DOI：10.13973/j.cnki.robot.240329 **CSTR**：32165.14.robot.240329

基于深度强化学习的四旋翼无人机双向推力控制

李晓信，刘志宏，王冠政，王祥科
（国防科技大学智能科学学院，湖南 长沙 410073）

摘 要：目前四旋翼无人机控制技术主要采用电机正向推力的思路，限制了双向转动电机的应用潜力。为了提高四旋翼无人机的机动性，扩展其动作空间，实现机动飞行和快速控制，提出一种基于深度强化学习的四旋翼无人机双向推力控制方法。该方法首次将深度强化学习与四旋翼无人机双向推力控制相结合，在双向推力的旋翼无人机动力学模型的基础上，设计基于深度强化学习的神经网络控制器，控制无人机底层的 4 个电机的期望推力，实现了无人机在剧烈运动状态下的快速悬停。同时，开展了大姿态角、大速度、大角速度等剧烈运动状态下四旋翼无人机稳定悬停控制仿真实验，实验结果表明，与现有使用正向推力的控制器相比，本文提出的双向推力控制器执行动作更平滑、无人机状态波动更小、控制时间更短、鲁棒性更强，能够有效提升无人机的控制效果。

关键词：四旋翼无人机；深度强化学习；神经网络控制器；双向推力；底层控制

Bidirectional Thrust Control of a Quadrotor with Deep Reinforcement Learning

LI Xiaoxin，LIU Zhihong，WANG Guanzheng，WANG Xiangke
(*College of Intelligence Science and Technology*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: Current control technologies for quadrotors mainly adopt the approach of utilizing positive thrust from the motor, limiting the application potential of bidirectional motors. To improve the maneuverability of quadrotors, expand their action space, and achieve agile flight and rapid control, a bidirectional thrust control method for quadrotors based on deep reinforcement learning is proposed. This method combines deep reinforcement learning with bidirectional thrust control for quadrotors for the first time. Based on the dynamic model of quadrotors with bidirectional thrust, a neural network controller based on deep reinforcement learning is designed to control the expected low-level thrust of 4 motors, realizing rapid hovering under extreme conditions. In addition, simulations of quadrotor stable hovering control under extreme conditions, such as large attitudes, high speeds, and high angular speeds, are conducted. The experimental results show that, compared to existing controllers using positive thrust, the proposed bidirectional thrust controller performs smoother actions, with smaller state fluctuations, shorter control times, and stronger robustness, effectively improving the control performance of the quadrotor.

Keywords: quadrotor; deep reinforcement learning; neural network controller; bidirectional thrust; low level control

旋翼无人机具有操作简单、机动性好以及可垂直起降等优点，是目前低空经济运行的主要载体之一，在航拍摄影、物流配送、城市治理、智慧农业、应急救援、地理测绘、能源电力等领域获得了广泛的应用[1-2]。

旋翼无人机是一个典型的多输入多输出、非线性、强耦合的欠驱动系统，通过 4 个电机的转动实现空间中 6 自由度的平移和旋转。对于旋翼无人机的控制，目前主流的研究和应用集中在传统控制方法上，主要可以分为 2 类，一类为与模型无关的控制算法，该类方法不依赖无人机动力学模型，如 PID 控制[3]，但该类方法严重依赖控制器参数的设置，往往取决于人类的先验经验，难以发挥出旋翼无人机的最大性能。另一类为基于无人机动力学模型设计的控制算法，如 MPC（模型预测控制）[4]、LQR（线性二次型调节器）控制[5] 等，此类方法严重依赖建模的精确性，存在很大的难度，在无人机的实际飞行控制中较难发挥出较好的效果。

随着人工智能技术的快速发展，深度强化学习方法也逐渐应用到旋翼无人机控制领域[6]。相比于传统控制方法，强化学习方法主要有 2 个优势。第一，基于强化学习的控制器在无人机与环境的交互中不断得到优化和完善，不依赖于精确模型，具有较强的鲁棒性。第二，大部分的传统控制器输出为无人机的上层控制量，如位置、速度、加速度等，再通过一步一步解算得到底层控制量，而基于强化

基金项目：国家自然科学基金（U23B2032）.
通信作者：刘志宏，zhliu@nudt.edu.cn 收稿/录用/修回：2024-12-12/2025-02-05/2025-02-07

学习的控制器能够实现端到端的底层控制，从而能够快速响应，在机动飞行场景下优势明显。

文 [7] 首次提出使用深度强化学习方法进行旋翼无人机的底层控制，实现了旋翼无人机的快速悬停，证明了使用神经网络控制无人机的可行性。在其基础上，文 [8] 设计了动态随机化方法，使用同一个神经网络控制 3 种不同类型的旋翼无人机，且都取得了较好的控制效果，证明了使用深度强化学习所得到的神经网络控制器具有强大的鲁棒性。苏黎世大学机器人与感知团队 [9-12] 在本领域取得了惊人的研究成果，使用深度强化学习进行无人机竞速任务，不仅取得了超于传统控制方法的效果 [13]，也在同台竞技中击败了人类顶尖穿越机选手 [14]。但是，目前所有使用神经网络对无人机进行的底层控制，都局限于正向推力，默认电机只能朝一个方向转动。

随着科技的发展，现有的旋翼无人机硬件设施已经能够支持电机的双向转动，从而能够支撑无人机双向推力的实现。这扩大了无人机的动作空间，如果能将其好好利用，必将极大提升无人机的机动性。另一方面，电机的双向推力，也能扩展旋翼无人机的使用场景，如倒飞、降落在较大倾斜度的斜面上、翻转坠落后的恢复、快速下降等等。

目前对于双向推力的旋翼无人机控制研究较少，且都局限于传统控制方法。文 [15] 首次提出使用单向电机和可变桨叶进行旋翼无人机的双向推力控制，实现了倒立飞行和半翻转悬停，证明了双向推力的可控性。随着旋翼无人机硬件设施的提升，文 [16] 首次提出使用双向电机和固定桨叶进行旋翼无人机的双向推力控制，分析了对称桨叶的特性，实现了旋翼无人机在大倾斜度斜面上的降落。在其基础上，文 [17] 提出了基于双向推力的旋翼无人机倒立机动飞行。更进一步的，文 [18] 提出了新的控制分配算法，优化了双向推力情况下的电机饱和问题。文 [19] 提出了基于双向推力的旋翼无人机微分平坦轨迹规划方法。然而，上述基于传统控制的双向推力方法，仍面临着高度依赖精确建模和参数设置的问题。

综上所述，目前使用深度强化学习对旋翼无人机进行控制的方法有很多，但主要集中在上层控制任务，如目标跟踪 [20]、编队飞行 [21]、轨迹跟踪 [22] 等等，限制了无人机的机动性。且现有强化学习方法普遍采用单向推力的电机－桨叶模型，无法使无人机发挥其全部潜力。

基于当前的研究现状，为了提高旋翼无人机的机动性，扩展其动作空间，实现无人机机动飞行和快速控制，本文提出一种基于深度强化学习的旋翼无人机双向推力控制方法。该方法首次将深度强化学习与旋翼无人机双向推力控制相结合，在双向推力的旋翼无人机动力学模型的基础上，设计基于深度强化学习的神经网络控制器，控制无人机底层的 4 个电机的期望推力，实现了无人机在剧烈运动状态下的快速悬停。

1 面向双向推力的四旋翼无人机动力学模型设计（Design of a dynamic model for quadrotor with bidirectional thrust）

1.1 无人机结构及坐标系

旋翼无人机的结构如图 1 所示，采用正 X 形机架，将其建模为由 4 个电机驱动的 6 自由度刚体，电机顺序和旋转方向在图中标出。

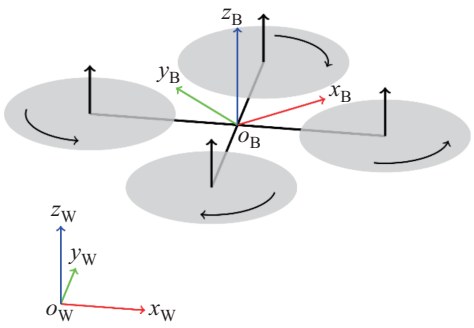

图 1 四旋翼无人机结构及坐标系

Fig.1 The structure and coordinate systems of quadrotor

选择世界坐标系 W（$o_W x_W y_W z_W$）与机体坐标系 B（$o_B x_B y_B z_B$）作为建立动力学模型的参考基准。

1.2 运动学方程

基于牛顿第二定律和牛顿－欧拉公式可得到无人机的运动学方程为

$$\dot{\boldsymbol{p}}_{WB} = \boldsymbol{v}_{WB} \tag{1}$$

$$\dot{\boldsymbol{p}}_{WB} = \frac{1}{2}\boldsymbol{\Lambda}(\boldsymbol{\omega}_B)\boldsymbol{q}_{WB} \tag{2}$$

$$\dot{\boldsymbol{v}}_{WB} = \boldsymbol{q}_{WB} \odot \boldsymbol{c} + \boldsymbol{g} \tag{3}$$

$$\dot{\boldsymbol{\omega}}_B = \boldsymbol{J}^{-1}(\boldsymbol{\eta} - \boldsymbol{\omega}_B \times \boldsymbol{J}\boldsymbol{\omega}_B) \tag{4}$$

其中，$\boldsymbol{p}_{WB} = [p_x, p_y, p_z]^{\mathrm{T}}$ 和 $\boldsymbol{v}_{WB} = [v_x, v_y, v_z]^{\mathrm{T}}$ 是无人机在世界坐标系下的位置和速度。为了避免欧拉角万向锁问题引发的歧义性，使用单位四元数 $\boldsymbol{q}_{WB} = [q_w, q_x, q_y, q_z]^{\mathrm{T}}$ 表示无人机的姿态，$\boldsymbol{\omega}_B = [\omega_x, \omega_y, \omega_z]^{\mathrm{T}}$ 为无人机机体坐标系下的角速度。无人机受到重力加速度 $\boldsymbol{g} = [0, 0, -g]^{\mathrm{T}}$ 和推力加速度 $\boldsymbol{c} = [0, 0, c]^{\mathrm{T}}$ 使其在空间中发生平移变化，受到转矩 $\boldsymbol{\eta}$ 使其在空间

中发生旋转变化。$\boldsymbol{J}=\mathrm{diag}(I_x,I_y,I_z)$ 对角矩阵为无人机的转动惯量矩阵。$\boldsymbol{\Lambda}(\boldsymbol{\omega}_B)$ 是一个斜对称矩阵，其展开为

$$\boldsymbol{\Lambda}(\boldsymbol{\omega}_B)=\begin{bmatrix}0&-\omega_x&-\omega_y&-\omega_z\\ \omega_x&0&\omega_z&-\omega_y\\ \omega_y&-\omega_z&0&\omega_x\\ \omega_z&\omega_y&-\omega_x&0\end{bmatrix} \tag{5}$$

1.3　受力分析

无人机的重力加速度取 $g=9.81\ \mathrm{m\cdot s^{-2}}$。机体坐标系下的推力加速度为

$$c=(f_1+f_2+f_3+f_4)/M \tag{6}$$

其中，M 为无人机的总质量。根据无人机的结构可得到其转矩为

$$\boldsymbol{\eta}=\begin{bmatrix}\dfrac{L}{\sqrt{2}}(f_1-f_2-f_3+f_4)\\ \dfrac{L}{\sqrt{2}}(-f_1-f_2+f_3+f_4)\\ K_\tau(f_1-f_2+f_3-f_4)\end{bmatrix} \tag{7}$$

其中，L 为无人机的机臂长度，K_τ 为桨叶的转矩推力比系数，f_i 为单个桨叶的推力。

1.4　双向推力电机－桨叶模型

为了实现旋翼无人机的双向推力控制，基于对称桨叶和双向无刷电机，创新地提出双向推力的电机－桨叶模型，根据叶素－动量理论，可以得到单个桨叶的推力为

$$f_i=K_\mathrm{f}\cdot\mathrm{sgn}(\Omega_i)\cdot\Omega_i^2 \tag{8}$$

其中，K_f 为桨叶的推力系数，$\mathrm{sgn}(\Omega_i)$ 为电机的旋转方向，Ω_i 为单个电机的转速。对于单个电机，将其转速 Ω 建模为 1 阶系统：

$$\dot{\Omega}=(\Omega_\mathrm{des}-\Omega)/K_\alpha \tag{9}$$

其中，K_α 为电机响应时间常数。

2　基于深度强化学习的神经网络控制器设计（Design of neural network controller with deep reinforcement learning）

2.1　马尔可夫决策过程

本文的最终目标是使用深度强化学习的方法，获得一个神经网络控制器 $\boldsymbol{\pi}$，控制无人机在剧烈运动状态下实现快速稳定悬停，这里的剧烈运动状态指大姿态角（俯仰、横滚、偏航角度显著超出正常飞行范围）、大速度、大角速度。该控制器采用端到端的控制方式，输入为无人机当前状态与目标状态之差，输出为无人机 4 个电机的期望推力。

本深度强化学习任务可建模为马尔可夫决策过程，用元组 (S,A,P,R) 来表示。无人机在当前的状态 $\boldsymbol{s}_t\in S$ 下，基于策略 $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)\in\boldsymbol{\pi}$，生成一个动作 $\boldsymbol{a}_t\in A$，这个动作导致无人机转移到下一个状态 $\boldsymbol{s}_{t+1}\in S$，状态转移函数为 $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)\in P$，同时无人机获得一个即时奖励 $r(\boldsymbol{s}_t,\boldsymbol{a}_t,\boldsymbol{s}_{t+1})\in R$。

无人机从初始状态开始，一直到最终结束飞行，所有时刻的状态、动作、奖励所组成的集合称为轨迹：

$$\tau=\{(\boldsymbol{s}_0,\boldsymbol{a}_0,r_0),\cdots,(\boldsymbol{s}_T,\boldsymbol{a}_T,r_T)\} \tag{10}$$

一条轨迹上的累计奖励称之为回报：

$$R(\tau)=\sum_{t=0}^{T}\gamma^t r_t \tag{11}$$

其中，$\gamma\in[0,1]$ 为折扣因子，用来权衡奖励的短期与长期收益。深度强化学习的目标是通过不断训练，获得一组最优的神经网络参数 π_θ^*，使其最大化期望回报：

$$\pi_\theta^*=\arg\max_{\pi}E_{\tau-\pi}\left(\sum_{t=0}^{T}\gamma^t r_t\right) \tag{12}$$

下面将详细展开介绍各部分的设计。

2.2　神经网络控制器设计

定义无人机的状态 $\boldsymbol{s}_t$ 为 12 维向量：

$$\boldsymbol{s}_t=[p_x,p_y,p_z,v_x,v_y,v_z,\varphi_z,\theta_y,\rho_x,\omega_x,\omega_y,\omega_z]^\mathrm{T} \tag{13}$$

其中，$[p_x,p_y,p_z]$ 为无人机在世界坐标系下的位置，$[v_x,v_y,v_z]$ 为无人机在世界坐标系下的速度，$[\varphi_z,\theta_y,\rho_x]$ 为无人机的姿态欧拉角，$[\omega_x,\omega_y,\omega_z]$ 为无人机的机体角速度。定义无人机的动作 $\boldsymbol{a}_t$ 为 4 个电机的期望推力：

$$\boldsymbol{a}_t=[f_{1\mathrm{des}},f_{2\mathrm{des}},f_{3\mathrm{des}},f_{4\mathrm{des}}]^\mathrm{T} \tag{14}$$

无人机最终收敛悬停的目标状态为 $\boldsymbol{s}_\mathrm{des}$。将无人机当前状态 $\boldsymbol{s}_t$ 与无人机目标状态 $\boldsymbol{s}_\mathrm{des}$ 的差值 $\Delta\boldsymbol{s}_t$ 作为控制器的输入，将无人机的动作 $\boldsymbol{a}_t$ 作为控制器的输出。

策略 $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)$ 即神经网络控制器，由全连接神经网络组成，其有 2 个隐藏层，每个隐藏层含有 64 个神经元，激活函数为 tanh。这样设计主要是为了在计算效率和控制能力之间取得平衡。全连接神经网络是神经网络的经典结构，具有强大的表达能力和较为简单的结构，它通过层间的全连接，能够

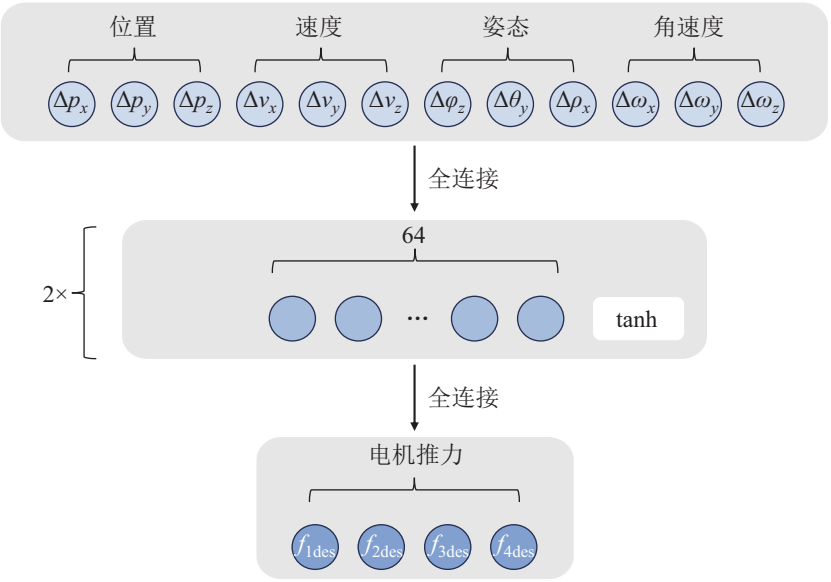

图 2 神经网络控制器结构

Fig.2 Structure of the neural network controller

学习到复杂的特征。2 层隐藏层属于较为浅层的神经网络结构，与更深的网络相比，其参数数量相对较少，因此计算成本较低。这种结构更适合旋翼无人机机载计算资源有限的场景。每层 64 个神经元的选择提供了适中的容量，可以学习到数据中的非线性特征，而不过度增加参数数量，避免过拟合风险。考虑到神经网络输出范围有正有负，采用激活函数 tanh 进一步增强了网络的非线性表达能力，使其能够适应飞行控制中的复杂动态行为。神经网络控制器结构如图 2 所示。

2.3 状态转移函数和奖励函数设计

无人机的动作 $\boldsymbol{a}_t=[f_{1\mathrm{des}},f_{2\mathrm{des}},f_{3\mathrm{des}},f_{4\mathrm{des}}]^{\mathrm{T}}$，通过式 (8) 双向推力的电机－桨叶模型的逆向推导，得到期望电机转速：

$$\Omega_{\mathrm{des}}=\mathrm{sgn}(f_{\mathrm{des}})\cdot\sqrt{\frac{f_{\mathrm{des}}}{K_{\mathrm{f}}}} \tag{15}$$

再通过旋翼无人机动力学模型对控制步长进行数值积分得到新的状态 $\boldsymbol{s}_{t+1}$。本深度强化学习的状态转移函数 $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)$ 为旋翼无人机在 3 维空间中的平移和旋转运动。

为了使无人机尽快稳定到悬停目标状态，设计奖励函数：

$$\begin{aligned}r_t=&\alpha_{\mathrm{p}}\cdot\|\Delta\boldsymbol{p}\|_2+\alpha_{\mathrm{v}}\cdot\|\Delta\boldsymbol{v}\|_2+\alpha_{\mathrm{o}}\cdot\|\Delta\boldsymbol{o}\|_2+\\&\alpha_{\omega}\cdot\|\Delta\boldsymbol{\omega}\|_2+\alpha_{\mathrm{c}}\cdot Q_{\mathrm{c}}+\alpha_{\mathrm{a}}\cdot Q_{\mathrm{a}}\end{aligned} \tag{16}$$

等号右侧各项依次为位置、速度、姿态、角速度误差惩罚项，以及坠毁惩罚与存活奖励项。当无人机坠毁时 Q_{c} 等于 1，Q_{a} 等于 0；当无人机存活时 Q_{c} 等于 0，Q_{a} 等于 1。起主要作用的是 $\alpha_{\mathrm{p}}\cdot\|\Delta\boldsymbol{p}\|_2$ 位置项与 $\alpha_{\mathrm{o}}\cdot\|\Delta\boldsymbol{o}\|_2$ 姿态项，它们能够引导无人机快速调整姿态，运动到期望的位置，是影响深度强化学习优化网络参数梯度方向的主要部分。$\alpha_{\mathrm{v}}\cdot\|\Delta\boldsymbol{v}\|_2$ 速度项和 $\alpha_{\omega}\cdot\|\Delta\boldsymbol{\omega}\|_2$ 角速度项的设计是为了减小无人机的动作和抖动，优化无人机的控制方式。坠毁惩罚项的设计是为了使无人机在训练开始阶段快速找到可行的控制方式，同时尽量避免出现安全问题。存活奖励项的设计是为了提升整个过程中的控制鲁棒性。

3 训练（Training）

3.1 深度强化学习算法

采用深度强化学习中的近端策略优化（PPO）算法进行神经网络的训练[23]。PPO 算法是目前主流的深度强化学习算法之一，通过在策略更新中引入限制机制，实现了稳定性与效率的平衡。它能够解决连续状态空间和动作空间中的策略优化问题，因易用性和良好的性能而得到广泛应用，也是 OpenAI 默认的强化学习算法。由于状态空间的维度高且无人机状态变动剧烈，本次任务对于 PPO 算法来说具有一定的挑战性，因此本文使用了随机初始化状态、并行训练这 2 个训练技巧，以便在训练过程中探索足够广的样本分布，并加快训练速度，以此来获得更好的神经网络控制效果。

3.2 随机初始化状态

为了使训练过程中探索的样本分布足够广，让训练得到的神经网络具有更好的鲁棒性，本文采

用随机初始化状态的方法。无人机目标悬停位置为 $[0,0,5]^{\mathrm{T}}$。在以目标位置为中心、边长为 2 m 的正方体空间内初始化无人机的位置；无人机初始姿态任意；无人机初始速度任意，单方向最大速度为 1 m·s^{-1}；无人机初始角速度任意，单方向最大角速度为 1 rad·s^{-1}。

3.3　并行训练

为了使训练过程中探索的样本足够多，加快训练速度、节省训练时间，本文采取并行训练的方法。在训练环境中同时开启 10 个线程。总共 100 架无人机并行进行采样，每条轨迹 τ 总仿真时长为 10.0 s，单步长为 0.02 s，共 500 步，无人机发生碰撞后将停止执行当前轨迹。每次迭代中系统共执行 50 000 步运算，设定总的迭代次数为 5 000。并行仿真训练效果如图 3 所示。

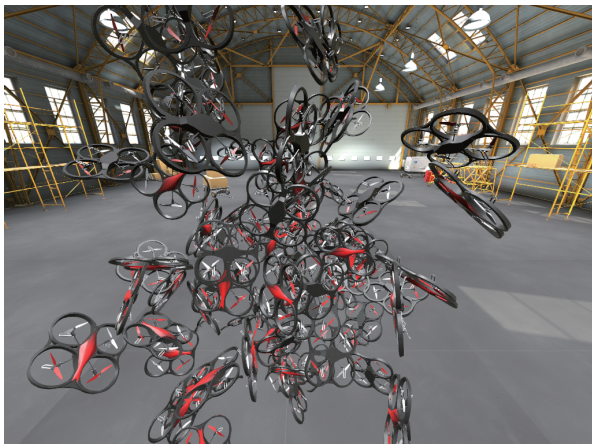

图 3　并行仿真训练效果图

Fig.3　Parallel simulation training renderings

3.4　训练环境及参数

在开源 3 维无人机仿真环境 Flightmare[24] 的基础上，加入了双向推力的电机－桨叶模型，同时进行了其他的一些修改和优化，得到新的仿真环境：FlightLxx。该仿真环境使用带输入的 4 阶 Runge-Kutta 方法对动力学模型中的方程进行数值积分，控制频率为 50 Hz。

FlightLxx 环境中的 PPO 算法实现是基于开源深度强化学习框架 Stable-Baselines 中的 PPO2，文中关于 PPO 算法所使用到的超参数如表 1 所示。

为了提升仿真实验结果的真实性与有效性，仿真模型参数值均来自旋翼原型机的实际测量值，如表 2 所示。

表 1　PPO 算法超参数

Tab.1　PPO hyperparameters

参数	数值
学习率	0.0003
折扣因子	0.99
优势估计权重	0.95
训练轮数	10.0
策略网络	MLP[64,64]
价值网络	MLP[64,64]
裁剪范围	0.2
熵系数	0.0
批量大小	1.0

表 2　四旋翼原型机参数

Tab.2　Parameters of quadrotor prototype

参数	数值
M	0.78 kg
L	0.125 m
I_x	2.3×10^{-3} N·m·s^{-2}
I_y	2.3×10^{-3} N·m·s^{-2}
I_z	3.6×10^{-3} N·m·s^{-2}
K_τ	0.01 m
K_f	1.5854 N·(r/min)$^{-2}$
K_α	0.033 s

本文用到的奖励函数系数值在表 3 中列出。

表 3　奖励函数系数

Tab.3　Coefficients of reward function

参数	数值
α_p	−0.02
α_o	−0.02
α_v	−0.0002
α_ω	−0.0002
α_c	−10.0
α_a	0.1

3.5　动作空间

所提出的创新方法为基于深度强化学习的 BTC（双向推力控制）神经网络控制器。根据所提出的双向推力电机－桨叶模型，设定控制器的输出推重比范围为 $[-2g,2g]$。

与之作为对比的是目前主流的控制方式[7]，OPTC（纯正向推力控制）神经网络控制器。根据主流的单向推力电机－桨叶模型，设定控制器的输出推重比范围为 $[0,2g]$。

结合安全考虑，在动作空间中设定了控制量的最大范围，限制了 4 个电机推力的安全阈值，以确保无人机安全飞行。控制器输出范围为唯一变量，其余参数和方法均相同，均已在上文列出。

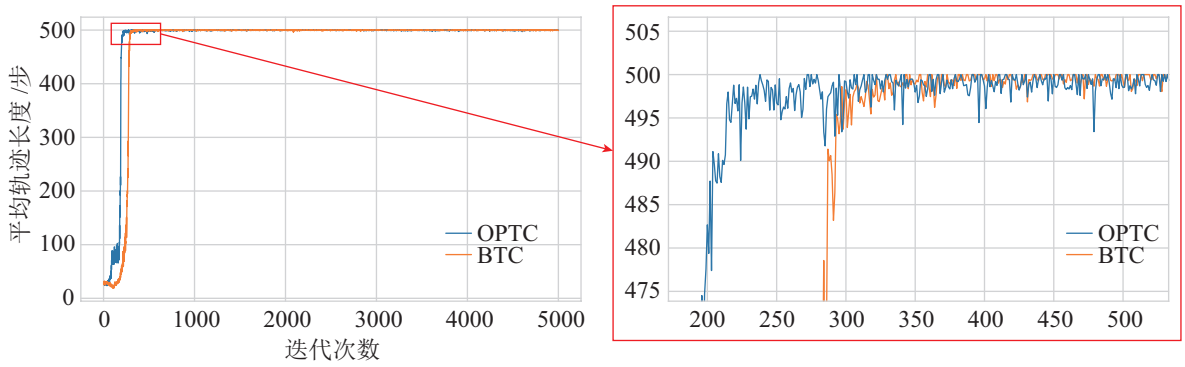

图 4 平均轨迹长度

Fig.4 Mean episode length

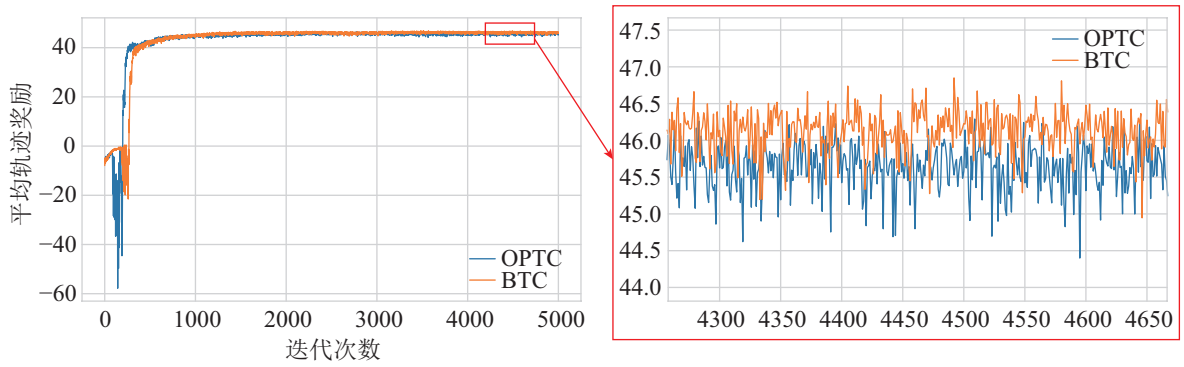

图 5 平均轨迹奖励

Fig.5 Mean episode reward

3.6 训练结果

训练过程平均轨迹长度见图 4，平均轨迹奖励见图 5。根据图 4，作为对比的 OPTC 方法学习训练收敛速度更快，这是因为其探索的动作空间较小，控制方式相对简单。但是，由图 5 可见，本文的 BTC 方法奖励值更高，表明其控制效果更好。

4 实验（Experiments）

4.1 实验设计

为了验证基于深度强化学习的 BTC 神经网络控制器的有效性和优越性，设计了剧烈运动状态下的旋翼无人机稳定悬停控制实验，选取 4 个典型的场景作为实验场景，分别为大姿态角、大速度、大角速度、以及混合三者的剧烈运动状态，并与 OPTC 神经网络控制器[7]进行对比。

仿真实验部署在桌面计算机上，其硬件配置为处理器 Intel i9-12900K、显卡 GeForce RTX 3090，操作系统为 Ubuntu 22.04。仿真实验中无人机的动力学模型参数如表 2 所示，仿真时长为 10.0 s，控制频率为 50 Hz。仿真实验的具体步骤为：1) 打开仿真环境，输入无人机的初始化状态；2) 初始化无人机，打开 BTC 神经网络控制器，对无人机进行剧烈运动状态下的稳定控制，记录实验过程数据；3) 初始化无人机，打开 OPTC 神经网络控制器，对无人机进行剧烈运动状态下的稳定控制，记录实验过程数据；4) 绘制数据对比图。

4.2 大姿态角悬停控制实验

选择具有代表性的大姿态角——180° 横滚角。初始化无人机的姿态为横滚 180°，其余状态量均与期望悬停状态相同。无人机初始化后神经网络控制器开始控制。无人机的运动过程如图 6 所示，图 6(a)～(c) 为无人机的位置，图 6(d)～(f) 为无人机的速度，图 6(g)～(i)为无人机的姿态欧拉角，图 6(j)～(l)为无人机的机体角速度，图 6(m)～(p)为神经网络控制器的输出动作。

由图中黄色线条可以看出，在 BTC 控制器的作用下无人机在 2 s 左右就能快速收敛稳定，而蓝色线条的 OPTC 控制器需要将近 4 s，前者的效果明显优于后者；由图 6(m)～(p)可以看出，BTC 控制器的输出动作更小，控制方式更优；由图 6(a)～(c)可以看出，在 BTC 控制器的作用下，无人机的状态波动更小。在无人机姿态调整的过程中，BTC 控制器能够大幅减小位置量的波动，这在实际飞行中能够增强无人机的鲁棒性和安全性。

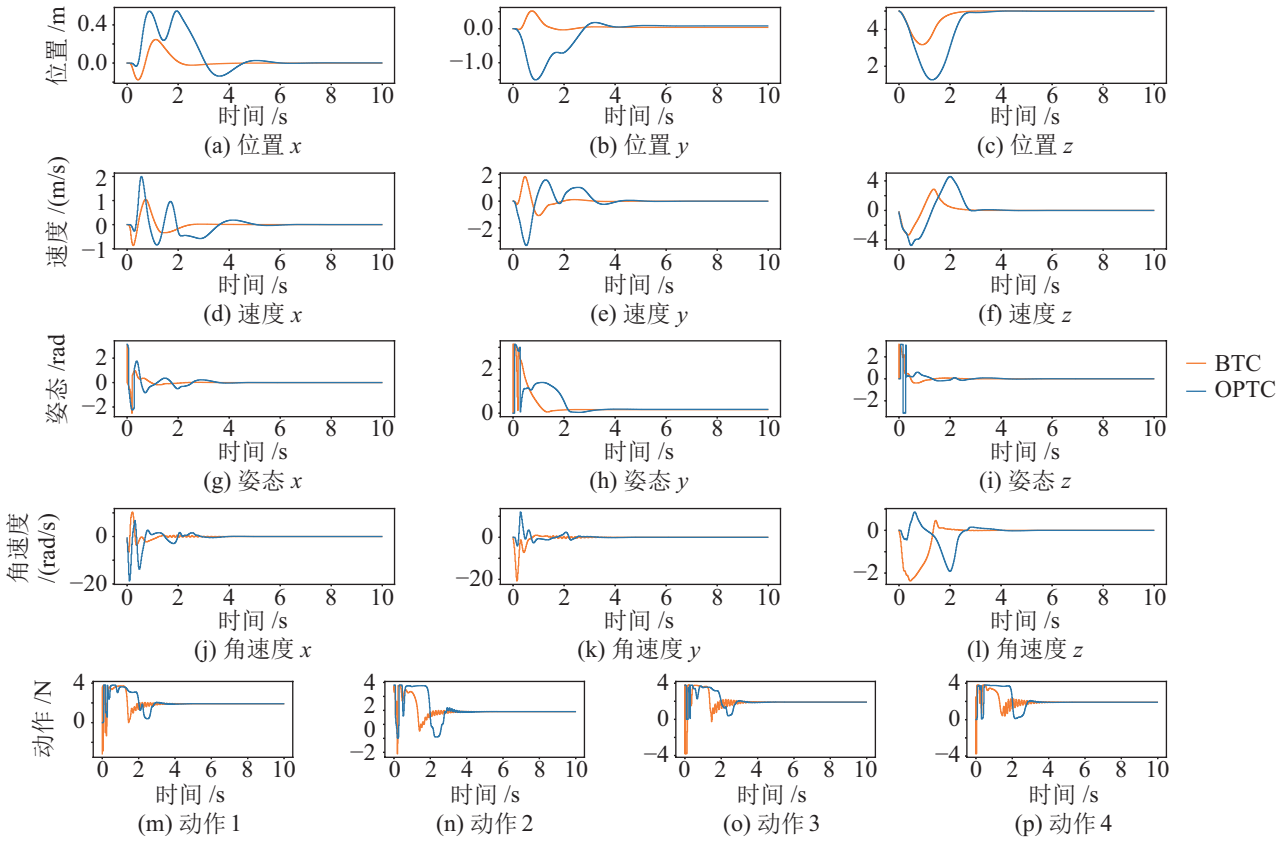

图 6　大姿态角悬停控制实验数据

Fig.6　Data of hover control experiment from large attitudes

4.3　大速度悬停控制实验

选择具有代表性的大速度——前向 5 m/s。初始化无人机速度为 [5,0,0] m/s，其余状态量均与期望悬停状态相同。无人机的运动过程如图 7 所示。

BTC 控制效果与 OPTC 接近，都能实现较好的控制，但由图 7(m)～(p)可以看到 BTC 控制器的输出动作更加平滑，抖动更少。

4.4　大角速度悬停控制实验

选择具有代表性的大角速度——机体角速度 3 个方向都为 3π rad·s^{-1}，其余状态量均与期望悬停状态相同。无人机的运动过程如图 8 所示。

蓝色线条终止于 1.5 s 左右，这表明此初始状态超过了 OPTC 控制器的能力范围，在此剧烈运动状态下控制器无法收敛，无人机坠毁。而黄色线条能够快速收敛稳定。由图 8(m)～(p)观察，BTC 控制器在初始阶段输出了大量反推力，动作空间的增大使其能够采取更多样的控制方式，拥有更好的机动性，从而在复杂情况下实现无人机的稳定，提升了其鲁棒性。

4.5　剧烈运动状态悬停控制实验

选择具有代表性的剧烈运动状态——大姿态角、大速度、大角速度三者混合。初始化无人机处于翻转姿态、前向速度为 5 m·s^{-1}、机体角速度 3 个方向都为 3π rad·s^{-1}，位置与期望悬停位置相同。

无人机的运动过程如图 9 所示，由蓝色线条可以看出此状态下 OPTC 控制器能力不够，无人机坠毁。而黄色线条展示的本文 BTC 方法能较好地控制无人机，使其在剧烈运动状态下稳定。

无人机的位姿变化如图 10 所示。无人机的状态及其变化都非常剧烈，说明本文的 BTC 控制器具有较强的鲁棒性。

从根本上而言，无人机动作空间的增大，能够使其机动性增强，控制策略也更加丰富，只要能够利用好这种优势，必将取得更好的控制效果。

5　结论（Conclusion）

面向旋翼无人机大姿态角、大速度、大角速度等剧烈运动状态下的稳定悬停问题，为了提高无人机的机动性、控制方法的鲁棒性和强实时性，首次提出基于深度强化学习的旋翼无人机双向推力控制方法。该方法使用基于深度强化学习的神经网络控制器，以无人机当前状态与无人机目标状态的差值作为控制器的输入，通过端到端的方式直接输出无人机底层的 4 个电机的期望推力，实现了无人机在

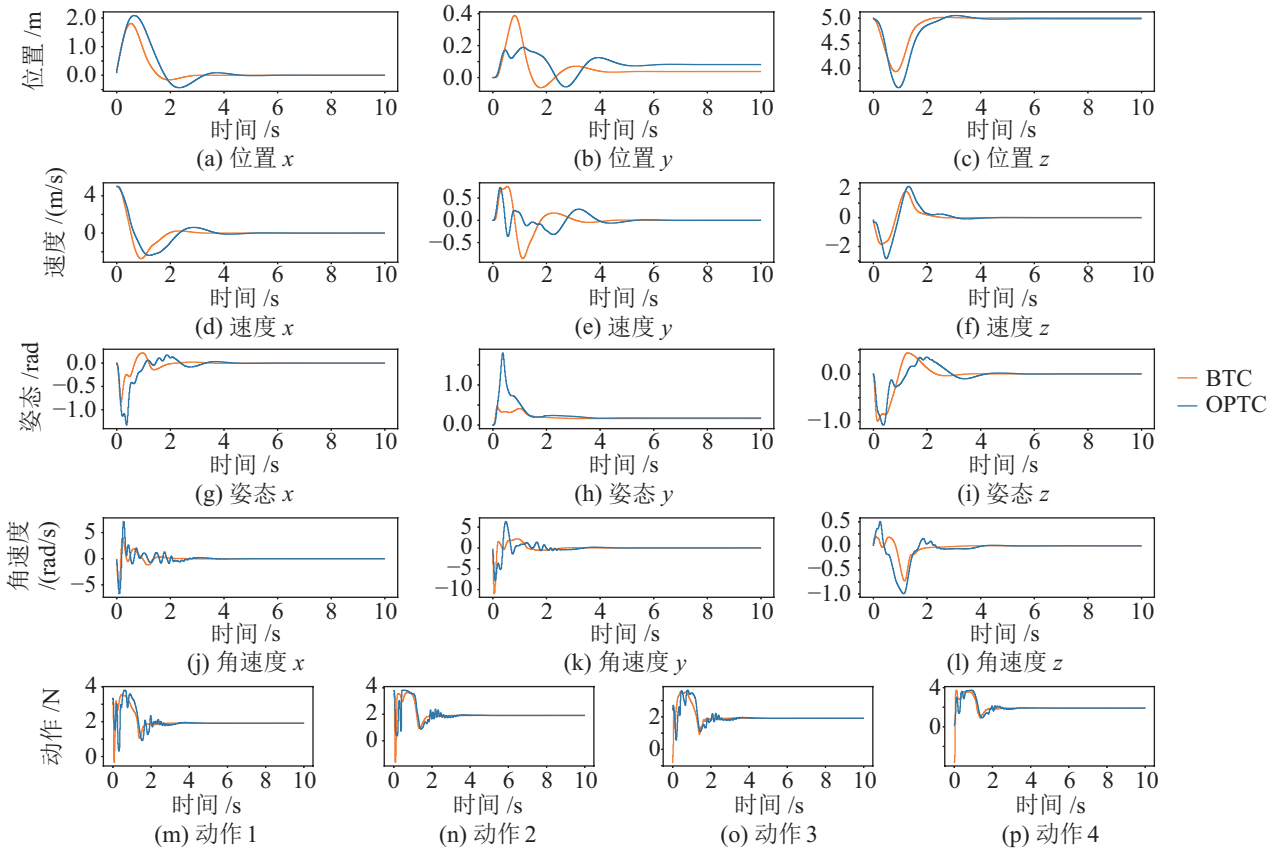

图 7 大速度悬停控制实验数据

Fig.7 Data of hover control experiment from high speeds

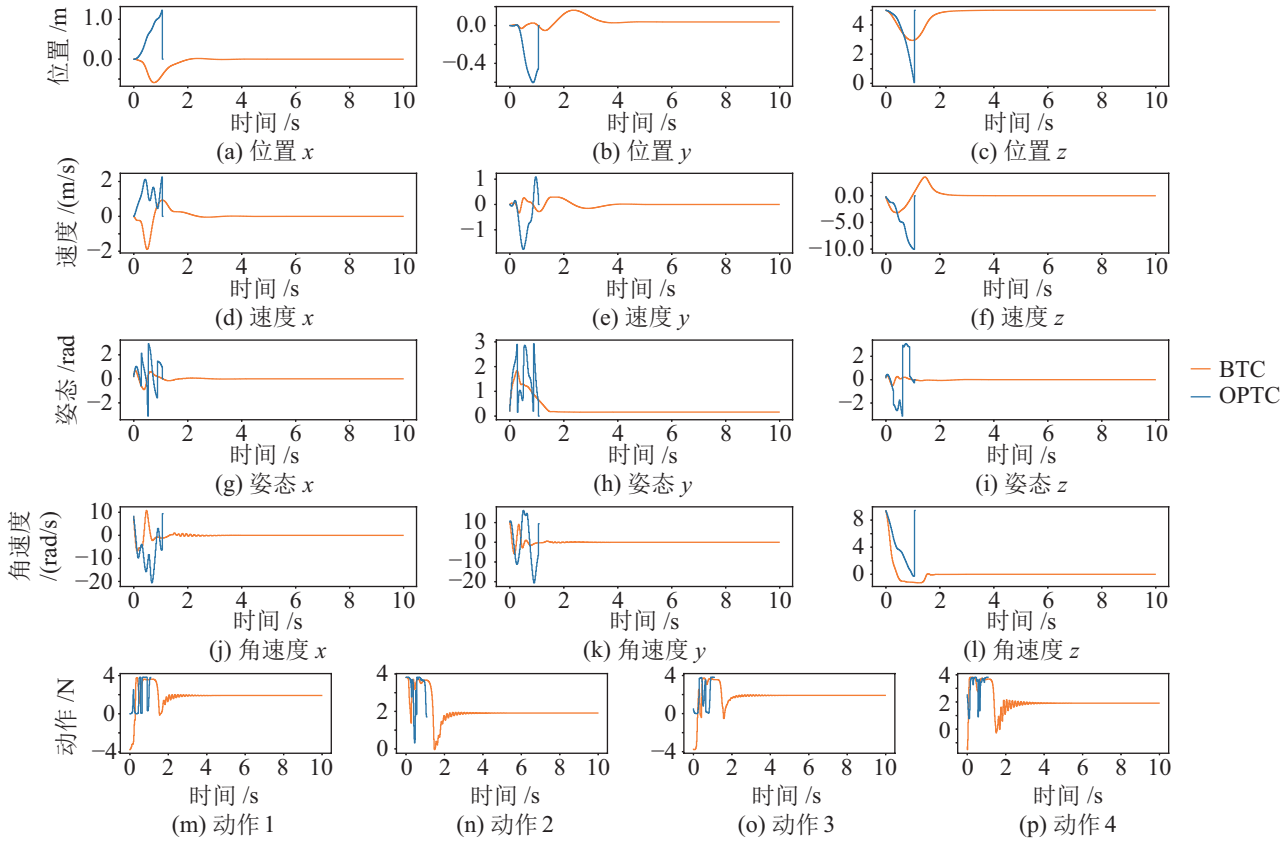

图 8 大角速度悬停控制实验数据

Fig.8 Data of hover control experiment from high angular velocities

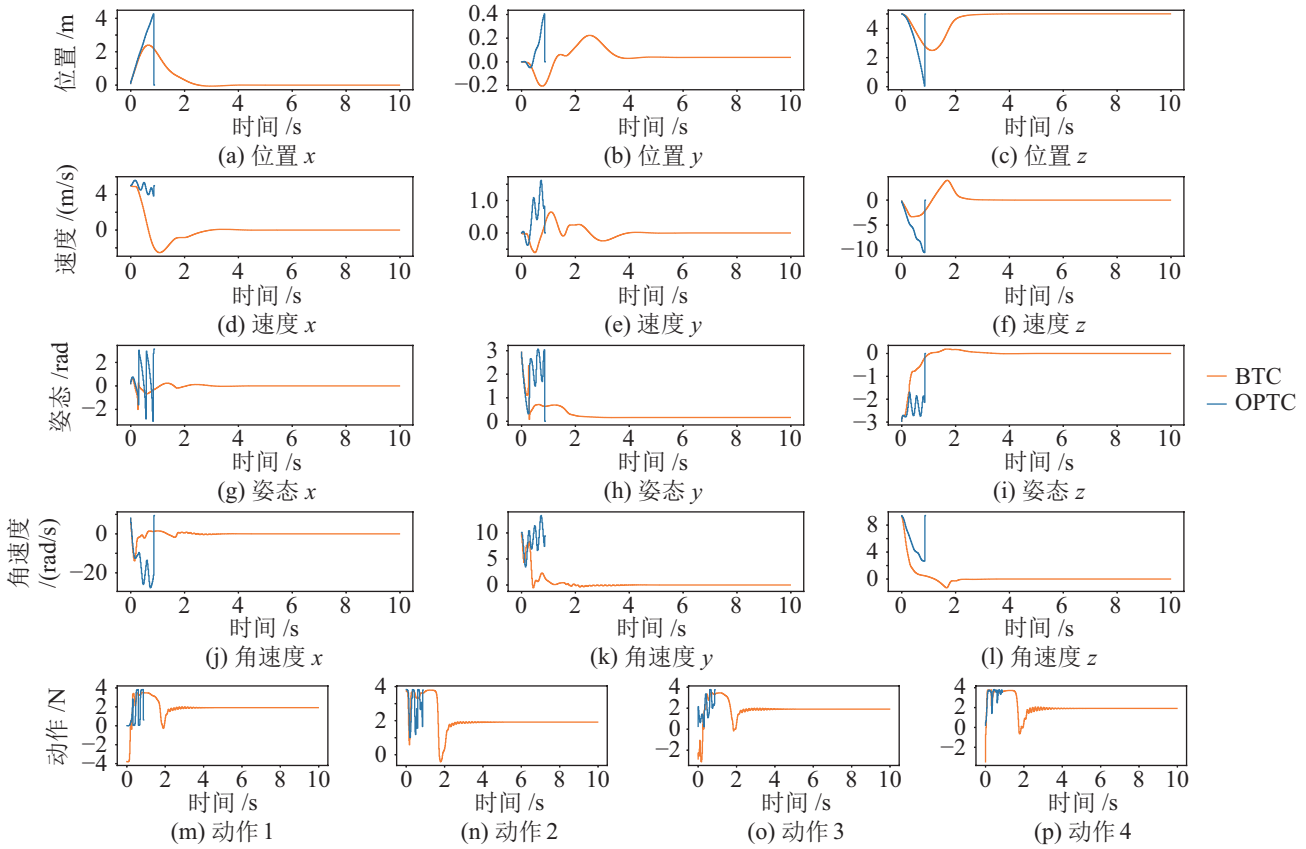

图 9　剧烈运动状态悬停控制实验数据

Fig.9　Data of hover control experiment from extreme conditions

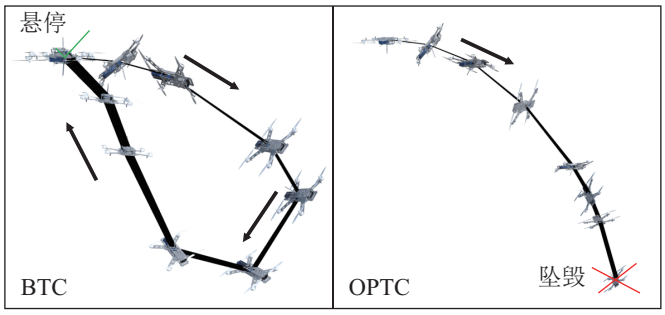

图 10　剧烈运动状态悬停控制实验中四旋翼无人机的位姿变化

Fig.10　Quadrotor position and attitude in hover control experiment from extreme conditions

剧烈运动状态下的快速悬停，能够有效扩展旋翼无人机的动作空间，增强其机动性和控制鲁棒性。仿真实验表明，在剧烈运动状态下，该方法具有比现有单向推力控制器更平滑的动作控制、更小的状态波动、更短的控制时间和更强的鲁棒性。本文所有代码均开源，且录制了相关实验视频。后续工作将主要围绕实物部署展开，研究如何缩小从仿真到实物实验的差距，包括硬件平台的搭建、实际参数的测量、无人机动态随机化、安全强化学习等等。

数据可用性声明

支撑本研究的科学数据已在中国科学院科学数据银行 ScienceDB 平台公开发布，访问地址为 https://www.doi.org/10.57760/sciencedb.j00003.00034。

参考文献（References）

[1] 张通．微型旋翼无人机自主飞行及应用[M]．5 版．北京：科学出版社，2024.
ZHANG T. Autonomous flight and application of micro quadrotors[M]. 5th ed. Beijing: Science Press, 2024.

[2] 陈谋，马浩翔，雍可南，等．无人机安全飞行控制综述[J]．机器人，2023，45(3)：345-366.
CHEN M, MA H X, YONG K N, et al. Safety flight control of UAV: A survey[J]. Robot, 2023, 45(3): 345-366.

[3] MEIER L, HONEGGER D, POLLEFEYS M. PX4: A node-based multithreaded open source robotics framework for deeply embedded platforms[C]//IEEE International Conference on

Robotics and Automation. Piscataway, USA: IEEE, 2015: 6235-6240.

[4] REMERO A, SUN S S, FOEHN P, et al. Model predictive contouring control for time-optimal quadrotor flight[J]. IEEE Transactions on Robotics, 2022, 38(6): 3340-3356.

[5] FAESSLER M, FALANGA D, SCARAMUZZA D. Thrust mixing, saturation, and body-rate control for accurate aggressive quadrotor flight[J]. IEEE Robotics and Automation Letters, 2017, 2(2): 476-482.

[6] 陈佳盼，郑敏华．基于深度强化学习的机器人操作行为研究综述[J]．机器人，2022，44(2)：236-256.
CHEN J P, ZHENG M H. A survey of robot manipulation behavior research based on deep reinforcement learning[J]. Robot, 2022, 44(2): 236-256.

[7] HWANGBO J, SA I, SIEGWART R, et al. Control of a quadrotor with reinforcement learning[J]. IEEE Robotics and Automation Letters, 2017, 2(4): 2096-2103.

[8] MOLCHANOV A, CHEN T, HONIG W, et al. Sim-to-(multi)-real: Transfer of low-level robust control policies to multiple quadrotors[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2019: 59-66.

[9] SONG Y L, STEINWEG M, KAUFMANN E, et al. Autonomous drone racing with deep reinforcement learning[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2021: 1205-1212.

[10] LOQUERCIO A, KAUFMANN E, RANFTL R, et al. Learning high-speed flight in the wild[J]. Science robotics, 2021, 6(59). DOI: 10.1126/scirobotics.abg5810.

[11] KAUFMANN E, BAUERSFELD L, SCARAMUZZA D. A benchmark comparison of learned control policies for agile quadrotor flight[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2022: 10504-10510.

[12] PENICKA R, SONG Y L, KAUFMANN E, et al. Learning minimum-time flight in cluttered environments[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7209-7216.

[13] SONG Y L, REMORO A, MULLER M, et al. Reaching the limit in autonomous racing: Optimal control versus reinforcement learning[J]. Science Robotics, 2023, 8(82). DOI: 10.1126/scirobotics.adg1462.

[14] KAUFMANN E, BAUERSFELD L, LOQUERCIO A, et al. Champion-level drone racing using deep reinforcement learning[J]. Nature, 2023, 620(7976): 982-987.

[15] MICHINI B, REDDING J, URE N K, et al. Design and flight testing of an autonomous variable-pitch quadrotor[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2011: 2978-2979.

[16] MAIER M. Bidirectional thrust for multirotor MAVs with fixed-pitch propellers[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018. DOI: 10.1109/IROS.2018.8593836.

[17] JOTHIRAJ W, MILES C, BULKA E, et al. Enabling bidirectional thrust for aggressive and inverted quadrotor flight[C]//International Conference on Unmanned Aircraft Systems. Piscataway, USA: IEEE, 2019: 534-541.

[18] JOTHIRAJ W, SHARF I, NAHON M. Control allocation of bidirectional thrust quadrotor subject to actuator constraints[C]//International Conference on Unmanned Aircraft Systems. Piscataway, USA: IEEE, 2020: 932-938.

[19] MAO K, WELDE J, HSIEH M A, et al. Trajectory planning for the bidirectional quadrotor as a differentially flat hybrid system [C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2023: 1242-1248.

[20] ZHANG W, SONG K, RONG X W, et al. Coarse-to-fine UAV target tracking with deep reinforcement learning[J]. IEEE Transactions on Automation Science and Engineering, 2019, 16(4): 1522-1530.

[21] MA C, CAO Y Z, DONG D B. Reinforcement learning based time-varying formation control for quadrotor unmanned aerial vehicles system with input saturation[J]. Applied Intelligence, 2023, 53(23): 28730-28744.

[22] WEN G X, HAO W, FENG W W, et al. Optimized backstepping tracking control using reinforcement learning for quadrotor unmanned aerial vehicle system[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 52(8): 5004-5015.

[23] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[DB/OL]. (2017-07-20) [2024-12-01]. https://arxiv.org/abs/1707.06347.

[24] SONG Y L, NAJI S, KAUFMANN E, et al. Flightmare: A flexible quadrotor simulator[C]//Proceedings of the 2020 Conference on Robot Learning. PMLR, 2021: 1147-1157.

作者简介：

李晓信（1997 –），男，硕士生。研究领域：无人机控制，深度强化学习。

刘志宏（1986 –），男，博士，副教授。研究领域：无人机集群，无人机控制，深度强化学习。