International Journal of Powertrains

ISSN online: 1742-4275 - ISSN print: 1742-4267
https://www.inderscience.com/ijpt

Energy consumption optimisation for unmanned aerial vehicle based on reinforcement learning framework

Ziyue Wang, Yang Xing

DOI: 10.1504/IJPT.2024.10057473

Article History:

Received:	03 October 2022
Last revised:	28 April 2023
Accepted:	03 May 2023
Published online:	16 April 2024

Energy consumption optimisation for unmanned aerial vehicle based on reinforcement learning framework

Ziyue Wang* and Yang Xing

Department of Aerospace,
Cranfield University,
College Rd, Wharley End,
Bedford MK43 0AL, UK
Email: ziyue.wang.907@cranfield.ac.uk
Email: Yang.X@cranfield.ac.uk
*Corresponding author

Abstract: The average battery life of drones in use today is around 30 minutes, which poses significant limitations for ensuring long-range operation, such as seamless delivery and security monitoring. Meanwhile, the transportation sector is responsible for 93% of all carbon emissions, making it crucial to control energy usage during the operation of UAVs for future net-zero massive-scale air traffic. In this study, a reinforcement learning (RL)-based model was implemented for the energy consumption optimisation of drones. The RL-based energy optimisation framework dynamically tunes vehicle control systems to maximise energy economy while considering mission objectives, ambient circumstances, and system performance. RL was used to create a dynamically optimised vehicle control system that selects the most energy-efficient route. Based on training times, it is reasonable to conclude that a trained UAV saves between 50.1% and 91.6% more energy than an untrained UAV in this study by using the same map.

Keywords: power consumption; machine learning; reinforcement learning; RL; trajectory optimisation; Q-Learning; energy efficiency; path planning.

Reference to this paper should be made as follows: Wang, Z. and Xing, Y. (2024) 'Energy consumption optimisation for unmanned aerial vehicle based on reinforcement learning framework', *Int. J. Powertrains*, Vol. 13, No. 1, pp.75–94.

Biographical notes: Ziyue Wang is currently a PhD student in Aerospace at Cranfield University. She received her BEng (honours) in Aerospace Engineering at University of Nottingham Ningbo China in 2021 and MSc in Autonomous Vehicle Dynamics and Control at Cranfield University in 2022.

Yang Xing is the Course Director of MSc Applied AI at Cranfield University. He serves as an Associate Editor in *IEEE Trans. on Intelligent Vehicles*, and Review Editor in *Frontiers in Mechanical Engineering*. He was a Guest Editor-in-Chief/Editor in *IEEE Internet of Things Journal*, *Mechanical Systems and Signal Processing*, and *IEEE Intelligent Transportation Magazine*, etc. He was a Session Chair/Co-chair on IEEE SWC 2023, IEEE MFI 2022, IEEE SMC 2020, IFAC Workshop on CPHS 2020, and IEEE IV 2018. He won the Best Workshop/Special Session Paper on IEEE IV 2018 and Best Paper on China National Intelligence Technology Conference 2019.

1 Introduction

1.1 Background

Unmanned aerial vehicles (UAVs) are mainly used for short-term signal transmission (Witik et al., 2011) and delivery (Witik et al., 2011) for commercial and public applications. The issue of carbon emissions has received significant global attention in recent years. Due to the widespread use of UAVs and the development of smart cities and transportation, connected UAVs are expected to spend more energy on communication and swarm collaboration to meet the demand for safe, intelligent, and opportunistic air transportation (Witik et al., 2011). Therefore, it is essential to regulate energy consumption during the operation of UAVs.

In general, autonomous vehicles have a shorter travel time than regular automobiles due to payload limitations and battery life (Witik et al., 2011). The flying time of small recreational UAVs is typically between 15 and 30 minutes (Aljohani et al., 2021). Batteries provide the drone with a power source and usable energy and play a significant role in determining its durability (Aljohani et al., 2021). Additionally, payload, wind resistance, and obstacles can limit the duration of UAV flights (Hong et al., 2021). In the past, most studies focused mainly on hardware to increase the operational duration (Yildirim et al., 2014; Duan et al., 2020; Elkerdany et al., 2020), and therefore, the efficiency of motors today exceeds 90% (Hong et al., 2021) owing to hardware breakthroughs. Accordingly, it is crucial to devise a strategy to boost the drone's energy efficiency without compromising its technology. Some researchers focused on the photovoltaic electricity management system (PPMS), which offers additional electric energy from the battery, and this method increased the working duration to 54.1 minutes (Jung et al., 2019). Verbeke et al. (2014) built a quadcopter-mounted hexacopter with two arms and large propellers, and UAV flight time improved by 58%. Knowing that previous researchers focused primarily on enhancing hardware, it was decided to focus on software development in this paper. Reinforcement learning (RL) could be one of the most effective methods for improving energy efficiency if external variables do not change, based on previous experience (Duan et al., 2020). Therefore, in this study, the optimal flight path will be determined based on the impact of mission objectives, environmental conditions, and system performance, resulting in an increase in power efficiency based on the RL framework.

By using RL, this study demonstrates the planning of a maximum energy-efficient path for a single UAV. The study assumed that the drone will fly in an indoor environment without wind resistance at a low and steady speed. Therefore, the optimal energy problem can be transformed into an optimal trajectory planning problem. The overall technique follows three steps. First, the trajectory planning model for the UAV was trained with RL and deep determinist policy learning framework for optimal trajectory for joint path planning and energy optimisation. Finally, the trained model with energy consumption optimisation is compared to that of the untrained model with the same parameters. The investigation reveals that trained UAVs conserve between 50.1% and 91.6% more energy than their untrained counterparts.

2 Literature review

UAVs are extensively used in industries, particularly for short-distance, low-quality transportation, such as aerial photography and signal transmission. However, due to the limitations of the payload and its own size, the flight duration is typically limited to no more than 30 minutes, especially for miniature recreational drones that can only fly for 10 to 15 minutes (Chan and Kam, 2020). Therefore, some researchers have focused on increasing the flight duration of drones.

Several previous researchers have focused on hardware enhancements to extend the working time of drones. Jung, for instance, used a PPMS to supply solar electricity to the UAVs, resulting in an operational time extension to 54.14 minutes (Jung et al., 2019). They also designed an autonomous, independent charging mechanism that can enhance the efficiency of constant charging in the open air (Jung et al., 2019). The study of Seoul National University involved optimal analysis and advanced design of UAVs, including propeller aerodynamic analysis, frame structure analysis, and electrical system analysis (Kim et al., 2018). As a results, they achieved a 30% increase in the drone's hovering duration, from 31 minutes to 40 minutes (Kim et al., 2018). Verbeke et al. (2014) developed a hexacopter consisting of two arms with a large propeller attached to a quadcopter frame. This design led to a 58% improvement in UAV flying time. Additionally, some researchers have developed methods for evaluating the flight time of UAVs, excluding studies on extending flight time.

Clearly, previous researchers have focused primarily on hardware breakthroughs, and the working period of UAVs has greatly risen. Although several researchers have addressed the issue of the working period, there has been a dearth of studies on energy consumption during the flight, especially using the RL-based method.

According to the study conducted by Chan and his team, they used a carbon strand composite material and designed the propulsion system components to reduce the power consumption of an Octa-rotor UAV (Chan and Kam, 2020). Their study demonstrated that the proposed approach can generate an accurate prediction with an error rate of less than 7.4% (Chan and Kam, 2020), and this method can be applied in UAV design and flight path planning. Duan's team employed a neural network-based model to estimate the power consumption of a drone in order to maximise flight control (Duan et al., 2020). This model provides an accurate and adaptable prediction, which optimises the flight time of a drone by calculating its energy consumption during the flight (Duan et al., 2020). Hong took into account real-time interactions with the surrounding environment, where the drone must avoid obstacles such as other drones or barriers (Hong et al., 2021). The advanced TD3 model conducts energy-efficient route planning at the edge-level drone, and the overall energy consumption of in-flight drones with online path planning is approximately 106% of the total energy consumption of in-flight drones with offline path planning (Hong et al., 2021).

To enhance communication quality while reducing the overall energy consumption of the network, Ajiohani et al. (2021) presented a distributed RL-based energy-efficient framework for UAV networks with limited energy under jamming attacks. This architecture allows each relay UAV to autonomously determine its transmit power based on previous state-related data, without knowledge of the moving trajectory of other relay UAVs or the jammer, and this technique reduces energy usage by 22.8% compared to the traditional method (Aljohani et al., 2021). Additionally, Yildirim et al. (2014) have proposed an energy-efficient UAV-assisted IoT network, in which a low-altitude

quad-rotor UAV provides a mobile IoT device data gathering service. They have introduced an optimisation framework that concurrently optimises the UAV's trajectory, devices; association, and transmit power allocation at each time slot, while ensuring that each device meets a specified data rate limit. Their numerical results validate the research and provide various insights into the optimal UAV trajectory (Yildirim et al., 2014). Compared to the particle swarm optimisation algorithm, the proposed technique reduced the overall energy consumption of all devices by 6.91%, 8.48%, and 9.94% in 80, 100, and 120 available time slots of UAV, respectively (Yildirim et al., 2014).

According to the previous studies, it is known that even though some researchers focused on the RL-based method to maximise energy efficiency, they did not combine energy consumption with path planning. Therefore, this study aims to improve power efficiency through optimal trajectory planning. To achieve this, the study uses the DJI Mavic 3 as a commercial example of a drone's battery specifications, with all data obtained from the DJI UK official website.

The primary work of this study can be separated into two main objectives. Firstly, design and construct a mathematical model for energy-efficient optimisation of dynamical UAV motion control based on mission objectives, environmental conditions, and system performance. Secondly, to validate the designed autonomous control system, an extensive model evaluation will compare the power consumption of the untrained model with the trained model by conducting a power consumption experiment for commercial UAVs.

2.1 Contribution

Based on the limitation of the existing studies, the following difficulties were encountered. Firstly, the route was clear of obstacles. Secondly, data such as speed, flying angle, and SOC were never collected, resulting in UAV crashes and abrupt power loss. Lastly, due to random flights, the start and end points were not well-defined.

The issues mentioned above translate into the main purpose of this study, which is to design, construct, and test an autonomous vehicle control system that is dynamically tuned to optimal energy efficiency based on mission objectives, environmental circumstances, and system performance. The mean contributions of this study can be summarised as follows.

- First, theoretical knowledge and a detailed literature review were proposed to identify the main limitations of exiting UAV path planning for joint energy consumption optimisation. Then, an efficient RL-based learning framework for joint energy consumption and path planning optimisation was proposed.
- In the second phase, the RL and DDPG models were included in the UAV model, which is trained to compute the optimal trajectory. This model can output the UAV's speed and flying angle when determining if the UAV is in a stable cruise. This phase investigated how to select the ideal path by teaching the drone while it functions; SOC would be considered later.
- During the third stage, the battery model was included in the reward function of the RL-based UAV model, enabling real-time monitoring of the battery's SOC. In addition, by comparing the SOC of trained and untrained UAVs, it is feasible to establish the model was designed using RL.

2.2 Paper organisation

The remainder of this paper is organised as follows.

Section 2 provides details on the methodology used in this study, which focuses on the fundamental design of RL for energy-aware path planning of the UAV.

Section 3 describes the process of designing the autonomous control system. The first step is the conceptual design of the model and problem formulation. Next, the simulation environment is developed, and a mathematical model is proposed that can determine the ideal trajectory using RL and Q-learning.

Section 4 discusses the implementation for comparison. The optimal trajectory and Simulink results of the developed autonomous control system are presented, along with a comparison of performance that focuses on the trajectory and energy consumption of untrained and trained models.

The final section concludes this study and highlights the final results and contributions.

3 Methodology

3.1 Policy design for the RL model

This section elaborates on the main techniques employed in this study, including RL and Q-Learning.

RL is a subfield of ML that deals with how an intelligence agent can maximise cumulative rewards in a given environment. Unlike other ML techniques, RL training requires mapping the environment to an action. Therefore, actions are chosen to maximise the value of rewards obtained from the environment, and the most effective strategies are learned through trial and error (Lyu et al., 2022). The diagram below illustrates the essential components of the RL model.

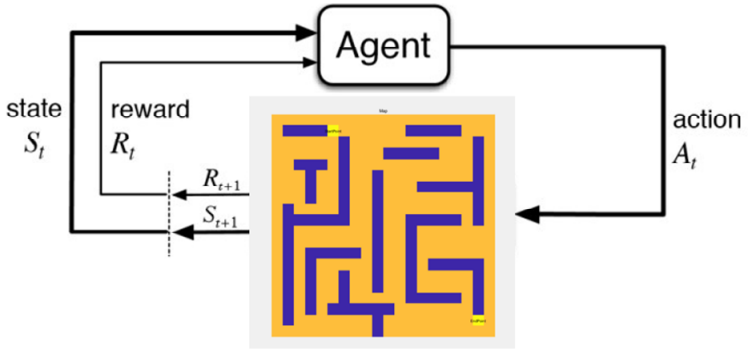
The RL model can be thought of as the agent performing actions in the environment, while receiving feedback from the states. As shown in the previous section, the core RL model consists of five key components: agent, environment, action, state, and reward.

Several key considerations of the design of the RL model in this study are as follows.

- The agent receives an observation and a reward at the end of a time step, and then sends an action to the environment.
- The environment is the physical world in which the agent operates. In this study, it is the map where the UAV works.
- State describes the current condition of the agent.
- Reward indicates the feedback that the agent receives from the environment.
- The policy determines the action for each RL model. In this study, the drone aims to maximise energy efficiency.

RL is defined as the process of learning "how to maximise advantage." Instead of mandating which tasks must be performed, the RL model must identify which behaviours will produce the highest score (Lyu et al., 2022). Frequently, actions impact not only current rewards but also future rewards and subsequent situations (Lyu et al., 2022).

Figure 1 RL model (see online version for colours)



Based on the above summary of RL, it is evident that RL can be one of the optimal strategies for maximising energy efficiency in this study. Due to the time constraints of this study, Q-learning, a sub-method of RL, will also be examined. Q-Learning is a model-free RL paradigm for determining the value of an action in each state. This technique can allocate random rewards and transitions without adaptation (Tan et al., 2021) and without the need for an environment model. This study applies Q-Learning to RL in order to adapt the policy to unknown settings.

3.2 Action selection strategy

The previous section discussed the method for selecting the optimal path with RL. This part describes the strategy for detecting the SOC of the battery of the drone with the RL-based path planning framework.

This study estimates the battery SOC with the extended Kalman Filter (EKF) approach. The Kalman Filter (KF) is an effective recursive filter that estimates the state of a dynamic system from a sequence of noise-free observations (Yang and Li, 2016). The basic KF only applies to Gaussian-distributed systems, but the EKF can be used for nonlinear dynamic systems in time. EKF is a suboptimal filter (Yang and Li, 2016) since its fundamental concept is linearising a nonlinear system before applying a KF. Since the starting SOC of the battery is unknown, EKF is required to calculate the battery's SOC. This section will discuss EKF.

When the state or measurement equation is nonlinear, EKF is commonly utilised. The EKF truncates the Taylor expansion of the nonlinear function by linearising the first order and ignoring the remaining higher-order components. Thus, the nonlinear problem is turned into a linear one, allowing the nonlinear system to be filtered using the Kalman linear filtering approach (Liu et al., 2007).

The EKF is a simple nonlinear approximate filtering algorithm for cases where the equations of motion or observation are not linear (Yang and Li, 2016). To simplify computation; the EKF linearises the equations of motion/observation using a first-order Taylor decomposition. KF and EKF share the same algorithmic structure, both describe the posterior probability density in Gaussian form, and both are obtained by computing a Bayesian recursive formula (Liu et al., 2007). The most significant difference is that the EKF's state transfer matrix (state information at the previous moment) and observation

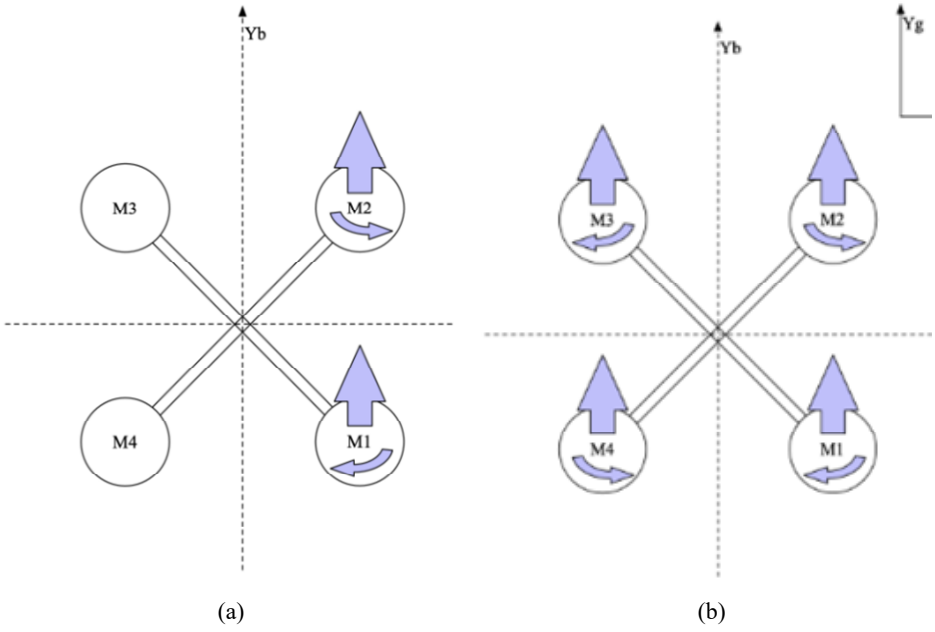
matrix (one-step prediction) are both Jacobi matrices of state information when calculating the variance (Liu et al., 2007).

4 System model

4.1 Problem formulation

This part focuses on the conceptual design of the model and the problem formulation derivation. It is anticipated that the quadcopter drone will be utilised in this study. The free body diagram (FBD) is illustrated in Figure 1.

Figure 2 FBD of UAV (a) shows the FBD of the single propeller and (b) shows the FBD of two coordinate systems



(a) (b)

Note: Upward-pointing arrows indicate lift and rotating arrows indicate torque.

Source: Chan and Kam (2020)

The assessment of a single propeller is the first step. Generally, the lift and torque generated by each of the UAV's four propellers are comparable. For the purpose of this study, it is assumed that each propeller generates the same amount of lift and torque. When a single propeller rotates, as shown for M1, M2, M3, and M4, an upward lift force F and a rotational moment M are generated. The direction of the left force is perpendicular to the propeller, pointing upwards, and the direction of the rotational moment is assumed to be parallel to the location plane of the drone and perpendicular to the arm.

The second phase involves establishing two distinct coordinate systems: one for the ground, represented by X_g–Y_g, and one for the UAV, represented by X_b–Y_b.

The overall lift in the z-direction during vertical take-off of the UAV can be stated as follows:

$$F_{z_b} = F + F + F + F = 4F \tag{1}$$

where, F_{Zh} means the overall lift in the z-direction.

Assume uniform linear motion along the x and y axes with no external forces to simplify the concept. F_{x_h} stands for the overall lift in the x-direction while F_{y_h} is on behalf of the overall lift in the y-direction. Therefore

$$F_{x_b} = 0 \tag{2}$$

$$F_{y_b} = 0 \tag{3}$$

In the third phase, the roll angle ϕ, pitch angle θ, and heading angle ψ, as well as air resistance, are analysed.

$$\ddot{\phi} \ -\frac{C_d\dot{\phi}}{I_x} \tag{4}$$

$$\ddot{\theta} = -\frac{C_d\dot{\theta}}{I_y} \tag{5}$$

$$\ddot{\psi} = -\frac{C_d\dot{\psi}}{I_z} \tag{6}$$

where, C_d is the coefficient of drag.

To evaluate these equations, however, the UAV coordinate system is used. Convert them to the Earth's coordinate system and take gravity into account.

$$\begin{bmatrix} F_{x_g} \\ F_{y_g} \\ F_{z_g} \end{bmatrix} = \begin{bmatrix} cos\theta cos\psi & sin\phi sin\theta\psi - cos\phi sin\psi & cos\phi sin\theta cos\psi + sin\phi sin\psi \\ cos\theta sin\psi & sin\phi sin\theta cos\psi + cos\phi sin\psi & cos\phi sin\theta sin\psi - sin\phi cos\psi \\ -sin\theta & sin\phi cos\theta & cos\phi cos\theta \end{bmatrix} \begin{bmatrix} F_{x_b} \\ F_{y_b} \\ F_{z_b} \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} F_{x_g} \\ F_{y_g} \\ F_{z_g} \end{bmatrix} = \begin{bmatrix} cos\phi sin\theta cos\psi + sin\phi sin\psi \\ cos\phi sin\theta sin\psi - sin\phi cos\psi \\ cos\phi cos\theta \end{bmatrix} 4F - \begin{bmatrix} 0 \\ 0 \\ mg \end{bmatrix} \tag{8}$$

It follows from Newton's second law that,

$$\ddot{x}_g = \frac{(cos\phi sin\theta cos\psi + sin\phi sin\psi)4F}{m} \tag{9}$$

$$\ddot{y}_g = \frac{(cos\phi sin\theta sin\psi - sin\phi cos\psi)4F}{m} \tag{10}$$

$$\ddot{z}_g = \frac{(cos\phi cos\theta)4F}{m} - g \tag{11}$$

Use the following formula to get the overall energy consumption of a single drone.

$$E_{total} = E_{motor} + E_{communication} + E_{external} \tag{12}$$

- E_{motor}, it is the amount of energy required to convert electrical energy into mechanical energy and gravitational energy. This energy consumption is typically proportional to the UAV's speed, altitude, and duration of the operation.
- $E_{communication}$ indicates the amount of energy expended during message reception and transmission. Compared to the energy required by the engine, this study's communications consumption is little.
- $E_{external}$ include additional energy expansion of the battery and energy generated by air resistance, and it can be deemed constant. In the framework of the mathematical model, this proportion can be subtracted immediately as a constant number, such that the actual available energy equals the capacity of the battery minus the percentage of energy.

$$E_{actual} = E - E_{external} \tag{13}$$

where, E represents the original capacity of the battery. E_{actual} represents the actual available energy. Therefore, $E_{total} \approx E_{motor} + E_{external}$ and $E_{actual} \geq E_{total}$. In addition, E_{motor} can be separated into three parts, $E_{takeoff}$, E_{hover} and $E_{landing}$.

$$E_{motor} = E_{takeoff} + E_{hover} + E_{landing} \tag{14}$$

As the drone used in this study runs at a constant and modest speed, take-off and landing are performed just once. Therefore, the preceding equation can be simplified as follows:

$$E_{motor} \approx E_{hover} \tag{15}$$

The relationship between time and energy can alternatively be described as

$$Energy[J] = Voltage[V] * Ch\arg e[mA * h] * 3.6 \tag{16}$$

$$Energy[J] = Power[W] * Time[s] \tag{17}$$

$$worktime[h] = \frac{batterycapacity[W * h]}{Voltage[V] * Current[A]} \tag{18}$$

In conclusion, the design of the Simulink model is determined by the preceding equations.

3.2 Optimal trajectory model

This section introduces the design of the RL model for the UAV, including the reward function. In the first stage, only the trajectory will be examined. The design of the UAV model is shown in Figure 4, illustrates how to train a DDPG model and create UAV flight paths.

The environment model will be constructed using the UAV model, which will also be described in the next section. The model's variables are initialised, and as the take-off of an UAV differs from that of a standard aircraft, there is no taxiing movement during take-off; hence, the heading angle and angle of attack are irrelevant (Hong et al., 2021). It is assumed that the drone can launch under stable conditions.

The sample interval and simulation duration are specified. The sample period affects the timing of action and reward, which influences the total duration of learning (Zhang et al., 2022). The average processing time for each episode is determined by the duration of the entire simulation (Hong et al., 2021). The subsequent stage is to develop an integrated model. To train an intelligent body for the UAV model, create an integrated model with a ready-to-train block of RL intelligent bodies using the exiting function.

As shown in the RL model in Figure 1, the main components of the model are the state, action, and reward function. In this study, the state is the real-time location and battery SOC of the UAV, while the action aims to choose the optimal trajectory to move.

The reward function is a crucial component of RL and has a significant impact on its performance. In this paper, a non-parametric reward function is proposed, which avoids the time-consuming weighting parameters and complexity of traditional reward functions in RL (Yan et al., 2022). As mentioned earlier, the SOC of the battery is an important parameter of the reward function. Specifically, after each operation, a higher SOC will result in a larger reward compared to a lower SOC. Therefore, the non-parametric reward function could be defined as R = SOC.

The first step in this stage is to specify the names for the observation and action specification and to restrict the range of the thrust action. This allows for more efficient training and prevents potential crashes (Zhang et al., 2022). The agent, observation, and action blocks were combined at this level, and a fixed random number generator was configured to increase repeatability (Wang et al., 2021).

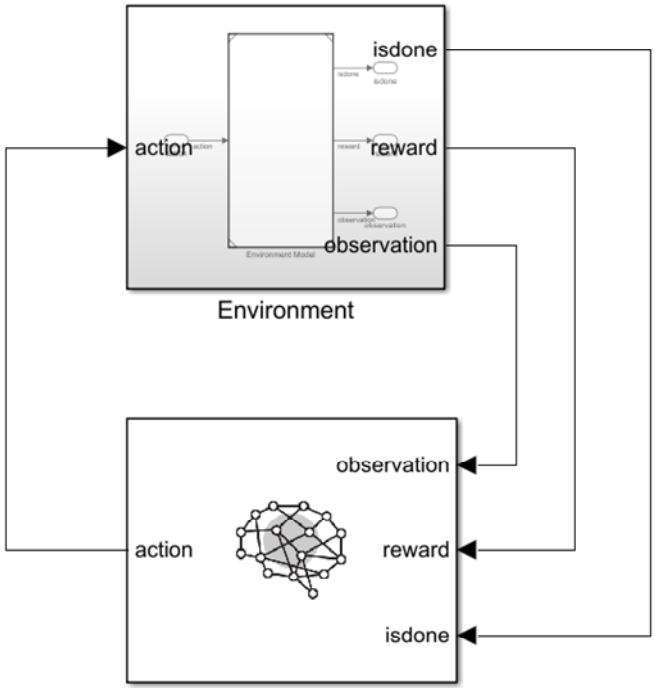
DDPG estimates the long-term reward of observation or action using a critic value function model. To create a critic, a deep neural network with two inputs (observations and actions) and one output (Hu and Zhang, 2022) is first developed. This can be used to predict deterministic policies and maximise total reward by a single step of off-policy policy modification (Hu and Zhang, 2022). Before the development of DDPG, the problem of continuous actions in RL was often solved by discretising the continuous actions and then applying RL. However, DDPG makes it possible to directly predict continuous action (Hu and Zhang, 2022), which is the approach used in this study.

Additionally, the actor's representation is generated using the provided neural network and choices, and the agent's observation and action criteria must also be specified (Hu and Zhang, 2022). An alternate code can be used for this operation. Using actors, the DDPG agent determines the action to take. First, a deep neural network with one input, i.e., observation, and one output, i.e., action, is formed to generate an actor (Khodabakhsh et al., 2018). To build a DDPG agent, the DDPG intelligence body parameters must be provided. Finally, the actor's representation is used to produce the agent.

Before training an agent, training choices must be established. The maximum number of learning episodes in each session has been established to be 1,000, with each episode lasting a maximum of the ceil time step. Then, training is discontinued when the agent's average total reward for 10 consecutive sessions exceeds 450. After many rounds of training, 450 was chosen as the judgment criterion since it represented the most efficient cumulative reward. G Training stop when the agent receives an average cumulative reward greater than 450 over 10 consecutive episodes, indicating that the agent can drive the UAV to the goal position. Additionally, the intelligence of each episode is tripled when the overall reward reaches 450. It should be noted that RL model training is a computationally challenging and time-consuming procedure.

To validate the performance of the trained model, UAV simulation models were run in the environment. The following Simulink model depicts the design of the agent. The model consists of five basic components, as seen in the image below: the action module, the environment module, the accomplishment module, the reward module, and the observation module. The action module enables the drone to maximise its energy efficiency. Details on the environment module will be provided later. The achievement module holds previously processed information. The reward module conveys the environment's response to the agent for subsequent delivery. Finally, upon completion of a particular action, the observation module collects data by observing the environment's status (Khodabakhsh et al., 2018).

Figure 3 RL model



Notes: The environment is the map where the UAV works. Reward indicates the input the agent will get from the environment. the action is to select the optimal path planning.

The structure of the module is represented by the following model in Figure 3. The thrust is the module's first input, which propels the drone forward. The thrust parameters are sent to the saturation dynamic block, which provides an output signal whose value is restricted by the saturated input value up or down (Sun et al., 2021). It guarantees that succeeding blocks can recognise the data at the beginning of the block. The initial data processed by the saturation dynamic block is then delivered, respectively, to the reward block and the dynamics block. The output parameters of the dynamics block are the new position of the drone, new flight angles, velocity, acceleration, and the wrapped pitch angle. The wrapped pitch angle reflects the adjusted flight angle of the drone after

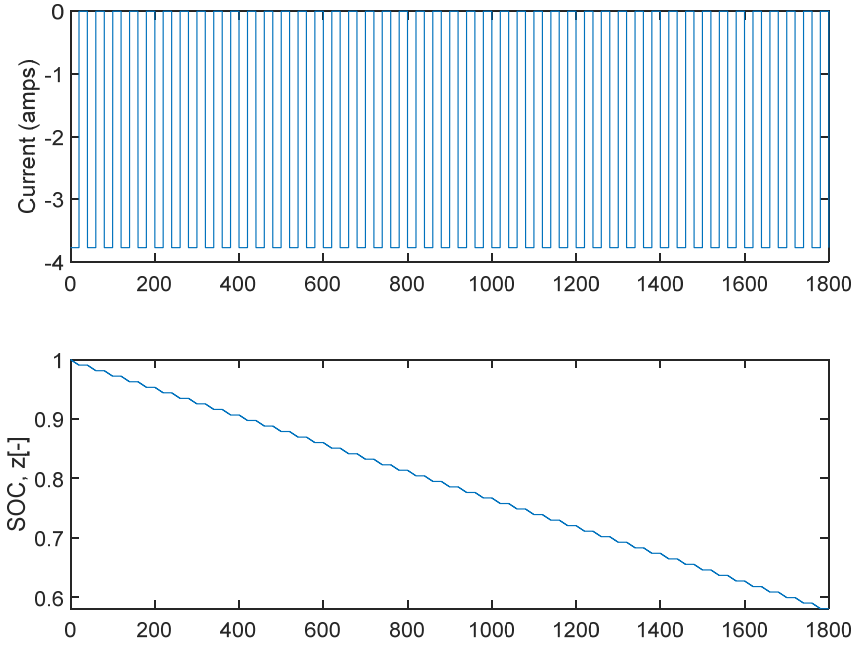
training. The position coordinates represent the new position of the UAV, and other parameters will be used throughout the observation block.

The modified position was sent to the bound block, and the output parameters consisted of the finalised data. Additionally, the reward block receives the wrapped pitch angle, position, derivation of flight angles, and constraint data. These environment input parameters are handled by the reward block and delivered to the agent in the upcoming episode (Sripad and Viswanathan, 2021). The Simulink model's process should match the UAV mathematical model.

The bound block is used to limit the range of mobility of the drone. By restricting the x and y coordinates, the drone's flying range is confined within a specific map. The following reward block is the primary block of the system, which delivers the agent the environment block's feedback. This block is mostly used to calculate the parameters the agent can accept (Apuroop et al., 2021).

The final stage of developing the RL model is to create an integrated model. To train the agent for the UAV model, the code provided can be used to generate an integrated model with a block of RL intelligence that is ready for training.

Figure 4 Current and SOC of battery (see online version for colours)



3.3 Energy consumption model

Accurate calculation of the SOC is crucial in the field of UAV batteries to ensure the safety of charging and discharging, optimise battery performance, and extend the range and service life of UAV. One of the most important determinants of whether the RL model can maximise the UAV's energy efficiency is the accuracy of the SOC estimate for the drone's battery. (DJI Official, 2022). Therefore, developing, creating, and testing an

accurate Simulink model of the UAV battery is the most crucial component of producing the SOC.

The model demonstrates the current and SOC's adaptability. While the change in current and SOC corresponds to the previous formulas, the idea must also be updated because the drone's operating environment cannot be simple. For example, the Coulomb counting technique calculates the SOC in real-time from the current value, while the internal resistance method relies solely on the impedance characteristics of the cell to calculate the SOC (DJI Official, 2022).

In this study, only the simulation model of SOC is considered, which means that an ideal condition for the initial state of the battery was used, and it can be an easier metric to evaluate the dynamics of SOC, for example in perception evaluation. Therefore, the initial battery SOC was chosen as 1 but indeed can be any suitable value in the real world. The change of the current and SOC are shown in Figure 4.

In addition, the battery is a complex, nonlinear, time-varying system, making it difficult to represent the variation of battery parameters under actual operating conditions (Xin, 2020). The method given in the preceding section utilises straightforward metrics to determine SOC. Some signal estimation-based systems employ the battery's SOC as an interference signal (Xin, 2020). Therefore, EKF could be well-applied to the estimated SOC.

4 Implementation for comparison

4.1 Simulation results and discussion

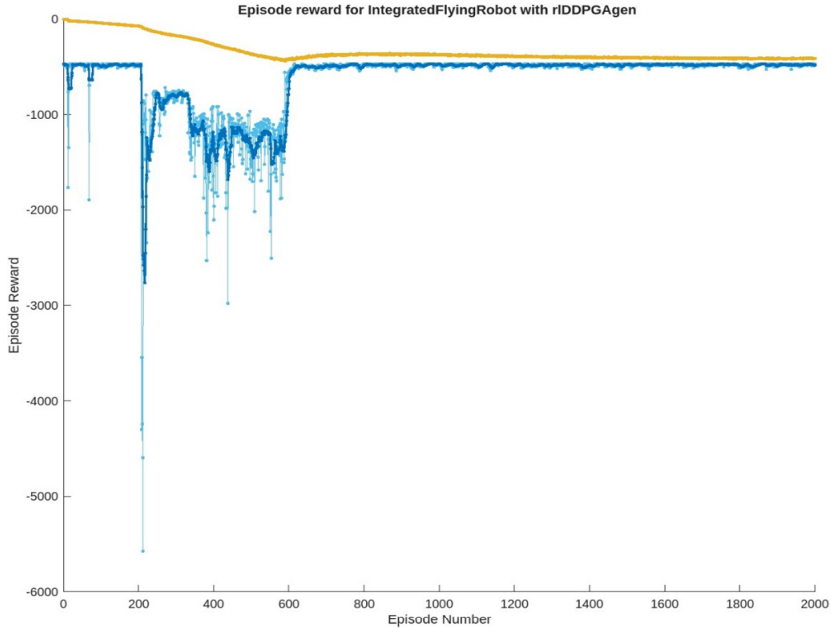
The process of RL in this study can be understood as a scoring game where points are deducted for choosing a longer path and added for choosing a shorter path. This approach trains the drone to select the optimal path in any environment.

Figure 5 shows the training results of the RL model. The total number of training episodes depicted in Figure 5(a) was 2,000. However, as the graph shows, the difference remained constant at over 400 episodes. As seen in Figure 5(b), the total number of training episodes was eventually reduced to 1,000. Thus, the Simulink model can learn how to pick the optimal trajectory and reduce the total time spent on training, thereby increasing the effectiveness of the training.

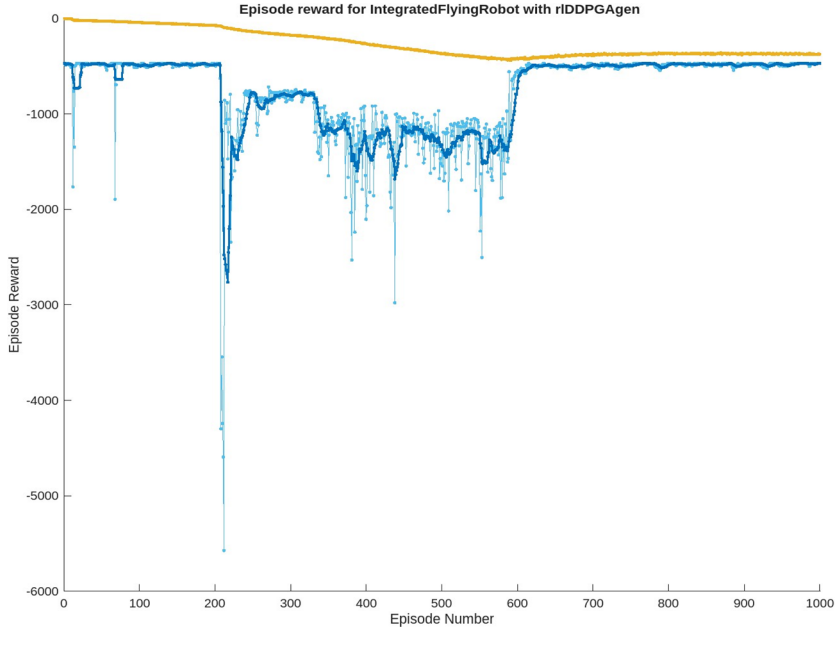
4.2 Optimal trajectory

The original Simulink model has been enhanced using Q-Learning to improve the visibility of the training outcomes. The Q-Learning algorithm is a model-independent off-policy RL method. Based on the value of each completed step, the algorithm progresses to the next phase (Sun et al., 2021). Q-Learning is a value-based RL approach in which Q represents the expected reward of taking an action in a certain state at a given time. The environment will send input to the agent in response to its actions; hence, the algorithm's basic premise is to generate a Q-value table of states and actions and then select the action that will provide the highest reward based on the Q-value (Sripad and Viswanathan, 2021).

Figure 5 Learning results (a) learning results 1 (b) learning results 2 (see online version for colours)



(a)

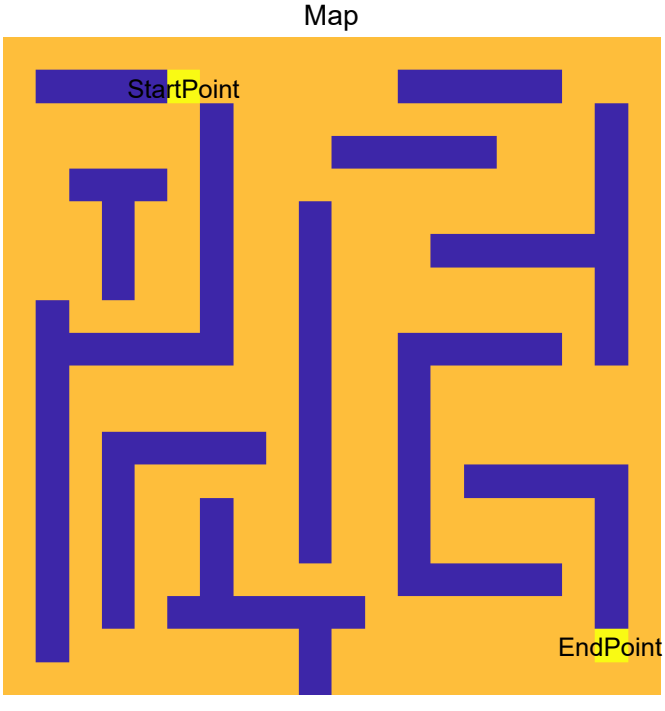


(b)

Notes: The light blue line represents the prize for each episode, whereas the dark blue line is the average payment for all episodes. The yellow line represents episode Q0.

Figure 6 is a map of the area where the UAV is operating, including any obstructions.

Figure 6 Trajectory map of the UAV (see online version for colours)



Notes: The light-yellow points stand for the start point and the end point of the UAV. The Purple line means the obstacle on the map. The deep yellow lump is the place where available for flying.

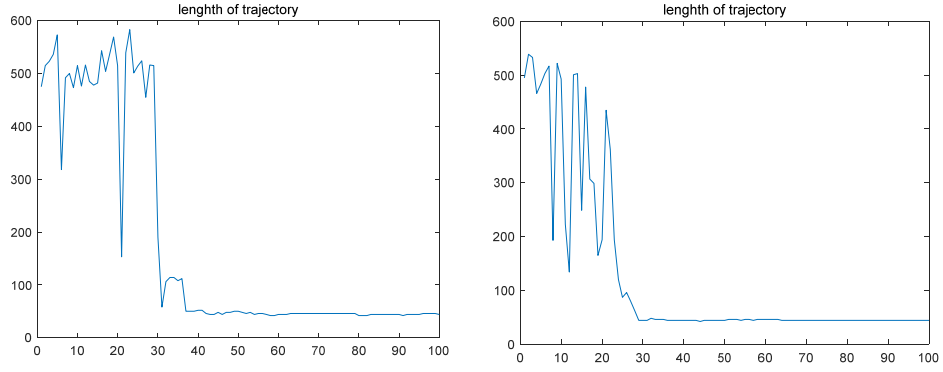
Q-learning is essentially a greedy algorithm, which means that if the action with the highest expected reward is chosen at each step, it may fail to explore other potential actions during the training phase or fall into "local optimality" (Khodabakhsh et al., 2018), where it fails to select the optimal trajectory. Therefore, coefficients are chosen to ensure that the agent has a chance of acting optimally and a certain probability of taking random actions (Chan and Kam, 2020). To avoid small loops, the paths taken are stored in a memory bank.

By setting the reward value for diagonal movement to 2, the drone is prevented from moving, for example, up and then left and then down rather than moving directly to the left. This value is determined based on the relative distance between the two map frames. The following figures illustrate the results of RL on the trajectory.

Comparing the above (a) and (b) figures in Figure 7, it could know that the maximum and minimum length of trajectory in (a) is larger than that of (b).

The length of trajectories under various learning techniques is depicted in Figure 8. By comparing them to Figure 7, it was feasible to conclude that the trajectory became shorter after 250 learning events.

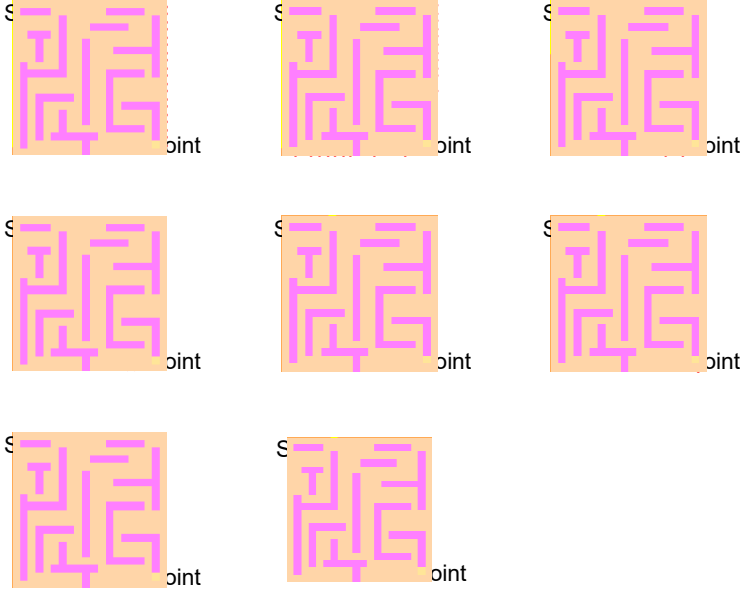
Figure 7 Length of trajectory (a) length of trajectory 1 (b) length of trajectory 2 (see online version for colours)



(a) (b)

Notes: The horizontal coordinate represents the number of training episodes, and the number on the coordinate multiplied by 10 is the total number of training episodes. The vertical coordinate represents the length of the moving trajectory.

Figure 8 Comparison of trajectory (see online version for colours)



Notes: The pink line means the obstacle on the map. the orange lump is the place where available for flying, and the yellow cube with a red point stands for the working trajectory of the UAV.

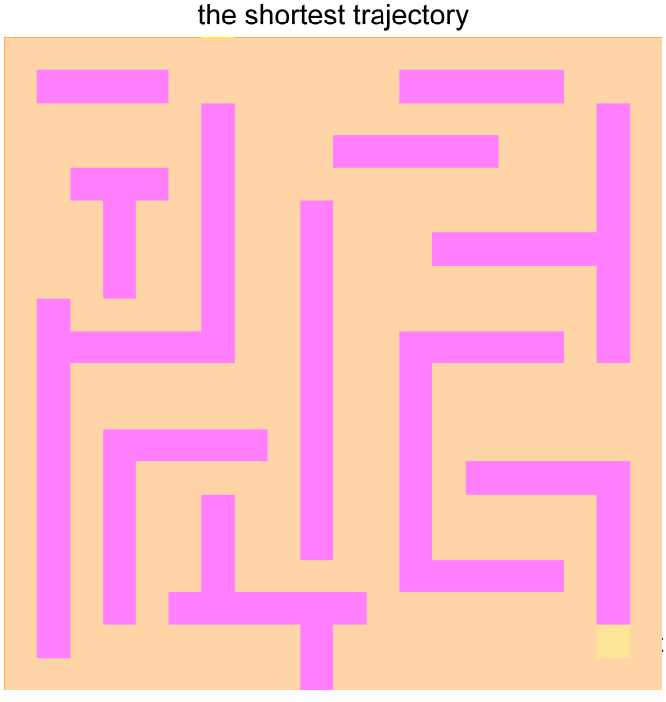
Figure 8 displays a comparison of trajectories. The trajectory when the drone was initially untrained is represented in the upper left corner, while the trajectory with the shortest training is shown in the second column and third row. As demonstrated in the graph above, the trajectory does not degrade as the number of training episodes increases during

the learning process. This is also consistent with the graphed data on episode reward and trajectory length. All the evidence suggests that the learning outcomes were accurate.

Figure 9 displays the Simulink model's shortest path, which is the optimal trajectory for optimising energy efficiency.

According to the simulation results, the RL-based UAV area coverage route planning framework developed in this study allows the UAV to cover the mission area in a static environment with a small number of steps. However, in a dynamic environment with moving obstacles, the framework for UAV area coverage route planning based on deep RL still enables the UAV to cover the mission area in a limited number of steps without colliding with moving objects.

Figure 9 The shortest trajectory (see online version for colours)



4.3 Performance comparison

In the previous discussion, the SOC of the battery and UAV trajectories were analysed individually. It is known that learning can be used to choose the optimal trajectory, and the SOC of the battery decreased nonlinearly during operation. However, the previous discussion only addressed the UAV's initial and final learning stages. Therefore, in this paragraph, untrained trajectories and SOC are compared with trained trajectories and SOC.

After adding the battery model to the reward block, the model was run, and the results were comparable to those previously presented. After about 15 minutes of learning, the optimal path can be chosen. The simulation results show that the RL-based UAV area coverage route planning framework developed in this study enables the UAV to cover the

mission area in a static environment with a short number of steps. However, in a dynamic environment with moving obstacles, the framework for UAV area coverage route planning based on deep RL still enables the UAV to cover the mission area in a limited number of steps without colliding with moving objects. As shown in Figure 8, the upper left corner represented the trajectory when the drone was initially untrained, while the second column and third row represent the shortest training trajectory. Therefore, it is evident that the shortest path can be chosen by using RL.

The study of optimising the performance of UAVs has a long history and has generated a theoretical framework. In recent years, the optimal design of energy-efficient trajectories for long-flight-duration drones has attracted substantial interest (Aljohani et al., 2021) due to emerging energy issues. Thus, the SOC of the drone's battery is a critical indicator of whether it has chosen the optimal flight path. The data demonstrate that the trained drones utilise less energy. After several rounds of RL, it was determined that a trained UAV saves between 50.1% and 91.6% more energy than an untrained UAV on the same map.

5 Conclusions

In this study, a dynamically optimised vehicle control system capable of selecting the optimal path was developed. Then, an EKF-based mathematical model for the battery SOC estimation under varying UAV trajectories was developed. The two control systems were combined to evaluate the proposed control system's potential to satisfy the study's objectives. After training on a specified map with obstacles, the results showed that the UAV can calculate the optimal movement path and jointly optimise energy efficiency. The dynamically optimised vehicle control system that can achieve maximum energy efficiency depending on mission objectives, environmental conditions, and system performance. The overall system shows the efficiency of combining RL-based path planning and battery SOC monitoring model to achieve joint optimisation for the UAV status under an ideal simulation environment.

In the future, more challenging scenarios will be used for comprehensive testing purposes. To do further exploration, the method used in this study will be verified using increased obstacles and larger site areas. Additionally, to improve the efficiency of energy consumption optimisation, Explainable RL might be useful.

References

Aljohani, T., Ebrahim, A. and Mohammed, O. (2021) 'Real-Time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and Markov chain model', *Electric Power Systems Research*, March, Vol. 192, p.106962, DOI: 10.1016/J.EPSR.2020.106962.

Apuroop, K., Le, A., Elara, M. and Sheu, B. (2021) 'Reinforcement learning-based complete area coverage path planning for a modified hTrihex robot', *Sensors*, Vol. 21, No. 4, p.1067, DOI: 10.3390/s21041067.

Chan, C. and Kam, T. (2020) 'A procedure for power consumption estimation of multi-rotor unmanned aerial vehicle', *Journal of Physics: Conference Series*, Vol. 1509, No. 1, p.12015, DOI: 10.1088/1742-6596/1509/1/012015.

DJI Official (2022) *DJI – Official Website* [online] https://www.dji.com/uk (accessed 7 June 2022).

Duan, D., Wang, Z., Wang, Q. and Li, J. (2020) 'Study on integrated optimization design method of high-efficiency motor propeller system for UAVs with multi-states', *IEEE Access*, Vol. 8, pp.165432–165443, DOI: 10.1109/ACCESS.2020.3014411.

Elkerdany, M.S., Safwat, I.M., Yossef, A.M.M. and Elkhatib, M.M. (2020) 'A comparative study on using brushless DC motor six-switch and four-switch inverter for UAV propulsion system', in *Proc. 12th International Conference on Electrical Engineering*, pp.58–61.

Hong, D., Lee, S., Cho, Y.H., Baek, D., Kim, J. and Chang, N. (2021) 'Reference to this paper should be made as follows: energy-efficient online path planning of multiple drones using reinforcement learning', in *IEEE Transactions on Vehicular Technology*, Vol. 70, No. 10, pp.9725–9740, October 2021, DOI: 10.1109/TVT.2021.3102589.

Hu, D. and Zhang, Y. (2022) 'Deep reinforcement learning based on driver experience embedding for energy management strategies in hybrid electric vehicles', *Energy Technology*, Vol. 10, No. 6, p.2200123, DOI: 10.1002/ente.202200123.

Jung, S., Jo, Y. and Kim, Y.J. (2019) 'Flight time estimation for continuous surveillance missions using a multirotor UAV', *Energies*, March, Vol. 12, No. 5, pp.1–15.

Khodabakhsh, S., Fard, B.M. and Bagheri, A. (2018) 'Energy consumption analysis for the limit cycle walking biped robots', *2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM)*, Tehran, Iran, pp.159–165, DOI: 10.1109/ICRoM.2018.8657529.

Kim, H., Lim, D. and Yee, K. (2018) '*Development of a comprehensive analysis and optimized design framework for the multirotor UAV', in 31st Congress of the International Council of the Aeronautical Sciences*, Belo Horizonte, Brazil, September 2018.

Liu, B., Ma, X-C., Hao, C-P., Hou, C-H. and Li, M. (2007) 'A bearings-only-tracking framework based on the EKF and UKF combined algorithm', *2007 International Symposium on Intelligent Signal Processing and Communication Systems*, Xiamen, pp.184–187, DOI: 10.1109/ISPACS.2007.4445854.

Lyu, L., Shen, Y. and Zhang, S. (2022) 'The advance of reinforcement learning and deep reinforcement learning', *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, Changchun, China, pp.644–648, DOI: 10.1109/EEBDA53927.2022.9744760.

Sripad, S. and Viswanathan, V. (2021) 'The promise of energy-efficient battery-powered urban aircraft', *Proceedings of the National Academy of Sciences*, Vol. 118, No. 45, DOI: 10.1073/pnas.2111164118.

Sun, M., Xu, X., Qin, X. and Zhang, P. (2021) 'AoI-energy-aware UAV-assisted data collection for IoT networks: a deep reinforcement learning method', in *IEEE Internet of Things Journal,* Vol. 8, No. 24, pp.17275–17289, 15 December, DOI: 10.1109/JIOT.2021.3078701.

Tan, B., Peng, Y. and Lin, J. (2021) 'A local path planning method based on Q-learning', in *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, pp.80–84, DOI: 10.1109/CONF-SPML54095.2021.00024.

Verbeke, J., Hulens, D., Ramon, H., Goedemé, T. and De Schutter, J. (2014) 'The design and construction of a high endurance hexacopter suited for narrow corridors', in *2014 International Conference on Unmanned Aircraft Systems (ICUAS)*, May 2014, pp.914–920.

Wang, W., Lv, Z., Lu, X., Zhang, Y. and Xiao, L. (2021) 'Distributed reinforcement learning based framework for energy-efficient UAV relay against jamming', in *Intelligent and Converged Networks*, Vol. 2, No. 2, pp.150–162, June, DOI: 10.23919/ICN.2021.0010.

Witik, R.A., Payet, J., Michaud, V. et al. (2011) 'Assessing the life cycle costs and environmental performance of lightweight materials in automobile applications', *Composites Part A*, November, Vol. 42, pp.1694–1709, DOI: 10.1016/J.COMPOSITESA.2011.07.024.

Xin, X. (2020) 'A novel state of charge estimation method for ternary lithium batteries based on system function and extended Kalman filter', *International Journal of Electrochemical Science*, Vol. 15, pp.2226–2242, DOI: 10.20964/2020.03.47.

Yan, F., Wang, J., Du, C. and Hua, M. (2022) 'Multi-objective energy management strategy for hybrid electric vehicles based on TD3 with non-parametric reward function', *Energies*, Vol. 16, No. 1, pp.1–16, December, DOI: 10.3390/en16010074.

Yang, S. and Li, H. (2016) 'Application of EKF and UKF in target tracking problem', in *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp.116–120, DOI: 10.1109/IHMSC.2016.25.

Yildirim, M., Polat, M. and Kürüm, H. (2014) 'A survey on comparison of electric motor types and drives used for electric vehicles', in *Proc. 16th Int. Power Electronics and Motion Control Conference and Exposition*, pp.218–223.

Zhang, L., Celik, A., Dang, S. and Shihada, B. (2022) 'Energy-efficient trajectory optimization for UAV-Assisted IoT networks', in *IEEE Transactions on Mobile Computing*, Vol. 21, No. 12, pp.4323–4337, 1 December, DOI: 10.1109/TMC.2021.3075083.