Energy-Optimal Flight Strategy for Solar-Powered Aircraft Using Reinforcement Learning With Discrete Actions

WENJUN NI1,2, DI WU1, AND XIAOPING MA1
1Institute of Engineering Thermophysics, Chinese Academy of Sciences, Beijing 100190, China
2University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Di Wu (wudi@iet.cn)

This work was supported by the Foundation of the Special Research Assistant of Chinese Academy of Sciences under Grant E0290A0301.

ABSTRACT The low efficiency of photovoltaic cells limits the energy absorption of high-altitude long-endurance (HALE) solar-powered unmanned aircraft vehicles (UAVs), which dramatically weakens the capacity for long-endurance missions. Therefore, finding a method to extend the flight duration with finite solar energy drives extensive research. The present work introduces a method that applies a deep reinforcement learning (DRL) framework to generate an energy-optimized flight strategy for HALE solar-powered aircraft. The neural network controller is designed to realize autonomous flight navigation by giving commands of thrust, attack angle, and bank angle. A mission area with a radius of 5 km is assumed to test the RL controller performance. The simulation results show that the RL controller leads to a 28 % increase in the battery SoC after a 24-hour flight, which indicates that a controller based on the RL framework might be a potential method for solving the solar-powered UAV trajectory planning problem. Aiming to explore the applicability of the RL controller, a sustained flight test is implemented. The results show that a 39-day endurance flight is achieved by the RL controller, which is 50% higher than the base case with a steady flight trajectory.

INDEX TERMS High-altitude long-endurance aircraft, flight strategy optimization, automatic navigation, reinforcement learning, photovoltaic cell.

I. INTRODUCTION

Research on high-altitude long-endurance (HALE) solar-powered unmanned aerial vehicles (UAVs) has received considerable attention in recent years mainly due to its promising future in various applications such as long-endurance intelligence, surveillance, and reconnaissance (ISR) [1], [2]. However, the poor capabilities of the photovoltaic cell and the rechargeable battery lead to a continuous increase in aircraft scale to meet the 24-hour flight energy requirement, which strongly obstructs the design of HALE solar-powered aircraft [3]. Therefore, researchers began to find alternative methods to help HALE UAVs absorb and store more energy. Since the solar energy absorption varies with the time and sunlight incident angle, the trajectory optimization becomes a potential methodology to enhance the endurance of HALE aircraft, which mainly includes two approaches: first, path planning aiming to maximize solar energy absorption by optimizing the incident angle between the sunlight and the solar panels; second, the energy management strategy that stores solar energy as gravity potential during the daytime.

The earliest research on solar-powered UAV flight path planning was implemented by Klesh and Kabamba [4], [5] with a detailed discussion on the influence of flight bank angle and speed when the solar-powered UAV is cruising in an optimal path. A nondimensional parameter that power ratio is proposed to predict the qualitative features of the aircraft energy-optimal state. Klesh et al’s results showed that a perpetual endurance flight can be achieved when the power ratio exceeds a specific threshold. However, only the simplest flight path, such as flying from the starting point to the destination, was applied in their work. Therefore, it is still worth exploring whether their method can be extended to a more complicated scenario.
Aiming to achieve continuous tracking to moving ground targets. Huang et al. [6] proposed an integrated model and studied energy-optimal path planning using a particle swarm optimization (PSO) algorithm. Numerical results showed that the proposed method provided the possibility for solar-powered UAVs to achieve longer target tracking. Unfortunately, the longitudinal angles were not considered in their solar energy absorption calculation, which means that their method can only be applied when the UAV is in a level flight.

Spangelo and Gilbert [7], [8] explored path planning when aircraft flies in three-dimensional space and estimated the effect of longitudinal motion on energy absorption. By allowing aircraft to change speed and altitude, an increase in solar energy absorption up to 30% is achieved compared to the steady case with constant speed and altitude. However, Spangelo et al. fixed the aircraft’s spatial location on the surface of the vertical cylinder, which actually distorted the three-dimensional space to a two-dimensional surface.

Martin et al. [9] expanded Spangelo et al.’s work by allowing the aircraft to fly inside the vertical cylinder rather than just on the surface. Therefore, a more complex flight path can be made, which leads to better solutions. A non-linear model predictive control (MPC) formulation was applied to maximize the total stored energy, \( E_{\text{tot,stored}} = E_{\text{battery}} + E_{\text{potential}} \) with a receding-horizon approach. The results showed the maximum \( E_{\text{tot,stored}} \) is increased compared to the steady-state trajectory. Similar research was further conducted by Marriott et al. [10], and the computational efficiency of the path optimization process improved by adopting a greedy dynamic programming algorithm with buffering. However, both studies applied an optimization process with a limited forward horizon to plan the trajectories, which cannot provide end-to-end control instructions due to the optimization process.

In addition to making an energy-optimal path, some researchers tend to introduce the energy management strategy (EMS) into solar-powered aircraft flights. A novel EMS that partly storing solar energy as gravitational potential in the daytime was proposed by Gao et al. Their results showed that an additional energy surplus up to 23.5 % can be achieved with the proposed EMS in one 24-hour flight [11]. An attempt to reduce the rechargeable battery weight by utilizing gravitational potential was further implemented [12]. Gao et al. believed the solar-powered UAV endurance might be greatly improved by energy storage using gravitational potential.

Based on Gao et al.’s work, a developed energy management strategy considering uncertain weather conditions was investigated by Lee and Yu [13].

Wang et al. [14] considered the installation angle of solar panels in an energy absorption model. The flight state and control variables were estimated by the Gauss pseudospectral (GPS) method. Through comprehensive adjustments to flight attitude and altitude, the aircraft received an 18 % increase in the total stored energy after a 24-hour flight. Furthermore, a mission-oriented design for three-dimensional path planning using a coupled method with GPS and colony algorithms was presented to maximize the mission effectiveness in Wang et al.’s work [15].

At present, the trajectory planning or energy management of solar-powered aircraft based on energy optimization mainly focuses on improving real-time performance and utilizing energy day and night. However, high-accuracy optimization algorithms are often complex and have high energy consumption, while the aircraft needs to adjust its flight state in real time according to the predicted target position [16], [17]. Additionally, some high-efficiency online methods, such as MPC, still require the receding-horizon optimization process [18]. Recently, reinforcement learning (RL) has received considerable attention in the aeronautic community due to its real-time end-to-end control performance and adaptability to an unknown environment. Relying on the deep neural network, a transformation from dynamic trajectory generation to an end-to-end controller can be achieved, which reduces the computational cost and provides real-time navigation or guidance command functions based on realizing long-term objective optimization. Several relevant works are described below.

To increase the endurance of unmanned aircraft, Woodbury et al. studied the flight strategy when a glider went through the ascending thermals. Reinforcement learning was applied to generate the reference bank angle command for leading the glider close to the ascending thermals. Therefore, a circling trajectory can be established to help gliders gain energy from the atmosphere. Three variables, the distance to the thermal, azimuth to the thermal relative to aircraft heading, and the aircraft bank angle were chosen to describe the state in RL. The numerical results showed that the RL-based controller can consistently navigate the aircraft to the ascending thermals and the online glider trajectory generation can be implemented by low-computational-burden table lookup processing in a static state-action value table [19]. Reddy et al. further developed Woodbury et al.’s work using an RL model-free framework. In addition to the bank angle, Reddy et al. noted that the local vertical wind acceleration and rollwise torque were the key variables as navigational cues. The results showed that a longer endurance was achieved with the conversion between gravitational potential and kinetic energy [20].

Bohn et al. first adopted the proximal policy optimization (PPO) algorithm to generate UAV flight controller commands directly [21]. The Skywalker X8 fixed-wing UAV was used as a test aircraft in the simulated turbulent wind field. Up to seven variables related to the UAV flight were selected to establish the state in RL. With a well-trained neural network, it was observed that the RL control framework has several superior characteristics, such as excellent anti-interference ability, precision and quick response compared with the traditional PID algorithm.

Bellemare et al. applied the distributional QR-DQN algorithm to realize stagnation point maintenance of high-altitude stratospheric balloons by determining the wind field use
The RL framework was applied to create a high-performing flight controller. The data self-correcting system was introduced in their work to prevent the effect of imperfect data. A 39-day controlled experiment over the Pacific Ocean was implemented, and the experiment results showed that the RL-based controller outperformed the conventional control algorithm.

The aforementioned studies indicated that RL can process high-dimensional data by optimizing long-term goals and learning information, which has great potential in autonomous path planning and end-to-end aircraft control. Therefore, an alternative method might be to combine solar-powered UAV path planning with EMS as an integral flight strategy. However, few works have reported the application of machine learning to solar-powered HALE unmanned aircraft.

Given this lack, a double deep Q-network (DDQN) with dueling architecture is chosen as the reinforcement learning algorithm for solar-powered aircraft trajectory planning for the following reasons. First, the basic Q-learning algorithm and the modified algorithms have very strong robustness and can be applied to a variety of tasks, such as video games [33] and quadrotor control [34]. Second, a deep Q-network algorithm has been used to guide the path of air vehicles and has achieved very good results [22], [35]. To our knowledge, this is the first time that the reinforcement learning method has been completely applied to HALE solar-powered aircraft flight trajectory planning. The present work makes the following contributions:

(1) The reinforcement learning framework is innovatively applied to the three-dimensional energy-optimal flight strategy of HALE solar-powered UAVs.

(2) Aiming at the mission that the UAV is restricted to specific regions, a neural network controller is designed by the model-free reinforcement learning method to realize automatic navigation. The flight commands of thrust, attack angle and bank angle are directly generated by the proposed controller.

(3) The proposed controller does not require a precomputed feasible trajectory via a motion planning algorithm and can achieve a millisecond response.

(4) A sustained flight test is first implemented to explore the applicability of the RL controller in divergent flight conditions. Compared to the base case with a constant speed and altitude, the present work shows an additional 50 percent flight endurance in sustained flight.

The remaining part of the paper proceeds is organized as follows. The dynamic simulation models are described in Section II. The configuration of a model-free DRL algorithm and the key design decisions are presented in Section III. Then, in Section IV, a level flight base case is described, and the training process is presented on this basis. The controller is evaluated considering 24-hour flight and sustained flight cases. Finally, Section V presents the final comments and suggestions for further work.

II. MODELS

In the air, HALE aircraft can absorb solar energy from the environment and convert it into electric energy through the photovoltaic cells, which is allocated to the propulsion and avionics systems or stored in rechargeable batteries. The interconnection of these subsystems on HALE solar-powered aircraft is shown in Fig. 1.

In this section, the numerical models used in the present work are introduced in five parts: aircraft dynamics and kinematics model, solar irradiance model, energy absorbed model, energy consumption model and energy storage model, respectively.

A. AIRCRAFT DYNAMICS AND KINEMATICS MODEL

For trajectory optimization problem research, the focus is on the aircraft’s macroscopic characteristics during the flight. Therefore, a simplified mass point model can be used. Assuming the aircraft is not affected by wind and sideslip angle during the flight and ignoring the propulsion system installation angle, Fig. 2 shows the aircraft body frame.
By analyzing the forces on the aircraft, the dynamics and kinematics equations of the aircraft can be expressed in the following differential form. A detailed derivation can be found in [23]:

\[
\begin{align*}
\frac{dV}{dt} &= \frac{1}{m}(T \cos \alpha - D - mg \sin \gamma) \\
\frac{d\gamma}{dt} &= \frac{mV}{mV \cos \gamma}[(T \sin \alpha + L) \cos \phi - mg \cos \gamma] \\
\frac{d\psi}{dt} &= \frac{mV \cos \gamma}{mV \cos \gamma}[(T \sin \alpha + L) \sin \phi] \\
\frac{dx}{dt} &= V \cos \gamma \cos \psi \\
\frac{dy}{dt} &= V \cos \gamma \sin \psi \\
\frac{dz}{dt} &= -V \sin \gamma \\
\end{align*}
\]

(1)

where \(x, y\) and \(z\) are the aircraft positions in the earth-fixed inertial frame \(O_gX_gY_gZ_g\), which point to the orientations of north, east, and down, respectively. \(V\) is the aircraft velocity, \(\alpha\) is the attack angle measured between the \(O_bX_b\) axis and the \(V\) vector, and \(\gamma\) is the flight path angle. \(\psi\) is the yawing angle, which is equal to the heading angle because there is no sideslip. \(\phi\) is the bank angle, \(m\) is the aircraft mass, and \(g\) is the acceleration of gravity.

Another Euler angle, pitch angle \(\theta\), is calculated by the following formula:

\[
\theta = \gamma + \alpha
\]

(2)

In addition to gravity, the aircraft is subject to aerodynamic drag force, lift force and thrust generated by the propulsion system, whose directions are shown in Fig. 2: the thrust is aligned with the \(O_bX_b\) axis, the drag is parallel to the negative velocity vector, and the lift is perpendicular to the velocity vector.

Then, lift force \(L\) and drag force \(D\) can be calculated by

\[
\begin{align*}
L &= \frac{1}{2} \rho V^2 C_L S \\
D &= \frac{1}{2} \rho V^2 C_D S
\end{align*}
\]

(3)

where \(\rho\) is the air density around the aircraft that is calculated according to the 1976 US Standard Atmospheric Model [24]. \(C_L\) and \(C_D\) are the lift and drag coefficients. \(S\) is the wing area. In this work, \(C_L\) and \(C_D\) are fitted to a set of Reynolds number \(Re\) and attack angle \(\alpha\) equations, where the Reynolds number is calculated according to the local density and flight velocity [25]. The fitted curves of the lift coefficient and drag coefficient are shown in Fig. 3, and some aerodynamic coefficients under certain working conditions are listed in Table 2.

### Table 2. Aerodynamic coefficients under certain working conditions.

| Operating condition 1 | \(Re = 1089880\) | \(Re = 305058\) |
|-----------------------|------------------|------------------|
| \(\alpha\)\(^{-4\circ}\) | \(\alpha = 0\circ\) | \(\alpha = 4\circ\) | \(\alpha = 4\circ\) |
| \(C_L\) | 0.056 | 0.402 | 0.741 | 0.041 | 0.377 | 0.711 |
| \(C_D\) | 0.022 | 0.012 | 0.021 | 0.026 | 0.015 | 0.024 |

where \(\rho\) is the air density around the aircraft that is calculated according to the 1976 US Standard Atmospheric Model [24]. \(C_L\) and \(C_D\) are the lift and drag coefficients. \(S\) is the wing area. In this work, \(C_L\) and \(C_D\) are fitted to a set of Reynolds number \(Re\) and attack angle \(\alpha\) equations, where the Reynolds number is calculated according to the local density and flight velocity [25]. The fitted curves of the lift coefficient and drag coefficient are shown in Fig. 3, and some aerodynamic coefficients under certain working conditions are listed in Table 2.

### B. SOLAR IRRADIANCE MODEL

According to [26], the solar energy that can be absorbed by HALE aircraft in the daytime consists of two parts: (1) direct solar irradiance, which is directly radiated to the photovoltaic cells installed on the upper wing surface, and (2) diffused...
The solar flux through the aircraft photovoltaic cells, which is converted into electric energy, can be calculated based on the geometric relationship between the normal vector of the solar cell plane and the incidence vector of sunlight [29]. For a photovoltaic panel with an area of $S_{PV}$, its external unit normal vector in the aircraft body coordinate system is represented as $[x_i, y_i, z_i]^T$; then, the solar flux through this cell can be obtained by

$$\Phi_i = I_{tot} S_{PV} \cos \langle n_s, n_{pm} \rangle$$

(4)

where $n_s$ is the unit vector pointed from the photovoltaic cell to the sun and $n_{pm}$ is the external normal unit vector of the photovoltaic module plane, as shown in Fig. 4.

They are given by

$$n_s = [-\cos \alpha \gamma \sin \theta, -\cos \alpha \gamma \cos \theta, -\cos \alpha \sin \gamma]^T$$

$$n_{pm} = R_{gb} [x_i, y_i, z_i]^T$$

(5)

(6)

where

$$R_{gb} = \begin{bmatrix} \cos \psi & -\sin \psi & 0 \\ \sin \psi & \cos \psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi & \cos \phi \end{bmatrix}$$

(7)

denotes the coordinate transformation matrix from the aircraft body coordinate system to the inertial coordinate system, using shorthands $\alpha \gamma$ and $\alpha \sin \gamma$ to represent $\cos(\alpha \gamma)$ and $\sin(\alpha \gamma)$. Then, the total input energy power converted from solar flux $P_{solar}$ is

$$P_{solar} = \sum_i \eta_{MPPT} \eta_{PV} \Phi_i$$

(8)

where $\eta_{MPPT}$ is the efficiency of the maximum power point tracking (MPPT) controller and $\eta_{PV}$ is the efficiency of the photovoltaic cells.

**D. ENERGY CONSUMPTION MODEL**

Solar-powered aircrafts rely on propellers to provide forward thrust to overcome aerodynamic drag and some necessary avionics systems need to consume energy. Hence, during the period of all-day flight, the energy consumption is composed of the above two parts [30]. In this work, a simplified propulsion system model and constant avionics power are taken to calculate the power required $P_{required}$ [31].

$$\begin{cases} P_{prop} = \frac{TV}{\eta_{prop} \eta_{mot}} \\ P_{required} = P_{prop} + P_{acc} \end{cases}$$

(9)

where $P_{acc}$ denotes avionics power, $P_{prop}$ denotes the power of the propulsion system, $\eta_{prop}$ denotes the propeller efficiency and $\eta_{mot}$ denotes the motor efficiency.

**E. ENERGY STORAGE MODEL**

Currently, solar-powered UAVs mostly use Li-based batteries, which have the advantages of high specific energy and stable charging and discharging processes [14]. The battery pack on the plane can either power all the systems or store excess solar energy for use at night. SoC is used to estimate the remaining available energy in the battery compared to the full-charge state [32].

$$\text{SoC} = \frac{E_{\text{battery}}}{E_{\text{battery,max}}} \times 100\%$$

(10)

Eq. 10 states the definition of SoC, where $E_{\text{battery}}$ is the temporal stored energy of the battery, $E_{\text{battery,max}}$ is the maximum battery pack capacity, which ignores capacity loss due to cyclic charge and discharge and is set to a constant value in the present work.
TABLE 3. Formulas for calculating solar irradiance [27], [28].

| Name                              | Formulas                                                                 |
|-----------------------------------|--------------------------------------------------------------------------|
| The day number of the year        | $n_d$                                                                    |
| Radius of the earth               | $R_{earth}$                                                              |
| Local longitude                   | $L_{eq}$                                                                 |
| Altitude                          | $h$                                                                      |
| Apparent solar time               | $H_a = H_{cl} + (L_{eq} - L_{eq})/15 + E_{cl}/60$                        |
|                                  | $E_{cl} = 0.0172 + 4.28 \cos B - 7.35 \sin B - 3.34 \cos 2B - 9.732 \sin 2B$ |
|                                  | $B = 360(n_d - 1)/365$                                                  |
| Hour angle                        | $\theta_h = 15^\circ \times (H_s - 12)$                                |
| Solar declination angle           | $\delta_s = 23.45 \sin [360(n_d + 284)/365]$                           |
| Solar altitude angle              | $\alpha_s = \sin^{-1}(\sin \phi_{hi} \sin \delta_s + \cos \phi_{hi} \cos \delta_s \cos \theta_h)$ |
| Solar azimuth angle               | $\sin \gamma_s = \frac{\cos \delta_s \sin \theta_h}{\cos \alpha_s}$ |
| Extraatmospheric solar irradiance | $I_{on} = 1367 \left[1 + 0.033 \cos \left(360n_d/365\right)\right]$ |
| Direct solar irradiance           | $I_{dir} = I_{on} \exp \left\{ -\frac{C}{D^{0.674h/30}} \right\}$, where $C = 0.357 \exp \left(\frac{h}{7}\right), D = \sin \left(\frac{\alpha_s + \alpha_{dep}}{1 + \alpha_{dep}/90^\circ}\right)$ |
| Diffused solar irradiance         | $I_{diff} = 0.08 I_{dir} \exp(-h/7)$                                    |
| Total available solar flux        | $I_{tot} = I_{dir} + I_{diff}$                                           |

SoC can be updated by the following Eq. 11 and here, the differences in battery charging and discharging processes are not considered:

$$SoC = \frac{E_{battery,0} - \int P_{battery} dt}{E_{battery,max}}$$  (11)

where $E_{battery,0}$ is the initial electric energy in the battery pack and $P_{battery} = P_{solar} - P_{required}$ is the net battery power. For battery charging, $P_{battery}$ is positive because the energy consumed in flight is less than the energy converted by the photovoltaic cells; otherwise, it is negative. When the battery is fully charged $P_{battery} = 0$. Therefore, the net energy power $P_{net}$ is defined below, and it is not less than zero if there is the possibility to make a long flight:

$$P_{net} = P_{solar} - P_{battery} - P_{required}$$  (12)

In addition to batteries, HALE aircraft can store gravitational potential energy $E_{potential}$ through elevation that can be released by gliding down to maintain the airspeed required for flight at night. In this article, for purposes of the net energy expression, only the gravitational potential energy relative to the initial height $h_0$ is considered:

$$E_{potential} = mg(h - h_0)$$  (13)

In summary, through establishing the above five mathematical models, the aircraft state can be solved in real time through a numerical solution method in the simulation process, which provides a detailed description of the environment in the reinforcement learning framework for the aircraft trajectory optimization problem. The pseudocode of the calculation flow is shown in Algorithm 1. Detailed parameters and some physical constraints of the studied HALE solar-powered aircraft are listed in Table 4 and Table 5, respectively.

III. REINFORCEMENT LEARNING METHODOLOGY

The objective is to utilize reinforcement learning to train a neural network controller offline and to provide control commands to achieve superior flight endurance. However, since the stability of the solar UAV is greatly affected by the pitch angle, its climbing process can be determined by the thrust and angle of attack. Therefore, the designed RL controller is
tasked with outputting the control commands of bank angle, thrust and attack angle online according to the flight states. Here, it is assumed that the inner-loop controller is ideal and can automatically track its control instructions. At each time step, the controller receives an immediate reward, and its goal is to establish a control law that maximizes the sum of future discount rewards.

### A. DOUBLE Q-NETWORK WITH DUELING ARCHITECTURE

DDQN with a dueling architecture algorithm belongs to the model-free off-policy reinforcement learning methods. Considering the readability and completeness, the DDQN algorithm and dueling architecture are introduced briefly.

The Q-learning algorithm aims to map each state-action pair \((s, a)\) to its corresponding value \(Q(s, a)\) and select the action corresponding to the maximum \(Q(s, a)\) according to the strategy; one of which is that under a given strategy \(\pi\), the value of an action in state \(s\) can be indicated as:

\[
Q_\pi(s, a) = \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t = s, a_t = a, \pi]
\]

(14)

where \(\gamma \in [0, 1]\) is a discount factor that trades off the importance of immediate and later rewards. By selecting the action with the highest optimal value \(Q_\pi = \max_a Q_\pi(s, a)\) in each state, it is easy to derive the optimal strategy from the optimal value.

With the expansion of the state space, the traditional table lookup method cannot meet the requirements, so the DQN algorithm [33] is proposed in which a multilayer neural network is used to approximate the optimal action-value function \(Q(s, a)\). Once the current state \(s\) is given, these networks evaluate the \(Q\) values for all action state pairs. However, it is not necessary to estimate the value of each candidate action for all states. Wang et al. [38] proposed the dueling structure to improve the DQN performance. This structure separates the state value estimation function \(V(s)\) and action advantage function \(A(s, a)\), which represents a relative measure of the importance of each action in that state, and recombines them through a special aggregating operation to approximate the Q-value of each action:

\[
Q(s, a; \theta) = V(s) + A(s, a) - \frac{1}{|A|} \sum_{a'} A(s, a')
\]

(18)

where \(|A|\) is equal to the size of the designed action set.

In the DQN and DDQN algorithms, once the current state \(s\) is given, these networks evaluate the \(Q\) values for all action state pairs. However, it is not necessary to estimate the value of each candidate action for all states. Wang et al. [38] proposed the dueling structure to improve the DQN performance. This structure separates the state value estimation function \(V(s)\) and action advantage function \(A(s, a)\), which represents a relative measure of the importance of each action in that state, and recombines them through a special aggregating operation to approximate the Q-value of each action:

\[
Q(s, a; \theta) = V(s) + A(s, a) - \frac{1}{|A|} \sum_{a'} A(s, a')
\]

(18)

where \(|A|\) is equal to the size of the designed action set.

### B. STATE SPACE

The state space consists of the position, attitude, local clock time, battery and action information, so at time step \(t\), the state can be represented by

\[
s_t = [x_t, y_t, h_t, \phi_t, \theta_t, \psi_t, V_t, \alpha_t, T_t, H_{cl}, \text{SoC}]
\]

(19)

The ODEINT package in Python SciPy is employed to solve the differential equations in Section II and update the state [9]. The smaller the simulation step size of the formulation integration taken, the more accurate the attitude obtained. However, to reduce the number of interactions between agents and the environment in each episode of reinforcement learning and reduce the calculation time, the aircraft state is observed once every \(n\) iteration steps, and the corresponding command is selected, which is retained in the next skipped iteration practice steps [40]. This process is illustrated visually in Fig. 5.

The components of the state vector are expressed in different units and have different dynamic ranges. Therefore, the convergence process may fluctuate when the neural network is updated based on gradient descent. To improve the training speed and accuracy, all elements are normalized to the range of 0 to 1 by Eq. 20 so that the neural network does
Algorithm 1 Pseudocode for the HALE Aircraft Simulation Environment

**Input:** state $s_t$, instruction action $u_{cmd}$

for iterationstep $= 1$ to $n$ do

- Calculate aerodynamic lift $L$ and drag $D$ by Eq. 3
- Update current control variable $u$ by Eq. 23
- Update attitude and location by Eq. 1-2 using the Runge-Kutta method
- Calculate total available solar flux $I_{tot}$ using the equations in Table 3
- Calculate total input energy power $P_{solar}$ using Eq. 4-8
- Calculate power required $P_{required}$ using Eq. 9
- Update battery state SoC using Eq. 11 and potential energy $E_{potential}$ using Eq. 13

end for

**Output:** state $s_{t+1}$

Algorithm 2 Double DQN With Dueling Architecture

Initialize replay buffer memory $D$ to capacity $N$

Initialize the network parameters $\theta$

Initialize the target network parameters $\theta^{-} \leftarrow \theta$

for episode $= 1$ to $M$ do

- Initialize $s_1$
- Initialize preprocessed sequence $g_1 = g(s_1)$

for timestep $= 1$ to $T$ do

- Select a random action $a_t$ with probability $\epsilon$ or $a_t = \arg\max_a Q(g(s_t), a; \theta)$
- Execute action $a_t$
- Update flight state by Algorithm 1
- Observe the reward $r_t$ and $s_{t+1}$
- Preprocess $g_{t+1} = g(s_{t+1})$ and rescale
- Store transition tuple $(g_t, a_t, r_t, g_{t+1})$ in $D$, and replace the oldest tuple if $||D|| > N$

- Sample a minibatch size $N_m$ tuples $(g_j, a_j, r_j, g_{j+1})$ from $D$ randomly
- Set $y_j = \begin{cases} r_j, & \text{if episode terminates} \\ r_j + \gamma Q(g_{j+1}, \arg\max_{a_{j+1}} Q(g_{j+1}, a_{j+1}; \theta^{-}); \theta), & \text{otherwise} \end{cases}$
- Update network parameters $\theta$ by performing a gradient descent step with loss:
  $$L = \frac{1}{N_m} \sum_j ||y_j - Q(g_j, a_j; \theta)||^2$$
- Update the target network: $\theta^{-} \leftarrow \tau \theta + (1 - \tau)\theta^{-}$
- Set $s_{t} \leftarrow s_{t+1}$

end for

end for

FIGURE 5. Schematic diagram of the preprocessed sequence.

not have to deliberately learn this scale [21].

$$f(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$ (20)

In addition, because the policy is a feedforward network with no memory, the input vector of each time step $t$ is composed of the values of the previous several time steps to facilitate the acquisition of dynamics.

C. ACTION SPACE

In this work, the action space of the controller is three dimensional, consisting of the commanded increment, thrust $T_{cmd}$, attack angle $\alpha_{cmd}$ and bank angle $\phi_{cmd}$. From Eq. 1, the aircraft can change the attitude and flight altitude by these three commands. Because the DDQN with dueling architecture needs discrete action space, it is designed to have five basic values for the change in each command, as shown in Table 6, which means 125 action pairs can be chosen by the controller. To achieve more accurate control in some situations and to be able to change attitude quickly, numerical values of symmetry are selected to guarantee not to exceed physical limits during the response.
TABLE 6. Values for the change in each command.

| Command Type | Alternative values |
|--------------|--------------------|
| $T_{cmd}$ | $-10N$ | $-1N$ | $0N$ | $+1N$ | $+10N$ |
| $\alpha_{cmd}$ | $-0.5^\circ$ | $-0.1^\circ$ | $0^\circ$ | $+0.1^\circ$ | $+0.5^\circ$ |
| $\phi_{cmd}$ | $-0.5^\circ$ | $-0.1^\circ$ | $0^\circ$ | $+0.1^\circ$ | $+0.5^\circ$ |

D. REWARD

In reinforcement learning, the reward function directly affects the training results. Therefore, it is very important to design a reward function that is reasonable and has realistic physical implications. Generally, the goal of solar-powered UAV trajectory optimization during high-altitude navigation is to achieve longer flight endurance or obtain more energy than the initial condition after a 24-hour flight under the premise of meeting mission constraints and system dynamics. Although environmental uncertainty and other factors affect flight time, the total energy available in the system is still the main factor affecting flight time [11]. Therefore, the reward function constructed for the HALE solar-powered aircraft is shown in Eq. 21.

\[
r_t = \begin{cases} 
E_{\text{battery}} + \kappa E_{\text{potential}}, & \text{Normal flight phase} \\
-100, & \text{Crash}
\end{cases} 
\]  

(21)

where $E_{\text{battery}}$ is the temporal energy stored in the battery and $E_{\text{potential}}$ is the gravitational potential energy relative to the initial height calculated by Eq. 13. $\kappa$ is a weight coefficient used to amplify the influence of potential energy and is equal to 0 at night to save battery power:

\[
\kappa = \begin{cases} 
\frac{E_{\text{battery,max}} - E_{\text{battery,min}}}{E_{\text{potential,max}}}, & \alpha_s > 0 \\
0, & \alpha_s \leq 0
\end{cases} 
\]  

(22)

To be closer to the situation of a real flight, the rules are made that when the state of the aircraft is beyond the constraint range, such as angle of attack exceeding the specified maximum or the HALE aircraft flies out of the specified altitude, the aircraft is judged to crash, and a larger penalty is given.

E. THE INNER-LOOP RESPONSE MODEL

To concentrate on the autonomous navigation problem, it is assumed that an ideal low-level controller can track the control command $u_{cmd}$ within the constant response time range after the control commands are given. So $\phi$, $T$, $\alpha$ in Eq. 1 can be calculated by Eq. 23 until the next control commands are chosen.

\[
u(t) = \begin{cases} 
u_{(t_0)} + \frac{u_{cmd}}{t_a}(t - t_0), & t < t_a \\
u_{(t_0)} + u_{cmd}, & t \geq t_a
\end{cases} 
\]  

(23)

where $u(t)$ represents the response value to the command at time $t$, $u_{cmd}$ is the control command given at $t_0$ and $u_{(t_0)}$ is the initial value. $t_a$ is the constant response time range, which means that after this time, the specified command can be fully tracked.
to lead to a minimum thrust power requirement at a level flight state. Similar base cases were also used in [9] and [14] to compare their optimized trajectories but with different radius constraints and flight test locations.

The minimum thrust power when the aircraft cruises on a circular trajectory can be estimated by solving the following optimization problem.

\[
\text{minimize } P_{\text{required}} \\
\text{with respect to } T, \alpha, \phi, V, h \\
\text{subject to } \frac{dV}{dt} = 0, \quad \frac{d\gamma}{dt} = 0, \quad \frac{dz}{dt} = 0, \quad R = R_{\text{max}} \tag{24}
\]

Since the base case has a steady circular trajectory, the initial conditions are consistent with the results of Eq. 24, as shown in Table 7. On this basis, the 24-hour flight data were calculated, and its three-dimensional trajectory is shown in Fig. 9. It can be found that the final SoC was 27.1% after a 24-hour flight, which is lower than the initial SoC of 30%, and the altitude was maintained at 15 km, which is the same as the initial altitude. The 24-hour flight data of base-case indicates the total stored energy of aircraft $E_{\text{tot,stored}}$ decreases with the initial conditions. This case provides a baseline for the subsequent comparisons.

C. NETWORK ARCHITECTURE AND HYPERPARAMETER SETTING

The network architecture adopted is shown in Fig. 10: there are 2 fully-connected layers with 512 units, and then the dueling network splits into two streams of fully-connected layers. The value streams have a fully-connected layer with 256 units, and the advantage streams have a fully-connected layer with 128 units. The final hidden layers of the value and advantage streams are both fully-connected with the value stream having one output and the advantage of as many outputs as valid actions.

The discount factor is set to $\gamma = 0.99$, and the Adam optimizer [41] is chosen for training the neural network with a learning rate of 0.0001 and minibatch size of 64. For the soft target update coefficient $\tau = 0.001$. The replay buffer size is $10^6$. The above hyperparameter settings are referenced from [42]. In addition, the exploration rate $\epsilon$ linearly decreased from 0.7 to 0.1 over a period of 1 million steps then fixed.

The initial conditions for each episode are consistent with the counterparts of the base case shown in Table 7.

![FIGURE 7](image_url) The total available solar flux at different altitudes above Beijing in the summer solstice.

![FIGURE 8](image_url) 3-D trajectory of the 24-hour flight generated by the base case.

![FIGURE 9](image_url) Altitude and SoC curves of the 24-hour flight generated by the base case.

| Parameter | Description | Value |
|-----------|-------------|-------|
| $t_0(h)$  | Initial time | 5     |
| $t_f(h)$  | Final time  | 5 (Next day) |
| $[x_0, y_0]^T$ (km) | Initial position | $[5, 0]^T$ |
| $h_0$(km) | Initial height | 15    |
| $V_0$(m/s) | Initial velocity | 21.8 |
| $\gamma_0(°)$ | Initial flight path angle | 0     |
| $\psi_0(°)$ | Initial heading angle | 90    |
| $\alpha_0(°)$ | Initial attack angle | 7.9   |
| $\phi_0(°)$ | Initial bank angle | 0.6   |
| $T_0$(N) | Initial thrust | 45.6  |
| $\text{SoC}_0$ | Initial state of battery package | 0.3   |
| $t_a(s)$ | Response time for inner-loop | 2     |
FIGURE 10. The adopted architecture of the DDQN with dueling.

FIGURE 11. Total reward of every test episode, which is smoothed by a moving average.

The flight data are updated every 1 second, and the RL controller generates a command every 20 seconds. Therefore, there are at most 4,320 timesteps per episode. When a 24-hour flight is completed or the aircraft is determined to crash, the episode restarts with the initial conditions.

The test episode is run every 1 million timesteps, and the total reward of the episode is visualized during the training, as shown in Fig. 11. Using the models and constraints described above, the final controller can be obtained in approximately 140 million steps. The trained model takes 1 millisecond to output flight commands on a laptop computer with an i5-10210U CPU when tested. This means that the RL controller can reasonably be expected to operate in real flight. Additionally, the calculation time of the single-core CPU required by different online algorithms in predicting the 30-minute-horizon trajectory is selected to further compare the real-time performance of the method [10]. The results are shown in Table 8 with the same sampling time achieved with Marriott’s work, which shows that the computational time greatly decreases by using the end-to-end control capabilities of the RL controller.

D. 24-HOUR FULL TRAJECTORY OF RL-CONTROLLED AIRCRAFT

The well-trained RL controller is employed to test the performance in a 24-hour full trajectory. The controller selects an action at intervals of 10 s and 20 s. The comparison of two cases is shown in Table 9. The relative differences between the maximum stored energy and the final SoC in the 24-hour trajectory are -0.17% and -0.04%, respectively, so the sampling time is set at 20 s for the presentation of the results. At the same time, the longer sampling time can also use the remaining CPU computing time to process other aircraft tasks, such as reconnaissance data processing and communication relay task assignment.

The altitude and SoC curves of the 24-hour flight are shown in Fig. 12(a). The power consumption time histories of the RL-controlled aircraft propulsion system, base-case propulsion system and avionics system are presented in Fig. 12(b). It is observed that the RL controller learned a strategy that can be divided into five intuitive stages: preclimb, climb, high-altitude cruise, glide descent and low-level hover. Complete three-dimensional flight trajectories are presented in Fig. 13. Moreover, typical trajectories of the first four stages are shown in Fig. 14.

1) STAGE1: PRECLIMB

The preclimb stage lasted from 5:00 to approximately 6:50. Since the total available solar flux increased with altitude, the RL controller decided to make full use of the battery’s remaining energy for a slight climb from 15 km to 17 km until SoC reached the minimum value of 0.205, which was

| Method                           | Compute Time |
|----------------------------------|--------------|
| Gradient [29]                    | 37.50 min    |
| MBC [9]                          | 12.03 min    |
| Greedy with buffering [10]       | 1.5 min      |
| RL controller                    | 1.6 s        |

TABLE 9. Comparison of different sampling times in the 24-hour flight test process.
very close to the limited state. This caused an increase in solar energy absorption due to a higher altitude. A dramatic power requirement was found simultaneously in the beginning of the preclimb stage, as shown in Fig. 12(b), which is consistent with the RL controller command. Subsequently, a short cruise at 17 km altitude was found, which induced a decline in propulsion power and gave a slight recharge to ensure the health of the battery SoC until the average solar flux was sufficient to continuously climb after 6:40. A closed view of the aircraft flight altitude and battery charging power can be seen in Fig. 17. In addition, Fig. 14(a) shows a typical trajectory for a circle during this process, and Table 10 shows the
RL-controlled aircraft had an orbital period of approximately 0.31 h with an orbital radius of 4,200 m, which was faster than the 0.4 h of the base case. This was due to the increase in flight speed at high altitudes. Fig. 15(b) shows the detail of the time-based comparison of solar energy absorption between the RL controller case and base case. It can be seen clearly that the RL-controlled aircraft absorbed more solar power in the majority of the preclimb stage. The preclimb stage shows the RL controller attempted to balance the solar energy absorption and flight energy cost even in an ultimate state. To the best of our knowledge, this is the first report of a preclimb stage, although it is reasonable intuitively.

2) STAGE2: CLIMB

The climbing stage lasted from approximately 6:50 to 12:10. It is shown in Fig. 12 that the aircraft decided to increase the
altitude rather than recharge the batteries, which implies that the RL controller understands the increase in flight altitude is more significant than recovering the battery SoC, and there is enough time to recharge during the remaining daytime. Therefore, a full altitude-priority strategy was generated. When the aircraft approaches 24 km, the aircraft decreased the climbing rate in the case of exceeding the ceiling of 25 km. A typical trajectory for a circle during the climbing stage is shown in Fig. 14(b), and a detailed time-based comparison of solar power received between the RL controller and base case is shown in Fig. 15(c). Consistent results with the preclimb stage show that the RL-controlled aircraft receives more solar energy than the base case by using a higher altitude with a fast circular motion. The orbital period decreased to 0.23 h with a radius of 3,981 m at approximately 8:00 due to a continuous increase in the flight speed. The aircraft stored the gravitational potential energy of 3.6 kWh and 60% SoC at the end of this stage.

3) STAGE3:HIGH-ALTITUDE CRUISE
The high-altitude cruise stage lasted from 12:10 to 18:30. The aircraft cruised in a circle trajectory between 24 km to 24.5 km altitude for more than 6 hours and the battery SoC achieved 100% at approximately 15:00, see Fig. 12. A typical trajectory for a circle during the high-altitude cruise stage is shown in Fig. 14(c), and a detailed time-based comparison of solar power received between the RL-controlled aircraft and base case is shown in Fig. 15(d). It can be seen that the RL-controlled aircraft gained an additional 0.3 kW boost in solar power absorption on average compared with the base case at approximately 12:00. Moreover, a slight climb was found after the battery SoC became full, which further implies that the RL controller attempted to take full advantage of solar power under a ceiling of 25 km though the solar power dramatically decreased compared to midday. After 17:00, slight declines in both flight altitude and SoC were found due to insufficient solar power. This hints that the RL controller began to turn its strategy from altitude priority to SoC priority.

4) STAGE4:GLIDE DESCENT
The glide descent stage was from 18:20 to 21:30. Due to a sustained decline in solar power, the RL controller fully turned to the SoC priority strategy to maintain the battery energy. The aircraft thrust decreased gradually to zero in this stage, and the aircraft hardly consumed any battery energy except the necessary power to keep avionics working, as shown in Fig. 16(a). In this stage, the aircraft fully released the gravitational potential stored in the daytime until it reached the minimum height limit of 15 km. A typical trajectory for a circle during this stage is shown in Fig. 14(d), and a detailed time-based comparison of solar power receiving is shown in Fig. 15(e). It can be seen that the RL controller gained more solar power in the majority of this stage compared to the base case due to a higher altitude. This helped to reduce the battery energy consumption caused by avionics.

5) STAGE5:LOW-ALTITUDE HOVER
When the aircraft reached the minimum altitude limit of 15 km, it flew around the outer circle of the mission radius until the sun came up the next day. It attempted to maintain optimal output power of 1.75 kW while maintaining altitude. Due to the use of discrete action commands, its power consumption showed a small oscillation, but the average was generally close to the minimum power of the base case, as shown in Fig. 16(b).

The last four stages were also found in Wang et al’s results [15] to maximize the net energy absorption in a 24-hour flight circle. However, in Wang’s research, the stages were predefined in the problem setting before applying the pseudospectral method with flexible parameters, which limits the type and number of stages. This might cause the preclimb stage to not be found.

Fig. 12(b) shows the 24-hour time histories of the RL controller thrust power, base-case thrust power and avionics power. Due to the discrete control command, the thrust power of the RL controller oscillated. However, it can be clearly seen from the smoothed curve that the thrust power of the RL controller was approximately 0 kW in the glide descent stage and consistent with the base case’s in the low-altitude hover stage.
Fig. 18 shows the comparisons of the battery SoC and total stored energy $E_{\text{tot,stored}}$ between the RL controller and base case. The results imply that a maximum total stored energy of 27.6 kWh was received by the RL controller due to the gravitational potential store, which is 15% higher than the base case’s total stored energy of 24 kWh. After a 24-hour flight, the total stored energy left was 8.34 kWh, which is 14% higher than the initial state of 7.2 kWh and 27.9% higher than the base case’s 6.5 kWh. The results strongly suggest that the neural network controller trained by reinforcement learning can automatically plan the flight path and support a long-endurance flight.

E. SUSTAINED FLIGHT RESULTS

Previous researchers have focused on flight strategy optimization in one 24-hour flight and many remarkable results have been reported based on the comparison between the trajectory optimization results and the control group [9], [14]. However, how well these fixed optimized flight strategies work when applied in divergent flight conditions, such as different days, without a reoptimization process, remains a question. To explore the applicability to unknown solar conditions, the RL controller trained in the aforementioned 24-hour flight was directly applied to a sustained flight test. The initial conditions were consistent with those in Table 6. A detailed comparison between the RL controller performances and base case in a sustained flight test were implemented.

The time histories of flight altitude and battery SoC are presented in Fig. 19. It can be seen that an entire 39-day enduring flight was reached from the departure date of the summer solstice. Five flight stages, preclimb, climb, high-altitude cruise, glide descent and low-altitude hover were observed, which was consistent with the 24-hour flight. When the temporal battery energy was sufficient to support the climb, the preclimb stage gradually faded away. Extreme values of maximum SoC, minimum SoC, maximum altitude, minimum
altitude and their reaching time in the sustained flight test are presented in Fig. 21. It is observed the maximum battery SoC reached 100% every day, but the arrival time delayed gradually from 14:45 to 16:00 with the increase in the day number, as shown in Fig. 21(a). However, opposite tendencies were found in the time histories of the minimum SoC and its arrival time, as shown in Fig. 21(b). This is due to a continuous decline in the intensity and duration of solar radiation. Moreover, Fig. 21(c-d) shows few changes were found in the maximum and minimum flight altitudes and their reaching times. However, the cruising time exceeding a 24 km altitude gradually decreased from 6.1 hours on the 5th day to 3.6 hours on the 35th day, and the time to reach 15 km altitude gradually increased earlier (see Fig. 20(a)). And the cruising time maintaining 100% SoC also decreased from 2.7 hours on the 5th day to 1.4 hours on the 35th day, see Fig. 20(b). This shows that the RL controller adjusts the flight strategy dynamically to meet the energy requirement in subsequent flight stages.

In addition, an aircraft with the base-case strategy crashed after 26 days because the battery SoC was lower than the protective value of 20%. Fig. 22(a) shows the total stored energy comparison between the strategy generated by the RL controller and the base case at 5:00 every day. An increase of 28% was observed on the 5th day, which indicates that the strategy by the RL controller is beneficial to maintain long flight. The total stored energy on the 25th day was 8.29 kWh, which is comparable to the 8.22 kWh of the 20th day. This implies that the RL controller attempted to improve its energy.
Table 11 shows the comparison of available solar energy at 20 km altitude in five different stages. The stages were divided according to the starting time of each stage on the first day. It was found the available solar energy in stage 4 and stage 5 is declined sharply due to an earlier sunset. Furthermore, a slight decrease of 3.5% was observed in stage 1 due to the later sunrise. All of these declines led to an insufficient power supply, which caused the crash to occur on the morning of the 40th day.

By combining Eq. 12 and integrating it over 24 hours, the relationship can be inferred between the energies:

\[ E_{\text{net}} = E_{\text{solar}} - \Delta E_{\text{battery}} - E_{\text{required}} \]  
\[ (25) \]

where \( \Delta E_{\text{battery}} \) denotes the subtraction of the battery energy at the end and the beginning of integral. The solar energy utilization ratio \( \eta \) can be defined as

\[ \eta = \frac{E_{\text{required}}}{E_{\text{net}} + E_{\text{required}}} \]  
\[ (26) \]

The physical meaning of the solar energy utilization ratio \( \eta \) is the aircraft’s capacity to use the entire 24-hour solar energy. A higher \( \eta \) indicates less solar energy is wasted.

In Fig. 22(b), it can be found that the RL controller can absorb more energy and arrive at a utilization factor of approximately 95%, which is only 65% of the base case. With this trained RL controller, the flight endurance of the aircraft increased by 50%.

V. CONCLUSION

In the present work, a deep reinforcement learning framework was applied to trajectory planning and automatic navigation of high-altitude long-endurance unmanned aerial vehicles. A station-keeping task was assumed, which requires flight within a radius of 5 km and an altitude region from 15 km to 25 km. The solar radiation model and the flight dynamics model were introduced. The numerical results show that the RL-controlled aircraft received more energy after the entire 24-hour flight, which was 27.9% higher than the base case. Subsequently, a sustained flight was implemented to test the applicability of the RL controller. An entire 39-day flight was achieved whereas the base case only flew 26 days. The presented work shows that the controller generated by reinforcement learning might be an alternative method for improving the HALE solar-powered UAV’s flight endurance.

However, double DQN with dueling architecture is an algorithm for discrete action space; hence, the flight control command may produce vibration. In future work, a continuous action space RL method might be adopted. Moreover, the wind field should be considered in the simulation. Since reinforcement learning aims at maximizing long-term cumulative return, the proportional relationship between potential energy and battery energy for the optimization process still needs to be further explored.

REFERENCES

[1] R. J. Boucher, “Sunrise, the world’s first solar-powered airplane,” J. Aircr., vol. 22, no. 10, pp. 840–846, Oct. 1985.
[2] P. Rajendran and H. Smith, “Review of solar and battery power system development for solar-powered electric unmanned vehicles,” Adv. Mater. Res., vol. 1125, pp. 641–647, Oct. 2015.
[3] J.-K. Shiu, D.-M. Ma, C.-W. Chiu, and J.-R. Shie, “Optimal sizing and cruise speed determination for a solar-powered airplane,” J. Aircr., vol. 47, no. 2, pp. 622–629, Mar. 2010.
[4] A. Klesh and P. Kabamba, “Energy-optimal path planning for solar-powered aircraft in level flight,” in Proc. AIAA Guid., Navigat. Control Conf. Exhibit, Aug. 2007, p. 6655.
[5] A. T. Klesh and P. T. Kabamba, “Solar-powered aircraft: Energy-optimal path planning and perpetual endurance,” in Proc. AIAA Guid., Navigat. Control Conf., Jun. 2009, vol. 32, no. 4, pp. 1320–1329.
[6] Y. Huang, H. Wang, and P. Yao, “Energy-optimal path planning for solar-powered UAV with tracking moving ground target,” Aeron. Sci. Technol., vol. 33, pp. 241–251, Jun. 2016.
[7] S. C. Spangelo and E. G. Gilbert, “Power optimization of solar-powered aircraft with specified closed ground tracks,” J. Aircr., vol. 50, no. 1, pp. 232–238, Jan. 2013.
[8] S. Spangelo, E. Gilbert, A. Klesh, P. Kabamba, and A. Girard, “Periodic energy-optimal path planning for solar-powered aircraft,” in Proc. AIAA Guid., Navigat., Control Conf., Aug. 2009, p. 6016.
[9] R. A. Martin, N. S. Gates, A. Ning, and J. D. Hedengren, “Dynamic optimization of high-altitude solar aircraft trajectories under station-keeping constraints,” J. Guid., Control, Dyn., vol. 42, no. 3, pp. 538–552, Mar. 2019.
[10] J. Marriott, B. Tezel, Z. Liu, and N. E. Stier-Moses, “Trajectory optimization of solar-powered high-altitude long-endurance aircraft,” in Proc. 6th Int. Conf. Control, Autom., Robot. (ICCAR), Apr. 2020, pp. 473–481.
[11] X.-Z. Gao, Z.-X. Hou, Z. Guo, J.-X. Liu, and X.-Q. Chen, “Energy management strategy for solar-powered high-altitude long-endurance aircraft,” Energy Convers. Manage., vol. 70, pp. 20–30, Jun. 2013.
[12] X.-Z. Gao, Z.-X. Hou, Z. Guo, R.-F. Fan, and X.-Q. Chen, “The equivalence of gravitational potential and rechargeable battery for high-altitude long-endurance solar-powered aircraft on energy storage,” Energy Convers. Manage., vol. 76, pp. 986–995, Dec. 2013.
[13] J.-S. Lee and K.-H. Yu, “Optimal path planning of solar-powered UAV using gravitational potential energy,” IEEE Trans. Aerosp. Electron. Syst., vol. 53, no. 3, pp. 1442–1451, Jun. 2017.
[14] S. Wang, D. Ma, M. Yang, L. Zhang, and G. Li, “Flight strategy optimization for high-altitude long-endurance solar-powered aircraft based on gauss pseudo-spectral method,” Chin. J. Aeronaut., vol. 32, no. 10, pp. 2286–2298, Oct. 2019.
[15] X. Wang, Y. Yang, D. Wu, Z. Zhang, and X. Ma, “Mission-oriented 3D path planning for high-altitude long-endurance solar-powered UAVs with optimal energy management,” IEEE Access, vol. 8, pp. 227629–227641, 2020.
[16] A. Alisharofa, H. Ghazzai, M. Yuksel, A. Kadri, and A. E. Kamal, “Trajectory optimization for multiple UAVs acting as wireless relays,” in Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops), Kansas City, MO, USA, May 2018, pp. 1–6.
[17] Z. Yuan, Y. Yang, Y. Hu, and X. Ma, “Channel-aware potential field trajectory planning for solar-powered relay UAV in near-space,” IEEE Access, vol. 8, pp. 143950–143961, 2020.
[18] J. Wu, H. Wang, Y. Huang, Z. Su, and M. Zhang, “Energy management strategy for solar-powered UAV long-endurance target tracking,” IEEE Trans. Aerosp. Electron. Syst., vol. 55, no. 4, pp. 1878–1891, Aug. 2019.

[19] T. D. Woodbury, C. Dunn, and J. Valasek, “Autonomous soaring using reinforcement learning for trajectory generation,” in Proc. 52nd AIAA Sci. Meeting, Jan. 2014, p. 0990.

[20] G. Reddy, J. Wong-Ng, A. Celani, T. J. Sejnowski, and M. Vergassola, “Glider soaring via reinforcement learning in the field,” Nature, vol. 562, no. 7726, pp. 236–239, Oct. 2018.

[21] E. Bohn, E. M. Coates, S. Moe, and T. A. Johansen, “Deep reinforcement learning attitude control of fixed-wing UAVs using proximal policy optimization,” in Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS), Jun. 2019, pp. 523–533.

[22] M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang, “Autonomous navigation of stratospheric balloons using reinforcement learning,” Nature, vol. 588, no. 7836, pp. 77–82, Dec. 2020.

[23] B. Ekin, “Dynamics of atmospheric flight,” Dyn. Atmos. Flight, vol. 69, no. 39, pp. 882–883, 1972.

[24] U. S. N. Oceanic, A. Administration, and U. S. A. Force, US Standard Atmosphere, National Oceanic and Atmospheric Administration, Washington, DC, USA, 1976, vol. 76, no. 1562.

[25] J. Wu, H. Wang, Y. Huang, M. Zhang, and Z. Su, “Solar-powered aircraft endurance map,” J. Guid., Control, =Dyn., vol. 42, no. 3, pp. 687–694, 2019.

[26] A. B. Jemaa, S. Rafa, N. Essounbouli, A. Hamzaoui, F. Hnaien, and U. S. N. Oceanic, A. Administration, and U. S. A. Force, US Standard Atmosphere, vol. 39, pp. 882–883, 1972.

[27] B. Keidel, “Auslegung und simulation von hochfliegenden, dauerhaft stationierbaren solardrohnen,” Ph.D. dissertation, Dept. Maschinenwesen, Technische Universität München, München, Germany, 2000.

[28] J. A. Duffie and W. A. Beckman, “Available solar radiation,” in Solar Engineering of Thermal Processes. New York, NY, USA: Wiley, 1980.

[29] H. Bolandhemmat, B. Thomsen, and J. Marriott, “Energy-optimized trajectory planning for high altitude long endurance (HALE) aircraft,” in Proc. 18th Euc Control Conf. (ECC), Jun. 2019, pp. 1486–1493.

[30] A. Alsaflani and T. Rahulan, “Weight estimation of a conceptual wing for a high altitude, solar powered unmanned aerial vehicle,” in Proc. 5th Airsc. Struct. Design Conf., 2016, pp. 1–4.

[31] M. Asselin, An Introduction to Aircraft Performance. USA: American Institute of Aeronautics and Astronautics, 1997.

[32] S. M. Rezvanizaniani, Z. Liu, Y. Chen, and J. Lee, “Review and recent advances in battery health monitoring and prognostics technologies for electric vehicle (EV) safety and mobility,” J. Power Sources, vol. 256, pp. 110–124, Jun. 2014.

[33] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, 2015.

[34] A. Anwar and A. Raychowdhury, “Autonomous navigation via deep reinforcement learning for resource constraint edge nodes using transfer learning,” IEEE Access, vol. 8, pp. 26549–26560, 2020.

[35] C. Montella, “Learning how to soar: Steady state autonomous dynamic soaring through reinforcement learning,” in Proc. AIAA Scitech Forum, Jan. 2020, p. 1848.

[36] C. J. C. H. Watkins, “Learning from delayed rewards,” Ph.D. dissertation, Psychol. Dept., Univ. Cambridge, Cambridge, U.K., 1989.

[37] H. van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-learning,” 2015, arXiv:1509.06461. [Online]. Available: http://arxiv.org/abs/1509.06461

[38] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, “Dueling network architectures for deep reinforcement learning,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 1995–2003.

[39] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” 2015, arXiv:1509.02971. [Online]. Available: http://arxiv.org/abs/1509.02971

[40] M. Bellemare, J. Veness, and M. Bowling, “Investigating contingency awareness using Atari 2600 games,” in Proc. AAAI 26th Conf. Artif. Intell., 2012, vol. 26, no. 1, pp. 864–871.

[41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] C. Yan, X. Xiang, and C. Wang, “Towards real-time path planning through deep reinforcement learning for a UAV in dynamic environments,” J. Intell. Robot. Syst., vol. 98, pp. 297–309, Sep. 2019.

WENJUN NI received the bachelor’s degree in energy and power engineering from Shandong University, China, in 2018. He is currently pursuing the Ph.D. degree in aircraft design with the Institute of Engineering Thermophysics, Chinese Academy of Sciences, Beijing, China. His research interests include UAV systems, intelligent control of aircraft, and applications of reinforcement learning in solar-powered aircraft systems.

DI WU received the bachelor’s and master’s degrees in aeronautical and astronautical propulsion engineering from Northwestern Polytechnical University, in 2011 and 2014, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, in 2018. From 2018 to 2021, he served as an Assistant Professor with the Institute of Engineering Thermophysics, Chinese Academy of Sciences, Beijing, China. His research interests include low Reynolds number aerodynamics and aeroacoustics.

XIAOPING MA received the bachelor’s, master’s, and Ph.D. degrees in aircraft design from North-west Polytechnical University, China, in 1983, 1988, and 2008, respectively. He has been engaged in the research and development of UAV systems for a long time and has participated in various types of UAV developments, serving as the Deputy Chief Engineer and the Chief Engineer. He has published more than 20 articles related to UAV designs, among which five were retrieved by EI/SCI/ISTP/SSCI. His research interests include UAV systems, aircraft and structure design, launching and recovering techniques, and system flight tests. He serves as a Senior Member of the Aviation Society of China. He has received the first prize and the second prize of the National Science and Technology Progress, the second prize of the National Commission of Science, Technology and Industry for National Defense, and the first prize and the second prize of the Ministerial Awards.