基于线性滤波器的四旋翼无人机强化学习控制策略

华和安[①] 方勇纯*[①] 钱 辰[①] 张雪涛[②]

[①](南开大学人工智能学院 天津 300350)

[②](大连理工大学智能机器人实验室 大连 116024)

摘 要：针对四旋翼无人机(UAVs)系统，该文提出一种基于线性降阶滤波器的深度强化学习(RL)策略，进而设计了一种新型的智能控制方法，有效地提高了旋翼无人机对外界干扰和未建模动态的鲁棒性。首先，基于线性降阶滤波技术，设计了维数更少的滤波器变量作为深度网络的输入，减小了策略的探索空间，提高了策略的探索效率。在此基础上，为了增强策略对稳态误差的感知，该文结合滤波器变量和积分项，设计集总误差作为策略的新输入，提高了旋翼无人机的定位精度。该文的新颖之处在于，首次提出一种基于线性滤波器的深度强化学习策略，有效地消除了未知干扰和未建模动态对四旋翼无人机控制系统的影响，提高了系统的定位精度。对比实验结果表明，该方法能显著地提升旋翼无人机的定位精度和对干扰的鲁棒性。

关键词：四旋翼无人机；智能控制；强化学习；未知干扰

中图分类号：V279; TP273 **文献标识码**：A **文章编号**：1009-5896(2021)12-3407-11

DOI: 10.11999/JEIT210251

Reinforcement Learning Control Strategy of Quadrotor Unmanned Aerial Vehicles Based on Linear Filter

HUA He'an[①] FANG Yongchun[①] QIAN Chen[①] ZHANG Xuetao[②]

[①](*College of Artificial Intelligence*, *Nankai University*, *Tianjin* 300350, *China*)

[②](*Intelligent Robotic Laboratory*, *Dalian University of Technology*, *Dalian* 116024, *China*)

Abstract: In this paper, based on linear filter, a deep Reinforcement Learning (RL) strategy is proposed, then a novel intelligent control method is put forward for quadrotor Unmanned Aerial Vehicles (UAVs), which improves effectively the robustness against disturbance and unmodeled dynamics. First of all, based on linear reduced-order filtering technology, filter variables with fewer dimensions are designed as the input of the deep network, which reduces the exploration space of the strategy and improves the exploration efficiency. On this basis, to enhance strategy perception of steady-state errors, the filter variables and integration terms are combined to design the lumped error as the new network input, which improves the positioning accuracy of quadrotor UAVs. The novelty of this paper lies in that it is the first intelligent approach based on linear filtering technology, to eliminate successfully the influence of unknown disturbance and unmodeled dynamics of quadrotor UAVs, which improves the positioning accuracy. The results of comparative experiments show the effectiveness of the proposed method in terms of improving positioning accuracy and enhancing robustness.

Key words: Quadrotor Unmanned Aerial Vehicles (UAVs); Intelligent control; Reinforcement Learning(RL); Unknown disturbance

1 引言

随着智能机器人技术的发展，旋翼无人机受到了越来越多的关注[1–8]。旋翼无人机具有垂直起降、定点悬停、低空低速飞行、机动性强、响应迅速等优势，被广泛运用于军事侦察、地形探索、货物运输、空中监控等。然而，由于旋翼无人机的欠驱动、 开环不稳定和非线性等性质，设计有效的控制算法充满了挑战。与此同时，旋翼无人机的应用场景复杂多样，其中不乏充满未知干扰的环境。因此，设计智能有效的控制策略，使其能在干扰环境中执行任务，具有重大的实际和理论意义。目前，已经报道了许多有意义的研究结果，本文将这些结

收稿日期：2021-03-26；改回日期：2021-10-20；网络出版：2021-10-27

*通信作者： 方勇纯 fangyc@nankai.edu.cn

基金项目：国家自然科学基金(61873132, 61633012)

Foundation Items: The National Natural Science Foundation of China (61873132, 61633012)

果分为两类：基于模型的方法和基于学习的方法。

前者的基本思想是充分利用已知的动力学模型，并在此基础上设计控制器，然后对闭环系统进行稳定性分析[9–17]。基于这一思想，应用非线性控制技术，研究人员设计出许多有效的控制方法，其中主要包括反步法、滑模方法、容错控制、鲁棒控制、基于观测器的设计等等。具体而言，研究人员注意到旋翼无人机的动力学模型可以变换为严格的反馈形式，在文献[14]中提出了基于反步法的控制器设计，并利用Lyapunov理论证明了系统的稳定性。考虑到系统可能存在电机损坏的故障，在文献[17]中针对三旋翼无人机提出了一种非线性鲁棒容错控制方法。尽管研究人员们已经提出了许多行之有效的控制方法，仍然有以下问题亟待解决：(1)现有的控制方法通常依赖于准确的系统动力学模型。但是在实际系统中，很难非常精确地建模所有的系统动态并且整定所有模型参数；(2)基于模型的控制方法对于外部扰动的鲁棒性通常有待提高，当存在外部未知干扰、未建模动态和未知参数时，控制的效果往往没有保证；(3)尽管一些先进的非线性控制方法考虑了模型和环境的不确定性，但是通常也需要对这些不确定性做假设，在此基础上设计的控制器往往也是在控制精度和鲁棒性之间权衡。

随着人工智能技术的发展，基于深度学习的控制技术显示出了巨大的适应性和灵活性。强化学习作为典型的代表，已经成功地解决了许多的控制问题[18–25]。但是，对于旋翼无人机这种复杂的非线性系统而言，设计有效的强化学习方法仍是一个相当具有挑战性的问题。具体而言，针对四旋翼的视觉伺服问题，在文献[18]中采用了Q学习(Q-learning)来调整增益，并通过模糊方法来调整学习速率。在文献[20]中提出了深度Q学习(Deep Q-Network, DQN)算法，该深度强化学习算法在Atari游戏中达到了人类的水平。然而，DQN算法仅能解决离散和高维观测空间类型的问题，而无法解决存在连续和高维动作空间的问题。文献[21]提出了确定性策略梯度(Deterministic Policy Gradient, DPG)算法来解决连续动作空间中的强化学习问题。更进一步，研究人员在文献[22]中提出深度确定性策略梯度(Deep DPG, DDPG)算法，成功地完成了20多个模拟的物理任务。与其他的强化学习方法不同，DDPG可以在连续和高维动作空间中有效地生成最佳动作。最近，为了解决旋翼无人机在移动平台上的着陆问题，在文献[23]中，研究人员将DDPG算法用于训练着陆策略。在文献[24]中，为了减小四旋翼无人机系统的稳态误差，在DDPG算法中引入了积分补偿器。仿真和实验结果表明，无人机系统的定位精度有所提高。

目前为止，大部分强化学习的方法仅仅停留在仿真阶段，如何设计出可以在实际系统中应用的强化学习算法仍然需要进一步的研究。就强化学习策略在旋翼无人机上的应用而言，仍然有许多问题需要解决，具体总结如下：首先，由于旋翼无人机是多入多出的欠驱动、非线性系统，强化学习策略的搜索空间大，策略收敛慢；其次，由于强化学习策略对稳态误差不敏感，在有干扰的环境中定位精度较差。

针对以上问题，基于线性降阶滤波技术，本文提出一种智能高效的强化学习算法。具体而言，通过将端到端的深度强化学习和基于模型的控制器相结合，设计了一种协同高效的智能控制器。首先，为了加快强化学习策略的收敛速度，设计了全新的滤波器变量以减小输入空间的维度，提高探索效率，保证本文方法更适合在实时的旋翼无人机系统中应用。其次，通过引入位置积分信号，增强了策略对稳态误差的感知，提高了系统的定位精度。最后，将所提出的算法应用在自制的实验平台上，成功地提升了系统在强干扰环境下的控制性能。本文的贡献总结如下：

(1)针对四旋翼无人机系统，本文首次提出一种基于线性降阶滤波器的强化学习策略，提升了系统对未知干扰的抑制能力。

(2)与传统的控制方法相比，本文的控制策略不需要复杂的建模过程。控制器通过与系统实时交互，智能地做出控制决策，有效地提升旋翼无人机对干扰的鲁棒性，提高定位精度。

(3)通过设计降阶滤波器变量和引入积分器，提高了强化学习算法的探索效率，增强了策略对稳态误差的感知，使其更适合在实际系统中应用。

本文接下来的内容组织如下：在第2节中引入旋翼无人机的控制问题；在第3节中设计智能高效的深度强化学习策略；在第4节中针对提出的控制策略，设计训练-评价算法，训练策略的参数，设计对比实验，证明该方法的可行性和有效性；最后，在第5节中总结本文内容。

2 问题提出

2.1 旋翼无人机动力学模型

旋翼无人机和强化学习策略的交互过程如图1所示。

强化学习策略生成动作施加于旋翼无人机系统，并且从无人机得到反馈的状态；通过这种持续的动作-状态交互，学习到适应环境的最优控制策

略。在本文中，惯性系用F_I表示，连体系用F_B表示，其中心位置位于旋翼无人机的中心(如图1所示)。旋翼无人机的动力学模型表示为[9]

$$\dot{\boldsymbol{p}}=\boldsymbol{v}, m\dot{\boldsymbol{v}}+mg\boldsymbol{e}_3=f\boldsymbol{R}\boldsymbol{e}_3, \dot{\boldsymbol{R}}=\boldsymbol{R}\boldsymbol{\Omega}^{\times},$$
$$\boldsymbol{J}\dot{\boldsymbol{\Omega}}+\boldsymbol{\Omega}\times\boldsymbol{J}\boldsymbol{\Omega}=\boldsymbol{M} \tag{1}$$

其中，$\boldsymbol{p}=[x\,y\,z]^{\top}\in\mathbb{R}^3, \boldsymbol{v}\in\mathbb{R}^3$分别表示无人机在惯性系中的位置和速度；$m, g\in\mathbb{R}$分别表示无人机的质量和重力加速度；$\boldsymbol{e}_3=[001]^{\top}$表示竖直方向的单位向量；$\boldsymbol{J}\in\mathbb{R}^{3\times3}$表示无人机的转动惯量；$\boldsymbol{R}\in\mathbb{R}^{3\times3}$表示从连体系到惯性系的旋转矩阵；$\boldsymbol{\Omega}\in\mathbb{R}^3$表示无人机在连体系中的旋转角速度；$f\in\mathbb{R}$, $\boldsymbol{M}\in\mathbb{R}^3$分别表示无人机产生的升力和转动力矩；定义期望位置$\boldsymbol{p}_d=[x_d\,y_d\,z_d]^{\top}\in\mathbb{R}^3$，在此基础上，定位误差定义如下：$e_x=x-x_d, e_y=y-y_d, e_z=z-z_d$。本文所采用的模型包含系统的主要动力学特性，考虑到建模的复杂程度和实际需要，陀螺力矩等复杂动力学特性暂未考虑。

2.2 控制目标

本文致力于解决位置控制问题，即设计合适的控制力f，提高系统的定位精度。由于未知干扰、未建模动态和未准确整定参数等的影响，无人机的准确定位在现实中往往很难实现。为了解决该问题，本文的控制目标为：基于强化学习技术，设计合适的学习控制策略，实现无人机的准确定位。首先，由于系统动力学复杂，强化学习策略收敛慢；其次，由于传统的强化学习策略对稳态误差不敏感，往往很难提高无人机的定位精度。针对以上问题，本文设计的算法旨在同时提高强化学习的效率和生成算法对稳态误差的感知能力，从而完成无人机的准确定位控制。需要指出的是，四旋翼无人机是典型的欠驱动系统，控制量的个数小于系统自由度的个数，该类以少控多的问题很具挑战性。

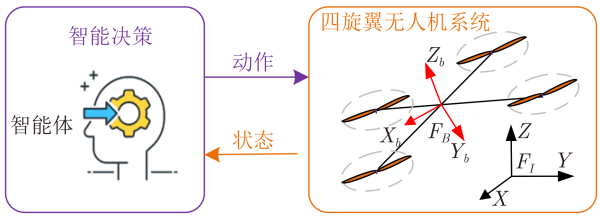


图 1 强化学习策略和旋翼无人机的交互

3 主要内容

本节的主要内容是设计强化学习算法，该算法的学习过程如图2所示。

具体而言，强化学习的网络采用演员-评论家(Actor-Critic)结构，并且引入目标网络和经验池(图2中的缓冲区)技术用以保证学习过程的稳定性和减少数据之间的相关性。其中，演员(Actor)是由μ参数化的神经网络A^{μ}；评论家(Critic)是由ω参数化的神经网络Q^{ω}。Actor通过调整策略参数μ，输出最优的动作；Critic在线估计Actor动作的价值来评价Actor的策略。通过训练生成的最优策略可在线学习，能主动适应各种外部环境和模型不确定性，提高旋翼无人机的鲁棒性和定位精度。

在针对高阶系统设计控制率时，为了降低设计难度，研究人员通常会使用线性滤波降阶的设计方法，即通过引入新的线性滤波器变量$\boldsymbol{r}=\dot{\boldsymbol{e}}+\boldsymbol{e}$，针对新的变量$\boldsymbol{r}$设计控制率，有效地降低了系统的阶次[13,17]。具体而言，设计以下线性滤波器变量作为强化学习网络的输入

$$r_x=e_x+\beta_x\dot{e}_x, r_y=e_y+\beta_y\dot{e}_y, r_z=e_z+\beta_z\dot{e}_z \tag{2}$$

其中，$\beta_i\in\mathbb{R}, i=x,y,z$表示滤波增益。为了进一步提高系统的定位精度，本文在线性滤波器变量式(2)的基础上设计集总误差为

$$\eta_x=r_x+k_x\int e_x\mathrm{d}t, \eta_y=r_y+k_y\int e_y\mathrm{d}t,$$
$$\eta_z=r_z+k_z\int e_z\mathrm{d}t \tag{3}$$

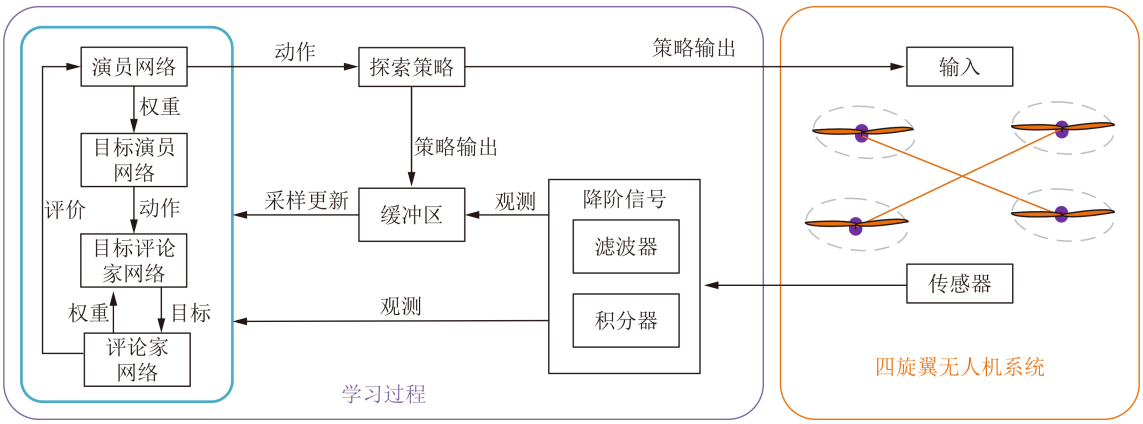


图 2 四旋翼无人机控制策略学习过程示意图

其中，$\eta_x,\eta_y,\eta_z\in\mathbb{R}$分别表示$x,y,z$ 3个坐标轴方向上的集总误差。$k_x,k_y,k_z\in\mathbb{R}$表示正的积分增益。通过引入位置积分信号设计集总误差式(3)，增强了策略对稳态误差的感知，有利于消除稳态误差，提高系统的定位精度。

注1 无人机位置部分的系统模型是2阶系统，其状态空间由系统的位置误差信号和速度误差信号组成，即$e_i,\dot{e}_i,i=x,y,z$组成的6维状态空间。传统的强化学习算法将在以上的6维状态空间中探索映射策略：每个时刻的控制量都在6维状态空间中生成。相比之下，本文基于线性滤波器式(2)提出的集总误差式(3)不仅将6维状态空间减少为3维，而且其中包含积分信号，增强了策略对稳态误差的感知。

基于信号式(3)，定义强化学习的输入信号和输出动作分别为：$\boldsymbol{s}_t=[\eta_x(t)\ \eta_y(t)\ \eta_z(t)]^{\top}$，$\boldsymbol{a}_t\in\mathbb{R}^3$。需要指出的是，强化学习需要和离散化后的系统交互，针对每个离散的时间步t设计，在文中下标t代表旋翼无人机和策略的交互时间点。本文的目标是实现旋翼无人机的准确定位，所以奖励函数设计为如下形式：$r_t=-e_x(t)^2-e_y(t)^2-e_z(t)^2$。基于该奖励函数，定义折扣奖励为

$$R_t=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots=\sum_{n=0}^{+\infty}\gamma^n r_{t+n} \tag{4}$$

其中，$\gamma\in(0,1)$表示折扣因子。奖励函数式(4)的优点是能准确地度量当前策略的定位性能，不足之处是该奖励机制依赖于误差信号的准确测量。该奖励机制中的参数γ表示未来的奖励在当前的权重，是表征策略短视还是远视的参数：当γ越接近0时，策略越在意短期奖励；越接近1，越关注长期奖励。本文通过权衡学习效率和跳出局部最优的能力，经过多次尝试后设置合适的γ。强化学习目标就是通过学习获得最大化折扣奖励的最优策略。首先，定义评价强化学习策略的性能指标函数$J(A^{\mu})$如下：$J(A^{\mu})=\mathbb{E}[R_1|A^{\mu}]$。根据确定性策略梯度定理，该性能指标函数对策略参数μ的梯度为

$$\nabla_{\mu}J(A^{\mu})=\mathbb{E}_{s\sim\rho^{A^{\mu}}}[\nabla_{\mu}A^{\mu}(s)\nabla_a Q^{A^{\mu}}(s,a)|_{a=A^{\mu}(s)}] \tag{5}$$

其中，$\rho^{A^{\mu}}$是遵循确定性策略的状态分布，按照该梯度方向更新参数μ，可以获得最大化折扣奖励的最优控制策略。但是式(5)中的动作值函数$Q^{A^{\mu}}$无法直接获得，需要由Critic网络进行估计。

注2 本文使用ω参数化的神经网络Q^{ω}估计真实的动作值函数$Q^{A^{\mu}}$。具体而言，通过时序差分算法对参数ω进行更新。时序差分误差$\delta_t=r_t+\gamma Q^{\omega}(s_{t+1},A^{\mu}(s_{t+1}))-Q^{\omega}(s_t,a_t)$表示目标值和Critic网络输出值$Q^{\omega}(s_t,a_t)$的差。通过最小化该误差，更新参数$\omega$，进而使得$Q^{\omega}$收敛到准确的评价。更多细节请见参考文献[20–22]。

根据式(5)和时序差分误差δ_t，按照梯度下降分别更新Actor和Critic网络的参数即可寻找到最优的策略。但是，对于旋翼无人机而言，由于系统的欠驱动、非线性和多变量性质，直接使用梯度下降算法往往探索效率低，策略收敛慢，难以保证学习过程的稳定性；此外，由于控制策略和无人机实时交互，相近时刻的交互数据关联性较大，往往不满足独立同分布，使用这些相关性大的数据更新网络参数不利于策略的收敛。为了减少训练数据之间的相关性和增加学习过程的稳定性，在训练过程中引入经验回放和目标网络技术。具体而言，经验回放技术是指构造一个足够大的回放缓冲区$\mathcal{B}$，该缓冲区由多个时间步的交互数据$(\boldsymbol{s}_t,\ \boldsymbol{a}_t,r_t,\ \boldsymbol{s}_{t+1}),t=1,2,\cdots$组成，缓冲区大小固定，每次有新的数据填充后，末尾的数据将被淘汰；在训练中，随机的从该缓冲区采样数据更新网络参数，即通过存储-采样的方式打破数据之间的关联性，提高训练效率。目标网络技术是指构造和Actor-Critic网络结构完全相同的网络(如图2所示)，以增加学习过程的稳定性。具体而言，在Critic更新时，其输出的值函数估计结果容易发生震荡，呈现出不稳定的行为。通过引入目标网络，利用平滑更新技术，避免了估计的值函数震荡，提高了旋翼无人机学习过程中的安全性。平滑更新的方式为

$$\omega'=(1-\tau)\omega'+\tau\omega,\mu'=(1-\tau)\mu'+\tau\mu \tag{6}$$

其中，$0<\tau\ll 1$表示平滑更新参数。通过选取较小的τ，目标网络的参数值会缓慢地跟踪学习到的策略，保证目标动作价值函数较为稳定，进而保证了学习过程的稳定。

注3 需要指出的是，本文引入缓冲区是为了打破数据之间的强相关性。由于训练数据是对控制系统的连续采样，相邻数据的关联性非常大。使用这样的数据训练会使得获得的策略过拟合到相近的数据上，而缺乏对整体系统动态的学习。从这方面来看，缓冲区越大，采样数据之间的关联性越小，也会更有利于学习策略的收敛。从另一方面来看，缓冲区越大，计算量也越大，也会导致遥远的交互数据加入到训练中，这些也是不利于训练的。所以，在考虑到数据的相关性和时效性，以及计算设备的运算能力后，经过大量的测试确定了本文的缓冲区长度。

注4 式(6)表示目标网络的参数更新，其中参

数τ直接决定了学习的稳定性。具体而言，Critic网络根据最小化损失函数δ_t更新参数，其中两部分都与当前的Critic网络Q^{ω}有关，进而导致Q^{ω}的更新会使得损失函数δ_t的波动很大，难以收敛。为了解决该问题，引入平滑变化的目标网络，保证学习的稳定性，即用$Q^{\omega'}$代替Q^{ω}，其中前者为使用式(6)平滑更新的目标网络。关于参数τ的选择：一方面，参数τ的取值应该远小于1，保证目标函数稳定；另一方面，过于小的τ也会减慢学习的速度。对于该参数的选取，要同时权衡学习稳定性和收敛的速度。本文在测试环节经过充分的测试，选定合适的参数τ。

在引入回放缓冲区和目标网络技术后，旋翼无人机的学习过程可以描述为：在每一个时间步t，旋翼无人机的状态为$\boldsymbol{s}_t$，Actor网络产生动作$\boldsymbol{a}_t$作用于无人机系统，系统返回下一个状态$\boldsymbol{s}_{t+1}$，并计算此时的奖励函数r_t，将交互信息$(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})$存入缓冲区$\mathcal{B}$，同时淘汰末尾的一组数据；在缓冲区随机选择$N$组数据$(\boldsymbol{s}_i, \boldsymbol{a}_i, r_i,\ \boldsymbol{s}_{i+1})$，$i = 1, 2, \cdots, N$用于更新策略参数。接下来给出本文的策略更新公式，首先定义Critic的损失函数为

$$L = \frac{1}{N}\sum_{i=1}^{N}[y_i - Q^{\omega}(\boldsymbol{s}_i, \boldsymbol{a}_i)]^2 \tag{7}$$

其中，y_i为目标网络输出的目标动作价值函数，表示如下：$y_i = r_i + \gamma Q^{\omega'}\left(\boldsymbol{s}_{i+1}, A^{\mu'}(\boldsymbol{s}_{i+1})\right)$，在此基础上按照梯度下降，Critic网络的参数更新为

$$\begin{aligned}&\omega_{t+1} = \omega_t + \alpha_{\omega}\nabla_{\omega}L,\\ &\nabla_{\omega}L = \frac{1}{N}\sum_{i=1}^{N}[y_i - Q^{\omega}(\boldsymbol{s}_i, \boldsymbol{a}_i)]\nabla_{\omega}Q^{\omega}\end{aligned} \tag{8}$$

其中，α_{ω}表示Critic网络的学习率。同理，按照梯度下降更新Actor网络的参数

$$\begin{aligned}&\mu_{t+1} = \mu_t + \alpha_{\mu}\nabla_{\mu}J(\mu),\\ &\nabla_{\mu}J(\mu) = \frac{1}{N}\sum_{i=1}^{N}[\nabla_{\mu}A^{\mu}(\boldsymbol{s}_i)\nabla_a Q^{\omega}(\boldsymbol{s}_i,\ \boldsymbol{a}_i)|_{a=A^{\mu}(\boldsymbol{s}_i)}]\end{aligned} \tag{9}$$

其中，α_{μ}表示Actor网络的学习率。为了保证旋翼无人机控制策略更新中的探索，本文将高斯噪声引入动作策略中：$\boldsymbol{a}_t^d = \boldsymbol{a}_t + \boldsymbol{d}_t$，其中$\boldsymbol{d}_t = [d_1\ d_2\ d_3]^{\mathrm{T}}$，$d_i, i = 1, 2, 3$分别服从$\mathcal{N}(0, \sigma_i^2)$的高斯分布，参数$\boldsymbol{\sigma} = [\sigma_1\, \sigma_2\, \sigma_3]$根据旋翼无人机的系统特性选择。

注5　在离散控制的强化学习中可以使用贪心策略探索动作空间，但在无人机等连续控制领域却不适用。为了使算法有足够多的试错尝试，本文的探索策略由探索噪声$\boldsymbol{d}_t$和动作策略$\boldsymbol{a}_t$组成。噪声大小的选择需要同时考虑探索效率和学习稳定性，过大的噪声往往使算法难以收敛，反之过小的噪声往往学习效率较低，而且较易陷入局部最优。本文在测试环节经过充分的测试，权衡学习效率和收敛性，选择合适的噪声值。

至此，本文的强化学习策略设计完成。通过在线更新深度网络的参数μ和ω，即可得到最大化折扣奖励的最优策略。就深度强化学习算法来说，策略的收敛依赖于有效探索。但是对于旋翼无人机的学习控制问题而言，连续系统的复杂非线性动态和高维的状态空间对学习策略的探索效率有较大的影响。针对该问题，为提高算法的探索效率，本文首先引入基于模型的几何控制器[26]作为系统先验部分，减轻复杂系统动态对学习策略的影响；其次提出集总误差式(3)将探索的状态空间由6维减少到3维，减小高维状态空间对学习策略的影响。在此基础上，本文的控制器设计为

$$f = (mg\boldsymbol{e}_3 - \boldsymbol{K}_1\boldsymbol{e}_1 - \boldsymbol{K}_2\boldsymbol{e}_2 + \boldsymbol{a}_t^d)^{\top}\boldsymbol{R}\boldsymbol{e}_3 \tag{10}$$

其中，$\boldsymbol{e}_1 = [e_x\, e_y\, e_z]^{\top}$，$\boldsymbol{e}_2 = [\dot{e}_x\, \dot{e}_y\, \dot{e}_z]^{\top}$分别表示系统的位置误差和速度误差；$\boldsymbol{K}_1, \boldsymbol{K}_2 \in \mathbb{R}^{3\times 3}$表示控制增益；$\boldsymbol{R}$和$\boldsymbol{e}_3$参见第2节中的定义。$mg\boldsymbol{e}_3 - \boldsymbol{K}_1\boldsymbol{e}_1 - \boldsymbol{K}_2\boldsymbol{e}_2$部分表示基于模型的几何控制器，$\boldsymbol{a}_t^d$部分表示强化学习策略输出的动作。

注6　与大部分现有方法不同，所提出的方法结合了基于模型控制器和基于学习策略的优点。其中几何控制器和强化学习策略相互补充，克服了几何控制器对干扰敏感和学习策略难以高效探索的缺点。由于四旋翼无人机系统的位置和姿态动力学相互耦合，本文在控制器式(10)中引入$\boldsymbol{R}\boldsymbol{e}_3$，用以实现准确的位置控制，同时保证整个系统的稳定性。就本文而言，在控制器的设计过程中没有引入相关的保守性步骤。因为依靠策略的泛化能力，该控制器可以智能地应对不同的工况，始终输出可靠的控制决策。

4　训练与实验结果分析

本节的主要目的是通过仿真训练使得强化学习网络能够快速学习到旋翼无人机系统的动力学特性，形成智能稳定的闭环系统，保证生成的策略可以在实际系统中安全应用，并将所提方法和现有智能控制方法对比，验证本文方法的有效性。本文共进行了4组测试。(1) 在第1组测试中，将根据表1中的算法训练本文所提策略，并和现有的先进智能控制方法对比。(2) 在第1组测试的基础上，在第2组测试中对比训练所得到的5种策略的控制性能，本组控制测试中的干扰包括两部分：第1部分是未建模动态和未整定的系统参数，其中主要包括未建模

的空气阻力模型和未准确整定的系统质量；第2部分是环境风扰动。(3) 在第2组对比的基础上，第3组测试将改变初始位置，以验证策略对不同初始位置的适应能力；并且在该组测试中额外加入两次突然的瞬时强干扰，以验证策略对强未知干扰的鲁棒性。本组测试中的干扰在第2组的基础上增加了瞬时强干扰。(4) 在以上3组测试的基础上，第4组测试将所提出的算法在自制的实验平台上验证其控制性能。本组测试中的未建模动态主要包括：电机的模型未知，桨叶的空气动力学未知，质量分布不均和参数测量误差等；外部的干扰主要是由风扇提供的未知风力干扰和电池电压波动等。

4.1 训练环境

本文的训练环境基于Ubuntu 16.04系统开发。具体而言，强化学习网络使用TensorFlow搭建，Actor网络的结构共4层：输入层的节点数为3，第2层和第3层的节点数都为128，输出层的节点数为3；Critic网络的结构共5层：输入层的节点数为6，第2, 3, 4层的节点数分别为30, 128, 128，输出层的节点数为1。无人机的物理模型来自RotorS Simlator[27]，型号为Pelican。在Gazebo软件中进行动力学交互，所有的通讯都是基于机器人操作系统(Robot Operating System, ROS)。本节的离线训练算法的参数选取如表2中所示，其中无人机的物理参数是根据实验平台的参数设置。值得一提的是训练中的控制增益是通过单独调节获得的。具体而言，在加入强化学习之前，本文先把几何控制器的参数调节到最佳，调节过程可参考比例微分(Proportional-Differential, PD)控制器。本文的仿真和实际实验中的控制信号频率均设为50 Hz。从实验结果来看，该频率能满足实际控制的需要，保证无人机的定位误差可以快速收敛。此外，在训练环境中加入固定大小的风作为扰动，其值为：[–0.6, 0.8, 0] m/s。在训练中，基于模型的控制器参数已经调整到最优，但是系统的参数m没有准确地补偿，作为系统的不确定性，以验证强化学习策略的抗不确定性能力。在每个回合的最后，会执行评价回合，用以评价这个回合的学习情况。在评价回合中，网络不进行训练，探索值为0，仅仅使用当前的策略完成控制任务，通过任务的完成情况评价策略的优劣。

表 1 强化学习训练-评价算法

随机初始化评论家和演员并且以相同的参数初始化对应的目标网络
初始化回放缓冲区

for $i = 1$ to 500 **do**
  随机初始化无人机位置，观测初始状态
  for $j = 1$ to 500 **do**
    根据控制器式(10)生成控制信号作用于无人机
    观测奖励值和下一状态
    将当前的交互数据保存在回放缓冲区中
    随机从缓冲区采样一组数据
    根据式(8)和式(9)更新评论家和演员网络
    根据式(6)更新目标网络
    if $j = 500$ **do**
      for $k = 1$ to 200 **do**
        测试当前策略
      end for
    end if
  end for
end for

4.2 测试1

为了充分验证本文所提策略的有效性，本文对比了确定性策略梯度方法(DDPG)[22,23]，积分补偿的确定性策略梯度方法(Deterministic Policy Gradient with Integral Compensator, DPGIC)[24]，本文所提框架式(10)下的确定性策略梯度方法(Geometric Control-DDPG, GC-DDPG)和积分补偿的确定性策略梯度方法(Geometric Control-DPGIC, GC-DPGIC)。为了保证对比的公平性，所有策略的训练参数和网络结构都相同。为了充分地展示该神经网络的训练过程，本文对每一种方法都做了10次重复训练，初始和期望位置分别设置为：[0, 1.2, 0.5] m，

表 2 系统的参数

参数	值	参数	值
m	1.6 kg	$k_i, i = x, y, z$	0.1, 0.1, 0.1
$\boldsymbol{J}$	diag[0.01, 0.01, 0.02] kg · m^2	$\boldsymbol{K}_1$	diag[3.8, 3.8, 3.5]
τ	0.001	$\boldsymbol{K}_2$	diag[5.0, 5.0, 4.5]
g	9.8 m/s^2	$\mathcal{B}$	10000
$\boldsymbol{\sigma}$	[0.1, 0.1, 1.0]	γ	0.95
$\beta_i, i = x, y, z$	0.1, 0.1, 0.1	N	64
α_ω	0.0001	α_μ	0.0001

[0, 0, 1.5] m。把所有评价回合的累计奖励和稳态误差记录下来，如图3(a)和图3(b)所示(为了更好地展示策略的收敛过程，本文展示了10500步以后的训练曲线，其中阴影部分表示10次训练的0.4倍标准差)。

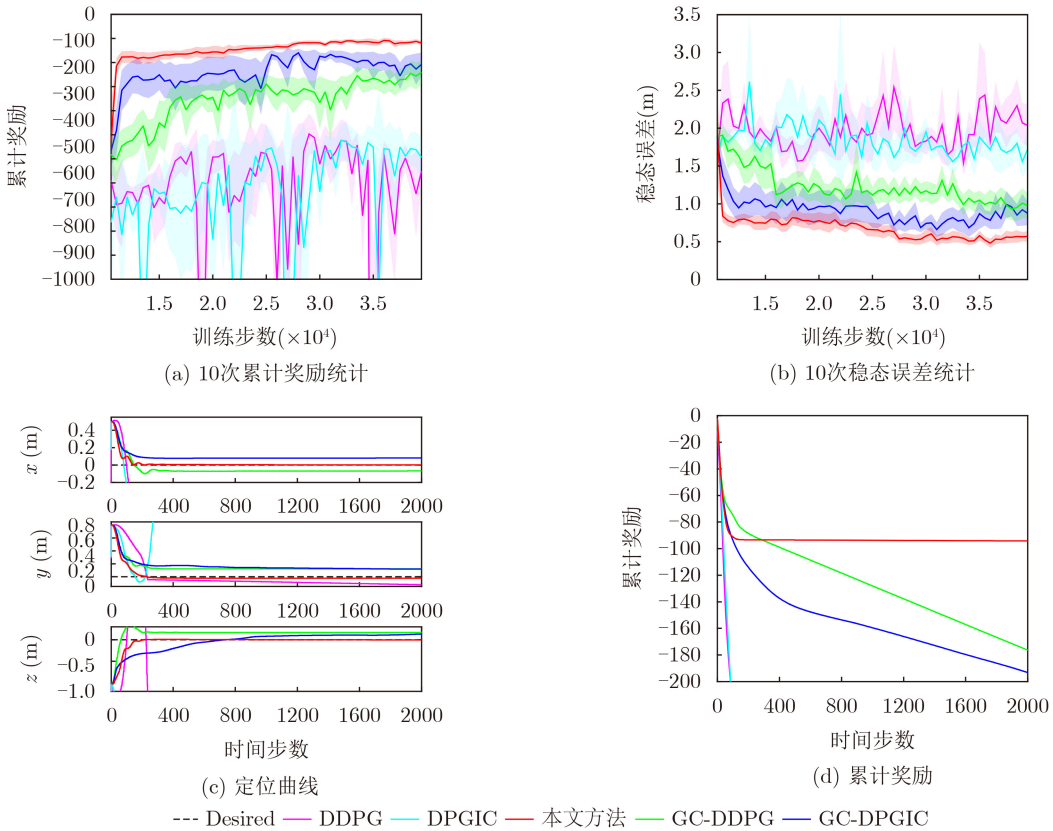
注7　需要指出的是，对于非线性系统的学习控制，网络初值的选取和优化目标函数往往与学习的性能有关。在本文中，为了保证对比的公平性，所有策略的训练参数、网络结构以及优化目标都相同。网络的初值都是随机产生，而且为了避免偶然因素的影响，充分地展示每一种方法的平均性能，本文对所有方法都做了10次重复训练，结果如图3(a)和图3(b)所示。

从图3(a)中可以看出，所提策略学习最快且效果最好。具体而言，所提方法的累计奖励函数在大约12000步训练以后就能探索到累计奖励较大的策略，并且随着训练次数的增加，系统的累计奖励也越来越大；相比之下，单纯使用DDPG和DPGIC策略很难在该环境下成功地学习到系统动态，这是由于系统的状态空间大，难以有效地探索，收敛到稳定的策略。相比于DDPG和DPGIC策略，GC-DDPG和GC-DPGIC策略有相对较好的学习效果。这是由于在本文的框架下，结合了基于模型的控制器，可以一定程度上增加系统的稳定性，使得学习的过程更稳定、高效。但是，由于GC-DDPG和GC-DPGIC策略的探索空间仍然是旋翼无人机的状态空间，系统动力学存在欠驱动、高度非线性、多变量等特点，强化学习算法的搜索空间大，策略收敛慢；累计奖励标准差较大，学习过程中的策略收敛会存在较大的波动。从图3(b)中可以看出，本文所提方法能最快地收敛到较小的稳态误差。综上所述，所提方法在策略收敛速度、学习效率、稳态误差、学习稳定性等方面都要优于对比方法。

以上的训练采用较大的探索值，使得网络参数能尽快地探索到最优策略附近，而且为了更好地展现训练的收敛步数，无人机的初始位置是固定的。为了进一步提高训练的充分性，在以上训练的基础上，本文进行了第2次训练。由于网络中已经学习到了系统的大部分动态，故本次训练采用较小的探索值 $\boldsymbol{\sigma}=$ [0.05, 0.05, 0.2]。训练的最大回合数和最大步数与上一次训练相同，并且此次训练的初始位置随机产生，以提高策略对不同初始位置的适应能力。训练所得的策略将在下面的小节中进行更详细的对比。

4.3　测试2

在本节中，为了验证所提方法的控制性能，本文对比了测试1中的5种智能方法。本次测试的初始位置和期望位置分别设置为：[0.5, 0.8, 0.5] m, [0, 0, 1.5] m。测试的误差曲线、累计奖励曲线如图3(c)、



(a) 10次累计奖励统计　(b) 10次稳态误差统计

(c) 定位曲线　(d) 累计奖励

图 3 5种学习方法的训练结果及测试2曲线图

图3(d)所示，此外本文统计了稳态区间(300～1800步)之间的最大稳态误差(Maximum Steady State Error, MSSE)和均方根误差(Root Mean Square Error, RMSE)，如表3所示。

从对比结果图3(c)中可以看出，在有外部干扰和存在系统不确定性的情况下，本文提出的强化学习策略能够将无人机快速准确地镇定到期望位置；其稳态误差是所有策略中最小的，说明本文的策略有较强的鲁棒性，瞬时和稳态性能较好。相比之下，DDPG和DPGIC策略由于没有成功地学习到系统动态，很难收敛到期望位置。GC-DDPG和GC-DPGIC策略可以完成控制任务，但是控制精度较差，系统的稳态误差没有完全消除。对比结果表明本文所提强化学习算法可以有效地提升旋翼无人机对干扰和模型不确定性的鲁棒性，提高系统的定位精度。从图3(d)中可以看出，本文所提策略的累计奖励最大，表明在该测试中本文策略要优于对比策略。从表3中可以看出，所提策略的最大稳态误差和均方根误差都最小，表明所提策略的稳态性能最好。该对比测试结果表明本文所提策略的定位精度高，稳态性能好。

4.4 测试3

在测试2的基础上，本节测试所提策略对瞬时强干扰的鲁棒性。为了验证策略对不同初始位置的适应能力，此次测试的初始位置和期望位置分别设置为：[–0.5, –0.8, 2.5] m, [0, 0, 1.5] m。在第800步和1600步时对系统X和Y 轴分别施加两次强风干扰，其大小为7 m/s,5 m/s，持续时间为1.5 s。5种智能控制方法的定位误差曲线和累计奖励曲线如图4(a)、图4(b)所示，表4统计了稳态区间(300～2500步)的最大控制误差和均方根误差。从图4(a)中可以看出，经过训练的强化学习策略能够将无人机快速准确地镇定到期望位置；在瞬时强干扰下，定位误差较小，并且在干扰消失后误差能够快速收敛。从图4(b)中可以看出，本文所提策略的累计奖励较大，在受到扰动时，累计奖励的波动较小，并且能够快速收敛，保证累计奖励始终较大，表明在该测试中本文所提策略要优于对比策略。从表4中可以看出，所提策略的最大稳态误差和均方根误差都较小，表明该策略对强干扰的鲁棒性好。该对比测试结果表明本文所提策略对外界干扰的鲁棒性强，能有效地抑制强干扰对系统的影响，提升了系统在强干扰环境下的控制性能。需要指出的是，由于系统的强耦合特性，在受到干扰时，对比策略在Z轴产生了较大的偏差。相比之下，本文所提的方法在Z轴几乎没有偏差，表现出对干扰更强的鲁棒性。

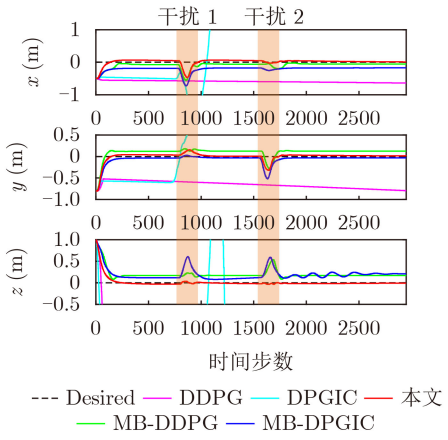
4.5 测试4

这一节将所提出的强化学习算法应用在实际的实验平台上，以验证其有效性。基于前3节的测试，对比的智能控制方法在收敛速度、控制精度、稳态误差和对干扰的鲁棒性等方面没有保证，安全性没有保障，应用于实际的无人机系统中存在安全隐患。故在本节中，将本文所提的方法和基于模型的几何控制器对比，以验证其在实际系统中的有效性。本文使用四旋翼无人机作为实验平台(如图5所示)，系统的状态由运动捕捉系统测量，所有的算法都是在机载电脑上实时处理。具体而言，四旋翼无人机的位置、速度、航向角和航向角速度信号由运动捕捉系统测量，通过WIFI发送至机载电脑；四旋翼无人机的其他姿态角和角速度信号都由机载惯性测量单元测量；在此基础上，基于强化学习的旋翼无人机系统实现实时反馈，并且不依赖任何外部计算设备。实验的初始位置和期望位置分别设置为：[0.2, 1.2, 1.0] m, [–0.2, 0.2, 1.2] m。本文中测试所用的机载电脑型号为：Intel STK2mv64cc，搭载酷睿m5处理器，基本频率1.1 GHz， 最大睿频2.8 GHz，运行内存4 GB。经测试该设备足以满足运行强化学习程序的要求，实现实时控制。

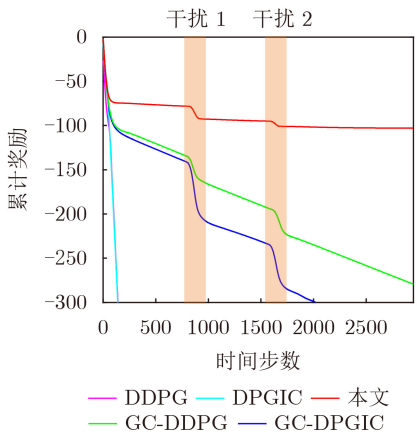
机载电脑和无人机的融合不会显著地影响无人机的工作效率。具体而言，本文中机载电脑和无人机的传输频率和传感器的采样频率基本一致，可以充分保证系统的工作效率。实验的曲线如图4(c) 、图4(d)所示，从中可以看出，尽管两种方法都能将无人机镇定到期望位置附近，但是几何控制器的鲁

表 3 测试2：最大稳态误差和均方根误差

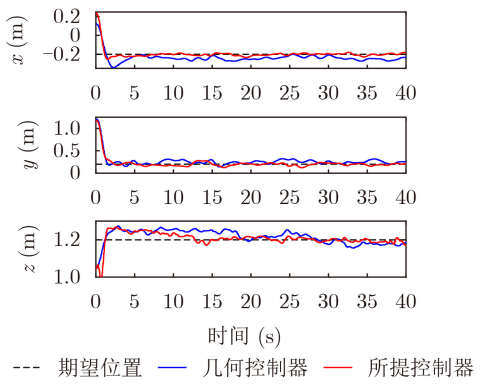
方法	X轴(m)		Y轴(m)		Z轴(m)	
	MSSE	RMSE	MSSE	RMSE	MSSE	RMSE
本文算法	0.0075	0.0035	0.0216	0.0212	0.0078	0.0034
DDPG	1.4904	1.3661	0.1140	0.0821	1.4002	1.4001
DPGIC	2.2356	2.0656	1.7060	1.6571	1.4003	1.3998
GC-DDPG	0.0716	0.0698	0.1297	0.1227	0.1740	0.1689
GC-DPGIC	0.0803	0.0788	0.1757	0.1439	0.2910	0.1114



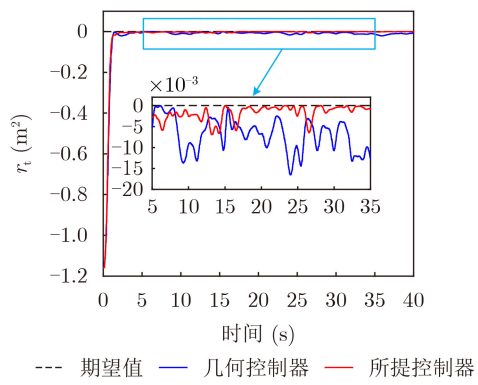
(a) 干扰下定位误差



(b) 干扰下累计奖励



(c) 硬件实验定位曲线



(d) 硬件实验奖励函数 r_t

图 4 测试3和测试4结果曲线图

表 4　测试3：最大稳态误差和均方根误差

方法	X轴(m)		Y轴(m)		Z轴 (m)	
	MSSE	RMSE	MSSE	RMSE	MSSE	RMSE
本文算法	0.4715	0.0887	0.3207	0.0647	0.0356	0.0203
DDPG	0.6239	0.5904	0.7480	0.6440	1.4002	1.4001
DPGIC	56.45	41.68	15.26	11.80	3.792	1.497
GC-DDPG	0.5543	0.1170	0.2918	0.1320	0.5382	0.1896
GC-DPGIC	0.7250	0.2228	0.5186	0.0912	0.6081	0.2014

图 5 实验平台

棒性较差，无人机在平衡点附近波动较大，特别是在X轴方向，其控制精度较差，这是由于基于模型的几何控制器对模型的依赖程度高，而实验中的未建模动态和外部干扰都会影响系统的定位精度。相比之下，本文所提策略能够抑制这些未建模动态和外部干扰对系统的影响，将无人机准确地镇定到期望位置，实现准确定位。稳态时(5～35 s，共30 s时长)的统计结果如表5所示。需要指出的是，在图4(c)中，Z轴方向的曲线收敛较X和Y轴慢，这是由旋翼无人机的动力学特性所决定的，即控制器必须准确的补偿无人机的重力才能实现准确的Z向定位，故精确补偿重力所消耗的时间会导致Z轴方向的曲线收敛较X和Y轴慢。

本文所提策略在以上4组测试中都取得了最佳

表 5 测试4：最大稳态误差和均方根误差

方法	X轴(m)		Y轴(m)		Z轴(m)	
	MSSE	RMSE	MSSE	RMSE	MSSE	RMSE
几何控制器	0.0730	0.0437	0.1221	0.0595	0.0670	0.0371
所提控制器	0.0340	0.0117	0.0802	0.0348	0.0535	0.0203

的控制效果。具体而言，本文设计的基于线性滤波的强化学习策略，有效提高了旋翼无人机对外界干扰和模型不确定性的鲁棒性，提高了旋翼无人机的定位精度。所有的测试都是在相同的条件下进行，充分说明了本文所提策略的优良性能。

5 结束语

针对具有未知外界扰动和未建模动态的四旋翼飞行器，本文提出一种基于线性滤波的强化学习策略，提升了系统对未知干扰的适应能力和定位精度。首先，设计线性滤波器变量作为强化学习网络的输入，减小了最优策略的探索空间，提高了探索效率。在此基础上，引入位置积分信号，增强策略对稳态误差的感知，提高系统的定位精度。通过训练生成的智能控制策略对外界扰动和未建模动态有很强的鲁棒性。对比结果表明，本文所提的强化学习算法可以有效地应对多种工况，提升了系统的控制性能。

参 考 文 献

[1] 张坤, 高晓光. 未知风场扰动下无人机三维航迹跟踪鲁棒最优控制[J]. 电子与信息学报, 2015, 37(12): 3009–3015.
ZHANG Kun and GAO Xiaoguang. Robust optimal control for unmanned aerial vehicles' three-dimensional trajectory tracking in wind disturbance[J]. *Journal of Electronics & Information Technology*, 2015, 37(12): 3009–3015.

[2] 宋大雷, 齐俊桐, 韩建达, 等. 旋翼飞行机器人系统建模与主动模型控制理论及实验研究[J]. 自动化学报, 2011, 37(4): 480–495. doi: 10.3724/SP.J.1004.2011.00480.
SONG Dalei, QI Juntong, HAN Jianda, *et al*. Model identification and active modeling control for rotor fly-robot: Theory and experiment[J]. *Acta Automatica Sinica*, 2011, 37(4): 480–495. doi: 10.3724/SP.J.1004.2011.00480.

[3] 孟祥冬, 何玉庆, 韩建达. 接触作业型飞行机械臂系统的力/位置混合控制[J]. 机器人, 2020, 42(2): 167–178.
MENG Xiangdong, HE Yuqing, and HAN Jianda. Hybrid force/position control of aerial manipulators in contact operation[J]. *Robot*, 2020, 42(2): 167–178.

[4] 王诗章, 鲜斌, 杨森. 无人机吊挂飞行系统的减摆控制设计[J]. 自动化学报, 2018, 44(10): 1771–1780.
WANG Shizhang, XIAN Bin, and YANG Sen. Anti-swing controller design for an unmanned aerial vehicle with a slung-load[J]. *Acta Automatica Sinica*, 2018, 44(10): 1771–1780.

[5] 甄子洋. 舰载无人机自主着舰回收制导与控制研究进展[J]. 自动化学报, 2019, 45(4): 669–681.
ZHEN Ziyang. Research development in autonomous carrier-landing/ship-recovery guidance and control of unmanned aerial vehicles[J]. *Acta Automatica Sinica*, 2019, 45(4): 669–681.

[6] 赵太飞, 宫春杰, 张港, 等. 一种无人机集群安全高效的分区集结控制策略[J]. 电子与信息学报, 2021, 43(8): 2181–2188. doi: 10.11999/JEIT200601.
ZHAO Taifei, GONG Chunjie, ZHANG Gang, *et al*. A safe and high efficiency control strategy of unmanned aerial vehicles partition rendezvous[J]. *Journal of Electronics and Information Technology*, 2021, 43(8): 2181–2188. doi: 10.11999/JEIT200601.

[7] 李瑞涵, 王耀南, 谭建豪. Nesterov加速梯度无人机姿态融合算法[J]. 机器人, 2018, 40(6): 852–859.
LI Ruihan, WANG Yaonan, and TAN Jianhao. Attitude fusion algorithm of UAV based on Nesterov accelerated gradient[J]. *Robot*, 2018, 40(6): 852–859.

[8] 高杨, 李东生, 程泽新. 无人机分布式集群态势感知模型研究[J]. 电子与信息学报, 2018, 40(6): 1271–1278. doi: 10.11999/JEIT170877.
GAO Yang, LI Dongsheng, and CHENG Zexin. UAV distributed swarm situation awareness model[J]. *Journal of Electronics & Information Technology*, 2018, 40(6): 1271–1278. doi: 10.11999/JEIT170877.

[9] ZHENG Dongliang, WANG Hesheng, WANG Jingchuan, *et al*. Toward visibility guaranteed visual servoing control of quadrotor UAVs[J]. *IEEE/ASME Transactions on Mechatronics*, 2019, 24(3): 1087–1095. doi: 10.1109/TMECH.2019.2906430.

[10] ZHANG Xuetao, FANG Yongchun, ZHANG Xuebao, *et al*. A novel geometric hierarchical approach for dynamic visual servoing of quadrotors[J]. *IEEE Transactions on Industrial Electronics*, 2020, 67(5): 3840–3849. doi: 10.1109/TIE.2019.2917420.

[11] MAHONY R and HAMEL T. Image-based visual servo control of aerial robotic systems using linear image features[J]. *IEEE Transactions on Robotics*, 2005, 21(2): 227–239. doi: 10.1109/TRO.2004.835446.

[12] LIU Hao, ZHAO Wanbin, ZUO Zongyu, *et al*. Robust control for quadrotors with multiple time-varying uncertainties and delays[J]. *IEEE Transactions on*

Industrial Electronics, 2017, 64(2): 1303–1312. doi: 10.1109/TIE.2016.2612618.

[13] HUA He'an, FANG Yongchun, ZHANG Xuetao, *et al*. Auto-tuning nonlinear PID-type controller for rotorcraft-based aggressive transportation[J]. *Mechanical Systems and Signal Processing*, 2020, 145: 106858. doi: 10.1016/j.ymssp.2020.106858.

[14] ZUO Zongyu and MALLIKARJUNAN S. L_1 adaptive backstepping for robust trajectory tracking of UAVs[J]. *IEEE Transactions on Industrial Electronics*, 2017, 64(4): 2944–2954. doi: 10.1109/TIE.2016.2632682.

[15] LV Zongyang, LI Shengming, WU Yuhu, *et al*. Adaptive control for a quadrotor transporting a cable-suspended payload with unknown mass in the presence of rotor downwash[J]. *IEEE Transactions on Vehicular Technology*, 2021, 70(9): 8505–8518. doi: 10.1109/TVT.2021.3096234.

[16] TIAN Bailing, YIN Liping, and WANG Hong. Finite-time reentry attitude control based on adaptive multivariable disturbance compensation[J]. *IEEE Transactions on Industrial Electronics*, 2015, 62(9): 5889–5898. doi: 10.1109/TIE.2015.2442224.

[17] XIAN Bin and HAO Wei. Nonlinear robust fault-tolerant control of the tilt trirotor UAV under rear servo's stuck fault: Theory and experiments[J]. *IEEE Transactions on Industrial Informatics*, 2019, 15(4): 2158–2166. doi: 10.1109/TII.2018.2858143.

[18] SHI Haobin, LI Xuesi, HWANG K S, *et al*. Decoupled visual servoing with fuzzy *Q*-learning[J]. *IEEE Transactions on Industrial Informatics*, 2018, 14(1): 241–252. doi: 10.1109/TII.2016.2617464.

[19] 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. doi: 10.1109/LRA.2017.2720851.

[20] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529–533. doi: 10.1038/nature14236.

[21] SILVER D, LEVER G, HEESS N, *et al*. Deterministic policy gradient algorithms[C]. The 31st International Conference on Machine Learning, Beijing, China, 2014: 387–395.

[22] LILLICRAP T P, HUNT J J, PRITZEL A, *et al*. Continuous control with deep reinforcement learning[C]. Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2016: 1–14.

[23] RODRIGUEZ-RAMOS A, SAMPEDRO C, BAVLE H, *et al*. A deep reinforcement learning strategy for UAV autonomous landing on a moving platform[J]. *Journal of Intelligent & Robotic Systems*, 2019, 93(1/2): 351–366.

[24] WANG Yuanda, SUN Jia, HE Haibo, *et al*. Deterministic policy gradient with integral compensator for robust quadrotor control[J]. *IEEE Transactions on Systems, Man, and Cybernetics:Systems*, 2020, 50(10): 3713–3725. doi: 10.1109/TSMC.2018.2884725.

[25] WEI Qinglai, WANG Lingxiao, LIU Yu, *et al*. Optimal elevator group control via deep asynchronous actor-critic learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(12): 5245–5256. doi: 10.1109/TNNLS.2020.2965208.

[26] LEE T, LEOK M, and MCCLAMROCH N H. Geometric tracking control of a quadrotor UAV on SE(3)[C]. The 49th IEEE Conference on Decision and Control, Atlanta, USA, 2010: 5420–5425.

[27] FURRER F, BURRI M, ACHTELIK M, *et al*. RotorS-a Modular Gazebo MAV Simulator Framework[M]. KOUBAA A. Robot Operating System (ROS): The Complete Reference (Volume 1). Cham: Springer, 2016: 595–625.

华和安：男，1995年生，博士生，研究方向为旋翼无人机的智能控制与规划.

方勇纯：男，1973年生，教授，研究方向为非线性控制、机器人视觉伺服、无人机和桥式吊车等欠驱动系统控制.

钱　辰：男，1993年生，博士生，研究方向为扑翼飞行器和其他仿生机器人的设计和控制.

张雪涛：男，1992年生，副教授，研究方向为自主旋翼无人机的运动计划，视觉伺服，状态和干扰估计.

责任编辑：余　蓉