基于KPPO算法的四旋翼无人机飞行控制

武子怡[1]，高晓楠[1]，朱献超[1]，赵亮[2]，祝峰[3]，蔡磊[4]
（1. 河南工业大学人工智能与大数据学院，河南 郑州 450001；
2. 河南工业大学电气工程学院，河南 郑州 450001；
3. 电子科技大学基础与前沿研究院，四川 成都 610050；
4. 河南科技学院人工智能学院，河南 新乡 453003）

摘　要：针对近端策略优化（PPO）算法在四旋翼无人机飞行控制中存在的收敛速度慢和难以适应复杂环境问题，在PPO算法的基础上，融合了阈值触发的KL散度惩罚项与L2正则项构成的复合正则机制，提出了一种改进算法KPPO算法，并将其应用于四旋翼无人机飞行控制任务。通过四旋翼无人机飞行控制物理仿真验证，结果显示，在KPPO算法策略的驱动下，四旋翼无人机能够迅速达到策略收敛状态，并在复杂环境中做出正确的决策，从而显著提升了传统算法的性能及任务执行效率。研究结果证实了KPPO算法在优化四旋翼无人机飞行控制中的有效性。

关键词：近端策略优化算法；四旋翼无人机飞行控制；深度强化学习
中图分类号：TP13
文献标志码：A
doi: 10.11959/j.issn.2096-6652.202531

Quadrotor UAV flight control based on the KPPO algorithm

WU Ziyi[1], GAO Xiaonan[1], ZHU Xianchao[1], ZHAO Liang[2], ZHU Feng[3], CAI Lei[4]
1. School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450001, China
2. College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China
3. Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610050, China
4. School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang 453003, China

Abstract: To address the issues of slow convergence and the difficulty of adapting to complex environments in proximal policy optimization (PPO) algorithms for quadrotor UAV flight control, an improved algorithm was proposed, KPPO, which integrated a composite regularization mechanism composed of a threshold-triggered KL divergence penalty and an L2 regularization term based on the PPO framework. The proposed KPPO algorithm was applied to quadrotor UAV flight control tasks. Physical simulation validation of quadrotor UAV flight control demonstrates that under the guidance of the KPPO strategy, the quadrotor UAV rapidly achieves policy convergence and makes correct decisions in complex environments, thereby significantly enhancing the performance of conventional algorithms. Notably, the KPPO algorithm improves task execution efficiency, a key factor in quadrotor UAV operations. This reassures the audience of the effectiveness of the KPPO algorithm in improving quadrotor UAV flight control.

收稿日期：2025-03-21；**修回日期：**2025-06-13
通信作者：朱献超，xczhuiffs@163.com
基金项目：河南省自然科学基金青年项目（No.252300421806）；河南工业大学高层次人才引进项目（No.2022BS073）；河南省高校科技创新团队项目（No.25IRTSTHN018）；中原科技创新领军人才计划（No.254000510043）；河南省重点研发专项（No.241111110200）
Foundation Items: Natural Science Foundation of Henan (No.252300421806), Research Foundation for Advanced Talents of Henan University of Technology (No.2022BS073), Scientific and Technological Innovation Teams of Universities in Henan Province (No. 25IRTSTHN018), Zhongyuan Science and Technology Innovation Leadership Talent Program (No.254000510043), Henan Provincial Key R&D Special Project (No.241111110200)

Key words: proximal policy optimization, quadrotor UAV flight control, deep reinforcement learning

0 引言

随着智能无人系统技术的快速发展，机器人被广泛应用于工业检测[1]、应急救援[2]及军事侦察[3]等各类高难度场景，承担着高风险或复杂环境下的作业任务。在航空领域，具有自主目标搜索能力的无人机（unmanned aerial vehicle，UAV）[4-6]凭借其在航空遥感[7]、灾害检测[8]等领域的突出表现，近年来取得显著的技术突破。然而，面对复杂多变、充满不确定性的外部环境，现有飞行控制系统在环境适应性方面仍存在明显局限，难以应对突发扰动和环境的动态变化，这一瓶颈已成为制约无人机技术进一步发展的核心问题。在此背景下，如何优化控制算法架构，以提升系统的鲁棒性和容错能力，构建具有强大环境适应性的自主控制系统，已成为该领域亟待解决的关键技术难题。相关研究成果不仅对提升无人机在复杂环境下的自主决策能力具有重要的理论价值，同时在智能无人系统的深度应用方面也具有广泛的工程实践意义。

在无人机飞行控制技术发展的初期阶段，以PID（proportion integral differential）控制算法[9]为代表的非模型依赖型控制策略被广泛应用于姿态稳定控制领域。尽管该算法通过误差反馈机制能够实现较高的稳态精度，但其控制性能对增益参数具有显著的敏感性。现有的参数整定过程高度依赖工程经验的积累，需通过大量测试来实现参数优化。这种基于试错法的调试机制不仅效率低下，还导致系统鲁棒性不足和泛化能力受限。值得注意的是，近年来研究者们陆续提出了多种参数自适应算法。例如，Leung等[10]通过构建网格化参数调度框架，实现了基于局部线性化模型的非线性系统控制；Papadopoulos团队[11]则基于幅度最优理论，开发了PID控制器的频域自整定方法。然而，上述方法均需以动力学模型已知为设计前提，而在实际工程场景中，这一条往往难以满足，从而限制了其广泛应用。

在经典控制方法中，以线性二次型调节器（linear quadratic regulator，LQR）[12]为代表的基于模型的优化控制策略，通过构建系统动力学模型预测未来时域内的最优控制序列。尽管该方法存在局部收敛的风险，但在平衡点邻域内表现出良好的控制效果。然而，LQR算法严格依赖于线性化模型的精确性，且求解所得的控制律可能超出执行机构的动态约束边界，导致实际控制输入不可行。为此，Falanga等学者[13]引入了模型预测控制（model predictive control，MPC）框架。该框架通过在线更新状态观测数据并重构优化目标函数，能够动态补偿系统未建模动态及外部扰动，从而在模型失配与工况突变条件下仍保持鲁棒的跟踪性能。但需强调的是，无论是LQR还是MPC，其控制品质都依赖于高精度建模作为基础假设，才能有效减小控制误差。

在智能控制技术革新的背景下，强化学习（reinforcement learning，RL）作为无模型优化的核心范式，正逐步突破传统控制框架的局限。其通过智能体与环境的自主探索机制，结合深度神经网络的高维特征表征能力，在机器人领域得到了广泛应用[14-20]。研究表明，基于深度强化学习（deep reinforcement learning，DRL）的先进算法在飞行器姿态控制领域展现出显著优势。Koch团队[21]系统对比了深度确定性策略梯度（deep deterministic policy gradient，DDPG）、信赖域策略优化（trust region policy optimization，TRPO）及近端策略优化（proximal policy optimization，PPO）等算法的控制效能，实验数据证实其在响应速度与抗扰能力方面均超越了传统的PID控制算法。Hwangbo等[22]则通过构建从状态观测至电机推力的端到端映射模型，采用确定性策略优化实现了无人机底层动力系统的直接控制。值得关注的是，学者正着力优化算法收敛特性与鲁棒性。刘小雄等[23]基于深度Q网络（deep Q-network，DQN）实现了无人机多自由度位置控制器的参数优化；Haarnoja团队[24]引入最大熵强化学习理论，通过软动作−评价（soft actor-critic，SAC）算法显著提升了策略搜索效率，但该方法在高速运动场景中仍存在轨迹跟踪稳态误差增大的问题。尽管DDPG在动态目标持续跟踪任务中表现出色[25]，但其对突发障碍物及空域限制等复杂约束的实时响应机制尚未完善。对此，胡多修等[26]开发的自主引导−避障协同控制架构通过多模态感知融合与动态路径规划，有效提升了侦察任务全流程的自主性与安全性。

学者针对不同的应用场景提出了多样化的RL解决方案。梁吉团队[27]创新性地构建了基于端到端状态映射的自主控制架构，通过DRL直接建立了无人机全状态空间至执行机构输出的非线性关系，实现了不需要动力学先验知识的通用型控制策略。梁晨等[28]设计了基于DQN的决策框架，通过贪婪策略平衡探索与利用过程，有效规避了传统方法对模型参数标定与系统辨识的强依赖性。在复杂环境适应方面，黄号等[29]提出了改进型贪婪深度确定性策略梯度（greedy-DDPG）算法，通过在线贪婪动作筛选机制构建长机导航模型，在动态障碍物分布的稀疏奖励场景下显著提升了避障行为的泛化能力。孔飞等[30]开发了基于Actor-Critic架构的鲁棒强化学习算法：利用Critic网络构建策略性能函数，通过Actor网络在线求解满足输入受限条件的最优控制律，并结合干扰观测器技术增强了系统对时变扰动与局部故障的容错性。此外，江泰民团队[31]在PPO框架中引入动态误差补偿机制，提出了微分补偿强化学习（PPO-DC）算法，在固定翼无人机高精度轨迹跟踪任务中实现了快速收敛与稳态误差抑制的双重优化。这些研究不仅拓展了RL在无人机控制领域的应用范围，还为解决复杂环境下的控制问题提出了新的思路和方法。

本文针对DRL中的PPO算法进行改进，提出了融合阈值触发的KL散度惩罚项与L2正则项构成的复合正则机制的KPPO算法。该算法的核心创新在于引入阈值KL散度作为自适应正则化项，用于动态调整策略更新的信任区域，从而有效抑制策略迭代过程中可能出现的过度偏离现象，提升策略优化的稳定性。同时，KPPO将L2正则项与KL散度协同构成复合正则机制，以进一步增强策略网络参数的鲁棒性，从而显著提升了四旋翼无人机飞行控制系统在复杂环境下的稳态控制性能与策略泛化能力。仿真结果表明，在四旋翼无人机飞行控制的5项物理仿真任务中，KPPO算法相较于基准PPO算法和DDPG算法在收敛速度和任务完成度方面均表现出显著的性能优势。具体而言，KPPO算法不仅能够更快地收敛到局部最优策略，而且在复杂环境下展现出更强的适应能力，显著提升了任务完成率。这些实验结果充分验证了本文提出方法的有效性，为无人机飞行控制系统的RL优化提供了新的思路和技术支持。

1　背景知识

1.1　四旋翼无人机

1.1.1　飞行原理

四旋翼无人机通过调节4个旋翼的转速来实现对飞行姿态和位置的控制。其飞行原理基于牛顿-欧拉方程，通过控制旋翼产生的升力和力矩来实现姿态调整和位置移动。

（1）升力与姿态控制

每个旋翼通过旋转产生向上的升力$F_i=C_T\cdot\omega_i^2$，其中C_T是升力系数，ω_i是旋翼角速度。4个旋翼的总升力为：

$$F=F_1+F_2+F_3+F_4 \tag{1}$$

为了实现无人机的悬停状态，其总升力需与重力相平衡，即4个旋翼提供的合力需恰好抵消重力加速度产生的向下力。无人机的姿态控制主要包括3种形式（俯仰控制、横滚控制、偏航控制），其调节均依赖于各旋翼转速之间的精确差异。

- 俯仰控制：通过改变前后旋翼的转速差实现。增加前旋翼转速，减小后旋翼转速，使无人机向前俯仰。
- 横滚控制：通过改变左右旋翼的转速差实现。增加左旋翼转速，减小右旋翼转速，使无人机向右横滚。
- 偏航控制：通过改变对角旋翼的转速差实现。增加顺时针旋转旋翼的转速，减小逆时针旋转旋翼的转速，使无人机顺时针偏航。

（2）动力学与运动学模型

描述无人机在受到外力和外力矩时的速度和角速度变化的模型为动力学模型。位置动力学方程为：

$$\begin{cases} a_x=-\dfrac{F}{m}\left(\cos\psi\sin\theta\cos\phi+\sin\psi\sin\phi\right) \\ a_y=-\dfrac{F}{m}\left(\sin\psi\sin\theta\cos\phi-\cos\psi\sin\phi\right) \\ a_z=g-\dfrac{F}{m}\cos\phi\cos\theta \end{cases} \tag{2}$$

其中，ϕ、θ和ψ分别为滚转角、俯仰角和偏航角，a_x、a_y和a_z分别是四旋翼无人机在x、y、z方向上的加速度。

运动学模型是用于描述四旋翼无人机的位置和姿态变化。当无人机的姿态角较小时，扰动模

型为：

$$\begin{cases}\phi=\dfrac{1}{I_{xx}}\left(\tau_x+qr\left(I_{yy}-I_{zz}\right)-J_1 q\Omega\right)\\ \theta=\dfrac{1}{I_{yy}}\left(\tau_y+pr\left(I_{zz}-I_{xx}\right)+J_1 p\Omega\right)\\ \psi=\dfrac{1}{I_{zz}}\left(\tau_z+qp\left(I_{xx}-I_{yy}\right)\right)\end{cases}\tag{3}$$

其中，I_{xx}、I_{yy}、I_{zz}分别是绕x、y、z轴的转动惯量，τ_x、τ_y、τ_z分别是绕x、y、z轴的外力矩，p、q、r分别是绕机体坐标系的x、y、z轴的角速度，J_1是旋翼的惯性矩，Ω是旋翼的角速度。

1.1.2 四旋翼无人机的控制系统结构

四旋翼无人机的控制器是由传感器、控制单元（control unit）和执行机构组成的。

传感器包含惯性测量单元（IMU）、气压计（barometer）和GPS模块。IMU用于测量无人机的加速度和角速度，通常包括三轴加速度计和三轴陀螺仪，能够实时提供无人机的姿态和运动信息；气压计用于测量高度，通过检测大气压力的变化来估算无人机的相对高度；GPS模块用于提供地理位置信息，为无人机提供精确的经纬度和高度信息。

控制单元中的主控制器用于处理传感数据、运行控制算法和输出控制指令。在本实验中，主控制器通过KPPO算法控制四旋翼无人机的俯仰角、滚转角和偏航角，调整飞行姿态和速度，并且根据环境变化实时调整飞行高度和路径。KPPO算法通过与环境的交互，不断优化控制策略，使无人机在复杂环境中自主学习和适应。相比于依赖精确的系统模型和线性假设的传统控制方法，KPPO算法具有更强的适应性和灵活性，通过调整奖励函数来适应不同的任务需求。

执行机构由电机（Motor）和电机驱动器构成。电机驱动器通过PWM信号调节电机的转速，从而实现对旋翼转速的控制。电机是通过改变转速实现姿态调整。电机转速的控制计算式为：

$$\mathrm{Motor}_{\mathrm{right1}}=\mathrm{Thrust}_{\mathrm{cmd}}+\mathrm{Yaw}_{\mathrm{cmd}}+\mathrm{Pitch}_{\mathrm{cmd}}+\mathrm{Roll}_{\mathrm{cmd}}\tag{4}$$

$$\mathrm{Motor}_{\mathrm{left1}}=\mathrm{Thrust}_{\mathrm{cmd}}-\mathrm{Yaw}_{\mathrm{cmd}}+\mathrm{Pitch}_{\mathrm{cmd}}-\mathrm{Roll}_{\mathrm{cmd}}\tag{5}$$

$$\mathrm{Motor}_{\mathrm{right2}}=\mathrm{Thrust}_{\mathrm{cmd}}-\mathrm{Yaw}_{\mathrm{cmd}}-\mathrm{Pitch}_{\mathrm{cmd}}+\mathrm{Roll}_{\mathrm{cmd}}\tag{6}$$

$$\mathrm{Motor}_{\mathrm{left2}}=\mathrm{Thrust}_{\mathrm{cmd}}+\mathrm{Yaw}_{\mathrm{cmd}}-\mathrm{Pitch}_{\mathrm{cmd}}-\mathrm{Roll}_{\mathrm{cmd}}\tag{7}$$

其中，$\mathrm{Thrust}_{\mathrm{cmd}}$是总推力指令，用于控制无人机的垂直运动；$\mathrm{Yaw}_{\mathrm{cmd}}$是偏航力矩指令，用于控制无人机的偏航运动；$\mathrm{Pitch}_{\mathrm{cmd}}$是俯仰力矩指令，用于控制无人机的俯仰运动；$\mathrm{Roll}_{\mathrm{cmd}}$是横滚力矩指令，用于控制无人机的横滚运动。无人机结构如图1所示。

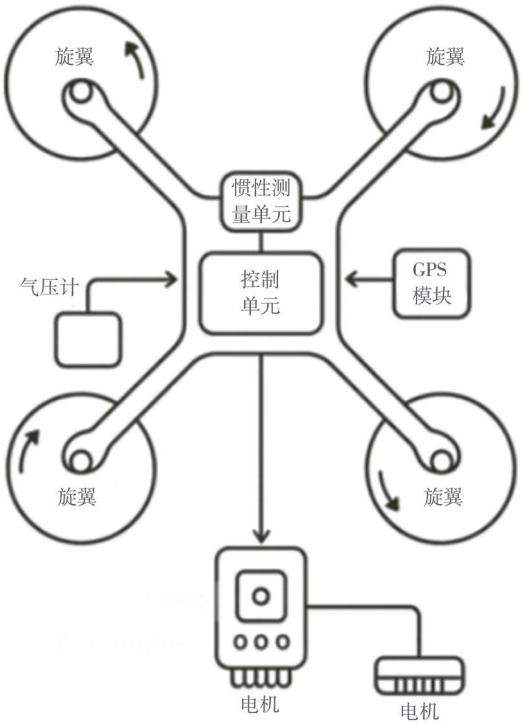


图1 四旋翼无人机结构

1.2 深度强化学习

深度学习由于其卓越的抽象与表示能力，常被用作RL中构建值函数和策略函数的重要工具，极大地拓宽了RL的应用领域。根据策略优化的方式不同，DRL可以分为基于价值的算法和基于策略的算法。

而本文主要讨论基于策略的算法中的PPO和DDPG。PPO有一个重要变种，即近端策略优化惩罚（PPO-penalty，PPO1）。与传统PPO方法不同的是，标准PPO在每次采样数据后仅进行一次参数更新，然后重新采样，PPO1使用θ^k与环境交互，采样这组数据之后可使θ更新很多次，用于最大化目标函数。因为引入了重要性采样机制，所以θ的更新次数不会影响正常训练。

除此之外，还有一个重要思想是自适应KL惩罚机制。具体而言，如果KL散度值过大，表明惩罚的项$\beta\mathrm{KL}\left(\theta,\theta^k\right)$没有发挥作用，此时需要$\beta$增大，而当KL散度值太小时，则意味着惩罚项过于强势，可能导致优化过程中偏重该项而忽略其他目标函数，使得θ和θ^k一样，因此需要适当减小β。基

于此，为实现各部分目标函数间的平衡，必须对KL惩罚系数β进行动态调整，下面为总结的自适应KL惩罚：$\mathrm{KL}\left(\theta,\theta^k\right)>\mathrm{KL}_{\max}$，增大$\beta$；$\mathrm{KL}\left(\theta,\theta^k\right)<\mathrm{KL}_{\min}$，减小$\beta$。

近端策略优化惩罚可表示为：

$$J^{\theta^k}(\theta)\approx\sum_{(s_t,a_t)}\frac{p_\theta\left(a_t|s_t\right)}{p_{\theta^k}\left(a_t|s_t\right)}A^{\theta^k}\left(s_t,a_t\right) \tag{8}$$

$$J^{\theta^k}{}_{\mathrm{PPO}}=\mathrm{J}^{\theta^k}(\theta)-\beta\mathrm{KL}\left(\theta,\theta^k\right) \tag{9}$$

与PPO算法不同，DDPG使用的是确定性策略梯度框架，由确定性策略网络（Actor）和一个价值网络（Critic）及其对应的目标网络构成。给定当前状态s，Actor的$\mu_\theta(s)$输出确定性动作a，并在训练时加入高斯噪声以增强探索：

$$a_{\mathrm{explore}}=\mu_\theta(s)+N\left(0,\sigma^2\right) \tag{10}$$

$$a=\mu_\theta(s) \tag{11}$$

Critic用于评估状态-动作对的价值。Critic的目标是最小化时序差分误差：

$$L_{\mathrm{Critic}}(\phi)=E_{(s,a,r,s')\sim D}\left[\left(r+\gamma Q_{\bar{\phi}}\left(s',\mu_{\bar{\theta}}(s')\right)-Q_\phi(s,a)\right)^2\right] \tag{12}$$

其中，D为经验回放池，γ为折扣因子，$\bar{\theta}$和$\bar{\phi}$分别为Actor和Critic的目标网络参数。状态s'是执行动作a后到达的新状态，r是执行a后环境给出的即时奖励。Critic的更新目标是最小化当前Q网络（参数为ϕ）与TD目标之间的均方差。TD目标使用目标网络$Q_{\bar{\phi}}$和目标策略$\mu_{\bar{\theta}}$生成。Critic的参数ϕ通过梯度下降得到。Actor的优化目标是最大化Critic对其输出动作的Q值评估，等价于最小化损失函数：

$$L_{\mathrm{Actor}}(\theta)=-E_{s\sim D}\left[Q_\phi\left(s,\mu_\theta(s)\right)\right] \tag{13}$$

为了提高训练稳定性，Actor和Critic均使用对应的目标网络$\left(\mu_{\bar{\theta}},Q_{\bar{\phi}}\right)$生成训练目标，并通过软更新策略平滑地跟踪主网络。其中τ为软更新系数。

$$\bar{\theta}\leftarrow\tau\theta+(1-\tau)\bar{\theta} \tag{14}$$

$$\bar{\phi}\leftarrow\tau\phi+(1-\tau)\bar{\phi} \tag{15}$$

2　KPPO算法

本文提出的KPPO算法在设计时借鉴了PPO1中关于KL散度惩罚的思想，但并未直接沿用，而是在此基础上进行了创新，引入了一种新的策略更新机制。具体而言，KPPO采用了一种门控惩罚策略，即仅在新旧策略之间的KL散度超过0.01时才施加惩罚，从而既能有效限制策略更新的幅度，又能保留必要的探索空间。同时，算法没有沿用传统PPO中的剪切方法，而是通过直接使用最小值约束来进行策略更新，以进一步提高策略迭代的稳定性。此外，为了防止参数剧烈波动和模型过拟合，KPPO还在损失函数中引入了L2正则化。最后，将该算法应用于四旋翼无人机飞行控制任务，以提高四旋翼无人机在复杂环境下的决策能力和执行稳定性。

在上述总体设计框架的基础上，本文进一步探讨KPPO算法的具体实现细节。KPPO算法通过计算KL散度来衡量新旧策略之间的变化程度，并在策略更新幅度过大时施加额外惩罚，以避免策略发生剧烈偏移，导致训练不稳定。其中D_{KL}是新旧策略的KL散度，用于衡量策略变化的幅度。∂是超参数，决定惩罚项的强度，δ是KL散度阈值，只有当D_{KL}>0.01才会受到惩罚。本文KL散度阈值被人为设定为0.01。过低的阈值会使策略更新过于保守，进而延缓收敛；过高的阈值又可能导致策略参数发生剧烈变动，从而引发训练不稳定。文献中普遍建议将该阈值设置0.01到0.03。为保证不同仿真任务间的可比性，本文所有仿真验证中均采用相同的超参数配置，未针对具体场景对该阈值进行额外调整。

$$D_{\mathrm{KL}}=E\left[\log\pi_{\theta_{\mathrm{old}}}\left(a_t|s_t\right)-\log\pi_\theta\left(a_t|s_t\right)\right] \tag{16}$$

$$L_{\mathrm{KL}}=\partial\cdot\max\left(D_{\mathrm{KL}}-\delta,0\right) \tag{17}$$

对策略网络中所有参数进行平方计算以生成L2正则化项，其主要目的是通过逐步衰减参数权重，有效抑制模型过拟合现象，并在确保训练过程中参数更新平稳的同时，提升模型在复杂任务中对未知数据的泛化能力。

$$L_{\mathrm{L2}}=\sum_i\left\|\theta_i\right\|^2 \tag{18}$$

策略损失函数如下所示：

$$L_{\mathrm{policy}}=E\left[-\min\left(r_tA_t,A_t\right)\right] \tag{19}$$

其中，r_t代表新旧策略的概率比，反映了新策略相对于旧策略在采取相同动作时的倾向性变化；A_t为优势函数，用于衡量当前执行的动作a_t相较于平均策略的优劣；其中R_t表示从某个时间步t开始累积

至未来的折扣奖励；$V(s_t)$表示从状态s_t开始，执行π_θ策略后的预期长期收益。如果A_t>0，说明实际回报高于状态的平均预期，此时策略应该多选择该动作；反之，则应减少选择该动作。为了防止策略更新幅度超过预期，新策略对优势函数的加权结果与直接基于旧策略计算的结果进行比较，并取其较小值，以此来严格控制更新范围。L_{value}表示状态值函数损失，用于优化值函数$V(s_t)$的估计，使其更加接近真实回报R_t，确保策略的稳定学习。

$$A_t = R_t - V(s_t) \tag{20}$$

$$r_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \tag{21}$$

$$L_{\text{value}} = E\left[\left(V(s_t) - R_t\right)^2\right] \tag{22}$$

本文设计的总损失函数综合了策略损失、KL散度惩罚、价值损失以及L2正则化项，旨在通过多方面的优化提升智能体的性能。其中L2正则化的权重系数β用于调节正则化的强度，防止模型过拟合并提升其泛化能力。该损失函数的核心作用在于通过梯度下降法不断更新策略网络的参数，从而实现对策略的优化，使智能体在与环境进行交互时表现得更加高效和稳健。

$$L = L_{\text{policy}} + \beta L_{\text{L2}} + L_{\text{KL}} + L_{\text{value}} \tag{23}$$

本文将KPPO算法用于解决四旋翼无人机飞行控制问题，并在多个物理仿真任务中充分验证了其有效性。在仿真环境参数的设置中，飞行器的关键状态参数，包括位置、姿态角（横滚、俯仰、偏航）、线速度和角速度等，会被实时采集并输入算法中，作为生成控制决策的依据。这些状态参数为KPPO算法提供了四旋翼无人机当前飞行状态的全面描述，使其能够根据环境动态调整策略，提高飞行的稳定性和适应性。在动作空间中，KPPO算法通过输出4个连续值作为无人机4个旋翼的控制指令，实现对飞行状态的精准调控。这种连续动作空间的设计使算法能够精细地调节无人机的飞行姿态和运动轨迹，从而适应复杂的飞行任务需求。通过KPPO算法的优化，无人机能够在动态环境中实现稳定的飞行控制，同时表现出高效的任务执行能力和鲁棒性。

在悬停任务中，首先需要定义目标悬停点，并设置相应的阈值来判定是否到达目标区域。当无人机与目标悬停点之间的距离小于该阈值时，认为任务已完成；否则，系统会根据无人机当前位置与目标位置之间的误差（包括位置误差和姿态角度误差）动态调整奖励值。误差越大，奖励越低，从而促使无人机尽可能减小偏差，逐步接近既定目标位置并调整至理想姿态。此外，当无人机顺利进入目标区域时，还会给予额外奖励以鼓励其尽快稳定悬停；若发生失控情况，则会施加惩罚措施；在每个时间步均附加固定惩罚，以督促飞行器快速、准确地完成任务。

在不同任务下，奖励机制会根据具体目标的变化进行灵活调整，以确保无人机能够适应多样化的任务需求。总体而言，KPPO算法通过其自主探索和优势估计机制，使系统能够在不需要人工干预的情况下，通过最大化累计奖励自主学习到长期局部最优策略，从而有效应对复杂外部扰动。这对于高维、非线性控制系统的无人机飞行控制具有重要意义。

3 仿真验证与结果分析

3.1 仿真验证设置

为验证本文提出KPPO算法的有效性，本文设计了5种物理仿真任务进行验证，并与代表性的PPO和DDPG算法进行对比，以突出本算法在性能上的优越性。物理仿真验证在Windows操作系统环境下进行，硬件平台采用AMD Ryzen 5 PRO 4650U处理器（集成Radeon Graphics，主频2.10 GHz）及16 GB内存的服务器。

KPPO和PPO的网络拓扑结构是相同的。网络输入为环境状态向量。网络主体为两层全连接多层感知机（MLP），其隐藏层分别含256个和128个单元，激活函数均采用ReLU。若施行参数共享，则前两层为共享层，随后分别分支为Actor和Critic。Actor分支输出一个动作均值向量，并配合固定标准差构成高斯策略；Critic分支输出单个标量状态价值估计。

DDPG中Actor网络接收环境当前状态向量作为输入，首先通过两层全连接隐藏层（分别包含256与128个神经元，均采用ReLU激活）提取特征，随后在输出层生成一个动作均值向量。为了将动作值约束到[−1,1]区间，输出向量会经过tanh进行限幅。Critic网络则将状态与动作拼接后输入，沿用与Actor相同的两层全连接隐藏层结构，并在输出层产生一个标量，用于对当前状态−动作对的

价值进行评估。为提升训练稳定性，Actor与Critic各自配置了一份与之同结构的目标网络，在每次参数更新时通过软更新策略将主网络参数渐进式地同步至目标网络，为后续的回报估计提供更平滑、可靠的参考。

物理仿真验证过程中所有任务采用统一的参数设定（见表1），其中学习率、折扣因子、批量大小、KL散度约束范围等关键参数的选择旨在确保算法在不同任务场景下均能保持稳定性和收敛性，以保证物理仿真验证结果的可比性和可靠性。为排除参数差异带来的影响、便于横向对比，本文在所有仿真验证中均采用统一的超参数配置。

表1　KPPO算法训练参数

参数	值
小批量样本数量	64
训练轮次	20
折扣因子	0.99
学习率	0.000 1
优化器系数	（0.9，0.999）
任务解决奖励	300
额外奖励	1 000
惩罚值	−1
最大训练回合数	1 000
每回合的最大步数	400
模型更新时间步长	400
KL散度惩罚系数	0.2
L2正则化系数	1×10^{-5}
随机种子	10
状态维度	12
动作维度	4
KL散度阈值	0.01

表2列出了DDPG算法中关键的训练参数，重复参数不再赘述。

表2　DDPG算法训练参数

参数	值
学习率	0.000 1
折扣因子	0.99
目标网络软更新系数	0.001
经验回放采样批大小	64
动作探索噪声强度	0.1

3.2　物理仿真任务

为了验证KPPO算法的有效性，本文设计并开展了物理仿真飞行控制任务进行验证。本文进行的仿真验证基于大疆御Mavic 2无人机模型，采用的物理仿真环境由QuadDynamics类实现，该类基于经典刚体动力学，为四旋翼飞行器在仿真环境中提供了精确的状态更新模型。其状态向量：

$$\boldsymbol{s}=\left[x,y,z,\varphi,\theta,\psi,v_x,v_y,v_z,w_x,w_y,w_z\right]^{\mathrm{T}} \tag{24}$$

其包括位置、姿态角（俯仰角、横滚角、偏航角）、线速度和角速度，共12维。给定4个转子推力输入T，先通过旋转矩阵将机体推力$\left[0,0,\sum_i T_i\right]^{\mathrm{T}}$转换到地面系，并与重力叠加，得到线性加速度。

$$a=\frac{1}{m}\left(R\left[0,0,\sum_i T_i\right]^{\mathrm{T}}+mg\right) \tag{25}$$

$$T=\left[T_1,T_2,T_3,T_4\right] \tag{26}$$

同时，根据力矩τ和转动惯性I计算角速度α。位置与角度的离散更新采用二阶欧拉积分：

$$p_{t+1}=p_t+v_t\triangle t+\frac{1}{2}a_t\triangle t^2 \tag{27}$$

$$v_{t+1}=v_t+\alpha_t\triangle t \tag{28}$$

$$\phi_{t+1}=\phi_t+\omega_t\triangle t+\frac{1}{2}\alpha_t\triangle t^2 \tag{29}$$

$$\omega_{t+1}=\omega_t+\alpha_t\triangle t \tag{30}$$

$$\tau=\left[\left(T_4-T_3\right)l,\left(T_2-T_1\right)l,0\right]^{\mathrm{T}} \tag{31}$$

$$I=\left[I_x,I_y,I_z\right] \tag{32}$$

$$\alpha=I^{-1}\tau \tag{33}$$

上述计算式均依赖已知物理参数（质量m、转体惯量I、转子臂长l、重力加速度g）和初始状态。为了保证仿真稳定，位置超过设定边界时进行限制并施加惩罚。其中已知的物理参数见表3。

表3　物理参数

参数名	数值	单位
重力加速度	9.81	$\mathrm{m/s^2}$
时间步长	0.02	s
四旋翼质量	0.665	kg
转子臂长	0.105	m
转动惯量	[0.002 3,0.002 5,0.003 7]	$\mathrm{kg\cdot m^2}$

5种物理仿真任务具体如下。

（1）飞行器追踪任务

该任务的核心目标是通过控制四旋翼飞行器接近并保持在设定的目标位置，同时尽量减少与目标位置的偏差。任务通过QuadDynamics类模拟四旋

翼的动态，其中状态空间为 12 维，主要利用飞行器状态空间中的位置信息以及在各个方向的速度信息。飞行器的动作空间包含 4 个控制输入，分别对应飞行器的推力和姿态控制。在奖励设计上，环境通过综合位置误差（计算飞行器与目标位置之间的距离误差）、速度误差（计算飞行器速度与目标速度之间的误差）和控制输入的惩罚来计算奖励函数，奖励函数如下：

$$\text{reward}=-\left(\text{err}_{\text{d}}+\text{err}_{\text{v}}\right) \tag{34}$$

$$\text{err}_{\text{d}}=w_1\cdot\left(\text{dist\_err}+\text{z\_err}\right) \tag{35}$$

$$\text{err}_{\text{v}}=w_2\cdot\left(\text{v\_err}+\text{v\_z\_err}\right) \tag{36}$$

其中，w_1 和 w_2 是位置和速度误差的权重参数，dist_err 是距离误差，z_err 是垂直方向的误差，v_err 和 v_z_err 分别是水平方向和垂直方向的速度误差。此外，当飞行器成功达到目标位置并且在设定的阈值范围内时，会获得额外的奖励。目的达成的条件是飞行器的状态误差小于阈值，且速度误差同样满足要求。任务在每一个时间步中，会根据控制输入更新状态，并根据当前状态计算奖励。任务终止条件包含目标达成或步数超限。

（2）飞行器悬停任务

该任务旨在训练智能体控制飞行器在三维空间内保持稳定的悬停，并精确到达指定的目标位置。状态空间和动作空间不变。该任务主要利用飞行器状态空间中的三维位置和姿态角。奖励设计上，环境结合位置误差（pos_err）和姿态误差（angle_err）来评估飞行器的表现。位置误差通过飞行器当前位置与目标位置的欧几里得距离计算，姿态误差通过飞行器的姿态角的大小来衡量。具体奖励函数如下：

$$\text{reward}=-\left(w_1\cdot\text{pos\_err}+w_2\cdot\text{angle\_err}\right) \tag{37}$$

其中，w_1 和 w_2 为调整位置误差和姿态误差权重的超参数。此外，当飞行器成功到达目标并保持稳定时，给予额外的奖励，奖励大小与目标达成次数成正比。任务的终止条件为飞行器达到目标位置或步数超限。

（3）飞行器高度控制任务

该任务旨在训练智能体控制飞行器的高度，使其能够接近并保持目标高度。状态空间和动作空间不变，奖励函数根据飞行器的高度误差（z_err）来引导其学习目标高度，具体奖励函数如下：

$$\text{reward}=-w_1\cdot\text{z\_err} \tag{38}$$

其中，w_1 是用于调节高度误差对奖励影响的权重参数。完成任务之后同样有额外的奖励促进目标达成。

（4）飞行器速度控制任务

该任务旨在训练智能体控制飞行器的速度，使其能够接近并保持目标速度，状态空间和动作空间不变。奖励函数根据飞行器的速度误差（speed_err）引导学习目标速度。具体计算式如下：

$$\text{reward}=-w_1\cdot\text{speed\_err} \tag{39}$$

其中，w_1 是用于调节速度误差对奖励影响的权重参数。当飞行器的速度误差小于设定的阈值时，给飞行器额外的奖励，并且随着达成目标次数的增多，奖励会逐步增大。

（5）飞行器追踪多目标任务

该任务旨在训练智能体控制飞行器，同时跟踪多个目标的位置。其中状态空间和动作空间不变，奖励函数设计的核心是引导飞行器同时接近所有目标，飞行器根据与每个目标的距离来计算奖励。奖励函数为：

$$\text{reward}=-w_1\cdot\sum_{\text{i}}\text{dist\_err}_{\text{i}} \tag{40}$$

其中，w_1 用于调节距离误差对奖励的影响。当飞行器成功接近所有目标时，给予额外的奖励，终止条件不变。

3.3 物理仿真验证结果分析

本文对 KPPO 算法与 PPO、DDPG 算法在奖励值方面进行了对比分析，结果如图 2~图 4 所示。在速度控制、高度控制和悬停控制 3 项任务中，KPPO 算法均展现出明显优势。

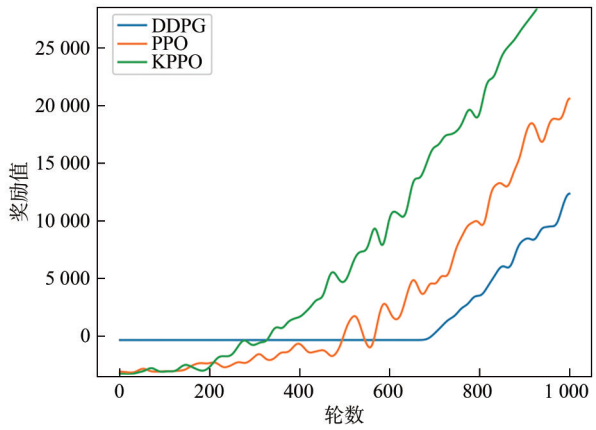


图 2 速度控制任务每轮奖励值对比

在图 2 所示的速度控制任务中，DDPG 在训练初期表现优越，但随着训练的深入，随机策略所具备的探索优势逐渐凸显，展现出在复杂环境中更强

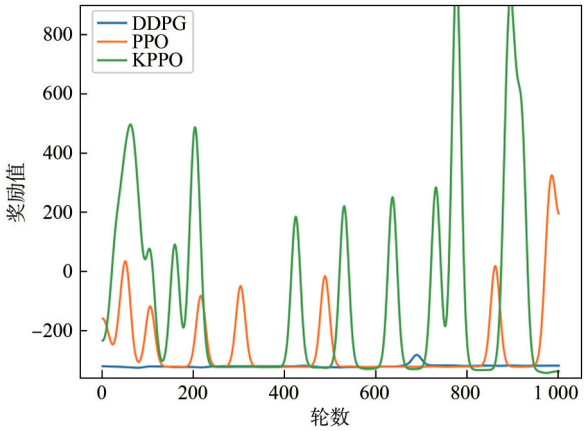


图3　悬停任务每轮奖励值对比

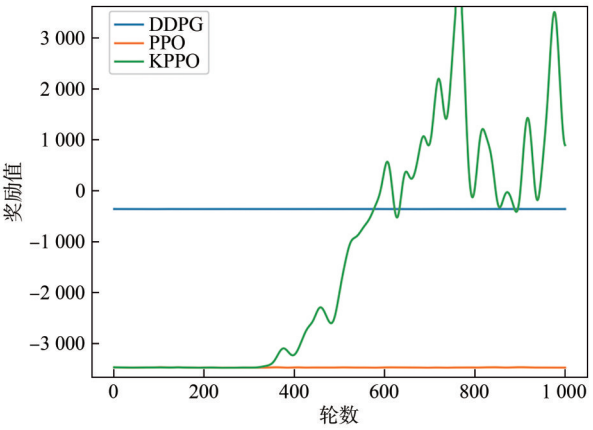


图4　高度控制任务每轮奖励值对比

的鲁棒性。在图3所示的悬停任务中，DDPG的奖励值波动相对较小，而PPO与KPPO的奖励值波动较大。这可能是由于悬停任务环境复杂，使得DDPG难以稳定学习有效策略进行充分探索。而在随机策略的加持下，KPPO能够快速尝试多种策略路径，从而不断优化其策略表现。图4为高度控制任务，其中DDPG与PPO的训练过程较为平稳，波动较小；而KPPO虽在训练初期处于劣势，但在约600轮后表现明显提升，展现出较强的后期优化能力。该现象主要源于DDPG采用的简单高斯噪声难以激发多样化的动作序列，导致智能体长期陷入局部最优，使得奖励值波动极小难以提升。此外，由于DDPG的目标网络与行为网络存在更新延迟，策略改进在训练早期表现出明显滞后性，降低了其在复杂环境中的适应能力。而PPO在训练初期策略尚未收敛时，剪切操作会过度抑制策略更新幅度，使智能体难以摆脱随机初始化的状态分布，造成奖励值提升缓慢，波动性较小。KPPO在训练初期KL散度较低，前期探索行为与PPO相似，而当训练推进至策略初具雏形、KL散度超过预设阈值后，引入的门控惩罚机制一方面可有效抑制策略过度偏移，另一方面加速其朝更优方向收敛，因而在约600轮后表现出显著性能提升。

综上所述，采用KPPO算法的飞行器能够迅速适应新环境并高效执行任务，充分体现了其较强的泛化能力和环境适应能力。

除了对单轮奖励进行分析之外，跨多轮的奖励平均值（avg_running_reward）也是评估训练过程的重要指标。单轮的奖励受随机初始化、环境状态和探索动作等因素的影响，通常波动较大且噪声显著，因此难以准确反映策略优化的整体趋势。而跨多轮奖励平均值通过对多个轮次的奖励取均值，有效平滑了随机波动，从而能够更可靠地刻画训练过程，并为评估学习策略的有效性提供坚实依据。

图5~图7展示了在3个环境中计算得到的跨轮奖励平均值，为深入分析和验证算法性能提供了详实的数据支持。从图5和图6可以看出，KPPO算法的跨多轮奖励平均值显著高于PPO和DDPG算法，在图7高度控制任务中前期KPPO处于劣势，但后期的策略逐步成型，在KL散度和L2正则化的加持下，约在600轮成功反超并处于爆发式增长的状态。

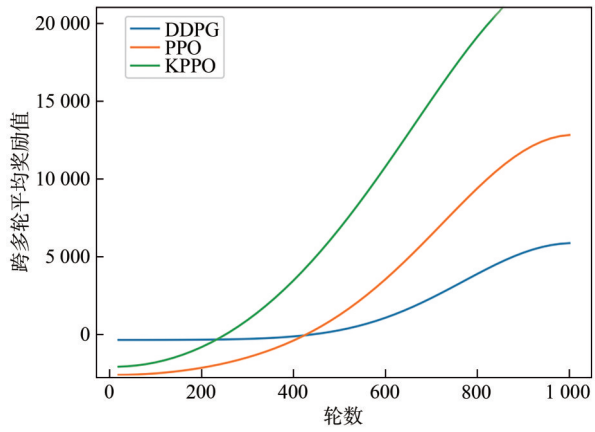


图5　速度控制任务跨多轮平均奖励值对比

这表明KPPO在多任务和复杂环境下能够更稳定地提升四旋翼无人机的控制性能和任务执行效率，同时有效降低由环境噪声和随机探索导致的性能波动。该结果进一步验证了KPPO算法在策略更新过程中更强的鲁棒性和泛化能力，使其在复杂环境下具备更优的学习稳定性和适应性。

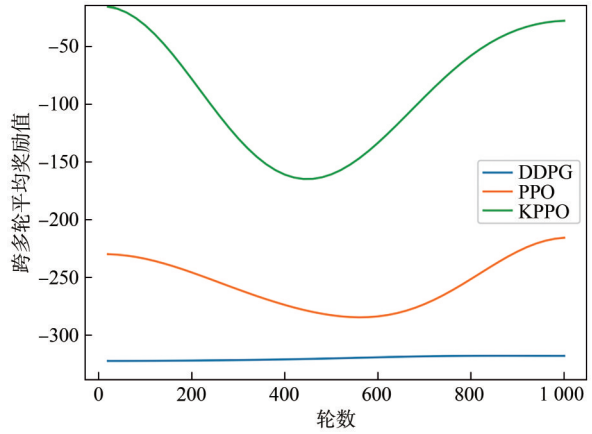


图6　悬停任务跨多轮平均奖励值对比

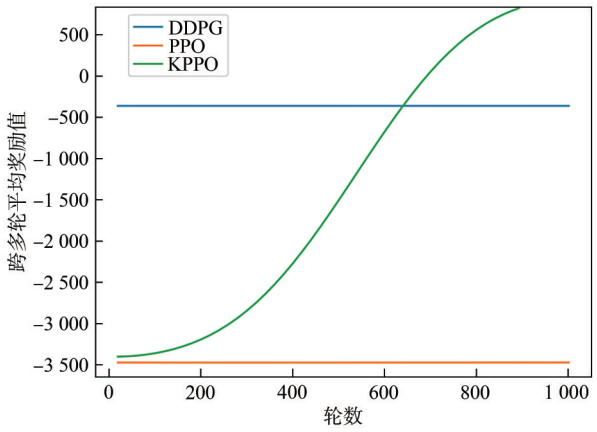


图7　高度控制任务跨多轮平均奖励值对比

在多目标追踪任务中，3 种算法的奖励值均波动性较小，说明在该任务中飞行器难以做出正确的决策，只能不断地尝试探索。从图 8、图 9 可以看出，DDPG 的奖励值要远远高于 PPO 和 KPPO 算法的奖励值。

造成这一现象的原因有两个，一方面，可能是因为 DDPG 属于离策略（off-policy）算法，可以在

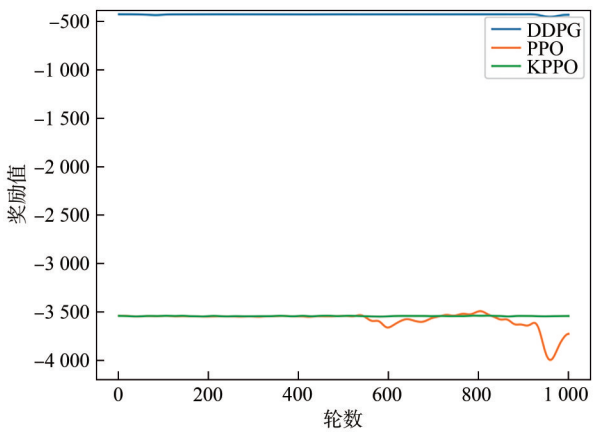


图8　追踪多目标任务每轮奖励值对比

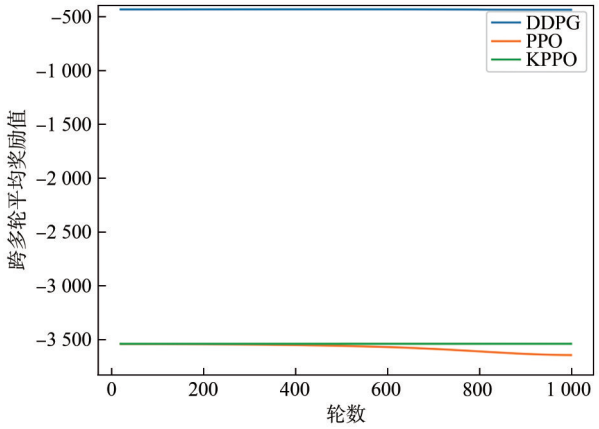


图9　追踪多目标任务跨多轮平均奖励值对比

经验回访缓冲区中反复使用同一批数据进行多次梯度更新，极大地提高了样本利用效率。而 PPO 和 KPPO 均为同策略（on-policy）算法，每一轮交互数据只能使用固定次数，后续即被丢弃，对样本的利用率较低。另一方面，可能是因为任务奖励函数对动作非常敏感，DDPG 构建的 Q-函数近似能够直接对动作价值进行精细评估与优化，而基于优势估计的 PPO 和 KPPO 需要先估计状态值，再结合概率比更新策略，步骤更多，收敛速度更慢。对比 KPPO 和 PPO 算法，尽管飞行器在某些情况下可能难以做出完全正确的决策，但 KPPO 算法的策略收敛稳定，能够维持一致的性能表现。而 PPO 算法的奖励值在后期呈现逐步下降的趋势，暗示其策略在复杂环境下变化较大，在复杂的环境下容易出现决策失误，从而导致奖励值降低。

此外，在复杂环境下，KPPO 算法展现出显著优势。如图 10 和图 11 所示，在飞行器追踪任务中，由于需要在多个维度上协调控制，对算法的鲁棒性和稳定性提出了更高要求。尽管在单轮的奖励对比中，KPPO 算法的优势可能不易直接显现，但跨多轮平均奖励值的对比结果清楚地表明，KPPO 算法驱动下的飞行器控制性能呈现出逐步上升的趋势，充分体现了其较强的抗干扰能力和环境适应性。而 DDPG 在多维度的追踪任务中，效果较差，大约在 500 轮和 1 400 轮出现奖励值下降的情况。

这可能是因为目标网络更新延迟的问题，当环境动态变化或者策略进行改进时，目标网络跟不上当前策略，导致短期内价值评估误差增大，继而使动作选择不再最优，出现奖励下滑的情况。除此之外，DDPG 的经验回放存储历史转移样本，随着策略不断更新，早期存储的质量较差的样本仍然被采

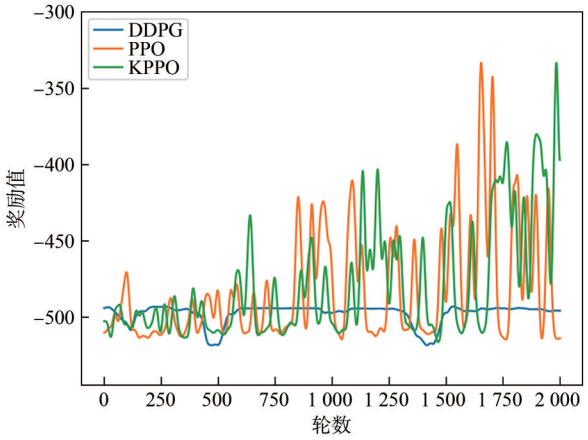


图10　追踪任务每轮奖励值对比

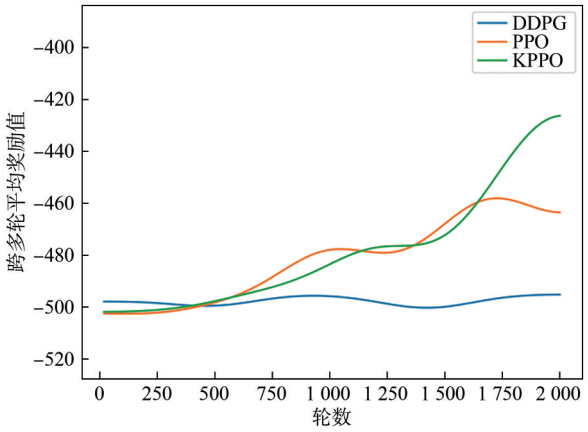


图11　追踪任务跨多轮平均奖励值对比

样并用于训练，干扰当前价值估计和策略改进，从而引起奖励的偶发下降。考虑到仿真环境的复杂性，为了更全面地评估算法性能，训练轮数被适当延长。仿真结果显示，在1 750轮训练之后，KPPO算法的奖励值出现了显著上升的趋势，这一现象进一步验证了KPPO算法在复杂场景下具备更高的适应性和稳定性，从而有效提升了四旋翼无人机飞行控制的整体性能。

表4和表5分别为3种算法的总奖励平均值与标准差在不同环境中的对比值。结合上述奖励对比图能够看出，在总平均奖励中，KPPO算法略胜一筹。

表4　3种算法总奖励平均值对比

任务	KPPO 平均值	PPO 平均值	DDPG 平均值
HeightControlEnv	−1 322.185	−3 472.665	−364.295
MultiTargetTrackingEnv	−3 542.908	−3 577.797	−432.059
SpeedControlEnv	**8 879.512**	3 242.066	1 504.431
HoverTaskEnv	**−90.670**	−255.520	−320.119
TrajTaskEnv	**−478.866**	−481.655	−497.700

表5　3种算法总奖励标准差对比

任务	KPPO 标准差	PPO 标准差	DDPG 标准差
HeightControlEnv	4 031.546	11.052	0.647
MultiTargetTrackingEnv	11.771	143.894	6.944
SpeedControlEnv	11 845.580	8 519.012	3 671.278
HoverTaskEnv	1 764.684	686.389	27.569
TrajTaskEnv	220.065	218.798	8.875

虽然在高度控制任务中的总奖励平均值低一些，但是当轮数增大时，前期KPPO的劣势将逐渐消失，能够反超DDPG。在表5的标准差对比中不难发现随机策略的波动性要大于确定性策略的波动，但对于较为复杂的无人机飞行控制随机策略能够进行更广泛的探索，学习更多可能的策略，寻找最合适的策略。确定性策略在复杂环境中难以发挥，限制其探索能力。

综上所述，本文设计的策略能够快速收敛，并且有效地提高飞行器执行任务的能力，更适合在飞行控制任务中使用。

图12展示了无人机在5种仿真飞行控制任务中的运动轨迹及其在x、y、z方向上随时间变化的曲线。在速度控制任务和悬停任务中，由于无人机均能在达到预定目标后即刻结束任务，故两者的横坐标均未达到最大步长400 000步。具体而言，在速度控制任务中，DDPG算法以最快速度完成目标，PPO最慢。而在悬停任务中，KPPO算法的收敛速度最快，DDPG与PPO则基本持平。上述现象产生的原因可能是DDPG在速度控制任务中能够快速利用丰富的速度误差梯度完成目标，然而在悬停任务中，由于策略易陷入局部最优，其收敛速度与PPO相近。相比之下，KPPO通过设定KL散度阈值，在策略更新中动态增强探索，并在后期阶段快速收敛，因而能够更快地实现悬停任务。

在高度控制任务、多目标追踪任务和单目标追踪任务中，无人机始终未能到达指定位置，直至达到最大步长400 000步仿真试验才终止。从曲线走势可见，KPPO与PPO在前期均表现出较强的探索行为——其轨迹波动幅度较大，以寻找目标位置；相比之下，KPPO的探索阶段持续时间更短，其更快地进入相对稳定的悬停状态，表明引入阈值触发KL散度能有效加速收敛。而DDPG的探索性较弱，绝大多数时间保持悬停状态，导致其整体收敛进程缓慢、对目标位置的搜索效率较低。上述现象产生

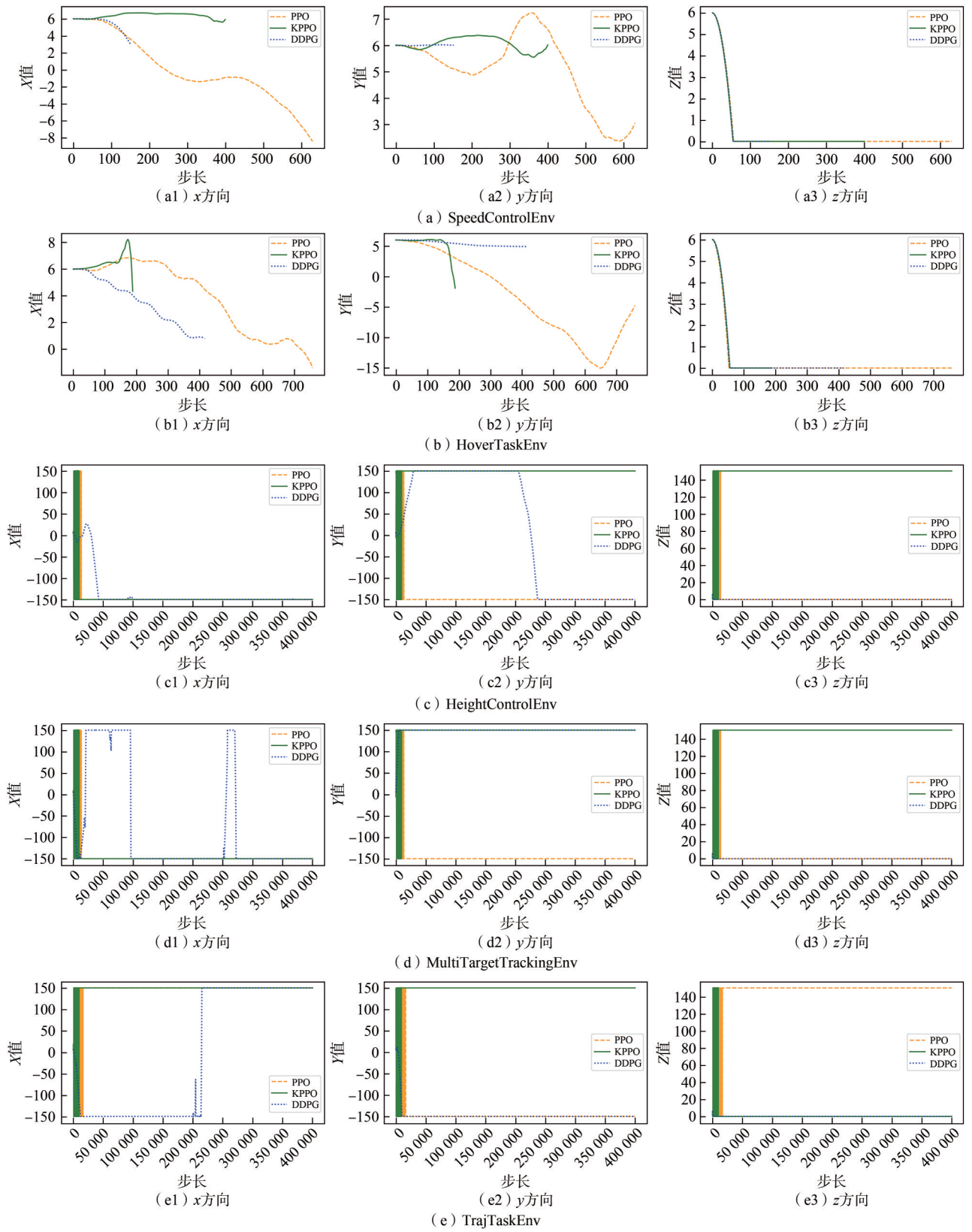


图 12　二维轨迹对比

的原因可能是DDPG依赖确定性策略梯度，其随机性较低，难以摆脱局部最优，因而常在试验早期即陷入悬停；PPO通过限制策略变动幅度以保证训练稳定，但也因此导致初期探索速度不足；KPPO在

PPO的基础上引入阈值触发的KL散度惩罚项与L2正则项构成的复合正则机制，在保证策略平稳更新的同时，提升了探索效率。

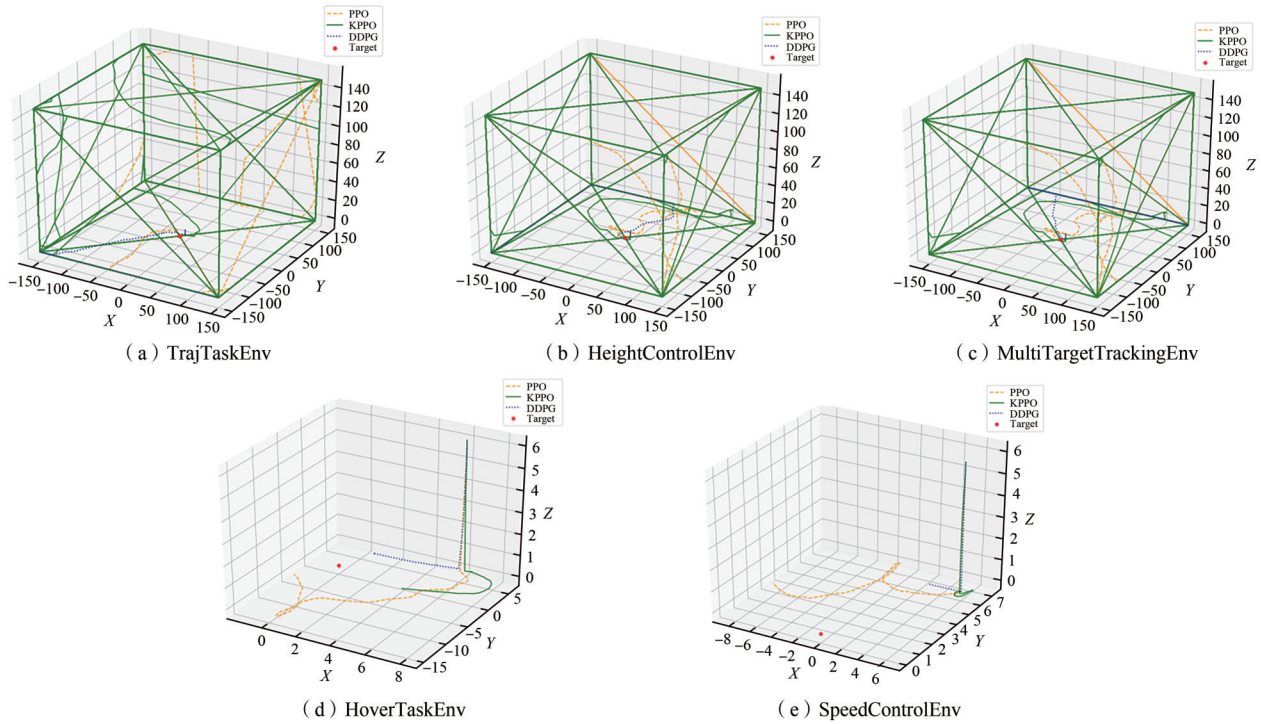
图13中5种仿真飞行控制任务的三维轨迹对比图清晰地展示了速度控制任务与悬停任务的快速收敛特性。在这两类任务中，3条算法曲线均在预设的空间容差范围内与目标点重合——仿真判定标准并非精确到达目标点，而是当飞行器进入目标点附近的允许半径内时，即视为任务完成。

相比之下，在高度控制任务、多目标追踪任务和单目标追踪任务中，由于飞行器始终无法进入既定容差区域，仿真运行直至最大步长400 000才终止。三维轨迹在有限空间内反复盘旋、交织，导致肉眼难以分辨哪种算法表现更优。此时结合图12能够更直观地比较各算法在不同坐标维度的偏差与收敛趋势，从而辅助评估算法的整体有效性。从对比结果可以看出，KPPO在这3种复杂任务中的表现优于DDPG和PPO，但其轨迹依然相对分散，未能收敛至目标区域。造成上述现象的主要原因可能是多目标追踪任务中目标点位置动态变化，对控制策略的实时调整要求极高；而单目标追踪与高度控制任务对精准定位与高度维持提出了双重挑战，虽然KPPO能够提升四旋翼无人机在复杂环境中的飞行稳定性和抗干扰能力，但尚缺乏自适应险情响应与多阶段规划能力。

4　结束语

本文提出的KPPO算法通过阈值触发KL散度对新旧策略进行约束，有效抑制了策略突变带来的不稳定性，增强训练过程的稳定性。同时，在四旋翼无人机飞行任务的回报函数中引入惩罚项，以加速任务执行并提升任务完成效率。此外，算法融合了L2正则化，与KL散度协同构成复合正则机制，对策略网络参数施加约束，从而进一步提升四旋翼无人机在复杂环境中的飞行稳定性和抗干扰能力。物理仿真任务表明，KPPO算法在提升四旋翼无人机飞行控制能力方面取得了显著成效，能够在复杂环境下实现更优的任务执行性能。

为进一步提升算法性能，后续工作将聚焦于提高数据利用率和增强模型自适应能力，以实现更稳定的模型训练和更鲁棒的策略表现。具体改进措施包括：引入优先经验回放（prioritized experience replay，PER）机制，优化数据采样策略，进一步加快模型的收敛速度；采用时间差分方法（TD_λ）估算回报，以解决传统蒙特卡洛方法中常见的高方差问题，使回报估计更加稳定，并提高训练过程的收敛效率。



（a）TrajTaskEnv　（b）HeightControlEnv　（c）MultiTargetTrackingEnv

（d）HoverTaskEnv　（e）SpeedControlEnv

图13　三维轨迹对比

参考文献：

[1] NGUYEN V N, JENSSEN R, ROVERSO D. Intelligent monitoring and inspection of power line components powered by UAVs and deep learning[J]. IEEE Power and Energy Technology Systems Journal, 2019, 6(1): 11-21.

[2] CHOUTRI K, LAGHA M, DALA L. A fully autonomous search and rescue system using quadrotor UAV[J]. International Journal of Computing and Digital Systems, 2021, 10(1): 403-414.

[3] 任博, 潘景余, 苏畅, 等. 不确定环境下的侦察无人机自主航路规划仿真[J]. 电光与控制, 2008, 15(1): 31-34, 46.
REN B, PAN J Y, SU C, et al. Autonomous path planning simulation of UAVs in uncertain environment[J]. Electronics Optics & Control, 2008, 15(1): 31-34, 46.

[4] ZHENG Y J, DU Y C, LING H F, et al. Evolutionary collaborative human-UAV search for escaped criminals[J]. IEEE Transactions on Evolutionary Computation, 2020, 24(2): 217-231.

[5] TIAN Y L, LIU K, OK K, et al. Search and rescue under the forest canopy using multiple UAVs[J]. The International Journal of Robotics Research, 2020, 39(10/11): 1201-1221.

[6] XING L J, FAN X Y, DONG Y X, et al. Multi-UAV cooperative system for search and rescue based on YOLOv5[J]. International Journal of Disaster Risk Reduction, 2022, 76: 102972.

[7] 晏磊, 廖小罕, 周成虎, 等. 中国无人机遥感技术突破与产业发展综述[J]. 地球信息科学学报, 2019, 21(4): 476-495.
YAN L, LIAO X H, ZHOU C H, et al. The impact of UAV remote sensing technology on the industrial development of China: a review[J]. Journal of Geo-Information Science, 2019, 21(4): 476-495.

[8] 赵静, 闫春雨, 杨东建, 等. 基于无人机多光谱遥感的台风灾后玉米倒伏信息提取[J]. 农业工程学报, 2021, 37(24): 56-64.
ZHAO J, YAN C Y, YANG D J, et al. Extraction of maize lodging information after typhoon based on UAV multispectral remote sensing[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(24): 56-64.

[9] ANG K H, CHONG G, LI Y. PID control system analysis, design, and technology[J]. IEEE Transactions on Control Systems Technology, 2005, 13(4): 559-576.

[10] NG T C T, LEUNG F H F, TAM P K S. A simple gain scheduled PID controller with stability consideration based on a grid-point concept[C]// Proceedings of the ISIE '97 Proceeding of the IEEE International Symposium on Industrial Electronics. Piscataway: IEEE Press, 2002: 1090-1094.

[11] PAPADOPOULOS K G, TSELEPIS N D, MARGARIS N I. On the automatic tuning of PID type controllers via the Magnitude Optimum criterion[C]//Proceedings of the 2012 IEEE International Conference on Industrial Technology. Piscataway: IEEE Press, 2012: 869-874.

[12] REYES-VALERIA E, ENRIQUEZ-CALDERA R, CAMACHO-LARA S, et al. LQR control for a quadrotor using unit quaternions: Modeling and simulation[C]//Proceedings of the CONIELECOMP 2013, 23rd International Conference on Electronics, Communications and Computing. Piscataway: IEEE Press, 2013: 172-178.

[13] FALANGA D, FOEHN P, LU P, et al. PAMPC: perception-aware model predictive control for quadrotors[C]//Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2018: 1-8.

[14] RAMEZANI DOORAKI A, LEE D J. An innovative bio-inspired flight controller for quad-rotor drones: Quadrotor drone learning to fly using reinforcement learning[J]. Robotics and Autonomous Systems, 2021, 135: 103671.

[15] GHOURI U H, ZAFAR M U, BARI S, et al. Attitude control of quadcopter using deterministic policy gradient algorithms (DPGA)[C]//Proceedings of the 2019 2nd International Conference on Communication, Computing and Digital systems. Piscataway: IEEE Press, 2019: 149-153.

[16] 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.

[17] LU J W, HAN L Y, WEI Q L, et al. Event-triggered deep reinforcement learning using parallel control: a case study in autonomous driving[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(4): 2821-2831.

[18] HAARNOJA T, BEN M R, LEVER G, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning[J]. Science Robotics, 2024, 9(89): eadi8022.

[19] ZHOU Y T, YANG J C, GUO Z W, et al. An indoor blind area-oriented autonomous robotic path planning approach using deep reinforcement learning[J]. Expert Systems with Applications, 2024, 254: 124277.

[20] ORR J, DUTTA A. Multi-agent deep reinforcement learning for multi-robot applications: a survey[J]. Sensors, 2023, 23(7): 3625.

[21] KOCH W, MANCUSO R, WEST R, et al. Reinforcement learning for UAV attitude control[J]. ACM Transactions on Cyber-Physical Systems, 2019, 3(2): 1-21.

[22] 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.

[23] 梁晨, 刘小雄, 张兴旺, 等. 基于强化学习的四旋翼无人机控制律设计[J]. 计算机测量与控制, 2021, 29(2): 71-75, 86.
LIANG C, LIU X X, ZHANG X W, et al. Design of control law for quadrotor UAV based on reinforcement learning[J]. Computer Measurement & Control, 2021, 29(2): 71-75, 86.

[24] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[J]. arXiv preprint, 2018, arXiv: 1801.01290.

[25] 何准, 董文瀚, 蔡鸣, 等. 基于DDPG的多旋翼无人机自主引导与跟踪方法[J]. 飞行力学, 2021, 39(2): 63-69, 76.
HE Z, DONG W H, CAI M, et al. Multi-rotor UAV autonomous guidance and tracking method based on DDPG[J]. Flight Dynamics, 2021, 39(2): 63-69, 76.

[26] 胡多修, 董文瀚, 解武杰. 无人机自主引导跟踪与避障的近端策略优化[J]. 北京航空航天大学学报, 2023, 49(1): 195-205.
HU D X, DONG W H, XIE W J. Proximal policy optimization for UAV autonomous guidance, tracking and obstacle avoidance[J]. Journal of Beijing University of Aeronautics and Astronautics, 2023, 49(1): 195-205.

[27] 梁吉, 王立松, 黄昱洲, 等. 基于深度强化学习的四旋翼无人机自主控制方法[J]. 计算机科学, 2023, 50(S2): 13-19.
LIANG J, WANG L S, HUANG Y Z, et al. Autonomous control method of quadrotor UAV based on deep reinforcement learning[J]. Computer Science, 2023, 50(S2): 13-19.

[28] 王伟, 吴昊, 刘鸿勋, 等. 基于深度强化学习的无人机姿态控制器设计[J]. 科学技术与工程, 2023, 23(34): 14888-14895.
WANG W, WU H, LIU H X, et al. Design of UAV attitude controller based on deep reinforcement learning[J]. Science Technology & Engineering, 2023, 23(34): 14888-14895.

[29] 黄号, 马文卉, 李家诚, 等. 未知环境下无人机编队智能避障控制方法[J]. 清华大学学报(自然科学版), 2024, 64(2): 358-369.

HUANG H, MA W H, LI J C, et al. Intelligent obstacle avoidance control method for unmanned aerial vehicle formations in unknown environments[J]. Journal of Tsinghua University (Science and Technology), 2024, 64(2): 358-369.

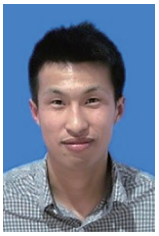
[30] 孔飞, 赵振根, 程磊, 等. 输入受限及干扰下固定翼无人机强化学习控制[J]. 电光与控制, 2024, 31(2): 21-28.
KONG F, ZHAO Z G, CHENG L, et al. Reinforcement learning control of fixed-wing UAV under input limitation and disturbance[J]. Electronics Optics & Control, 2024, 31(2): 21-28.

[31] 江泰民, 谭泰, 李辉, 等. 基于分层深度强化学习的六自由度固定翼无人机路径跟踪方法[J]. 计算机工程, 已录用, 2024: 0070197.
JIANG T M, TAN T, LI H, et al. Path following of 6-DOF fixed-wing UAV based on hierarchical deep reinforcement learning[J]. Computer Engineering, accepted, 2024: 0070197.

[作者简介]

武子怡（2004− ），女，河南工业大学人工智能与大数据学院在读，主要研究方向为强化学习。

高晓楠（2003− ），女，河南工业大学人工智能与大数据学院在读，主要研究方向为强化学习。



朱献超（1990− ），男，博士，河南工业大学人工智能与大数据学院讲师，主要研究方向为强化学习和机器学习。

赵亮（1978− ），男，博士，河南工业大学电气工程学院副教授，主要研究方向为深度学习、平行学习。

祝峰（1962− ），男，博士，电子科技大学基础与前沿研究院教授，主要研究方向为粗糙集和粒计算。

蔡磊（1979− ），男，博士，河南科技学院人工智能学院教授，主要研究方向为智能无人系统协同控制和机器人控制。