drones

MDPI

Article

PPO-Based Attitude Controller Design for a Tilt Rotor UAV in Transition Process

Rui Yang, Changping Du *, Yao Zheng, Huzhen Gao, Yuean Wu and Tianrui Fang

School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310027, China
* Correspondence: duchangping@zju.edu.cn

Abstract: The complex aerodynamic changes of the tilt-rotor UAV (TRUAV) in the transition process show strong nonlinearity, which brings a great impact on the stability of the vehicle attitude. This study aims to design a PPO-based RL controller for attitude control in the transition process. A reinforcement-learning PPO approach is used to learn the control strategy by interacting directly with the environment. And the reward function is designed and improved for the transition process. The performance of the proposed controller is tested and compared by simulation. The results show that the PPO algorithm is more suitable for the tilt-rotor transition process control than the A2C algorithm. Our proposed reward function improves the attitude control performance and the designed RL controller has good adaptability to changes in the takeoff weight, the diagonal wheelbase and the tilt rate. This study highlights the effectiveness and potential of reinforcement learning for tilt-rotor UAV transition process attitude control. These findings contribute to the advancement of autonomous flight systems by providing insights into the application of reinforcement learning algorithms. These results have important implications for the development of intelligent flight control systems and could guide future research in this area.

Keywords: UAV; flight control; tilt rotor; transition process; reinforcement learning

Citation: Yang, R.; Du, C.; Zheng, Y.; Gao, H.; Wu, Y.; Fang, T. PPO-Based Attitude Controller Design for a Tilt Rotor UAV in Transition Process. *Drones* **2023**, *7*, 499. https://doi.org/10.3390/drones7080499

Academic Editors: Mou Chen, Bin Jiang, Youmin Zhang, Zixuan Zheng and Shuyi Shao

Received: 1 July 2023
Revised: 18 July 2023
Accepted: 19 July 2023
Published: 31 July 2023

1. Introduction

Demand for unmanned aerial vehicle (UAV) delivery is on the rise from 2025 to 2050. UAVs fly point-to-point in a straight line through the air, which is less affected by terrain factors. Moreover, this is a distinct advantage for freight transport in remote areas and congested urban areas [1]. The tilt rotor aerial UAV (TRUAV) is a novel aircraft that combines two flight modes: copter mode and airplane mode. The versatile flight mode brings many application-level benefits. In the copter mode, the TRUAV can take off and land vertically and hover like a helicopter, which means the vehicle does not need a runway for takeoff and landing, broadening the application scenario. At the same time, it can have a fast cruise and large capacity like a fixed-wing aircraft in airplane mode. This means the vehicle can have long endurance. These are all beneficial to the TRUAV for cargo transportation applications. However, the new structural design introduces new technical difficulties, especially in the transition phase. As the tilt angle of front rotors changes, the airflow of the rotor interacts with the wing. At this phase, the aerodynamic characteristics of the TRUAV are complex and nonlinear.

The controller proposed in this paper focuses on the transition process. During the transition, the aerodynamic characteristics of the TRUAV are complex, with severe aeroelastic coupling and significant non-linearity. There are mature and stable control schemes for attitude control in both copter mode and airplane mode. The challenges in this work are the configuration change from copter mode to airplane mode and the dramatic aerodynamic change from stationary to constant-speed cruise. During the transition, the front motor tilt angle changes and the wing interacts with the propellers, which makes the aerodynamics complex and uncertain and brings difficulties to the control model design. At the same

time, the front motor tilt angle changes, making the TRUAV attitude control performance change, especially pitch attitude control. For the TRUAV, the change in takeoff weight also affects its attitude control performance. In summary, these are the difficulties of modeling and controller design.

Most control systems in aerospace applications are based on linearized systems designed by picking certain equilibria [2,3]. It is a common approach in TRUAVs to design different controllers for specific operating conditions. But then the design of controller switching or scheduling strategies is required. As a classical control theory, PID is the most widely used in engineering without accurate dynamics modeling [4]. Due to the underactuated property, the tracking of UAV attitude and velocity is generally achieved through a cascade structure [5–7]. According to the actual application scenario, the controller gain is adjusted to achieve the desired control effect. In the transition process, the direction of the front rotor thrust vector is changing, and a single cascade PID cannot adapt to the entire transition process and cannot meet the dynamic performance requirements of the TRUAV. Some scholars adopted the scheduled control approach to determine the possible flight range and flight envelope stability conditions of tilt-rotor UAVs based on trim point analysis [4], as shown in Figure 1. Due to the constraints of the actuator and lift generation, appropriate leveling points are selected to evaluate the stability corridor. The flight envelope is discretized into the required number of operating points, and an independent controller is developed for each operating point. These independent controllers are generally PID [8–10], LQR [11,12] and SMC [13]. As the flight configuration changes during the transition, the appropriate control law is loaded through the controller scheduling policies. In addition, the required controller switching may affect stability, if scheduling parameters change quickly concerning controller convergence. However, due to the limited number of linearized points, the overall control structure is limited to a specific region of the flight envelope [14].

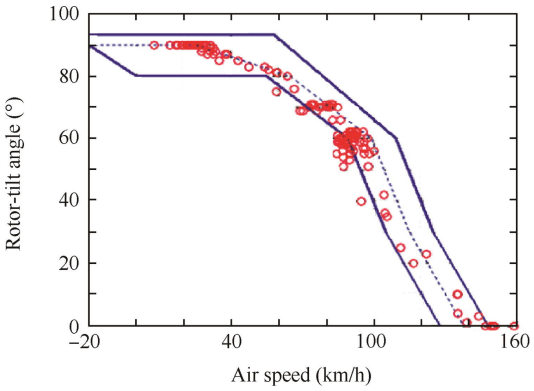

Figure 1. The tilt rotor aerial UAV (TRUAV) flight envelope [4].

Another approach to controller design is based on adaptive or nonlinear implementations of unified control methods that can handle nonlinear dynamics and cover the entire flight envelope, such as optimal control gain scheduling [15,16], dynamic inversion [17–19], robust control [20–22] and nonlinear model predictive control (NMPC) [23,24]. This method avoids the instability caused by the unreasonable selection of the equilibrium point. An improved back propagation (BP) neural network PID control algorithm is proposed by Peng et al. [25]. The method exploits the strong nonlinear mapping properties of the neural network to enhance the robustness of the tilt rotor vehicle transition process. Yu [26] and Yuksek [27] et al. used neural networks to compensate for dynamic inverse errors and external disturbances. The development of deep neural networks has led to a significant increase in the ability of reinforcement learning to be used for aircraft attitude control, particularly for highly nonlinear control problems. Koch et al. [28] developed GymFC, a reinforcement learning training environment for training an intelligent neural network attitude controller for UAV attitude control. The performance of the

intelligent attitude controller was compared with that of a PID controller in the GymFC environment. The results show that the intelligent neural network attitude controller has excellent performance. Xu et al. [29] directly applied a neural network controller to control hybrid UAVs. The designed neural network controller is robust and can take advantage of complex dynamics. The tilt rotor studied by Huo et al. [30] and Lee et al. [31] only makes structural changes while hovering. Huo et al. used deep reinforcement learning algorithms for trajectory tracking control of a trans-domain tilt-rotor robot control. Simulation results show that the neural network controller trained by the reinforcement learning algorithm is highly adaptable to complex environments. Lee et al. first attempted to apply reinforcement learning for low-level attitude control of a tilting multi-rotor. Reinforcement learning is increasingly used in the field of UAV control, while few relevant studies directly apply reinforcement learning to the TRUAV attitude control during the transition process.

In summary, scheduled control and unified control are two main types of control methods regarding the TRUAV, and their main features are presented separately. However, existing research has mainly focused on applying the pattern recognition capabilities of machine learning to improve these two types of controllers. Fewer studies have directly used reinforcement learning to interact with the environment to learn the optimal control strategy, especially for attitude control of the TRUAV during the transition process.

Therefore, the purpose of this paper is to design a controller using the reinforcement learning PPO algorithm and to design and improve the reward function for the TRUAV attitude control. The performance and adaptability of the proposed controller are tested through simulation. The structure of this study is as follows: the TRUAV transition process is modeled and simulated, followed by the reinforcement learning PPO algorithm and the design of the reward function for the transition process. This is followed by the analytical evaluation, which is divided into four parts, namely convergence analysis, the effect of improvement on the reward function, the adaptability of the proposed controller and the effectiveness of our proposed method compared to A2C, standard PPO algorithm.

2. Dynamic Model

Typical tilt rotor UAV models [14,32,33] are listed in Table 1, and the layout of the TRUAV in this paper is shown in Figure 2.

Table 1. Prototype of the tilt rotor aerial UAV (TRUAV).

Prototype	Dual-TRUAVs	Tri-TRUAV	Quad-TRUAV
Advantage	High flight efficiency Long range High speed	High speed Simple tilt structure Easy to control Low rotor/wing airflow disturbance	Large payload Simple tilt structure Easy to control Low rotor/wing airflow
Disadvantage	Complex tilting structure High rotor/wing airflow disturbance Difficult to control	Low power utilization High empty Weight	Low power utilization High empty Weight High cost

A flying wing layout is used with the four motors arranged at a distance from the wing to minimize mutual interference between the wing and the rotor. The center of gravity of the TRUAV is located in the center of the four motors. During the transition process, the tilt angle of two motors, r_1 and r_2, located in front of the wing, gradually tilts from 0° to 90°. The pull T_1 and T_2 are horizontal forward, and the rear motors r_3 and r_4 stop rotating at the end of the transition process.

The rigid body motion model of the TRUAV can be represented using Newton's Euler equation. The equation is as follows.

$$\begin{bmatrix} mI_3 & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} \dot{v}^B \\ \dot{\omega}^B \end{bmatrix} + \begin{bmatrix} \omega^B \times (m \cdot v^B) \\ \omega^B \times (J \cdot \omega^B) \end{bmatrix} = \begin{bmatrix} \vec{F}^B \\ \vec{M}^B \end{bmatrix} \tag{1}$$

where m is the mass, J is the inertia matrix (Ignoring the effect of rotors tilting), I_3 is the 3×3 identity matrix, v^B is the linear velocity, ω^B is the angular velocity, $\vec{F}^B = \begin{bmatrix} F_x^B & F_y^B & F_z^B \end{bmatrix}^T$ is the force and $\vec{M}^B = \begin{bmatrix} M_x^B & M_y^B & M_z^B \end{bmatrix}^T$ is the moment.

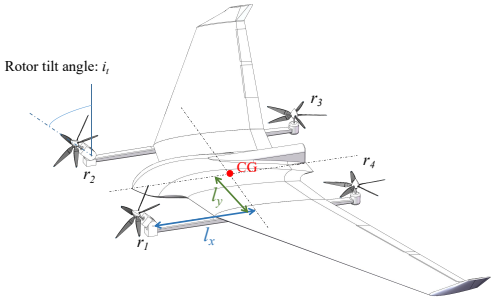

Figure 2. The layout of the TRUAV. The front motors r_1 and r_2 can rotate to the horizontal level compared to the fuselage. The rear motors r_3 and r_4 are fixed, and the thrust is vertically upward compared to the fuselage.

Moreover, $\vec{F}^B$ and $\vec{M}^B$ are the combined force and moment on the vehicle, consisting of the rotor aerodynamic force and moment, the wing aerodynamic force and moment and gravity. The equation is as follows.

$$\begin{bmatrix} \vec{F}^B \\ \vec{M}^B \end{bmatrix} = \begin{bmatrix} \vec{F}_G^B \\ 0 \end{bmatrix} + \begin{bmatrix} \vec{F}_T^B \\ \vec{M}_T^B \end{bmatrix} + \begin{bmatrix} \vec{F}_A^B \\ \vec{M}_A^B \end{bmatrix} \tag{2}$$

where $\vec{F}_G^B = \begin{bmatrix} 0 & 0 & -G \end{bmatrix}^T$ is gravity, $\vec{F}_T^B$ and $\vec{M}_T^B$ is the thrust and moment of the rotor in the body coordinate system B, $\vec{F}_A^B$ and $\vec{M}_A^B$ is the aerodynamic force and moment in the airframe coordinate system. The following equation can establish the rotor aerodynamic force and moment.

$$\begin{bmatrix} \vec{F}_T^B \\ \vec{M}_T^B \end{bmatrix} = A(i_t)T \tag{3}$$

where $T = \begin{bmatrix} T_1 & T_2 & T_3 & T_4 \end{bmatrix}^T$ is the propeller thrust and $A(i_t)$ is the 6×4 distribution matrix, as shown below:

$$A(i_t) = \begin{bmatrix} \sin i_t & \sin i_t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \cos i_t & \cos i_t & 1 & 1 \\ l_y \cos i_t - k \sin i_t & -l_y \cos i_t + k \sin i_t & -l_y & l_y \\ -l_x \cos i_t & -l_x \cos i_t & l_x & l_x \\ -l_y \sin i_t - k \cos i_t & l_y \sin i_t + k \cos i_t & -k & k \end{bmatrix} \tag{4}$$

where l_x is the distance between the rotor and the center of mass along the x-axis of the body, l_y is the distance between the rotor and the center of mass along the y-axis of the body, k is the propeller torque factor and i_t is the tilt angle.

The aerodynamic forces and moment are described by

$$\begin{bmatrix} \vec{F}_A^B \\ \vec{M}_A^B \end{bmatrix} = R_A^B \begin{bmatrix} C_D(\alpha) \\ C_Y \\ C_L(\alpha) \\ C_l \\ C_m \\ C_n \end{bmatrix} QS \tag{5}$$

where R_A^B is the transition matrix from the airframe coordinate system to the body coordinate system, $Q = \frac{1}{2}\rho V^2$ is the dynamic pressure, S is the wing area and $C_D(\alpha)$, C_Y, $C_L(\alpha)$, C_l, C_m, C_n are the drag coefficient, side force coefficient, lift coefficient, roll moment coefficient, pitch moment coefficient and yaw moment coefficient, respectively.

3. Method

The policy proximal optimization (PPO) algorithm is a model-free, on-policy method. PPO has high performance and low computational complexity. The policy network takes the observation state s as input and outputs the action a. For a continuous action space, the task of the policy network is to output the probability distribution from which the action is sampled. In this paper, the PPO algorithm is used to control the attitude of the TRUAV during the transition process accelerating from hover to cruise speed. To avoid divergence, it also uses the clip function to reduce the difference between old and new policies. The reward function is designed for the transition process according to the characteristics of the TRUAV.

3.1. Algorithm Framework

In this paper, a standard reinforcement learning structure is used for controller training. The algorithmic framework is shown in Figure 3. The control task involved in this paper conforms to a Markov decision process (MDP). The composition is (S, A, P, R, γ), where S is the state, A is the action and $P(s'|s,a)$ is the state transition function for the next state s′ given the current state s and the action a, $R(s,a)$ is the reward value, $\gamma \in [0,1)$ is the discount factor. The policy $\pi_\theta(a|s)$ gives the possibility of choosing action a at state s. Thus, the goal of the algorithm is to find the policy loss $J(\theta)$ that optimizes the parameters θ of the policy.

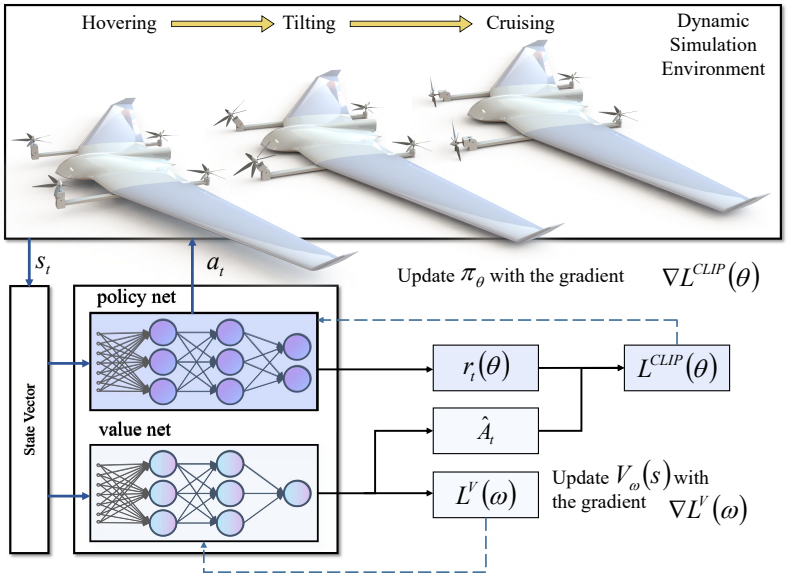

Figure 3. Schematic diagram of our process. In the simulation, the agent learns the continuous control of the attitude with the PPO algorithm during the transition process.

The policy gradient algorithm is based on the principle of policy gradient estimation and then computing the gradient for parameter θ updating. The gradient is estimated in a Monte Carlo method, and the policy loss J is obtained by the agent applying the policy to interact in the environment.

$$J(\theta) = E_{\tau\sim\pi_\theta(\tau)}\left[\sum_t R(s_t, a_t)\right] = E_{\tau\sim\pi_\theta(\tau)}\left[\sum_t R(\tau)\right] \tag{6}$$

In practice, these gradients are obtained by the gradient solver, which updates the parameters θ of the neural network by backpropagation.

$$\bigtriangledown_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}\left[\left(\sum_t^T \bigtriangledown_\theta \log \pi_\theta(a_t|s_t)\right) R(\tau)\right] \tag{7}$$

The critical challenge of the policy gradient is to reduce the variance of the gradient estimate in order to obtain a better policy. The generalized advantage estimation method (GAE) is utilized in the PPO to reduce the variance. The advantage function is built to reconstruct the reward signal.

$$\hat{A}_t = \sum_{t'>t} \gamma^{t'-t} r_{t'} - V_\omega(s_t) \tag{8}$$

The advantage function evaluates a behavior compared to other behavior available in the state, such that good behavior receives positive rewards and poor behavior has negative rewards. It is, therefore, essential to estimate the value function of the average reward. The value network is an independent supervised learning neural network estimated by collecting the reward value samples. Multiple agents are also used in PPO to collect samples simultaneously to increase the number of samples. The loss function of the PPO is as follows:

$$L^{CLIP}(\theta) = \hat{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right] \tag{9}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)}$ denotes the ratio between the old and new strategies. ε is a hyperparameter with a small value used to keep the new policy close to the old policy. The importance sampling method is used in the PPO algorithm to obtain the expectation of samples collected from the old policy, which improves sample utilization. The new policy is refined based on the old policy, and in this way, each sample can be used in multiple gradient ascent steps. As the new policy is refined, the two policies will deviate more, thus increasing the difference in estimates, so the old policy is periodically updated to match the new policy. The PPO algorithm reduces the computational complexity compared to the second-order optimization involved in the TRPO solution. The PPO uses a first-order optimization method with confidence intervals, which restricts the ratio between the old and new strategies $r_t(\theta)$ to the range $[1-\varepsilon, 1+\varepsilon]$ by clip.

3.2. Network Architecture

Because the PPO algorithm is based on the actor–critic method, the neural network used in this paper consists of policy and value networks, as shown in Figure 4. The policy network and the value network have the same structure, both have two fully connected layers with 128 units each, and the activation function is the Relu function. The output of the value network is the value of the current state vector s which is used to train the output of the policy network $a = (a1, a2, a3, a4)$. The action space of the policy network output is within a small symmetric range, such as $(-1, 1)$. Following this design increases versatility by limiting the action outputs to a maximum and minimum range, which can then be adapted to the action range of the actuator through a mix-and-match operation such as scaling. After the operation, we can find the speed of motor $r = (r_1, r_2, r_3, r_4)$.

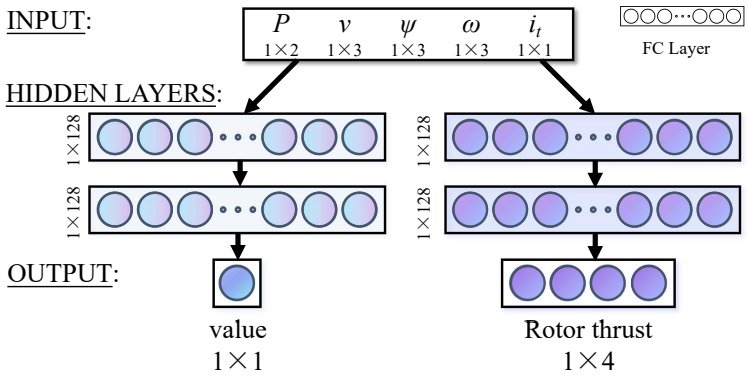

Figure 4. Neural network structure diagram. The neural network takes a twelve-dimensional vector s consisting of position $P = (y, z)$, velocity $v = (u, v, w)$, attitude angle $\psi = (\phi, \theta, \varphi)$, angular velocity $\omega = (p, q, r)$ and tilt angle (i_t) as input. The value network and the policy network respectively have two layers with 128 RULU units each. Value network output current state estimation value. The policy network outputs the amount of throttle $a = (a1, a2, a3, a4)$.

The input of the neural network is a 12-dimensional state vector s.

$$s = (y, z, u, v, w, \phi, \theta, \varphi, p, q, r, i_t) \in R^{12} \tag{10}$$

where (y, z) is position, (u, v, w) is velocity, (ϕ, θ, φ) is the Euler angle, (p, q, r) is the angular velocity and (i_t) is the current rotor tilt angle. To ensure the stability of the attitude of the vehicle during the transition, it is not only necessary to focus on the fact that the vehicle attitude angle (ϕ, θ, φ) needs to be at the desired value for a short period. It also needs to ensure that the velocity vector (u, v, w) is pointing forward, and the change in altitude and slip during the transition is not too severe. As the transition process is horizontal and forward, the x-axis position is constantly increasing, which has a large impact on the estimation error and the time required for training if not normalized. In this paper, the x-axis displacement has a small effect on attitude control, so the x-axis displacement is not used as a state input. Further, if the tilt angle is removed and only the pitch angle is used as a state input, the neural network must control both the motor output a and the tilt angle i_t. And this will increase the dimension of the action space and make learning the optimal strategy difficult in the absence of expert data.

3.3. Reward Function

The reward function is designed and optimized for the characteristics of the TRUAV transition process, given by the following Equation (11).

$$\begin{aligned} r_t = d + c_p\|\triangle P_t\| + c_{vx} \times \min(vx_t, v_{max}) + c_{vyz}\|\triangle v_t\| \\ + f(i_t) \times c_\Psi\|\triangle \Psi_t\| + c_\omega\|\triangle \omega_t\| + c_a\|a_t\| + c_{da}\|\triangle a_t\| \end{aligned} \tag{11}$$

where d is the survival reward; $\triangle P_t$ is the deviation from the y-axis and z-axis positions in the ground coordinate system; vx_t is the forward velocity and $\min(vx_t, v_{max})$ is a clap function to limit the maximum reward of the forward velocity. $\triangle v_t$ is the deviation from the y-axis and z-axis velocity in the ground coordinate system; $\triangle \Psi_t$ is the vehicle attitude angular deviation; $\triangle \omega_t$ is the vehicle attitude angular velocity deviation; a_t is the action output characterizing the vehicle energy consumption and $\triangle a_t$ is the amount of change from the previous moment of action representing action change magnitude.

The reward function in this paper focuses on guiding the vehicle forward with a positive x-axis velocity reward while maintaining flight stability and flight altitude. So speed-related rewards are divided into two parts, $c_{vx} \times \min(vx_t, v_{max})$ and $c_{vyz}\|\triangle v_t\|$. $\min(vx_t, v_{max})$ is a clap function to limit the maximum reward of the forward velocity.

A common problem in optimal control is the controller output constant oscillation between maximum and minimum. This causes unnecessary wear and tear on the actuators

in practice. This is not a significant problem in a simulator environment but is problematic in actual flight. In real aircraft, online training can have a significant impact on the stability of the aircraft due to perturbations in the ambient wind. The output of the PPO policy network is a Gaussian distribution of the action, which is sampled to obtain the output. The high variance of the distribution leads to high oscillations in the output actions. In this paper, $c_{da}\|\triangle a_t\|$ is introduced into the reward function as a penalty term to reduce the RL controller output jitter.

The aerodynamic force during the transition process is complex while the TRUAV accelerates from hover to cruise speed. Attitude and altitude considerations differ at the beginning and end of the TRUAV transition process. To reduce the deviation between attitude angles and desired values at the end of the process, a tilt-angle-related coefficient $f(i_t) = 1.0 + 0.5 * i_t/\pi$ is proposed, and the attitude error penalty term $c_\Psi\|\triangle\Psi_t\|$ introduces the variable coefficient.

3.4. Training

In this part, the controller of the TRUAV is trained with the PPO algorithm. This simulation environment is developed based on the open-source project framework named flightmare [34]. The agent interacts with the simulation environment for attitude controller training via reinforcement learning algorithms. To simulate a fully loaded situation, the maximum output thrust of each motor will be limited to a maximum thrust of 0.3 times the take-off weight of the vehicle. The goal of the training is to complete the transition process in 4 s with a simulation step length of 0.02 s, i.e., 200 steps. At the same time, the stability of the TRUAV attitude is ensured with the rotor tilt angle of the vehicle turning from 0° to 90°.

The training task is to perform attitude control of the UAV to complete the transition from copter to airplane mode under a random initial state, as shown in Table 2.

Table 2. Initial state random ranges and expectation values.

Parameter	Range
Initial Euler angle	$[\pm 30°, \pm 30°, \pm 30°]$
Expected Euler angle	$[0.0°, 0.0°, 0.0°]$
Initial speed	$[\pm 1.0, \pm 1.0, \pm 1.0]$ m/s
Expected speed	$[10.0, 0.0, 0.0]$ m/s
Initial position	$[\pm 1.0, 5.0 \pm 1.0]$ m
Expected position	$[0.0, 5.0]$ m [1]

[1] For position, focus only on the y-axis and z-axis.

Finally, the well-trained neural network controller was tested under different conditions, such as rotor tilt angle rate, rotor axis lengths and take-off weights, to verify its robustness and adaptability. Specific network hyperparameters are referenced in Table 3 below.

Table 3. Hyperparameters used in training.

Name	Value	Name	Value
d	0.1	lam	0.95
c_p	-0.002	gamma	0.92
c_{vx}	0.0009	n_step	200
c_{vyz}	-0.0003	learning_rate	1×10^{-5}
c_Ψ	-0.067	clip_param	0.1
c_ω	-0.0002	ent_coef	0
c_a	-0.0002		
c_{da}	-0.00002		

4. Results and Discussion

In this section, the convergence of the PPO algorithm for attitude control of TRUAV in the transition process is analyzed. Also, according to the characteristics of the TRUAV, the reward function is improved. There are two main points. One is to reduce the controller output jitter, and the other is to reduce the attitude angle deviation from desired values at the end of the transition process. Three simulations are reported to verify the adaptability of the designed RL controller during the transition process. In the first simulation, the adaptability of the RL controller to different takeoff weights was tested. In the second simulation, the adaptability of the RL controller to different diagonal wheelbases was tested. In the third simulation, the adaptability of the RL controller to different rotor tilt rates, i.e., rotor tilt strategies, was tested. Finally, the proposed method is compared with the A2C and standard PPO.

4.1. Convergence Analysis

Since the TRUAV model is highly nonlinear and tends to diverge during training, a lower learning rate and clip value were chosen. The average reward and the average episode length are shown in Figure 5. After 1500 iterations, the RL controller can keep the tilt-rotor UAV from falling during the transition process as shown in Figure 5b. Figure 5a shows that the maximum reward is reached after 5000 iterations.

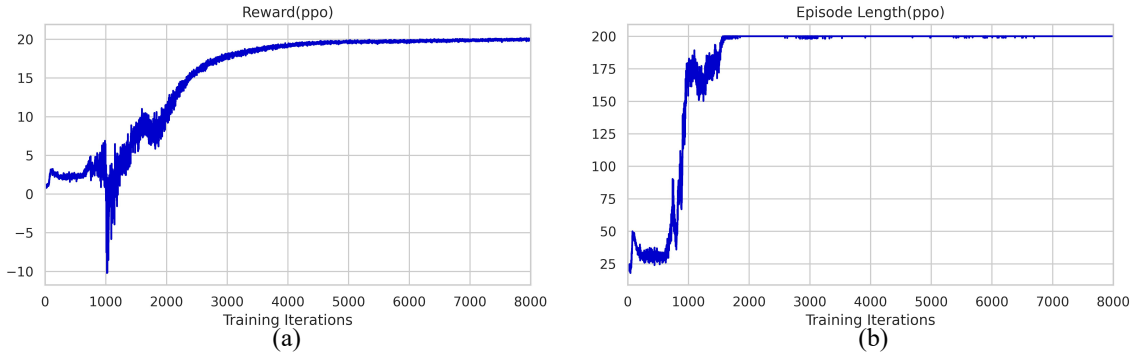

Figure 5. The training curve of the TRUAV: (**a**) Average accumulated reward; (**b**) Average steps for each episode.

A TRUAV simulation flight test is performed by applying the neural network parameters under different iterations with the same random initial state range of the TRUAV in Table 2 as in the training phase. To avoid losing generality, the simulation experiment was carried out 1000 times. The test results are shown in Figure 6. $Vx(m/s)$ is the x-axis speed of the TRUAV during the transition process. $Py(m)$ and $Pz(m)$ are the y-axis and z-axis positions of the TRUAV, respectively. $euler\_x(^\circ)$, $euler\_y(^\circ)$ and $euler\_z(^\circ)$ are the roll, pitch and yaw angle of the TRUAV, respectively. In Figure 6b, these are the outputs of the RL controller $a = (a1, a2, a3, a4)$.

The agent cannot complete the transition when the network is trained in 500 iterations, because it could fall at about 150 steps. After 1500 iterations, the TRUAV can complete the transition process and have a forward flight speed. However, the attitude control ability for roll and yaw is poor. In particular, the pitch angle changes greatly during the transition. The attitude angles have a large deviation from the desired values at the end of the transition process. And the motor throttle amounts of $a1$ and $a2$ are larger in 50 to 200 steps. After 2500 iterations, the attitude control effect is improved, especially the control effect of the pitch angle. The motor throttle amounts of $a1$ and $a2$ are improved. However, the $a3$ and $a4$ motor throttle amount oscillate more in 150 to 200 steps. After 8000 iterations, the performance of the attitude controller is improved significantly, especially the control effect of the pitch angle. The attitude angle is around the desired value, at the end of the

transition process. And it can be observed from the action output that the RL controller consumes less energy during the transition.

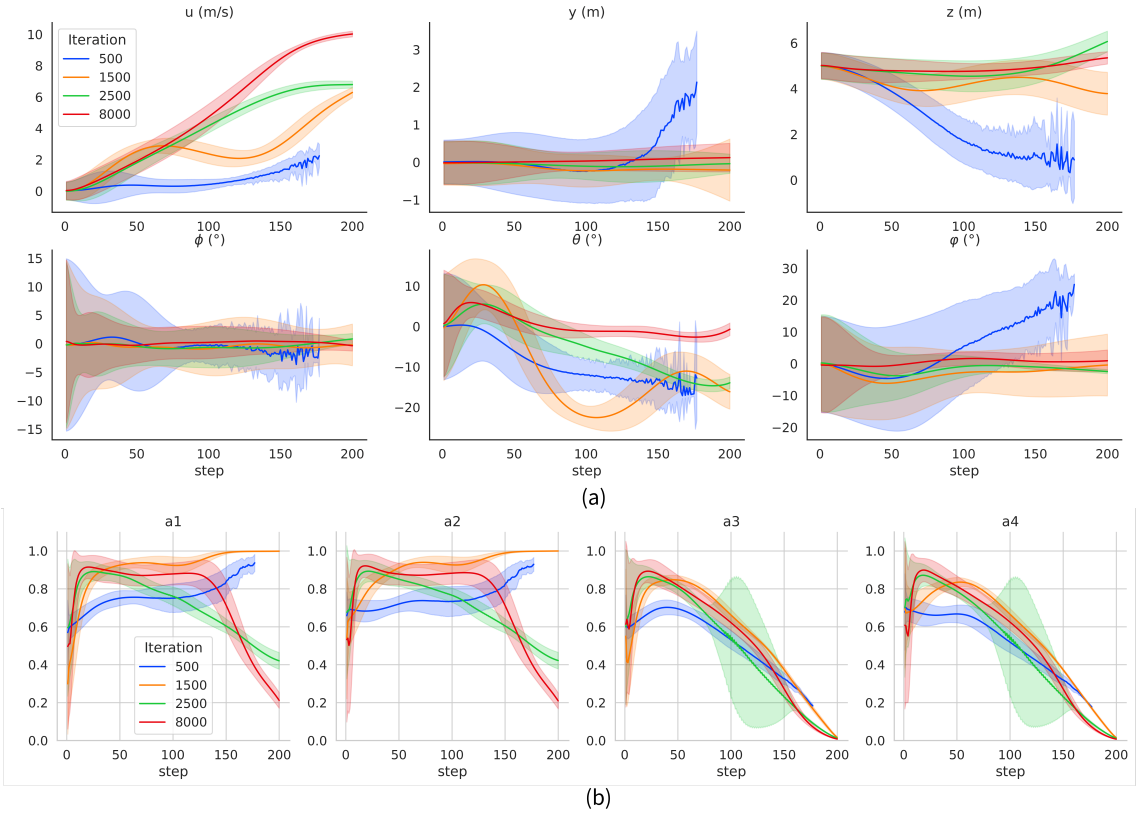

Figure 6. The simulation results of the TRUAV with different iterations: (**a**) $u(m/s)$ is the x-axis speed, $y(m)$ and $z(m)$ are the y-axis and z-axis positions, $\phi(^\circ)$, $\theta(^\circ)$ and $\varphi(^\circ)$ are the roll, pitch and yaw angle; (**b**) the motor throttle $a = (a1, a2, a3, a4)$.

4.2. Reward Function Improvements

In this paper, the reward function is modified to make it suitable for TRUAV transition process strategy learning. At the same time, the reward function is improved. There are two main points, the first point is to reduce the controller output jitter by adding the deviation from the previous time output into the reward function as a consideration. Another point is to reduce the deviation between attitude angles and desired values at the end of the transition process.

4.2.1. The Controller Output Jitter

This case compares the RL controller while training with or without the jitter cost $c_{da}\|\triangle a_t\|$ in the reward function. The test results are shown in Figure 7. The improved reward function results in lower output and a narrower variance band at the end of the transition process for the front motors $a1$ and $a2$. The RL controller outputs of the rear two motors $a3$ and $a4$ have a narrower variance band and significantly less oscillation in output at 100–150 steps. The improved reward function proposed in this paper has a narrower variance band, which can reduce the controller output jitter, especially for the output of the rear motors.

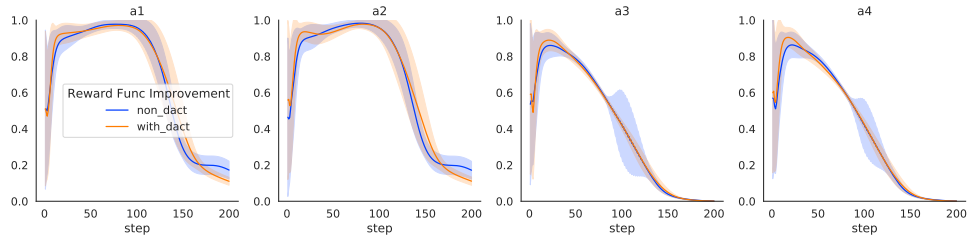

Figure 7. The figures are the output of the throttle $a = (a1, a2, a3, a4)$ while training with/without the jitter cost.

4.2.2. Attitude Angle Deviation from Desired Values

This paper proposes a coefficient related to the tilt angle of attitude angle deviation. In the reward function, the attitude angle deviation term $f(i_t) = 1.0 + 0.5 * i_t/\pi$ is multiplied by a gain that varies according to the rotor tilt angle. And it is used to reduce the deviation between attitude angles and desired values at the end of the process. The test results are shown in Figure 8. The TRUAV roll and yaw deviation from desired values are significantly reduced at the end of the transition process. Attitude considerations differ during the transition process. The transition process is very aerodynamically complex while the TRUAV accelerates from hover to cruise speed. Because the attitude angle changes and affects the angle of attack, the end of the transition process requires more attitude stability.

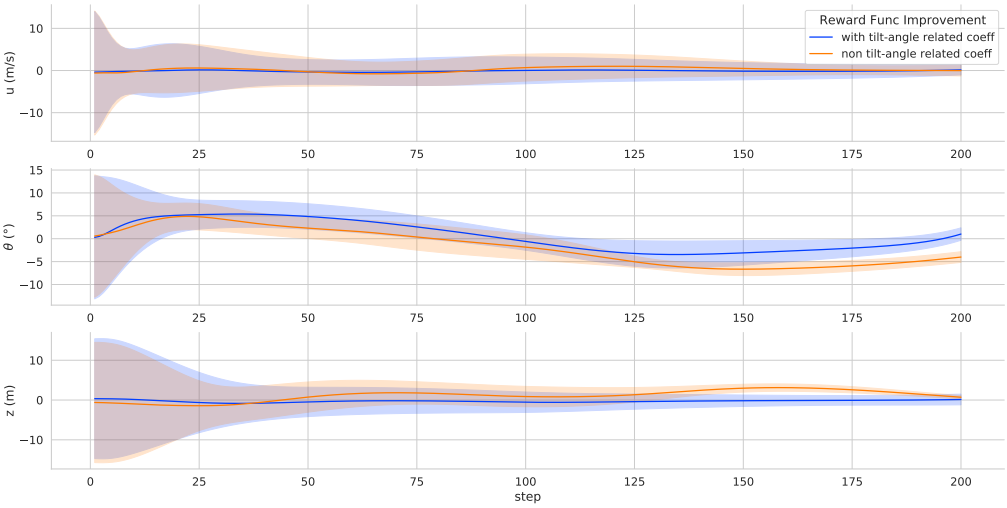

Figure 8. The performance of the UAV while training with/without the tilt-angle-related attitude cost.

4.3. Adaptability Test

To verify the adaptability and generalization of the RL controller trained in this paper, three simulations are performed. Two of them are to change the configuration of the TRUAV, take-off weight and diagonal wheelbase; the other is to change the rotor tilt rate during the transition process, i.e., the transition strategy.

4.3.1. Take-Off Weight Test

An enormous take-off weight means that more aerodynamic lift is required to maintain altitude. Different takeoff weights simulate the transition process under different operating conditions. Figure 9 shows that the controller can adapt to the take-off weight change using the trained neural network controller parameters. The RL controller completes the transition process by adjusting the speed and pitch to maintain the flight altitude. In the simulation environment, the adaptability of the TRUAV neural network controller was verified in varying take-off weights. There is no need to adjust the parameters, and the controller proposed in this paper can adapt to different takeoff weights.

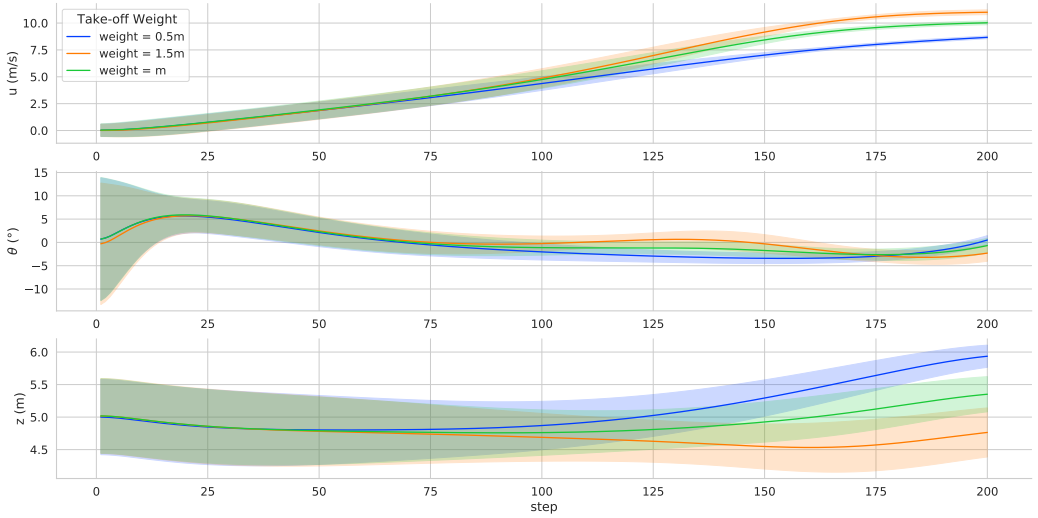

Figure 9. The simulation results of the TRUAV with different take-off weights. $u(m/s)$ is the x-axis speed, $\theta(^\circ)$ is the pitch angle and $z(m)$ is the altitude of the TRUAV during the transition process.

4.3.2. Diagonal Wheelbase Test

The simulation results are shown in Figure 10. There is a small change in the flight speed, pitch angle and altitude curves. Changes in the wheelbase directly affect the torque for attitude control.

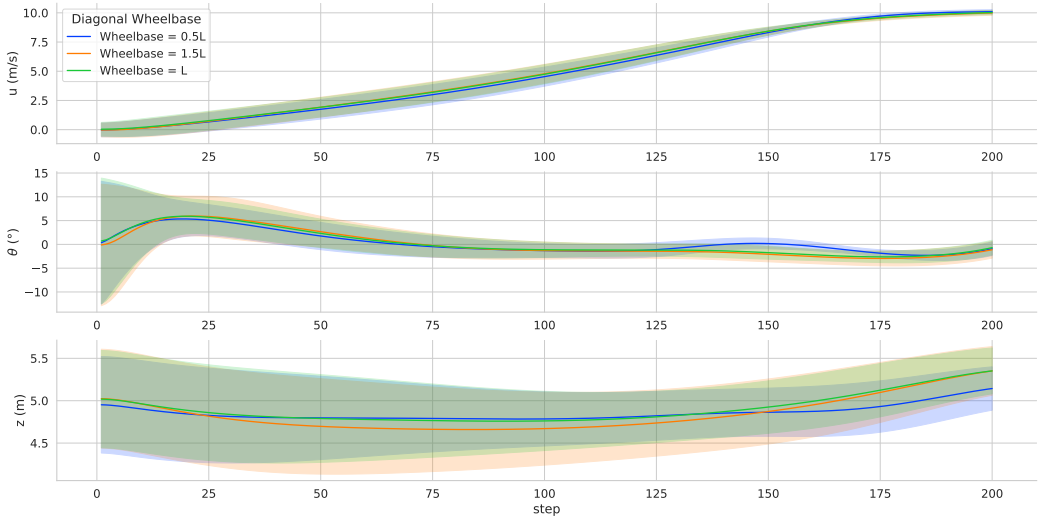

Figure 10. The simulation results of the TRUAV with different diagonal wheelbases.

When the wheelbase changes, it is a significant challenge to the controller for attitude control. From Figure 11, it can be found that decreasing the wheelbase has a more significant impact on the attitude control of the vehicle. The wheelbase decrease, and the motor throttle volume output jitters more. Such jitter may affect the TRUAV attitude control when subjected to external perturbations in practice. The adaptability of the neural network controller is verified by changing the diagonal wheelbase of the vehicle. In the actual application, if the diagonal wheelbase of the TRUAV changes, the actual flight data should be collected or the expert data should be used to update the neural network parameters.

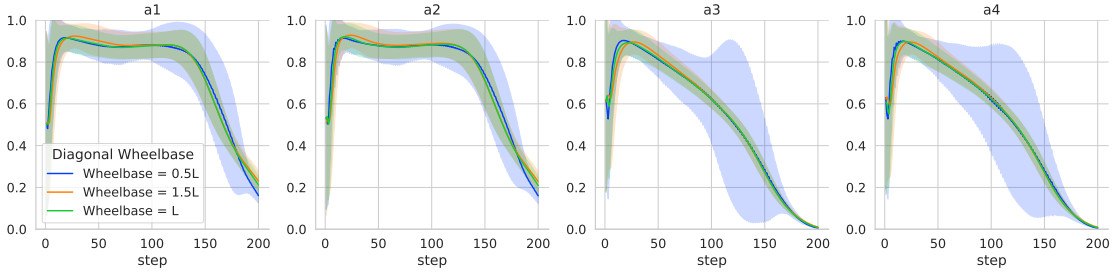

Figure 11. The simulation results of the TRUAV with different axis lengths. The output of the RL controller.

4.3.3. Rotor Tilt Angle Rate Test

The tilt angle is an important parameter in the TRUAV dynamic model. The change in the tilt angle rate directly affects the dynamic performance of the UAV. This variation affects the acceleration and flight attitude during the transition process. During RL controller training, the rotor tilt angle changes linearly. Three different tilt angle rate functions are selected to verify the adaptability of the RL controller, as shown below.

$$\begin{array}{lll} Linear & Function: & i(t) = 0.125 * \pi / t \\ Power & Function: & i(t) = 0.1984 * t^{2/3} \\ Sine & Function: & i(t) = 0.5 * \pi * sin(0.125 * \pi * t) \\ Quadratic & Function: & i(t) = 0.03125 * \pi * t^2 \\ & & 0 \leq t \leq 4 \end{array} \tag{12}$$

The controller is tested by varying the rotor tilt angle rate during the transition process. And the simulation results are shown in Figure 12. At different tilt rates, the desired speed can be achieved. And less attitude deviation between attitude angles and desired values can be achieved after the transition process. Results show that, under the condition of changing the tilt angle rate, the proposed controller can control the attitude of the TRUAV and has strong adaptability.

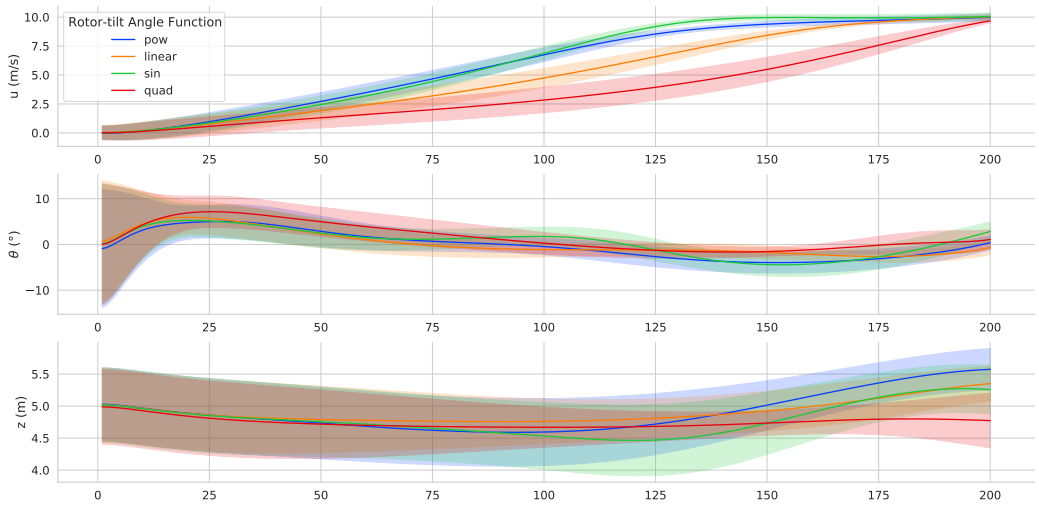

Figure 12. The simulation results of the TRUAV with different rotor tilt angle rates.

4.4. Performance

We compare our method with the standard PPO and the A2C algorithms in this case. The difference between standard PPO and the method proposed in this paper lies in the improvement of the reward function. The A2C algorithm involved in the comparison uses the same reward function and learning hyperparameters as the method proposed in

this paper. In particular, we compare the control performance of the three algorithms for attitude control. To avoid losing generality, each algorithm completes the transition process under the same random initial conditions in Table 2, and 1000 simulations are performed separately. The Euler angle of the TRUAV at the end of the transition process is collected and the simulation results are shown in Figure 13. From the figure, we can see that the standard PPO algorithm and the proposed method have better performance for the TRUAV attitude control during the transition process. However, under the same reward function and learning hyperparameters in Table 3, the A2C algorithm does not learn the attitude control of the transition process. At the end of the transition process, the TRUAV has a large pitch angle, and the TRUAV flies at a large angle of attack after the end of the tilt process. At the same time, the deviation of the roll angle and yaw angle from the desired value is large.

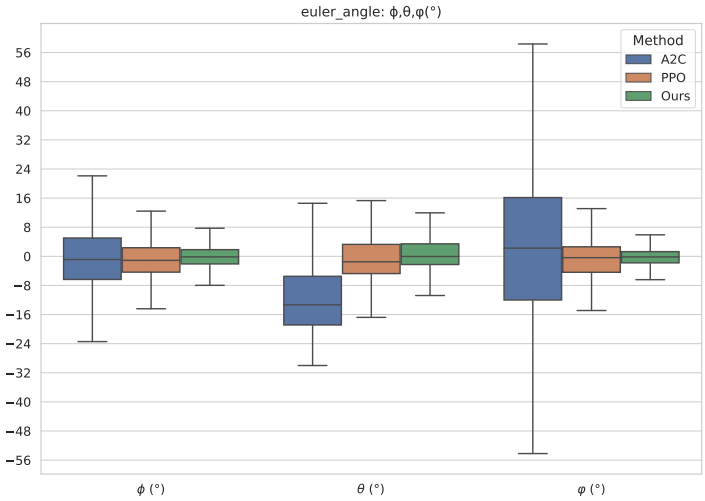

Figure 13. The attitude of the TRUAV with different methods at the end of the transition process (200th step). The standard PPO and our proposed method have a better performance at the end of the transition process compared with the standard A2C.

The energy consumption in the transition attitude control is also an important performance, as shown in Figure 14. It is found that the A2C algorithm consumes more energy in controlling attitude during the tilt-rotor transition process. The A2C controller outputs of the first two motors are almost in the state of maximum speed after 100 simulation steps, which further confirms that the TRUAV is flying at a large angle of attack at the end of the transition process. The results show that, although the RL controller obtained by A2C learning can complete the rotor tilt, it is in the high angle of attack flight state and cannot control the attitude well to reach the desired values.

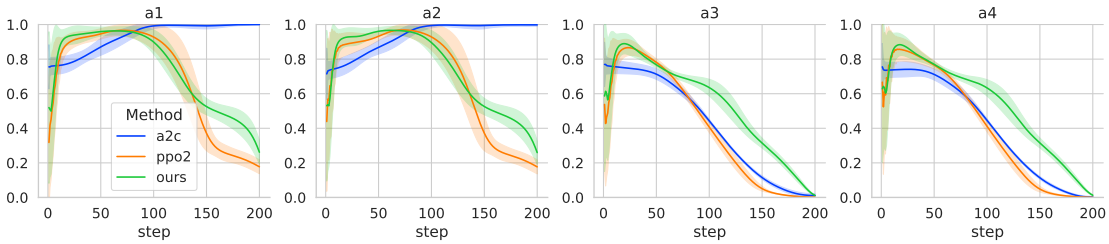

Figure 14. The outputs of the controllers with different methods. In terms of energy efficiency, the standard PPO and our proposed method have a better performance at the end of the transition process compared with the A2C. Our proposed algorithm reduces the deviation between attitude angles and desired values at a slightly larger energy cost compared with the standard PPO.

In Figure 15, we separately compare our method with the standard PPO algorithm. And it can be seen that our proposed method improves the attitude control performance of the TRUAV transition process significantly, which effectively reduces the deviation between attitude angles and desired values. In general, compared with the standard PPO algorithm, our proposed method has a higher performance in a deviation between attitude angles and desired values while consuming more energy in the second half of the transition process.

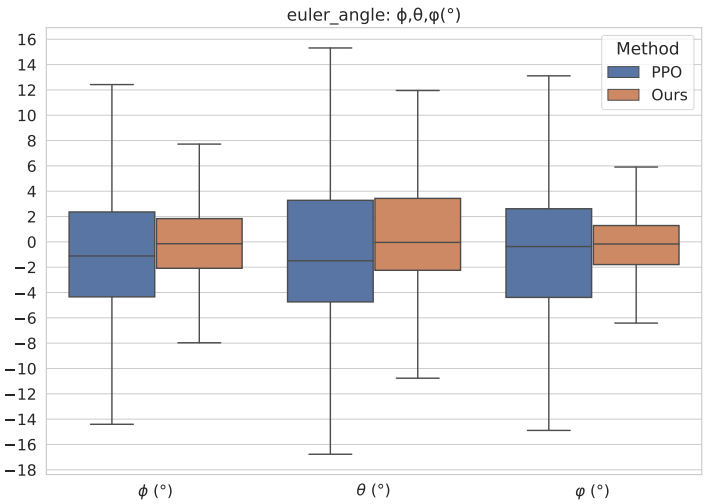

Figure 15. The attitude of the TRUAV with different methods at the end of the transition process. Compared with the standard PPO, our proposed method performs better at deviation between attitude angles and desired values.

5. Conclusions

In summary, the purpose of this study is to investigate the application of reinforcement learning in the design of the TRUAV controller in the transition process. Through simulation and evaluation, it is observed that the attitude control performance is significantly improved for the design of the reward function. And significant adaptability to changes in takeoff weight, axis distance and rotor tilt rate is demonstrated. These findings demonstrate the effectiveness of reinforcement learning in addressing flight control during the transition process of the TRUAV. The adaptive and self-learning properties of reinforcement learning enable the flight controller to adapt to flight conditions, thereby improving control performance. The reinforcement learning algorithm effectively learns the optimal control strategy by interacting with the environment. By successfully applying reinforcement learning to the flight control design of the TRUAV transition process, we have demonstrated its potential to improve the performance and adaptability of the flight controller. In the current study, the research was conducted solely in a simulated environment, which may not fully capture the complexities and uncertainties present in the real world such as measurement noise, environmental disturbances and possible system failures or anomalies. It is crucial to recognize that the outcomes obtained in a controlled simulation may not directly translate to practical applications without accounting for these uncertainties and challenges. In further research, the application of reinforcement learning in more complex scenarios in the TRUAV flight control deserves to be explored and refined.

Author Contributions: Conceptualization, C.D. and Y.Z.; methodology, R.Y.; software, R.Y.; validation, R.Y., H.G. and Y.W.; formal analysis, C.D., R.Y. and Y.W.; investigation, Y.W. and T.F.; resources, C.D.; data curation, R.Y.; writing—original draft preparation, R.Y.; writing—review and editing, C.D. and R.Y.; visualization, R.Y.; supervision, Y.Z. and C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ren, X.; Wang, J. Forecasting Traffic Volume of Urban Logistics Drones in Low-altitude Airspace. *J. Transp. Inf. Saf.* **2022**, *40*, 97–105.
2. Etkin, B.; Reid, L.D. *Dynamics of Flight: Stability and Control*; John Wiley & Sons: Hoboken, NJ, USA, 1995.
3. Stevens, B.L.; Lewis, F.L.; Johnson, E.N. *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
4. Liu, Z.; He, Y.; Yang, L.; Han, J. Control techniques of tilt rotor unmanned aerial vehicle systems: A review. *Chin. J. Aeronaut.* **2017**, *30*, 135–148. [CrossRef]
5. Maya-Gress, K.; Álvarez, J.; Villafuerte-Segura, R.; Romero-Trejo, H.; Bernal, M. A novel family of exact nonlinear cascade control design solutions for a class of UAV systems. *Math. Probl. Eng.* **2021**, *2021*, 2622571. [CrossRef]
6. Hermand, E.; Nguyen, T.W.; Hosseinzadeh, M.; Garone, E. Constrained control of UAVs in geofencing applications. In Proceedings of the 2018 26th Mediterranean Conference on Control and Automation (MED), Zadar, Croatia, 19–22 June 2018; pp. 217–222.
7. Li, S.; Lv, Z.; Feng, L.; Wu, Y.; Li, Y. Nonlinear Cascade Control for a New Coaxial Tilt-rotor UAV. *Int. J. Control Autom. Syst.* **2022**, *20*, 2948–2958. [CrossRef]
8. Nakamura, Y.; Arakawa, A.; Watanabe, K.; Nagai, I. Transitional flight simulations for a tilted quadrotor with a fixed-wing. In Proceedings of the 2018 IEEE International Conference on Mechatronics and Automation (ICMA), Changchun, China, 5–8 August 2018; pp. 1829–1836.
9. Shen, D.; Lu, Q.; Hu, M.; Kong, Z. Mathematical modeling and control of the quad tilt-rotor UAV. In Proceedings of the 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), Tianjin, China, 19–23 July 2018; pp. 1220–1225.
10. Bauersfeld, L.; Ducard, G. Fused-PID control for tilt-rotor VTOL aircraft. In Proceedings of the 2020 28th Mediterranean Conference on Control and Automation (MED), Saint-Raphaël, France, 15–18 September 2020; pp. 703–708.
11. Sun, Z.; Wang, R.; Zhou, W. Finite-time stabilization control for the flight mode transition of tiltrotors based on switching method. In Proceedings of the 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, China, 28–30 May 2017; pp. 2049–2053.
12. Liu, Z.; Theilliol, D.; Yang, L.; He, Y.; Han, J. Transition control of tilt rotor unmanned aerial vehicle based on multi-model adaptive method. In Proceedings of the 2017 International Conference on Unmanned Aircraft Systems (ICUAS), Miami, FL, USA, 13–16 June 2017; pp. 560–566.
13. Lin, H.; Fu, R.; Zeng, J. Extended state observer based sliding mode control for a tilt rotor UAV. In Proceedings of the 2017 36th Chinese Control Conference (CCC), Dalian, China, 26–28 July 2017; pp. 3771–3775.
14. Ducard, G.J.; Allenspach, M. Review of designs and flight control techniques of hybrid and convertible VTOL UAVs. *Aerosp. Sci. Technol.* **2021**, *118*, 107035. [CrossRef]
15. Ta, D.A.; Fantoni, I.; Lozano, R. Modeling and control of a tilt tri-rotor airplane. In Proceedings of the 2012 American Control Conference (ACC), Montreal, QC, Canada, 27–29 June 2012; pp. 131–136.
16. Yin, Y.; Niu, H.; Liu, X. Adaptive neural network sliding mode control for quad tilt rotor aircraft. *Complexity* **2017**, *2017*, 7104708. [CrossRef]
17. Yatsun, A.; Lushnikov, B.; Emelyanova, O. Motion control automation in the quadcopter convertiplane in a transient mode. In Proceedings of the 2018 International Russian Automation Conference (RusAutoCon), Sochi, Russia, 9–16 September 2018; pp. 1–6.
18. Yu, L.; He, G.; Zhao, S.; Wang, X. Dynamic inversion-based sliding mode control of a tilt tri-rotor UAV. In Proceedings of the 2019 12th Asian Control Conference (ASCC), Kitakyushu-shi, Japan, 9–12 June 2019; pp. 1637–1642.
19. Pan, Z.; Chi, C.; Zhang, J. Nonlinear attitude control of tiltrotor aircraft in helicopter mode based on ADRSM theory. In Proceedings of the 2018 37th Chinese Control Conference (CCC), Wuhan, China, 25–27 July 2018; pp. 9962–9967.
20. Hegde, N.T.; George, V.; Nayak, C.G. Modelling and transition flight control of vertical take-off and landing unmanned tri-tilting rotor aerial vehicle. In Proceedings of the 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 12–14 June 2019; pp. 590–594.
21. Chiappinelli, R.; Cohen, M.; Doff-Sotta, M.; Nahon, M.; Forbes, J.R.; Apkarian, J. Modeling and control of a passively-coupled tilt-rotor vertical takeoff and landing aircraft. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 4141–4147.

22. Cardoso, D.N.; Esteban, S.; Raffo, G.V. A nonlinear W∞ controller of a tilt-rotor UAV for trajectory tracking. In Proceedings of the 2019 18th European Control Conference (ECC), Naples, Italy, 25–28 June 2019; pp. 928–934.
23. Allenspach, M.; Ducard, G.J.J. Nonlinear model predictive control and guidance for a propeller-tilting hybrid unmanned air vehicle. *Automatica* **2021**, *132*, 109790. [CrossRef]
24. Bauersfeld, L.; Spannagl, L.; Ducard, G.J.; Onder, C.H. MPC flight control for a tilt-rotor VTOL aircraft. *IEEE Trans. Aerosp. Electron. Syst.* **2021**, *57*, 2395–2409. [CrossRef]
25. Peng, C.; Wang, X.M.; Chen, X. Design of tiltrotor flight control system in conversion mode using improved neutral network PID. *Adv. Mater. Res.* **2014**, *850*, 640–643. [CrossRef]
26. Yu, C.; Zhu, J.; Sun, Z. Nonlinear adaptive internal model control using neural networks for tilt rotor aircraft platform. In Proceedings of the 2005 IEEE Midnight-Summer Workshop on Soft Computing in Industrial Applications, Espoo, Finland, 28–30 June 2005; pp. 12–16.
27. Yuksek, B.; Inalhan, G. Transition Flight Control System Design for Fixed-Wing VTOL UAV: A Reinforcement Learning Approach. In Proceedings of the AIAA SCITECH 2022 Forum, Orlando, FL, USA, 23–27 January 2023; p. 0879.
28. Koch, W.; Mancuso, R.; West, R.; Bestavros, A. Reinforcement learning for UAV attitude control. *ACM Trans. Cyber-Phys. Syst.* **2019**, *3*, 1–21. [CrossRef]
29. Xu, J.; Du, T.; Foshey, M.; Li, B.; Zhu, B.; Schulz, A.; Matusik, W. Learning to fly: Computational controller design for hybrid uavs with reinforcement learning. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
30. Huo, Y.; Li, Y.; Feng, X. Memory-based reinforcement learning for trans-domain tiltrotor robot control. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2020; Volume 1510, p. 012011.
31. Lee, H.; Jeong, M.; Kim, C.; Lim, H.; Park, C.; Hwang, S.; Myung, H. Low-level pose control of tilting multirotor for wall perching tasks using reinforcement learning. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 9669–9676.
32. Hegde, N.T.; George, V.; Nayak, C.G.; Kumar, K. Design, dynamic modelling and control of tilt-rotor UAVs: A review. *Int. J. Intell. Unmanned Syst.* **2019**, *8*, 143–161. [CrossRef]
33. Misra, A.; Jayachandran, S.; Kenche, S.; Katoch, A.; Suresh, A.; Gundabattini, E.; Selvaraj, S.K.; Legesse, A.A. A Review on Vertical Take-Off and Landing (VTOL) Tilt-Rotor and Tilt Wing Unmanned Aerial Vehicles (UAVs). *J. Eng.* **2022**, *2022*, 1803638. [CrossRef]
34. Song, Y.; Naji, S.; Kaufmann, E.; Loquercio, A.; Scaramuzza, D. Flightmare: A Flexible Quadrotor Simulator. In Proceedings of the Conference on Robot Learning, Cambridge, MA, USA, 16–18 November 2020.