope's thermal control system can be automatically adjusted using the reinforcement learning method. This is the first-time that deep reinforcement learning has been applied to a space telescope thermal control system. This paper presents the design of two reinforcement learning agents with the same structure. The first agent type can automatically adjust the PID parameters online. The second reinforcement learning agent takes corresponding action according to previous learning experience in the case of an unknown disturbance and unknown temperature curve in the input and output data of the learning system. The results obtained by this study revealed that the novel thermal control strategy performs better than the RLPID, BPPID, and FuzzyPID.

The rest of this paper is organized as follows. Section II presents the thermophysical model of a space telescope established with Simscape in MATLAB. Section III presents the DRLTC process. Sections IV and V describe the simulation and experimental results, respectively. Finally, the conclusions drawn from this study are presented in Sec. VI.

II. Thermophysical Model Analysis of Space Telescopes

Before DRLTC can be applied to the thermal control of space telescopes, it is necessary to analyze the space telescope's thermophysical model, which is very complex. The node network method [19–21] is mainly used for modeling. The space telescope can be divided into various finite elements according to its characteristics, and each element is considered as an isothermal body and used as a node. The following heat balance equation is applied to each node:

$$m_i c_i \frac{dT_i}{dt} = Q_i + q_i + \sum_{j=1}^{N} D_{ji}(T_j - T_i) + \sum_{j=1}^{N} G_{ji}(T_j^4 - T_i^4) \quad (1)$$

where i and j are nodes; T_i is the thermal dynamic temperature value of node i; c_i is the specific heat capacity of node i; m_i is the quality of node i; t is time; G_{ji} is the radiative heat transfer coefficient from node j to node i; D_{ji} is the heat transfer coefficient from node j to the node i; Q_i denotes the external heat flux sources of node i; and q_i is the internal heat power for node i.

Based on the node network method, the space telescope installation box is divided into six planes and named as B_Face, D_Face, R_Face, E_Face, L_Face and T_Face, respectively, as shown in Fig. 1a, and each plane is divided into 24 cells, as shown in Fig. 1b is named LvBan. In addition, JRP is polyimide heating tablets and MLI is multilayer insulating material. Based on the network method, and using Simscape in MATLAB and Simulink [22,23], unit T41 was further divided into four units to build a thin plate unit body thermal model, as shown in Fig. 1c. After many validation cycles, the error between the thermophysical model established with Simscape in MATLAB and the finite element model constructed in UG/TMG was within 5%.

III. Thermal Control Strategy Based on DRLTC

A. Architecture of DRLTC Thermal Controller

Figure 2 shows the structure of the proposed DRLTC thermal controller based on the DDPG and PID. Because the thermal control of a space telescope can calculate the control amount only according to the deviation value of sampling, the continuous PID control algorithm cannot be directly used, and the discretization method is required. Therefore, the positional PID was adopted in this study [24] based on the concept of combining the DRLTC and discrete position PID controller expressed by Eq. (2).

$$\begin{aligned} u(t) &= K(t)x(t) + I(t) \\ &= k_P(t)x_1(t) + k_T(t)x_2(t) + k_D(t)x_3(t) + I(t) \\ &= k_P(t)e(t) + k_I(t)\sum_{j=0}^{k} e(j)Ts + k_D(t)\frac{e(t) - e(t-1)}{Ts} + I(t) \\ &= k_P(t)\left(e(t) + \frac{Ts}{Ts_I}\sum_{j=0}^{k} e(j) + \frac{Ts_D}{Ts}(e(t) - e(t-1))\right) + I(t) \end{aligned} \quad (2)$$

where

$$k_I = \frac{k_P}{Ts_I} \quad (3)$$

$$k_d = k_P Ts_D \quad (4)$$

$$t \approx kTs \quad (5)$$

$$\int_0^t e(t)\,\mathrm{d}t \approx Ts\sum_{j=0}^{k} e(jTs) = Ts\sum_{j=0}^{k} e(j) \quad (6)$$

$$\frac{de}{dt} \approx \frac{e(kTs) - e((k-1)Ts)}{Ts} = \frac{e(t) - e(t-1)}{Ts} \quad (7)$$

$$\begin{aligned} I(t) &= R(f) \\ &= R\left(\exp\left(\frac{2t^2}{\sum_{t=0}^{l}(e(t) - e(t-1))^2}\right)\right) \end{aligned} \quad (8)$$

where Ts is the sampling period; k is the sampling number, $k = 0, 1, 2, 3, \ldots$; $e(k-1)$ and $e(k)$ are the system error signals

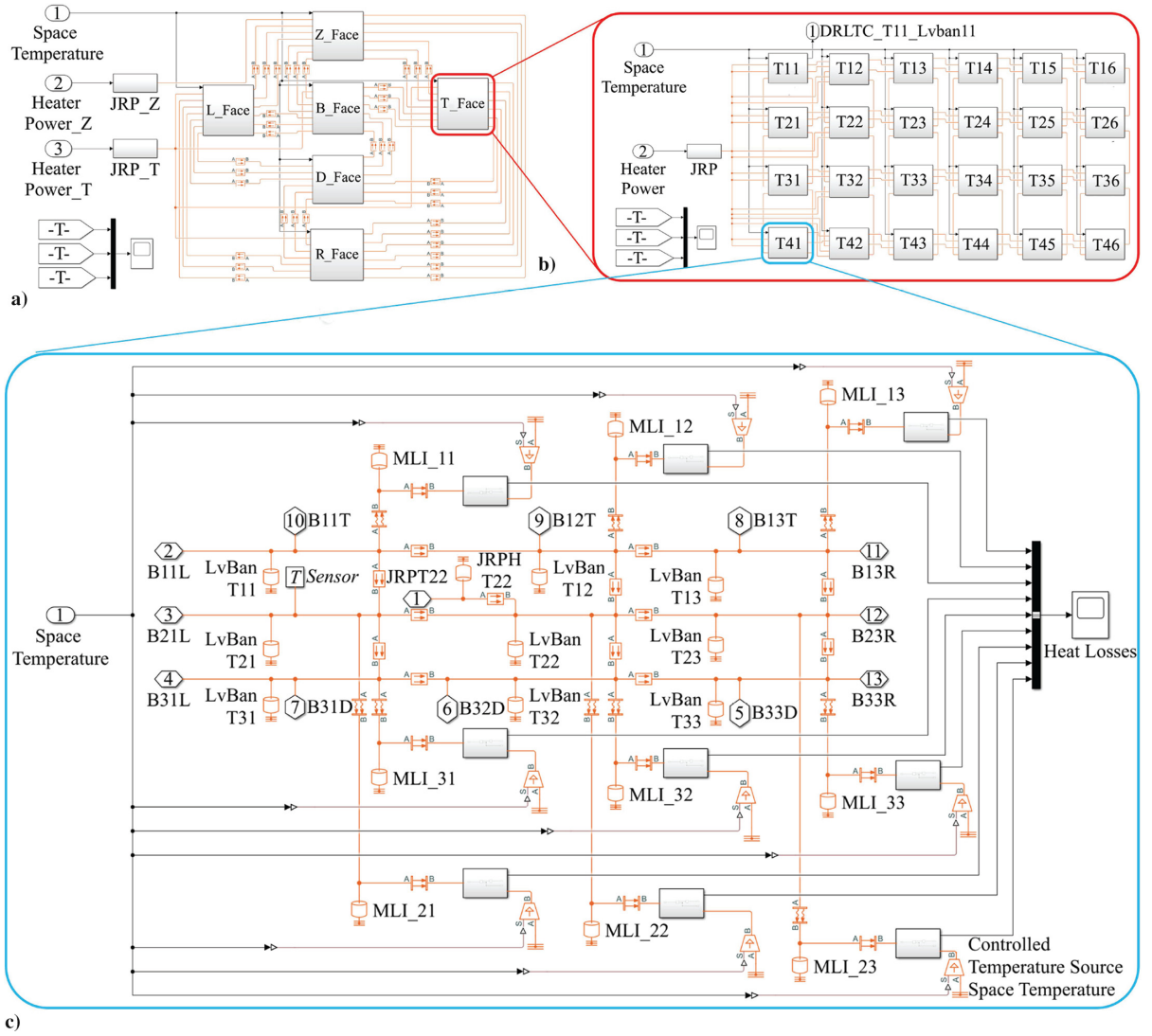

Fig. 1 Thermophysical model of space telescopes in Simulink.

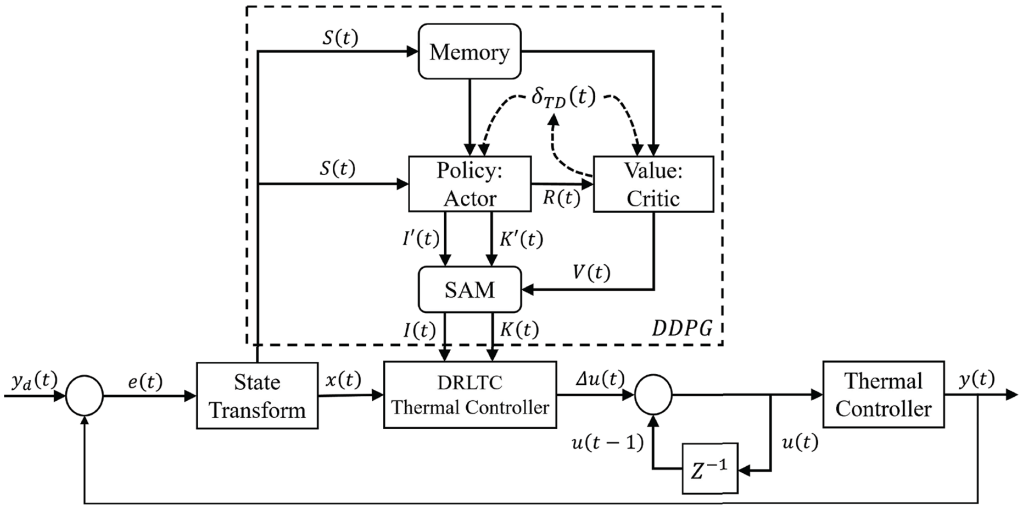

Fig. 2 Architecture of thermal controller based on DDPG.

obtained for samples $k-1$ and k, respectively; $K(t) = [K_P(t), K_I(t), K_D(t)]$ is a vector of PID parameters; $I(t)$ is the adaptive compensation of the PID output current according to the second reinforcement learning agent; and $R(f)$ is the current compensation function based on the value of the control error and the approximate value of the second reinforcement learning agent.

As shown in Fig. 2, $y(t)$ and $y_d(t)$ are the actual and expected system outputs, respectively. The system error $e(t) = y_d(t) - y(t)$ is converted to a system state vector $x(t)$ by a state converter, which is called the stochastic action modifier (SAM). The actor is used for policy estimation and for mapping the system state variables to the recommended PID parameters $K'(t)$ and compensation current $I(t)$. The output parameters of the actor are not directly involved in the design of the DRLTC thermal controller, and SAM corrects them based on the value function estimation information provided by the actor to obtain the actual PID parameters $K(t)$ and the compensation current $I(t)$. The critic receives the state vectors of the system from the replay memory and generates the cumulative reward R(t) from the external environment. Additionally, it evaluates the effect of decision making in every period of the DDPG and generates the temporal difference (TD) error (the internal reinforcement signal) δ_{TD} [25] and estimation function $V(t)$, where δ_{TD} is directly provided to the actor and the critic and used for updating their various parameters. At the same time, $V(t)$ is sent to the SAM and used to modify the actor's output.

B. Related Work on Reinforcement Learning

1. Deep Q-Learning Algorithms

Contrary to TD-Gammon and similar to online approaches, Mnih et al. [26] proposed a technique known as experience replay, wherein the agent's experiences are stored at each time-step $e_t = (s_t, a_t, r_t, s_{t+1})$ in a dataset $D = e_1, \ldots, e_N$ pooled over many episodes into a replay memory. Combined with Monte Carlo sampling and dynamic programming methods, the one-step prediction method is adopted to calculate the current state-value function, and the update formula of the Q-learning algorithm is obtained as follows:

$$Q(s, a|\theta) = Q(s_t, a_t|\theta) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a|\theta) - Q(s_t, a_t|\theta)] \tag{9}$$

where $\alpha \in [0, 1]$ is the learning rate and $\gamma \in [0, 1]$ is the discounting factor.

Because the state value of the thermal control system is always continuous, the original Q-learning based on the Q table will fail to process the continuous state. The deep Q-learning network (DQN) can learn the value function through a neural network. The network is trained offline by means of replay memory to minimize the correlation amongst the samples, and the target neural network is used to provide consistent targets during the temporal difference backups. The target neural network is denoted as θ^{Q-}, the approximate network function is denoted as θ^Q, and the loss function is expressed as follows:

$$L(\theta) = E([(r + \gamma \max_a Q(s, a; |\theta^{Q-}) - Q(s, a, |\theta^Q))^2]) \tag{10}$$

The target neural network uses a fixed parameter method to update each fixed number of steps. The gradient descent updating formula for DQN is expressed as follows:

$$\theta^Q_{i+1} = \theta^Q_i + \alpha[r + \gamma \max_a Q(s, a; |\theta^{Q-}) - Q(s, a, |\theta^Q)]\nabla Q(s, a, |\theta^Q) \tag{11}$$

2. Deterministic Policy Gradient Algorithms

Because the output of the DQN is discrete, this causes fluctuations in the optimization process and may even lead to convergence failure. Based on the Markov decision process, a single $R(\tau) = \sum_{i=0}^{H} R(t) = \sum_{i=0}^{H} R(s_i, u_i)$ represents the cumulative reward function of the thermal control strategy τ; $P(\tau, \theta)$ is defined as the probability of occurrence of the thermal control strategy τ; the objective function of reinforcement learning can be expressed as $U(\theta) = E(\sum_{i=0}^{H} R(s_i, u_i); \pi_\theta) = \sum_z P(\tau; \theta)R(\tau)$. In reinforcement learning, the objective is to learn an optimal parameter θ, which maximizes the expected return from the initial distribution defined as $\max_z U(\theta) = \max_\theta \sum_z P(\tau; \theta)R(\tau)$. The objective function is optimized using the gradient descent $\theta_{\text{new}} = \theta_{\text{old}} + \alpha\nabla_\theta U(\theta)$. In the equation, $\nabla_e U(\theta) = \nabla_\theta \sum_\tau P(\tau; \theta)R(\tau)$. The further consolidation is expressed as follows:

$$\nabla_\theta = \sum_\tau P(\tau; \theta)\nabla \log P(\tau; \theta)R(\tau) \tag{12}$$

The policy gradient algorithm achieves good results when dealing with continuous action space. According to the stochastic gradient descent, the performance gradient $\nabla_\theta J(\pi_\theta)$ can be expressed as follows:

$$\nabla_\theta J(\pi_\theta) = E_{x-\rho^\pi, a-\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s)Q^\pi(s, a)] \tag{13}$$

Konda et al. proposed an actor–critic, model-free algorithm based on deterministic strategy gradients [27,28]. The formula for the deterministic strategy gradients is expressed as follows:

$$\nabla_\theta J_\beta(\mu_\theta) = E_{s-\rho^\beta}[\nabla_\theta \mu^\theta(s)\nabla_a Q^\mu(s, a)|_{a=\mu^\theta(s)}] \tag{14}$$

Here, $\beta(a|s) \neq \pi_\theta(a|s)$ denotes a state behavior strategy. A neural network is used to approximate the action value function $Q^\omega(s|a)$ and the deterministic policy $\mu^s(s)$. The formula for updating the DDPG algorithm is expressed as follows:

$$\delta_t = r_t + \gamma Q^{\omega-}(s_{t+1}, \mu^e(s_{t+1})) - Q^\omega(s_t, a_t) \tag{15}$$

$$\omega_{t+1} = \omega_t + \alpha_\omega \delta_t \nabla_\omega Q^\omega(s_t, a_{it}) \tag{16}$$

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_{\theta^\mu} \mu^\theta(s_t)\nabla_a Q^\omega(s_t, a_t)|_{a=\mu^\theta(s)} \tag{17}$$

$$\theta^- = \tau\theta + (1 - \tau)\theta^- \tag{18}$$

$$\omega^- = \tau\omega + (1 - \tau)\omega^- \tag{19}$$

C. DRLTC Thermal Controller Scheme Based on DDPG

The framework of reinforcement learning agent 1 is constructed based on the actor–critic and PID thermal control system as the environmental object. The deterministic policy gradient algorithm is used to design the action network, and the DQN algorithm is used to evaluate the network. Thus, the self-tuning of the PID parameters is achieved.

$$s_t = [X(t - 1), e(t - 1), X(t), e(t), H(t + 1), H(t + 2)] \tag{20}$$

$$s_{i+1} = [X(t), e(t), X(t + 1), e(t + 1), H(t + 2), H(t + 3)] \tag{21}$$

$$r_{dk} = e^{-(1/2)\dfrac{[0.94(1) + 0.14]}{\sigma^2}} \tag{22}$$

As expressed by Eqs. (20) and (21), s_t denotes a set of reinforcement learning agent states; s_{t+1} denotes a set of agent states at the next time, which is obtained by the interaction of agent 1 with the environment. Additionally, in Fig. 3, $X(t + 1)$, $X(t + 2)$, and $X(t + 3)$ denote the temperature at the previous time, current time, and next time of the thermal control system, respectively; $H(t - 1)$, $H(t)$, and $H(t + 1)$ denote the control parameters of the thermal controller at next time, after two times, and after three times, respectively; $e(t - 1)$, $e(t)$, and $e(t + 1)$ are the system error of the previous time, current time, and next time, respectively; $\mu(s_t|\theta^\mu)$ is the action selected by the actor evaluation network according to states s_t; $\mu(s_{t+1}|\theta^{\mu^-})$ is the action selected by the actor target network according to state s_t. The reward function consists of the absolute value of

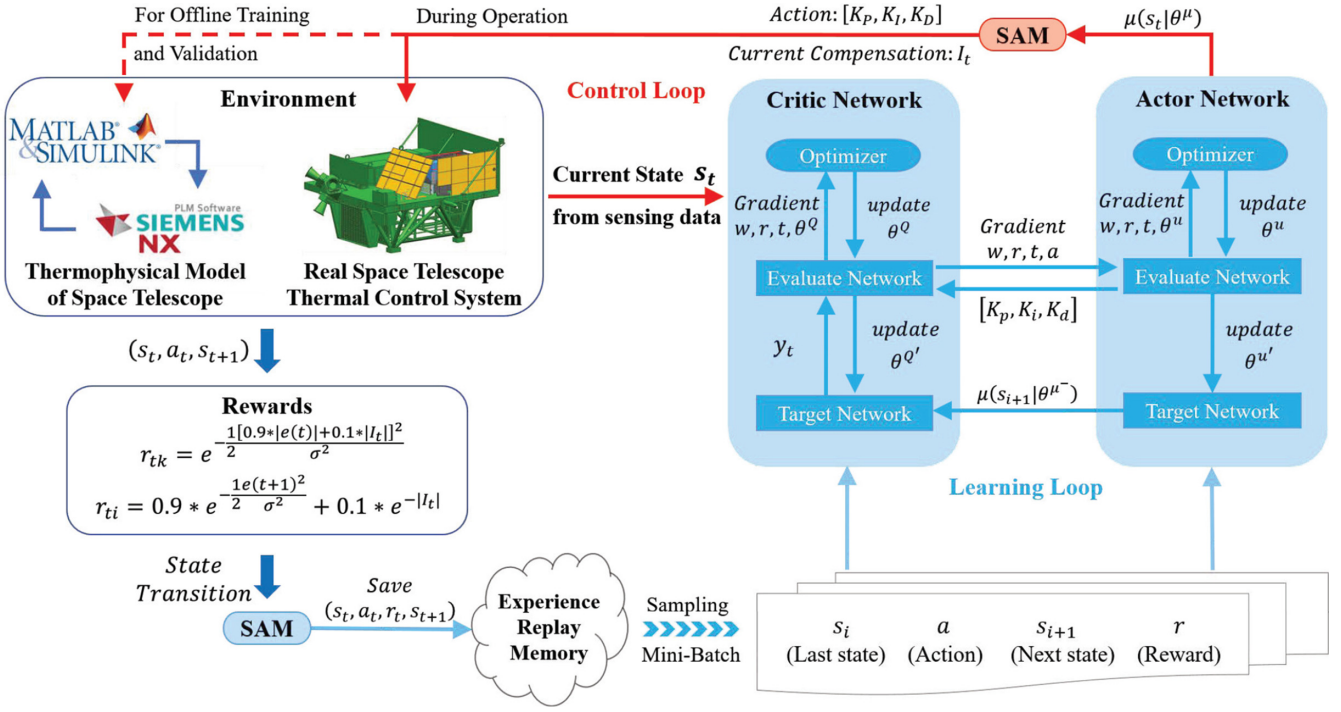

Fig. 3 Diagram of adaptive PID tuning strategy based on DRLTC.

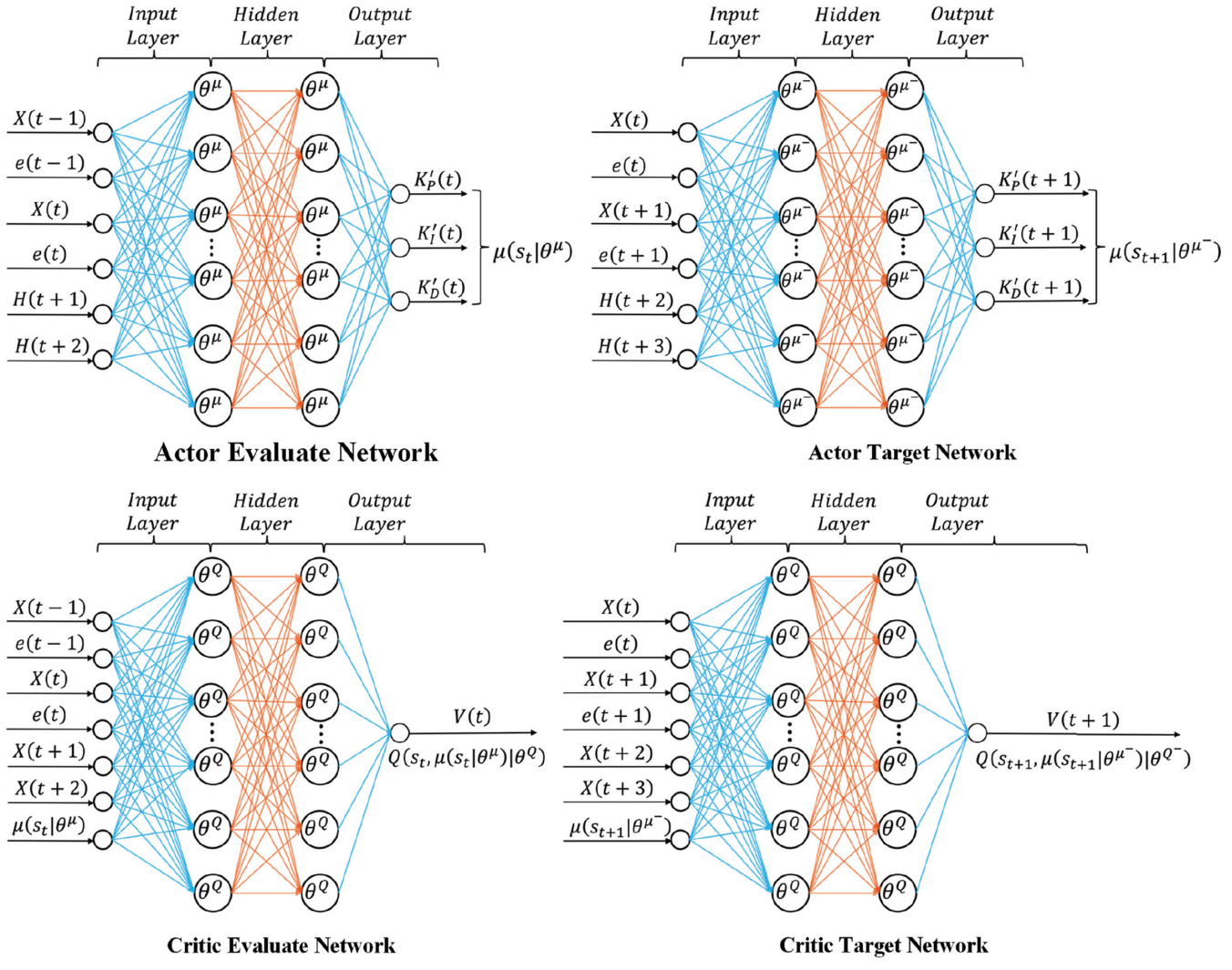

Fig. 4 Structural diagram of the actor–critic network.

the DRLTC thermal system output signal's systematic error. Then, the data are obtained by agent 1 through interaction with the environment and stored in memory M. The network update data were obtained by replaying the memory and training the neural network using the reward function expressed by Eq. (22). Based on the current state, the reinforcement learning agent 1 is obtained to select the appreciation $K(t) = [K_P, K_I, K_D]$, and the self-tuning of the PID parameters is implemented based on the DRLTC.

D. Adaptive Current Compensation Method Based on DRLTC

The framework of reinforcement learning agent 2 was constructed based on the actor–critic, similar to agent 1. Thus, the adaptive compensation of the output current of the DRLTC thermal control system was realized.

In Fig. 3, the current $[I_i]$ is compensated for the reinforcement learning agent 2 as a set of action variables μ. The network update data are obtained by replaying the memory and training the neural network using the reward function expressed by Eq. (22). Based on the current state, a reinforcement learning agent 2 is obtained to select the appreciation I_i, and the adaptive current compensation is implemented based on the DRLTC. The network update data are obtained by replaying the memory and training the neural network using the reward function shown in Eq. (23).

$$r_{id} = 0.9 * e^{-[1.d(i+1)^2/\sigma^2]} + 0.1 * e^{-1.1} \tag{23}$$

E. Design Scheme of Reinforcement Learning Agent

1. *Network Design Scheme of Actor*

Based on the deterministic policy gradient algorithm, two four-layer Radial basis function (RBF) neural networks [29] were established, namely, the actor evaluation and target networks, which have the same structure but different functions.

As shown in Fig. 4, the input layer of the actor evaluation network has six neurons, which form the set of the reinforcement learning agent state s_t. The action value $\mu(s_t|\theta^\mu)$ of the evaluation network is output. The inputs of the actor target network also have six neurons, which form the set of the reinforcement learning agent state s_{t+1}. The action value $\mu(s_{t+1}|\theta^-)$ of the target network is output. The actor network contains three hidden layers. The first hidden layer has 300 neurons, and the activation function of this hidden layer is Tanh. The second hidden layer also has 300 neurons, and its activation function is Tanh. The third hidden layer comprises 10 neural networks, and its activation function is Tanh. The neurons of each layer use L2 regularization to prevent neural network overfitting.

2. *Network Design Scheme of Critic*

Based on the DQN algorithm, two four-layer neural networks were established, namely, the actor evaluation and target networks, which have the same structure but different functions.

In Fig. 4, the critic evaluation network has the same structure and number of neurons as the critic target network. Both have seven neurons in the input layer, which is a set of reinforcement learning agent states s_t at the current time and s_{t+1} at the next time. The $\mu(s_t|\theta)$ and $\mu(s_t|\theta^{\mu^-})$ come from the output value of the actor evaluation network and target network, respectively. The output of the critic evaluation network and target network is $Q(s_t, \mu(s_t|\theta^\mu)|\theta^Q)$ and $Q(s_{t+1}, \mu(s_{t+1}|\theta^{\mu^-})|\theta^{Q^-})$, respectively. The critic network contains three layers. The number and activation function of the hidden layer neurons are also the same as those of the actor network.

F. Implementation Process of DRLTC Thermal Control Algorithm Based on DDPG

According to the above analysis, the entire DRLTC thermal control algorithm based on DDPG for spacecraft is presented in Algorithm 1.

Algorithm 1: DRLTC thermal control algorithm based on DDPG for spacecraft

Initialize parameters:
Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ;
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$;
Initialize replay buffer R;
for each episode **do**
Initialize state of system:
Initialize a random process $\mathcal{N}$ for action exploration;
Receive initial observation state s_1;
for each period **do**
Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise;
Execute action a_t and observe reward r_t and observe new state s_{t+1};
Store transition (s_t, a_t, r_t, s_{t+1}) in R;
Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R;
Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$;
Update critic by minimizing the loss based on the minimum mean square error:
$$Loss = (1/N)\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$$
Update the actor policy using the sampled policy gradient:
$$\nabla_{\theta^\mu} J \approx (1/N)\sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$
Update the target networks:
$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$
Evaluation Thermal Strategy:
Execute appreciation $K'(t) = [K_p, K_i, K_d]$ and $I'(t)$;
endfor
Target Thermal Strategy:
Execute control parameter $K(t) = [K_P, K_I, K_D]$ and $I(t)$;
endfor

IV. Simulation

To verify the effect of the thermal controller based on the DRLTC, the algorithm was applied to the temperature control of the main mirror support structure (MMS) of a space telescope. The relevant physical parameters of MMS are listed in Table 1. Because MMS has a high thermal coupling relationship with the main mirror, the temperature change of MMS directly affects the temperature stability of the main mirror. Hence, the temperature control requirements are very high (the precision is required to be 0.1°C). A multilayer material [30,31] with an emission coefficient of 0.69 and an absorption coefficient of 0.02 was coated onto the mirror to reduce the effect of the outside atmosphere on its internal temperature. The schematic diagram of the simulation environment is shown in Fig. 5.

To evaluate the performance of the DRLTC, RLPID, FuzzyPID, and BPPID under external random interference, the simulation experiment was divided into two heating stages. The total time of the experiment was 3600 s. From 0 to 1800 s, the temperature in the thermal coupling zone gradually increased from the initial

Table 1 Parameters of simulation environment

Parameter	Description	Value
Density	ρ_{MMS}, ρ_{MMI}	2637 kg/m^3
Quality	m_{MMS}	3.16 kg
	m_{MMI}	15.63 kg
Thermal conductivity	k_{MMS}, k_{MMI}	200 W/(m · K)
Specific heat capacity	c_{MMS}, c_{MMI}	904 J/(kg · K)
Volume	V_{MMS} (physical volume)	0.3 · 0.2 · 0.002 m^3
	V_{MMI} (surface volume)	0.3 · 0.21 · 0.12 m^3

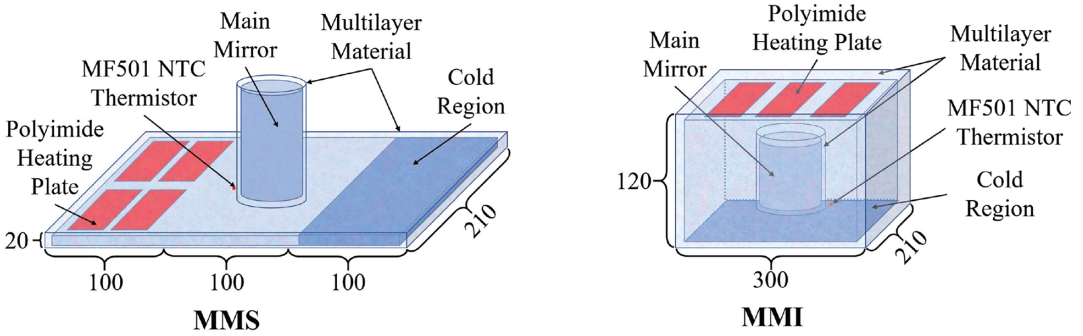

Fig. 5 Schematic diagram of simulation environment.

Table 2 Parameters of simulation environment

Parameter	Description	Value
P_{atm}	Atmospheric pressure	101.325 kPa
T_{inside}	Temperature inside the incubator	24.50°C
$T_{outside}$	Temperature outside the incubator	25.35 ± 0.3°C
P_{Heat}	Heating power	0–12.484 W
ε_{MLI}	Emission coefficient of multilayer	0.69
a_{MLI}	Absorption coefficient of multilayer	0.02

Table 3 Parameters of DRLTC

Parameter	Description	Value
α_A	Learning rate of actor	0.001
α_C	Learning rate of critic	0.02
ε	Tolerant error band	0.001
γ	Discount factor	0.97
N_a, N_c	Batch size of actor and critic	256
M_a, M_c	Memory capacity of actor and critic	50,000
S_a	Update steps of actor	15,000
S_c	Update steps of critic	13,000

temperature of 20.99–22°C, which served as the operating temperature in this stage and is called condition A. From 1801 to 3600 s, the temperature in the thermal coupling zone gradually increased to 23°C, which served as the operating temperature of this stage and is called condition B. Two random temperature disturbance signals were applied to conditions A and B, respectively. Additionally, the random temperature interference source at 1–2°C was applied to one side of the main mirror supporting structure to verify the robustness of the proposed control algorithm. An active temperature control loop was set on the opposite side of the main mirror, and the loop was controlled by the DRLTC. The relevant parameters of the simulation environment are listed in Table 2.

To analyze the control effect of the controller, the simulation results of the RLPID designed by Xiong, FuzzyPID designed by Carvajal, and BPPID designed by Chen were compared with the simulation results for the DRLTC, whose parameters are listed in Table 3. Additionally, the thermophysical model was developed with Simscape in Simulink, and the control algorithm was designed in MATLAB. The simulation results are presented in Fig. 6.

As can be seen, when the external disturbance changed, the FuzzyPID and BPPID had a small overshoot of approximately 0.05°C; the static error was only 0.01°C and there was almost no amplitude. The RLPID and DRLTC had approximately no overshoot and their static error was only 0.01 and 0.006°C, respectively. However, there existed a

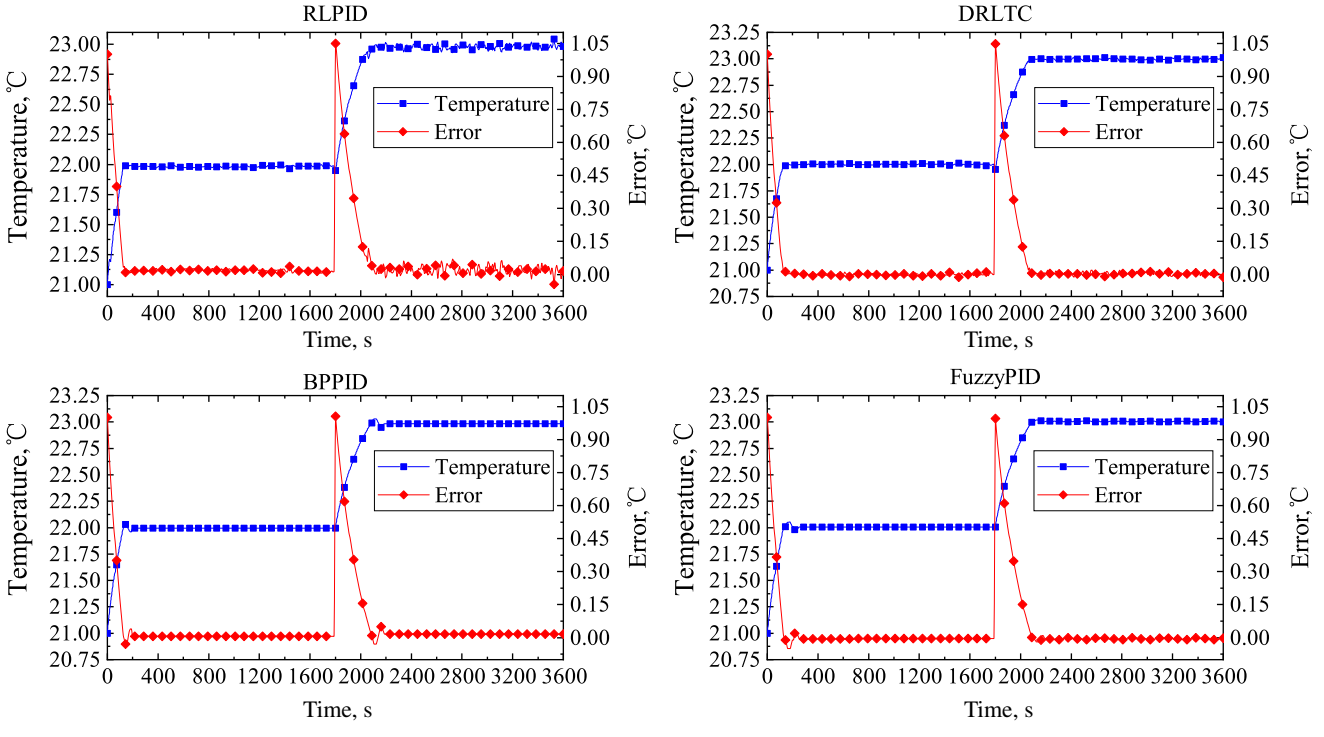

Fig. 6 Simulation results for MMS.

fluctuation of 0.04 and 0.06°C in the RLPID and DRLTC, respectively. As can be seen in Fig. 7, the duty cycle of FuzzyPID and BPPID was more stable. Therefore, as long as the parameters of the thermal control are reasonably optimized, the FuzzyPID and BPPID can control the precision and temperature fluctuation better than the RLPID and DRLTC. However, when the control object changed from the MMS to the main mirror installation (MMI), the simulation results changed dramatically. The simulated test environment was the same as that of MMS. Additionally, the random temperature interference signals were the same, and arranged on the main mirror installation surface of the MMI. The active temperature control loop was set on the opposite side of the main mirror installation surface of the MMI, and the loop was controlled by four thermal controller types. The relevant physical parameters of MMI are listed in Table 2. The control parameters of the DRLTC, RLPID, FuzzyPID, and BPPID were not modified in the MMS design.

Figures 8 and 9 show that the BPPID exhibited an enormous overshoot of 1.41°C and converge difficulty. The FuzzyPID exhibited an

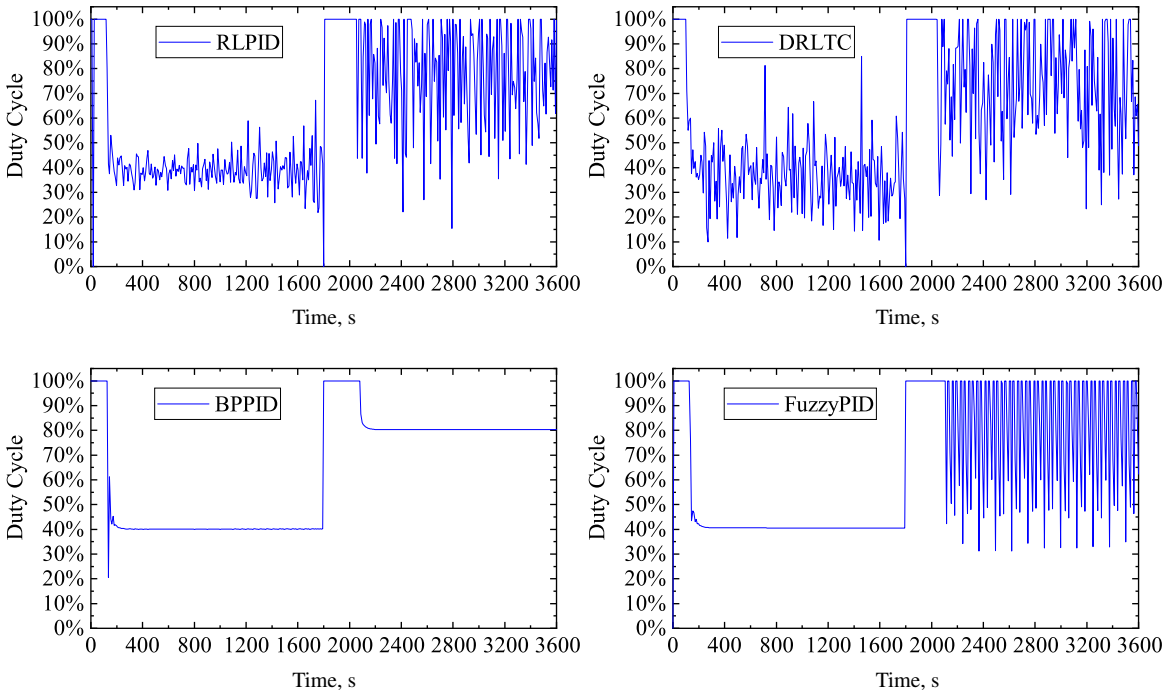

Fig. 7 Duty cycle of MMS in simulations.

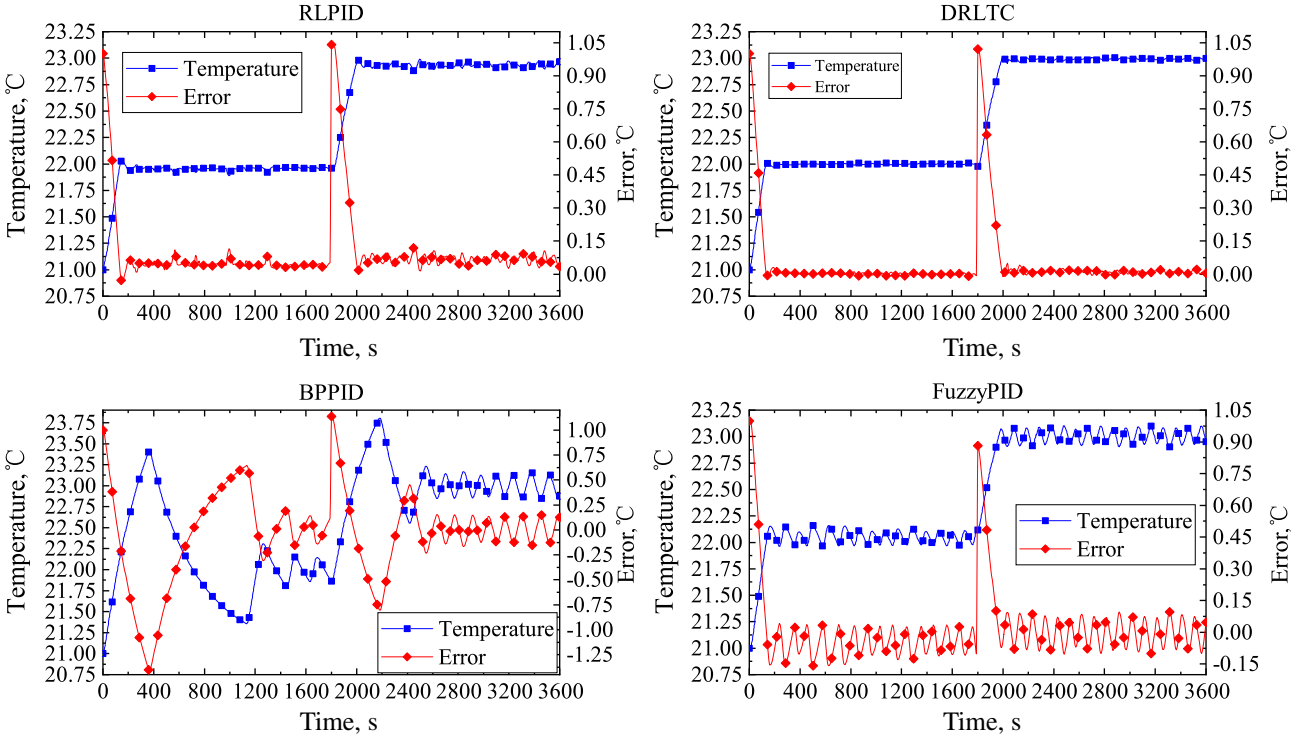

Fig. 8 Simulation results for MMI.

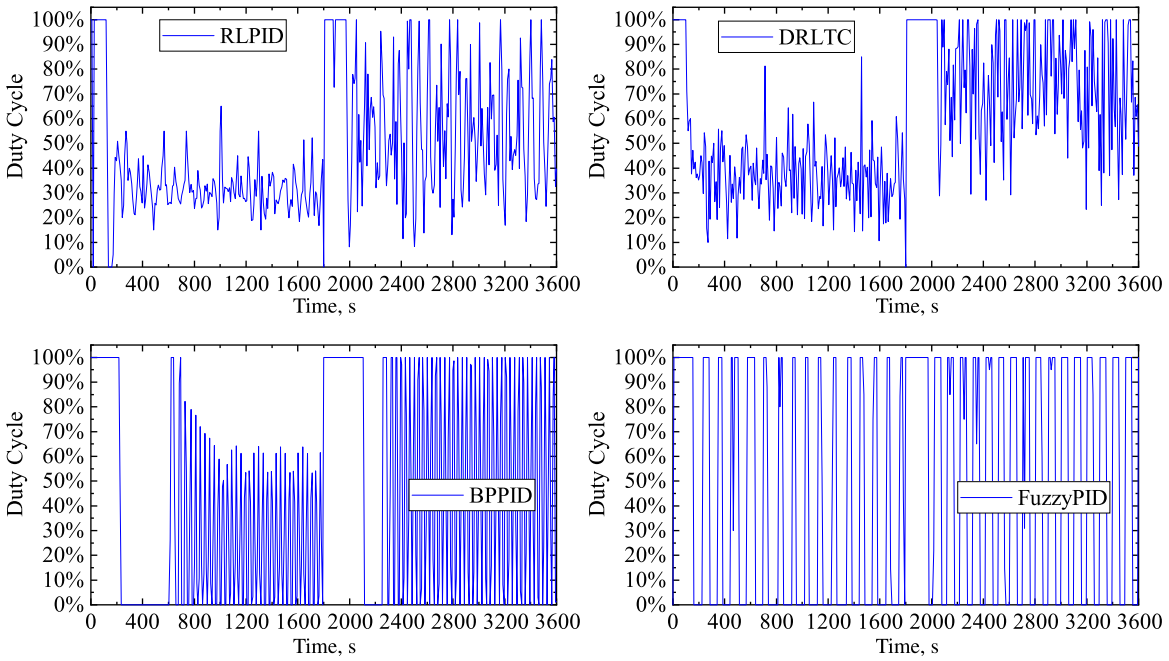

Fig. 9 Duty cycle of MMI in simulations.

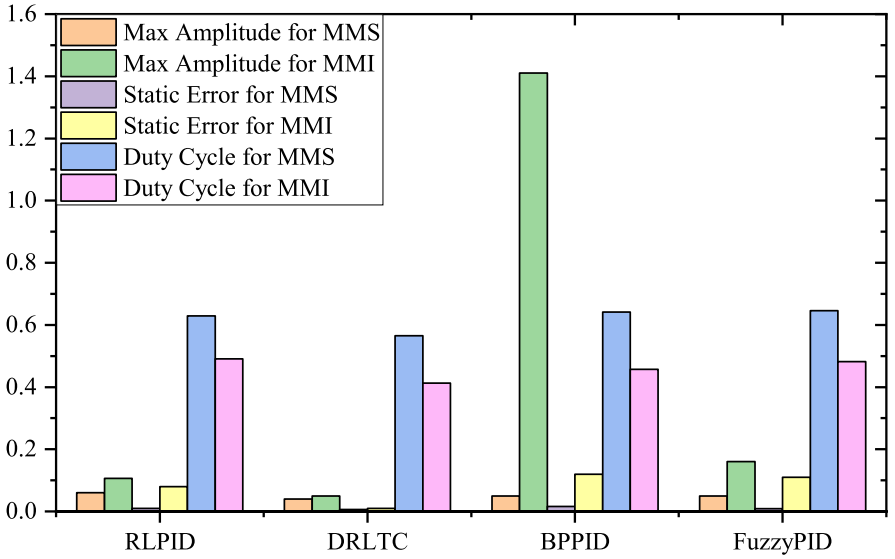

Fig. 10 Comparison and analysis of simulation results.

enormous fluctuation of 0.16°C. Both of them had a static error greater than 0.1°C. Finally, the RLPID achieved the convergence effect, but had an overshooting error of 0.1°C and a static error of 0.08°C, and there always existed a fluctuation of 0.03°C. However, the DRLTC had almost no overshoot, exhibited minor fluctuation, and had a smaller average duty cycle compared with the other four thermal controllers. As can be seen in Fig. 10, the static error of DRLTC was only 0.01°C. Compared with RLPID, BPPID, and FuzzyPID, the temperature control accuracy of the DRLTC improved by 87.5, 91.7, and 90.9%, respectively.

V. Experiments

We used the MMS and MMI of a space telescope, similar to the thermal experiment simulation (Table 1). The relevant parameters of the experimental and simulation environments are listed in Table 2. Because the accuracy of the sensors and actuators is very important for the precision thermal control of space telescope to measure the temperature changes more accurately, we used a negative temperature coefficient (NTC) thermistor for high-precision temperature measurements that can provide a resolution and accuracy of as low as 5 mK, called MF501 NTC, and then we placed them in some key locations to monitor the temperature changes in real time [32]. When the control object was MMS, the thermal coupling zone between the main mirror and the MMS was controlled, and the temperature changes in the heating zone and boundary zone of the hot and cold regions were monitored in real time by the MF501 NTC thermistor. When the control object changed from MMS to MMI, the heating zone was set on top of the MMS to prevent the polyimide heating plate from being too close to the main mirror, because high temperature during the heating process affects the image quality. As shown in Fig. 11, we placed the MMS and MMI in an incubator so that the experimental device could be insulated from the exterior. We changed the heating power of the heater by changing the voltage of the

polyimide heating plate, which was controlled by the programmable power supply controller of the DRLTC. To verify the robustness of each control method, two random temperature disturbance signals from 1 to 2°C were applied to the same position as that in the simulation.

The algorithm was developed in MATLAB, and LabView [33–35] was used for joint control. The experimental flow chart is shown in Fig. 11.

The thermal test temperature and power control program were established in LabView, and the temperature signal of the key node of MMS or MMI, which were collected in real time by the temperature data acquisition system, were transmitted to the control algorithm in MATLAB for analysis by the data exchanger. Then, the DRLTC algorithm made an intelligent decision and sent the control signal to the programmable power through the data exchange to control the temperature. As shown in Figs. 12 and 13, when the control object was the MMS, the temperature control effect of the four thermal controllers was very good. The DRLTC and RLPID had almost no overshoot and only a static error of 0.01°C, whereas the BPPID and

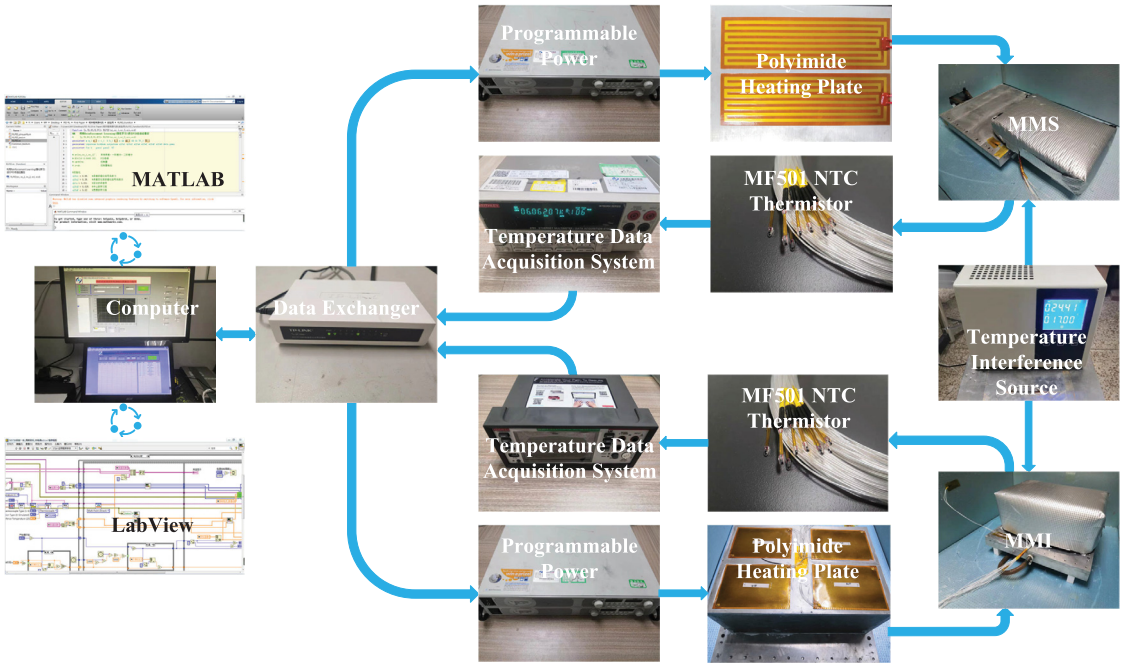

Fig. 11 Experimental flow chart.

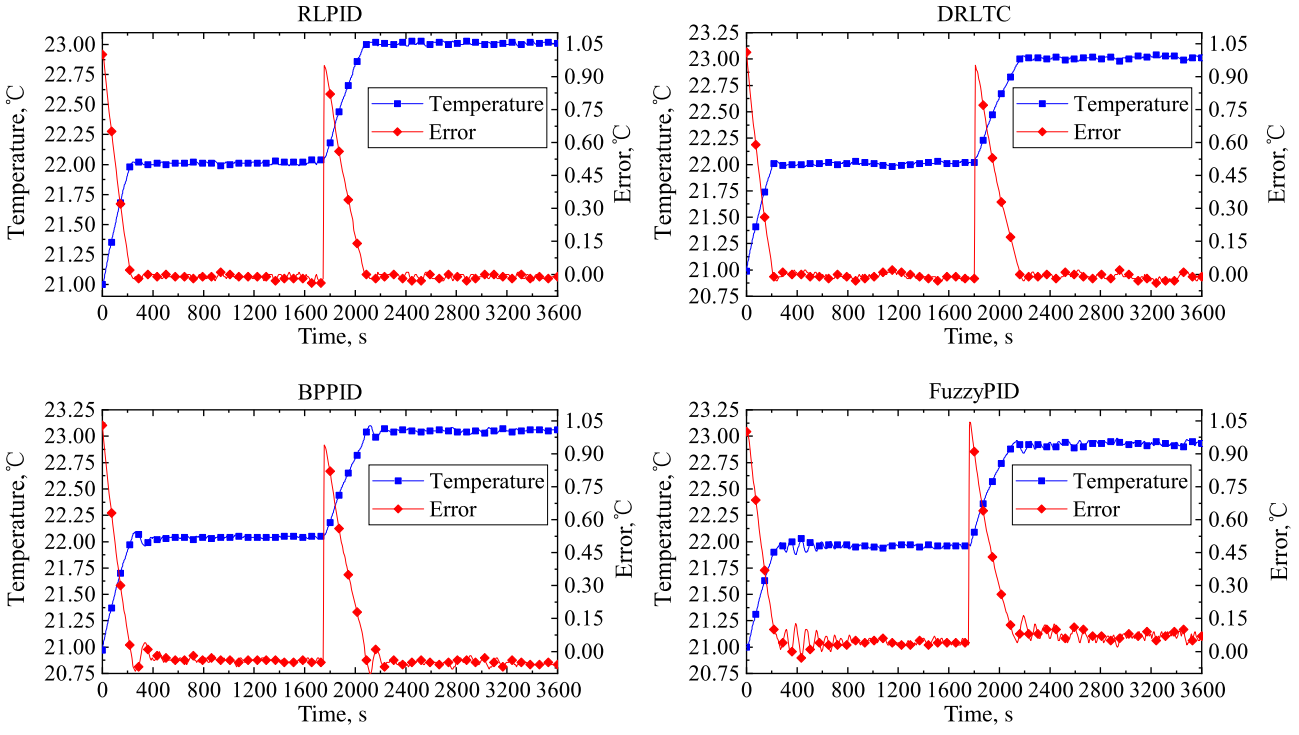

Fig. 12 Experimental results for MMS.

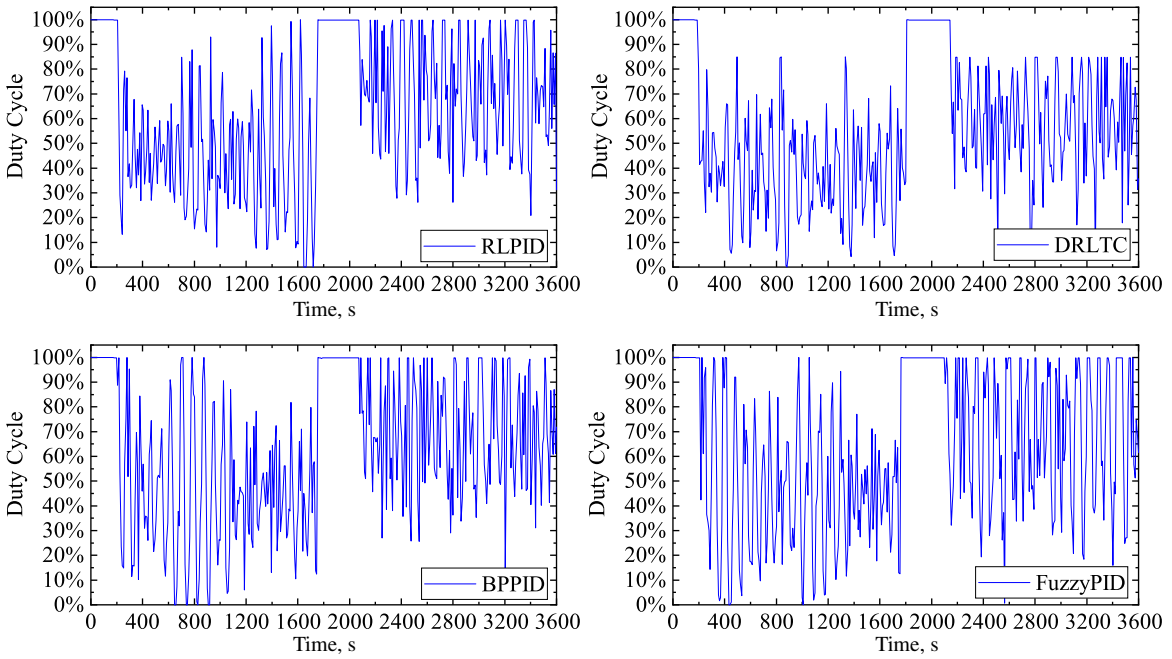

Fig. 13 Duty cycle of MMS in experiments.

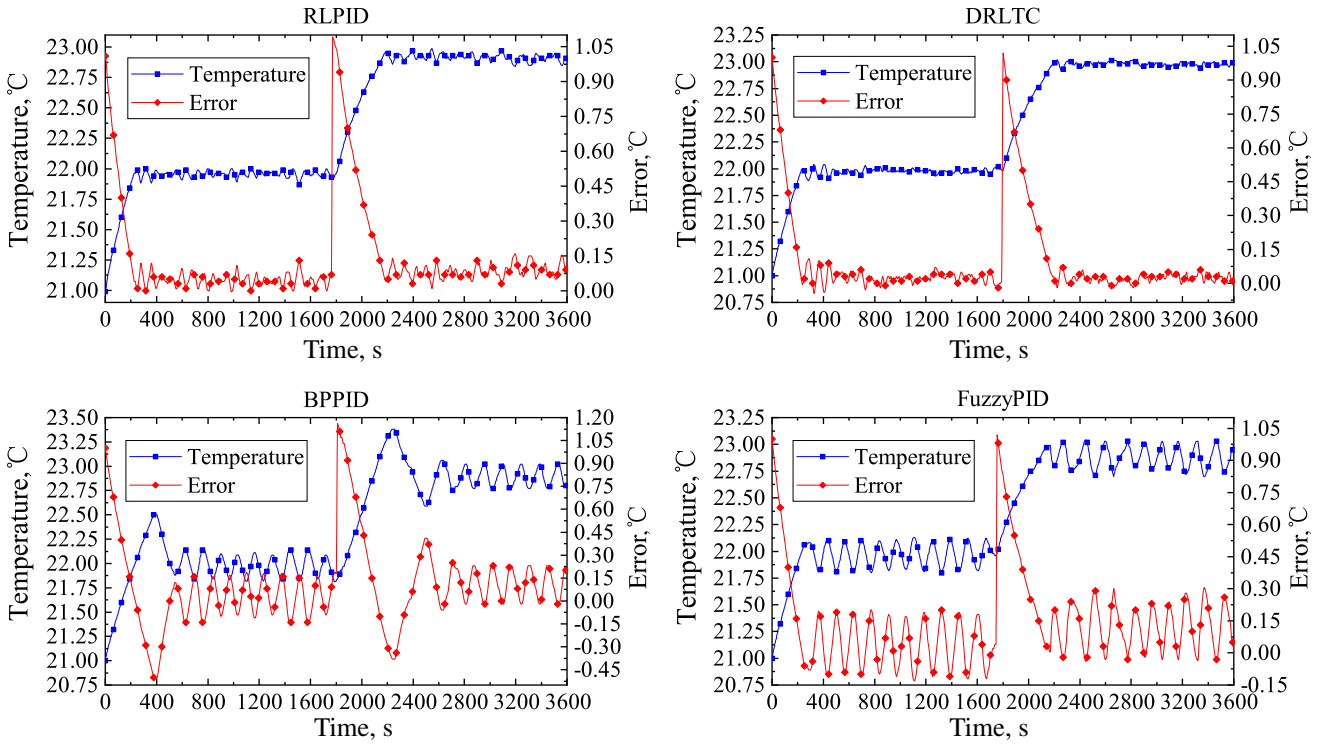

Fig. 14 Experimental results for MMI.

FuzzyPID had an overshoot of 0.1 and 0.16°C, respectively, with static errors of 0.05 and 0.07°C, respectively. Additionally, the average duty cycle of DRLTC was only 0.518, which is the smallest value among the four thermal controllers.

However, when the control object changed from MMS to MMI, as shown in Fig. 14, the BPPID had an overshoot of 0.5°C, had a static error of 0.2°C, and always exhibited convergence difficulty with a fluctuation of approximately 0.25°C. The FuzzyPID had a larger temperature amplitude range and a static error of 0.2°C. Although the maximum temperature fluctuation of the RLPID was only 0.15°C, the static error reached 0.1°C. However, the maximum temperature fluctuation of the DRLTC was only 0.11°C, and the static error was only 0.02°C. Compared with the RLPID, BPPID, and FuzzyPID, its temperature control accuracy improved by 80.2, 90.6, and 85.7%. Additionally, as shown in Figs. 15 and 16, the average duty cycle of the DRLTC was only 0.505, which is the smallest value among the four thermal controllers.

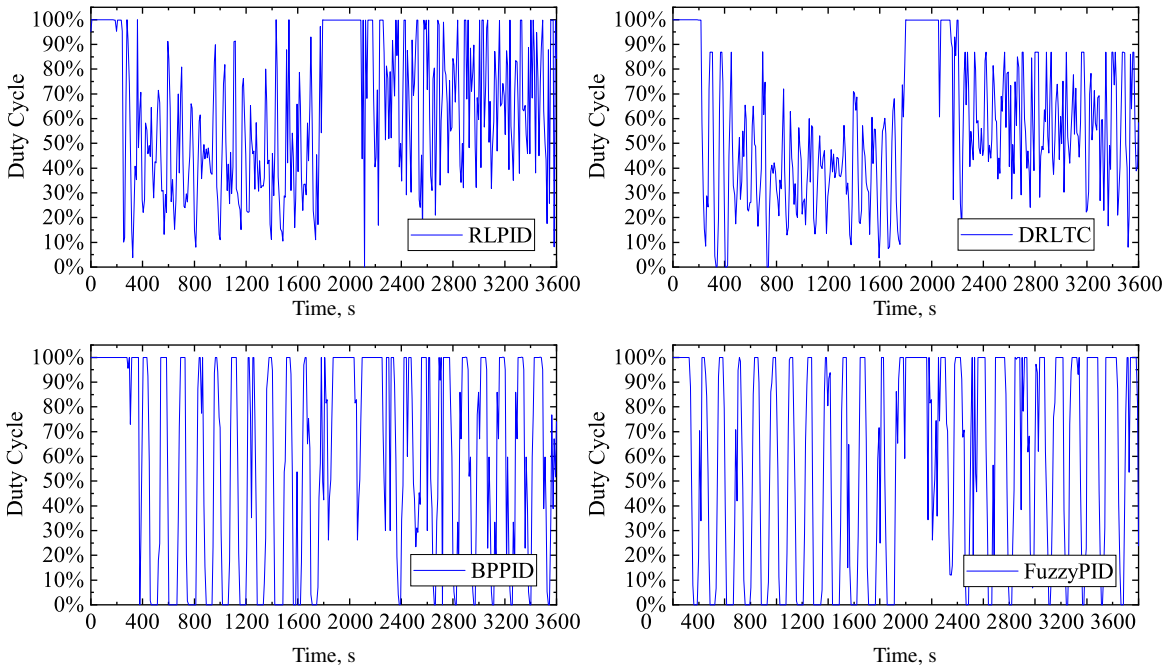

Fig. 15 Duty cycle of MMI in experiments.

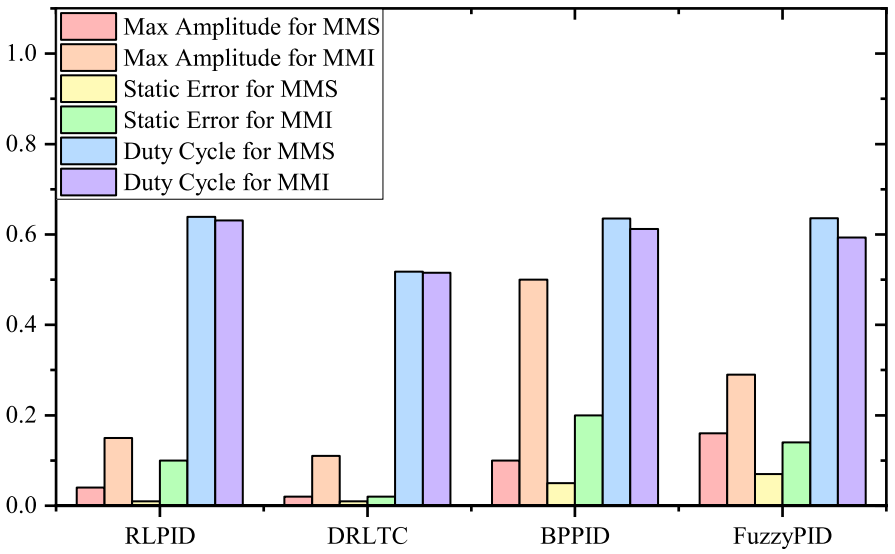

Fig. 16 Comparison and analysis of experimental results.

VI. Conclusions

This paper proposes the novel DRLTC thermal control strategy based on the DDPG for space telescopes, and a trained thermal controller based on DRLTC. This strategy can rapidly and accurately adjust the PID parameters according to an intelligent control object, and also perform online current compensation based on the thermal system. The control effect of the DRLTC was better compared with those of the RLPID, BPPID, and FuzzyPID. Additionally, it had shorter adjustment time, faster response speed, better resistance to interference, and better adaptation of the controlled object parameters. In particular, when the control object changed from MMS to MMI, the BPPID and FuzzyPID had to be optimized again to achieve good control accuracy. Although the RLPID achieved a good temperature control effect, it exhibited large temperature fluctuations and static error.

The theoretical and experimental results revealed that, when the control object was the MMS, the DRLTC achieved a temperature control precision of 0.01°C, and the steady-state error was reduced by 40.2, 62.5, and 33.3% in the simulation, and by 5.6, 80.6, and 85.7% in the experiment, compared with those of the RLPID controller, BPPID controller, and FuzzyPID, respectively. When the control object changed from MMS to the MMI, DRLTC achieved a temperature control precision of 0.02°C, and the steady-state error was reduced by 87.5, 91.7, and 90.9% in the simulation, and by 80.2, 90.6, and 85.7% in the experiment, compared with those of the other three thermal controllers. The DRLTC did not only improve the precision of temperature control, but also saved energy. Additionally, it had better reliability, more robustness, and better universality.

The DRTC thermal control algorithm is not only applicable to space telescopes, but also to all types of thermal controllers. However, the convergence of DRLTC is not particularly stable. Thus, it is necessary to further improve the convergence and stability of the deep deterministic policy gradient algorithm to improve the precision and stability of the control.

Acknowledgments

This study received funding from the National Natural Science Foundation of China (Grant No. 61605203) and the Youth Innovation

Promotion Association of the Chinese Academy of Sciences (Grant No. 2015173).

References

[1] Cheng, W.-L., Liu, N., and Wu, W.-F., "Studies on Thermal Properties and Thermal Control Effectiveness of a New Shape-Stabilized Phase Change Material with High Thermal Conductivity," *Applied Thermal Engineering*, Vol. 36, April 2012, pp. 345–352. https://doi.org/10.1016/j.applthermaleng.2011.10.046

[2] Xu, Z., Xu, M., Cheng, W., Peng, H., and Ding, Y., "High-Precision, Temperature Control Based on Grading-Structure and PID-Feedback Strategies," *Transactions of the Japan Society for Aeronautical and Space Sciences*, Vol. 61, No. 2, 2018, pp. 51–59. https://doi.org/10.2322/tjsass.61.51

[3] Carvajal, J., Chen, G., and Ogmen, H., "Fuzzy PID Controller: Design, Performance Evaluation, and Stability Analysis," *Information Sciences*, Vol. 123, Nos. 3–4, 2000, pp. 249–270. https://doi.org/10.1016/S0020-0255(99)00127-9

[4] Chen, J., and Huang, T.-C., "Applying Neural Networks to On-Line Updated PID Controllers for Nonlinear Process Control," *Journal of Process Control*, Vol. 14, No. 2, 2004, pp. 211–230. https://doi.org/10.1016/S0959-1524(03)00039-8

[5] Iruthayarajan, M. W., and Baskar, S., "Evolutionary Algorithms Based Design of Multivariable PID Controller," *Expert Systems with Applications*, Vol. 36, No. 5, 2009, pp. 9159–9167. https://doi.org/10.1016/j.eswa.2008.12.033

[6] Boubertakh, H., Tadjine, M., Glorennec, P.-Y., and Labiod, S., "Tuning Fuzzy PD and PI Controllers Using Reinforcement Learning," *ISA Transactions*, Vol. 49, No. 4, 2010, pp. 543–551. https://doi.org/10.1016/j.isatra.2010.05.005

[7] Lemmen, M., Kouwen, J., Koorevaar, F., and Pennings, N., "In-Flight Results of the Sciamachy Optical Assembly Active Thermal Control System," SAE TP 2004-01-2357, 2004. https://doi.org/10.4271/2004-01-2357

[8] Zhu, S., Zou, P., Meina, L. U., Zhang, A., Liu, Z., Qiu, Z., and Hong, J., "Temperature Control System Design of Infrared Detector Based on Bang-Bang and PID Control," *Infrared Technology*, Vol. 39, No. 11, 2017, pp. 990–995. https://doi.org/10.4271/2004-01-2357

[9] Choi, M. K., "Method of Generating Transient Equivalent Sink and Test Target Temperatures for Swift BAT," *2nd International Energy Conversion Engineering Conference*, AIAA Paper 2004-5686, 2004. https://doi.org/10.2514/6.2004-5686

[10] Choi, M. K., "Thermal Design to Meet Stringent Temperature Gradient/Stability Requirements of SWIFT BAT Detectors," *Collection of Technical Papers. 35th Intersociety Energy Conversion Engineering Conference and Exhibit (IECEC)*, Vol. 1, IEEE, New York, 2000, pp. 576–584. https://doi.org/10.1109/IECEC.2000.870806

[11] Choi, M., "Thermal Assessment of Swift Instrument Module Thermal Control System and Mini Heater Controllers After 5+ Years in Flight," *40th International Conference on Environmental Systems*, AIAA Paper 2010-6003, 2010. https://doi.org/10.2514/6.2010-6003

[12] Wang, X., Cheng, Y., and Sun, W., "Q Learning Based on Self-Organizing Fuzzy Radial Basis Function Network," *International Symposium on Neural Networks*, Springer, Berlin, 2006, pp. 607–615. https://doi.org/10.1007/11759966_90

[13] Moon, J. W., Jung, S. K., Kim, Y., and Han, S.-H., "Comparative Study of Artificial Intelligence-Based Building Thermal Control Methods—Application of Fuzzy, Adaptive Neuro-Fuzzy Inference System, and Artificial Neural Network," *Applied Thermal Engineering*, Vol. 31, Nos. 14–15, 2011, pp. 2422–2429. https://doi.org/10.1016/j.applthermaleng.2011.04.006

[14] Howell, M. N., and Best, M. C., "On-Line PID Tuning for Engine Idle-Speed Control Using Continuous Action Reinforcement Learning Automata," *Control Engineering Practice*, Vol. 8, No. 2, 2000, pp. 147–154. https://doi.org/10.1016/S0967-0661(99)00141-0

[15] Xiong, Y., Guo, L., Huang, Y., and Chen, L., "Intelligent Thermal Control Strategy Based on Reinforcement Learning for Space Telescope," *Journal of Thermophysics and Heat Transfer*, Vol. 34, No. 1, 2019, pp. 37–44. https://doi.org/10.2514/1.T5774

[16] Kaelbling, L. P., Littman, M. L., and Moore, A. W., "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, Vol. 4, May 1996, pp. 237–285. https://doi.org/10.1613/jair.301

[17] Szepesvári, C., "Algorithms for Reinforcement Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Vol. 4, No. 1, 2010, pp. 1–103. https://doi.org/10.2200/S00268ED1V01Y201005AIM009

[18] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D., "Continuous Control with Deep Reinforcement Learning," arXiv preprint arXiv:1509.02971, 2015.

[19] Richmond, J. A., "Adaptive Thermal Modeling Architecture for Small Satellite Applications," Ph.D. Thesis, Massachusetts Inst. of Technology, Cambridge, MA, 2010.

[20] Lyon, R., Sellers, J., and Underwood, C., "Small Satellite Thermal Modeling and Design at USAFA: FalconSat-2 Applications," *Proceedings IEEE Aerospace Conference*, Vol. 7, IEEE, New York, 2002, pp. 7–7. https://doi.org/10.1109/AERO.2002.1035316

[21] Kovács, R., and Józsa, V., "Thermal Analysis of the SMOG-1 PocketQube Satellite," *Applied Thermal Engineering*, Vol. 139, July 2018, pp. 506–513. https://doi.org/10.1016/j.applthermaleng.2018.05.020

[22] Parvaresh, A., Mohammadi, S. M. A., and Parvaresh, A., "A New Mathematical Dynamic Model for HVAC System Components Based on Matlab/Simulink," *International Journal of Innovative Technology and Exploring Engineering*, Vol. 1, No. 2, 2012, pp. 1–6.

[23] "MATLAB—MathWorks," The Mathworks, Inc., 2018, https://www.mathworks.com/.

[24] Åström, K. J., and Hägglund, T., *PID Controllers: Theory, Design, and Tuning*, Vol. 4, Instrument Soc. of America Research, Triangle Park, NC, 1995, pp. 200–270. https://doi.org/10.1109/TCST.2005.847331

[25] Sutton, R. S., and Barto, A. G., *Reinforcement Learning: An Introduction*, Vol. 13, 2nd ed., MIT Press, Cambridge, MA, June 2018, pp. 265–281. https://doi.org/10.1109/VLSIT.2018.8510680

[26] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M., "Playing Atari with Deep Reinforcement Learning," arXiv preprint arXiv:1312.5602, 2013.

[27] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M., "Deterministic Policy Gradient Algorithms," *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32, JMLR: W&CP, Beijing, China, June 2014.

[28] Konda, V. R., and Tsitsiklis, J. N., "Actor-Critic Algorithms," *Society for Industrial and Applied Mathematics*, Vol. 42, No. 4, 2000, pp. 1143–1166.

[29] Attaran, S. M., Yusof, R., and Selamat, H., "A Novel Optimization Algorithm Based on Epsilon Constraint-RBF Neural Network for Tuning PID Controller in Decoupled HVAC System," *Applied Thermal Engineering*, Vol. 99, April 2016, pp. 613–624. https://doi.org/10.1016/j.applthermaleng.2016.01.025

[30] Arpana, P., Raghavendra, K., Gavaskar, M., Venkatanarayana, M., and Amrit, A., "Transient Method for Estimating the Effective Emittance of Multilayer Insulation Blankets," *Journal of Thermophysics and Heat Transfer*, Vol. 30, No. 4, 2016, pp. 960–963. https://doi.org/10.2514/1.T4899

[31] Shi, J.-F., Wu, Q.-W., Chen, L.-H., and Yang, X.-W., "Review of Flight Tests for Multi-Layer Insulator Materials," *Chinese Optics*, Vol. 6, No. 4, 2013, pp. 457–469.

[32] Liu, G., Guo, L., Liu, C., and Wu, Q., "Evaluation of Different Calibration Equations for NTC Thermistor Applied to High-Precision Temperature Measurement," *Measurement*, Vol. 120, May 2018, pp. 21–27. https://doi.org/10.1016/j.measurement.2018.02.007

[33] Bozkaya, B., and Zeiler, W., "The Effectiveness of Night Ventilation for the Thermal Balance of an Aquifer Thermal Energy Storage," *Applied Thermal Engineering*, Vol. 146, Jan. 2019, pp. 190–202. https://doi.org/10.1016/j.applthermaleng.2018.09.106

[34] Moon, J. W., Yoon, Y., Jeon, Y.-H., and Kim, S., "Prediction Models and Control Algorithms for Predictive Applications of Setback Temperature in Cooling Systems," *Applied Thermal Engineering*, Vol. 113, Feb. 2017, pp. 1290–1302. https://doi.org/10.1016/j.applthermaleng.2016.11.087

[35] "LabVIEW—National Instruments," National Instruments, 2017, http://www.ni.com/zh-cn/shop/labview.html.