文章编号:1671-4598(2022)10-0088-07 DOI:10.16526/j.cnki.11—4762/tp.2022.10.015 中图分类号:TP181;V279.2 文献标识码:A

基于深度强化学习的无人机空中目标自主跟踪

杨兴昊[1,2], 宋建梅[1], 佘浩平[1], 吴程杰[1], 杨钦宁[3], 付伟达[3]
(1. 北京理工大学 宇航学院, 北京 100081; 2. 中国航空系统工程研究所, 北京 100012;
3. 航天东方红卫星有限公司, 北京 100094)

摘要: 针对空中对接任务中的目标自主跟踪问题, 提出了一种基于深度强化学习的端到端的目标跟踪方法; 该方法采用近端策略优化算法, Actor 网络与 Critic 网络共享前两层的网络参数, 将无人机所拍摄图像作为卷积神经网络的输入, 通过策略网络控制多旋翼无人机电机转速, 实现端到端的目标跟踪, 同时采用 shaping 方法以加速智能体训练; 通过物理引擎 Pybullet 搭建仿真环境并进行训练验证, 仿真结果表明该方法能够达到设定的目标跟踪要求, 且具有较好的鲁棒性。

关键词: 深度强化学习; 近端策略优化; 无人机; 目标跟踪; 端到端

Autonomous Tracking of UAV Aerial Target Based on Deep Reinforcement Learning

YANG Xinghao[1,2], SONG Jianmei[1], SHE Haoping[1], WU Chengjie[1], YANG Qinning[3], FU Weida[3]
(1. School of Aerospace Engineering, Beijing Institute of Technology, Beijing 100081, China;
2. China Aero Institute of System Engineering, Beijing 100012, China;
3. DFH Satellite Co., Ltd., Beijing 100094, China)

Abstract: Aiming at the problem of target autonomous tracking in the process of aerial docking, an end-to-end target autonomous tracking method based on deep reinforcement learning is proposed. In this method, the near end strategy optimization algorithm is adopted. The Actor network and Critic network share the network parameters of first two floors. The image captured by unmanned aerial vehicles (UAV) is used as the input of convolution neural network. The motor speed of rotor UAV is controlled by the strategy network to achieve the end-to-end autonomous target tracking. At the same time, the shaping method is used to accelerate the agent training. The simulation environment is built by the engine of the Pybullet, and the training verification is carried out. The experimental results show that the method can achieve the set target tracking requirements and has good robustness.

Keywords: deep reinforcement learning; proximal policy optimization; UAV; target tracking; end-to-end

0 引言

无人机具有机动性强、成本低廉等优点, 广泛应用于边境巡逻、目标打击、遥感测绘、农业植保、电力巡线等领域, 但由于受载荷限制, 其续航时间较短。若无人机能够在空中实现加油或更换电池等操作, 则可以有效地提高无人机的续航时间和机动性能, 同时减少燃油或电池的重量能让无人机承载更多有效载荷, 从而提高其综合作业能力, 因此空中对接在未来将成为无人机的基本技能。

实现空中对接前, 主动对接无人机需要实现对目标无人机的持续跟踪, 并保证两架无人机的相对位置在对接要求范围内, 主动对接无人机称为主动无人机, 目标无人机称为被动无人机。整个空中对接过程包括主动无人机对被动无人机的识别与相对位姿解算、主动无人机的精准对接控制两部分。具体过程为: 当两架无人机相对距离较远时, 通过 GPS 获取被动无人机的位置信息, 并控制主动无人机接近目标。当两架无人机距离较近时, 即当被动无人机清晰地出现在主动无人机机载摄像头拍摄图像中时, 采用视觉算法对被动无人机进行识别与相对位姿解算。或采用精度较高的差分 GPS 获取被动无人机的位姿信息后进行对接控制。然而对接过程中 GPS 信号容易受到干扰, 导致精度下降, 因此目前的空中对接任务通常采用 GPS 与视觉混合的方式实现。

在无人机识别与相对位姿解算方面, 传统的空中对接相对位姿解算过程中往往依赖于对特定锥套的识别, 常用的方法包括: 对锥套圆环的颜色进行改变[1-2]或在锥套上安装红外 LED 信标[3-4]。单尧等人[5]利用四旋翼无人机搭建自主空中加油演示验证平台, 通过将无人机拍摄的锥套图像进行二值化处理后进行椭圆拟合, 通过拟合椭圆与实际锥套尺寸进行比较以解算位置信息, 并通过基于位置的 PID 控制器控制无人机进行对接。刘爱超等人[6]将对接装置的颜色与形状两种特征进行结合以设计对接锥套, 首先利用

收稿日期:2022-03-08; **修回日期**:2022-04-17。

作者简介:杨兴昊(1997-),男,河北承德人,硕士研究生,主要从事深度学习、强化学习在航空航天方向的研究。
通讯作者:佘浩平(1978-),男,江西修水人,博士,副教授,主要从事飞行器制导与控制、智能飞行控制方向的研究。

引用格式:杨兴昊,宋建梅,佘浩平,等. 基于深度强化学习的无人机空中目标自主跟踪[J]. 计算机测量与控制,2022,30(10):88-94,102.

被对接无人机的 GPS/INS 信息进行粗略导航，随后再利用视觉图标的颜色与形状信息实现精确导航，该方法能够保证在较高的飞行速度下仍然具有较好的跟踪效果，但当视觉图标出现遮挡、反光等问题时其跟踪效果有所下降。

在无人机精确对接控制方面，王宏伦等人[7]考虑了对接过程中的气流干扰与对接装置的自由摆动，针对空中加油对接段的精确控制问题，设计了基于线性二次线调节器的参考轨迹发生器和轨迹跟踪控制器，实验结果表明该方法具有快速性和一定的抗干扰能力，最终对接跟踪误差在 0.2 m 以内。李大伟等人[8]针对空中加油过程中软管锥套会受气动干扰而产生不规则摆动的问题，以线性二次调节器比例积分型控制器作为稳定闭环，并加入自适应控制器，从而提高控制过程中的抗干扰能力。黄永康等人[9]针对空中对接过程中纵向轨迹跟踪控制的时间滞后问题，提出一种基于直接升力的控制器，采用非线性 L1 制导的方法，并基于 ESO 的动态逆方法设计飞控系统，以消除纵向轨迹跟踪的时间滞后，实现纵向轨迹的快速响应。朱虎等人[10]提出基于 L1 自适应动态逆的无人机自主空中加油对接跟踪控制方法，根据时标分离的原则，采用动态逆方法设计姿态回路控制器，并在回路中加入 L1 自适应系统补偿气流干扰和系统误差，该方法所设计的控制系统能够有效消除逆误差和气流干扰的影响。钱素娟等人[11]针对高速飞行中存在的对接装置振动问题，提出了基于辅助视觉的飞行器空中加油对接过程控制方法，通过统计飞行器空中加油辅助视觉图像出现共现的频率，计算对应图像在所有图像中的共现度，获取对应图像的权重，实现关键帧图像定位，运用图像中心点空间位置，建立近距空中加油时的尾流流场的气动影响数学模型，从而完成飞行器空中加油接口的定位。实验结果表明，在飞行震动较大的情况下该方法对接控制的准确度高于传统算法。

近年来，随着机器学习算法的飞速发展，深度学习与强化学习等智能算法也被应用到空中对接任务中。S. Sun 等人[12]采用深度学习的方法实现对目标锥套的检测，同时完成相对位姿的解算。王宏伦等人[13]进一步研究了无人机软管式自主空中加油过程中的精准对接控制，利用 CFD 仿真获取气动数据，随后采用深度学习的方法对气动数据进行曲面拟合以获取干扰模型，并用循环神经网络预测对接装置的运动规律，从而显著提高了自主空中加油的对接精度。张易明等人[14]针对空中对接中的位置估计问题，提出了深度学习与双目视觉相结合的定位方法，对 YOLOv4-Tiny 进行改进，并在其基础上建立基于投影算子的模型参考自适应控制器，仿真结果表明该方法满足对接要求。王浩龙[15]采用近端策略优化控制方法以被动无人机的位置、速度等信息作为神经网络的输入实现飞行器的自主跟踪与对接任务。

上述研究将被动无人机的位姿估计与主动无人机的控制问题分开考虑，而本文研究了基于深度强化学习的无人机空中目标自主跟踪方法，实现了位姿估计与控制一体化，为空中目标跟踪问题提出了端到端的解决方案。采用近端策略优化算法（PPO，proximal policy optimization），将无人机搭载的摄像头拍摄的图像作为卷积神经网络的输入，不需在被动无人机上设置特定的视觉标识即可实现对空中目标的自主跟踪。

1　空中目标跟踪问题描述

1.1　空中目标跟踪的坐标关系与动力学建模

空中目标跟踪是实现空中对接的重要环节，空中对接任务由两架无人机配合完成，主动无人机需跟踪被动无人机一段时间，保证其相对位置与姿态在可对接范围内，随后控制主动无人机完成空中对接。本文主要研究空中对接前的空中目标跟踪，两架无人机的相对位置关系如图 1 所示，主动无人机在被动无人机后方跟随飞行。

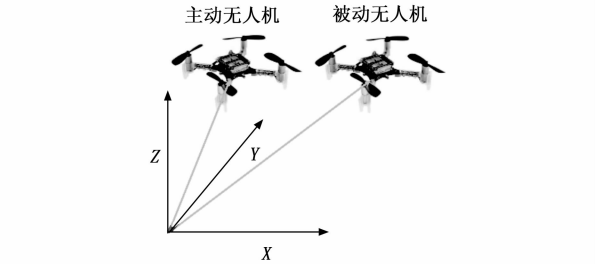


图 1　相对位置关系

主动无人机与被动无人机均为“X”型四旋翼无人机，其动力学方程为：

$$\begin{cases} \ddot{x} = \dfrac{(\cos\varphi\sin\theta\cos\psi + \sin\varphi\sin\psi)F}{m} \\ \ddot{y} = \dfrac{(\cos\varphi\sin\theta\sin\psi - \sin\varphi\cos\psi)F}{m} \\ \ddot{z} = \dfrac{\cos\varphi\cos\psi F - mg}{m} \\ \ddot{\varphi} = \dfrac{M_x + \dot{\theta}\dot{\psi}(I_y - I_z)}{I_x}L \\ \ddot{\theta} = \dfrac{M_y + \dot{\varphi}\dot{\psi}(I_z - I_x)}{I_y} \\ \ddot{\psi} = \dfrac{M_z + \dot{\theta}\dot{\varphi}(I_x - I_y)}{I_z} \end{cases} \tag{1}$$

式中，x、y、z 为无人机在世界坐标系下的位置坐标，φ、θ、ψ 分别为无人机的滚转角、俯仰角、偏航角，m 为无人机的质量，F 为无人机所受的合力值，I_x、I_y、I_z 为转动惯量，M_x、M_y、M_z 为螺旋桨升力产生的力矩。

本文选择“X”型四旋翼无人机，故其合力 F 与力矩 M_x、M_y、M_z 的计算方式为：

$$\begin{cases} F = \sum\limits_{i=1}^{4} F_i = \sum\limits_{i=1}^{4} k_F \omega_i^2 \\ M_x = -\dfrac{L}{\sqrt{2}}(F_1 - F_2 - F_3 + F_4) \\ M_y = -\dfrac{L}{\sqrt{2}}(F_1 + F_2 - F_3 - F_4) \\ M_z = \sum\limits_{i=1}^{4} M_i = \sum\limits_{i=1}^{4} (-1)^i k_M \omega_i^2 \end{cases} \tag{2}$$

式中，F_i 为各螺旋桨产生的升力，k_F 为螺旋桨升力系数，ω_i 为各电机转速，L 为电机到质心距离，M_i 为各螺旋桨产生的扭矩，k_M 为螺旋桨扭矩系数。

本文假设被动无人机沿世界坐标系的 x 轴正方向进行匀速直线飞行，其运动方程为：

$$\begin{cases}\dot{x}_{\text{target}} = v_{\text{target}} \\ \dot{y}_{\text{target}} = 0 \\ \dot{z}_{\text{target}} = 0\end{cases} \tag{3}$$

式中，x_{target}、y_{target}、z_{target} 为世界坐标系下被动无人机的坐标，为被动无人机初始速度。

1.2 无人机 PID 控制器与强化学习

PID 控制器具有算法简单、可靠性高的优点，因此在无人机控制中被广泛应用。作为无人机中的基本控制器，其结构如图 2 所示。

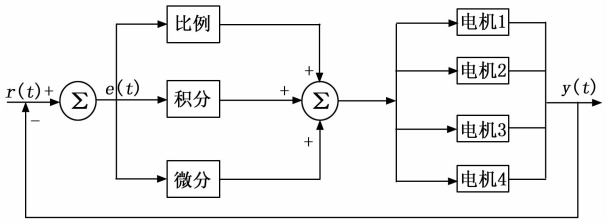


图 2 无人机 PID 控制器结构图

强化学习是研究如何使智能体在某一环境下获取最大奖励值的一类问题，该问题可以用马尔科夫决策过程表示。马尔科夫决策过程的基本元素包括：智能体、状态空间 S、动作空间 A、状态转移函数 $P(SP' \mid S, \alpha]$、奖励函数 R。智能体是在环境中进行学习的个体。状态空间 S 是对环境信息具体描述的集合，其中某一特定状态用 s 表示。智能体通常无法获得环境中的全部状态信息，因此智能体在环境下获得的部分状态信息也可用观测空间 O 进行描述。动作空间 A 是智能体在环境下能够完成的所有动作的集合，智能体的动作用 α 表示。状态转移函数 $P(SP' \mid S, \alpha]$ 为智能体在状态 s 下采取动作 α 后进入未来某一状态 s' 的概率。奖励函数 R 表示智能体在某一状态 s 下采取动 α 作后将获得多大的奖励值。

根据智能体所学习内容的差异，可分为基于策略的智能体和基于价值的智能体。基于策略的智能体直接学习策略函数 π，通过策略决定要采取的动作 α。策略函数 π 表示 t 时刻智能体在某一状态 s 下采取某一动作 α 的概率，即：

$$\pi(a \mid s) = P(a_t = a \mid s_t = s) \tag{4}$$

价值函数包括状态价值函数 $v^{\pi}(s)$ 和动作价值函数 $q^{\pi}(s, \alpha)$。状态价值函数 $v^{\pi}(s)$ 表示智能体在某种状态 s 时采用某种策略 π 后在未来能够获得多大的回报值 G。回报值 G 是指未来能够获得的所有奖励值 R 进行折扣后的和，即：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T \tag{5}$$

其中：G_t 表示 t 时刻获得的回报值，R_{t+1} 表示时刻获得的奖励值，T 表示结束时刻。γ 表示折扣因子，其取值范围在 0～1 之间，γ 取值越接近 0 时表示我们更加重视当前奖励，γ 取值为 1 时表示未来奖励与当前奖励同样重要。

状态价值函数 $v^{\pi}(s)$ 可表示为在状态 s 时采用策略 π 能够获得回报 G 的期望，即：

$$v^{\pi}(s) = E_{\pi}[G_t \mid s_t = s] \tag{6}$$

动作价值函数 $q^{\pi}(s, \alpha)$ 表示基于某种策略 π 时在某一状态 S 下采取某一动作 α 时所能获得回报 G 的期望，即：

$$q^{\pi}(s,a) = E_{\pi}[G_t \mid s_t = s, A_t = a] \tag{7}$$

策略函数 π、状态价值函数 $v^{\pi}(s)$ 与动作价值函数 $q^{\pi}(s, \alpha)$ 的关系可表示为：

$$v^{\pi}(s) = \sum_{a \in A} \pi(a \mid s) q^{\pi}(s,a) \tag{8}$$

基于价值的智能体通过学习价值函数从而隐式的学习策略，即从价值函数中推导出决定动作的策略。此时的策略 π_* 为采取价值函数取最大值时的动作，即：

$$\pi_* = \arg\max_{\pi} v^{\pi}(s) \tag{9}$$

或

$$\pi_* = \arg\max_{\pi} q^{\pi}(s,a) \tag{10}$$

此时需通过遍历所有状态 s 和动作 a 找到最大化的价值函数或动作价值函数。

状态空间 S 是对整个环境世界的完整描述，包含环境内的所有状态信息。而观测空间 O 是智能体对状态空间的部分描述，不一定包含所有信息。当智能体能够观测到全部状态信息时，称为完全可观测的，此时状态空间 S 与观测空间 O 等效。当智能体仅能观测到部分状态信息时，称为部分可观测的。本文中的智能体为四旋翼无人机，其无法感知环境中的全部状态信息，因此其基于观测空间 O 选取动作 a。下面依次介绍本文的观测空间 O 与动作空间 A 的设置。

观测空间 O：将主动无人机摄像头拍摄的图像作为强化学习的观测空间 O。与采用被动无人机位姿信息作为观测空间的方式不同，本文直接采用主动无人机摄像头拍摄所得的 RGB 图像作为观测空间。同时为加快训练速度，图像大小设置为 64×48。

动作空间 A：将经过 PID 控制输出的无人机期望速度作为强化学习的动作空间 A。在强化学习训练过程中的每一步均需要传递一个动作以控制智能体，本文中的动作空间为无人机在世界坐标系下的期望速度：

$$V = [v_x, v_y, v_z, v_M] \tag{11}$$

式中，v_x、v_y、v_z 为无人机在世界坐标系下单位速度向量沿 x 轴、y 轴、z 轴的分量，v_M 为期望速度的量纲。此处根据文献[16]中的方法通过包含位置与姿态控制的 PID 控制器将期望速度 V 转换为各电机转速 ω_i。

2 深度强化学习算法与网络结构

2.1 空中目标跟踪的深度强化学习描述

基于深度强化学习的空中目标跟踪任务框架如图 3 所示，主动无人机作为智能体根据从环境中获得的观测信息 O_t 产生动作 A_t 从而实现与环境的交互，环境受到智能体的动作影响后进入下一状态，同时智能体获得新的观测信息

O_{t+1}和奖励 R_{t+1}，智能体获得新的观测信息 O_{t+1}后产生新的动作 A_{t+1}，不断重复上述过程直至完成训练。

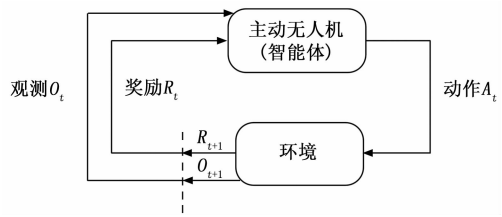


图 3　强化学习框架

2.2　近端策略优化

强化学习中常用的策略梯度算法存在不好确定学习率与步长的问题。当学习率或步长过大时，策略网络不断振动而无法收敛，当学习率或步长过小时，策略网络的训练时间又过长。近端策略优化算法通过控制新策略与旧策略的比值，从而限制新策略的更新幅度，使其训练过程更加稳定。

近端策略优化是一种演员－评论家方法（Actor-Critic）。将基于价值的智能体和基于策略的智能体进行结合，同时学习价值函数和策略函数并将学到的两种函数进行交互从而得到智能体的最佳行动，采用该方法能够有效地加快学习速度，取得更好的学习效果。在演员－评论家方法中，演员指的是策略函数，评论家指的是价值函数。当智能体训练完成后，策略函数用于决定智能体的实际动作，价值函数则不再起作用，其仅供训练时为策略函数打分，使其学习到能够尽可能获得更高汇报的策略函数。

此外，采用深度强化学习方法能够同时解决感知与控制问题。强化学习方法仅能够对维度较低的状态信息进行处理以实现对智能体的控制，而深度强化学习方法能够直接对高纬度信息进行处理从而实现对智能体的控制，即同时解决了感知与控制的问题，更加接近人脑的处理方式。

当被动无人机的 GPS 信息无法获取或其无明显对接标识供视觉算法解算位姿信息时，采用深度强化学习的方法将主动无人机拍摄图像作为观测空间以训练智能体能够同时解决感知与控制的问题，不再需要单独解算被动无人机的位置与姿态信息。因观测空间为图像信息，因此策略函数和价值函数均采用卷积神经网络，分别称为 Actor 网络和 Critic 网络。

近端策略优化算法在更新 Actor 网络时有两种方法：*KL* 惩罚和裁剪代理目标。下面分别介绍两种方法。

KL 惩罚的目标函数为：

$$L^{\mathrm{KLPEN}}(\theta)=\hat{E}_t\left[\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t\mid s_t)}\hat{A}_t-\beta KL[\pi_{\theta_{\mathrm{old}}}(\cdot\mid s_t),\pi_\theta(\cdot\mid s_t)]\right] \tag{12}$$

式中，θ 为 Actor 网络参数，π_θ 表示新策略，$\pi_{\theta_{\mathrm{old}}}$表示旧策略，$\hat{A}_t$ 为优势估计函数，$KL[\cdot]$ 表示 *KL* 散度，β 为 *KL* 散度权重系数。其中 *KL* 散度用于衡量新策略π_θ与旧策略 $\pi_{\theta_{\mathrm{old}}}$在状态 s_t 时采取动作间的差异，β可以进行动态调节。

裁剪代理目标的目标函数为：

$$L^{CLIP}(\theta)=\hat{E}_t[min(r_t(\theta)\hat{A}_t,clip(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t)] \tag{13}$$

式中，$r_t(\theta)$ 为新策略 π_θ 与旧策略 $\pi_{\theta_{\mathrm{old}}}$ 的比值，$clip(\cdot)$ 为裁剪函数，ε 为超参数。其中 $clip(\cdot)$ 裁剪函数表示当 $r_t(\theta)$ 小于时，函数输出值为 $1-\varepsilon$，当 $r_t(\theta)$ 大于 $1+\varepsilon$ 时，函数输出值为 $1+\varepsilon$，即将函数的输出值限定在 $1-\varepsilon$ 与 $1+\varepsilon$ 之间，如图 4 所示。其中 ε 需要人为调节。

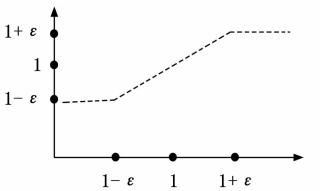


图 4　裁剪函数

2.3　空中目标跟踪的网络结构

观测空间 O 为主动无人机摄像头所拍摄图像，如图 5 所示，即网络输入为图像，图像为高维度信息需通过卷积神经网络进行特征提取，因此 Actor 网络和 Critic 网络均采用卷积神经网络。

图 5　对接无人机摄像头拍摄图像

文献［17］讨论了强化学习中“演员－评论家”类方法中 Actor 网络和 Critic 网络是否应分开的问题，结果表明当输入为图像等高维信息时，Actor 网络和 Critic 网络间的参数共享较为重要，能够有效提取特征，同时减少计算量。因此本文的 Actor 网络和 Critic 网络共享前几层的网络参数，其网络结构如图 6 所示。

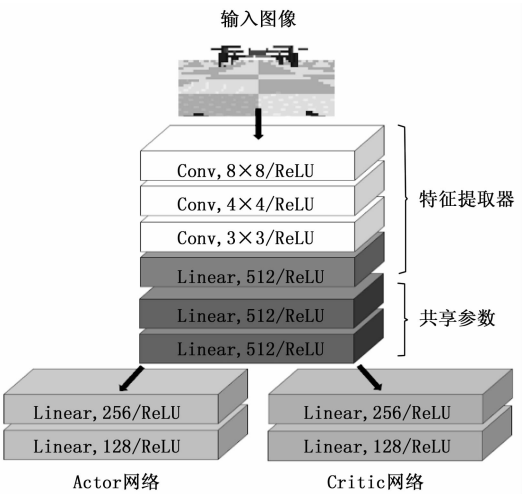


图 6　Actor-Critic 网络结构

网络主要由特征提取器和全连接神经网络两部分组成。特征提取器由三层卷积层构成，激活函数均为 ReLU 函数，

卷积核大小分别为 8×8、4×4、3×3，随后将卷积层输出数据拉平为一维数据，再经过节点数为 512 的线性层输入全连接神经网络。全连接神经网络首先由 Actor 网络和 Critic 网络共享的两层线性层构成，其节点数均为 512，随后网络进行分支，共享层的输出分别进入 Actor 网络和 Critic 网络。Actor 网络和 Critic 网络均由两层线性层构成，其节点数分别为 256 和 128，Actor 网络的输出为无人机 4 个电机的转速，Critic 网络的输出为该电机转速下的价值函数。

2.4 空中目标跟踪的奖励函数

强化学习中需要设置合适的奖励函数以使得智能体能够完成所期望的目标任务，但在空中目标跟踪任务中如果仅采用最终是否跟踪成功作为奖励函数则会导致奖励太过于稀疏，使得最终训练效果不佳、训练速度缓慢。文献[15]中通过利用 shaping 的方法加速智能体训练，有效地解决了环境中奖励稀疏的问题并成功实现了无人机目标跟踪。为加快智能体的训练速度，本文也采用 shaping 方法设计奖励函数，此时奖励函数为：

$$reward_t = shaping_t - shaping_{t-1} \tag{14}$$

式中，$reward_t$ 为 t 时刻的奖励函数，$shaping_t$ 为 t 时刻的 shaping 函数。即 t 时刻的奖励函数为 t 时刻的 shaping 函数值与时刻的 shaping 函数值的差。shaping 函数为：

$$shaping_t = -10\sqrt{(x_{\text{target}}-x_{\text{agent}})^2+(y_{\text{target}}-y_{\text{agent}})^2+(z_{\text{target}}-z_{\text{agent}})^2}+V_{\text{track}} \tag{15}$$

式中，x_{target}、y_{target}、z_{target} 表示被动无人机坐标，x_{agent}、y_{agent}、z_{agent} 表示主动无人机坐标。V_{track} 表示无人机完成达到空中目标跟踪要求时获得的奖励值，即：

$$V_{\text{track}}=\begin{cases}\text{if:} & |x_{\text{target}}-x_{\text{agent}}|<x_{\text{track}},\\ 10 & |y_{\text{target}}-y_{\text{agent}}|<y_{\text{track}},\\ & |z_{\text{target}}-z_{\text{agent}}|<z_{\text{track}}\\ 0 & \text{else}\end{cases} \tag{16}$$

式中，x_{track}、y_{track}、z_{track} 表示跟踪要求范围。

3 空中目标跟踪仿真实验

3.1 空中目标跟踪实验环境

本文基于文献[18]中提出的开源四旋翼强化学习仿真环境进行开发。基于 Pybuellt 搭建仿真环境，Pybullet 是基于 Bullet 进行物理仿真的 Python 模块。无人机模型选取“X”字型的四旋翼无人机 Crazyflie，其相关物理参数见表 1。

表 1 Crazyflie 物理参数

参数	数值
m/kg	0.027
L/m	0.039 7
k_f	3.16×10^{-10}
k_m	7.94×10^{-12}
I_x/(kg·m)	1.4×10^{-5}
I_y/(kg·m)	1.4×10^{-5}
I_z/(kg·m)	2.17×10^{-5}

同时为使摄像头拍摄图像中的被动无人机更易识别，将被动无人机进行适当倍数的放大。搭建的仿真环境如图 7 所示。

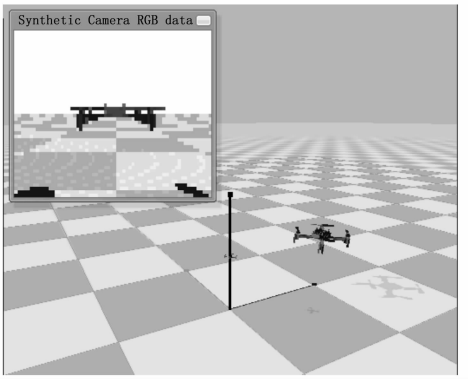


图 7 基于 Pybullet 的无人机跟踪仿真环境

硬件环境：GPU 为英伟达 2070 Super，CPU 为英特尔 i5-10400F。深度强化学习框架选择 Stable Baselines3，其是基于 Pytorch 和 Gym 的强化学习实现工具，集成了一系列强化学习经典算法，包括：A2C、DDPG、PPO、SAC、TD3 等。本文中强化学习算法选取近端策略优化（PPO）并采用裁剪代理目标的方式更新神经网络参数，相关超参数见表 2。

表 2 PPO 算法相关超参数

超参数	数值	超参数	数值
学习率	0.000 3	步数	1×10^{-7}
步长	2 048	衰减系数	0.99
批量	64	ε	0.2

仿真环境中被动无人机沿世界坐标系 x 轴的方向进行匀速直线飞行，其初始位置为：

$$\begin{cases}x_{\text{target}}=1\ \text{m}\\ y_{\text{target}}=0\\ z_{\text{target}}=0.5\ \text{m}\end{cases}$$

初始速度为：

$$\begin{cases}v_{x_{\text{target}}}=0.1\ \text{m/s}\\ v_{y_{\text{target}}}=0\\ v_{z_{\text{target}}}=0\end{cases}$$

主动无人机初始位置为：

$$\begin{cases}x_{\text{agent}}=0\\ y_{\text{agent}}=0\\ z_{\text{agent}}=0\end{cases}$$

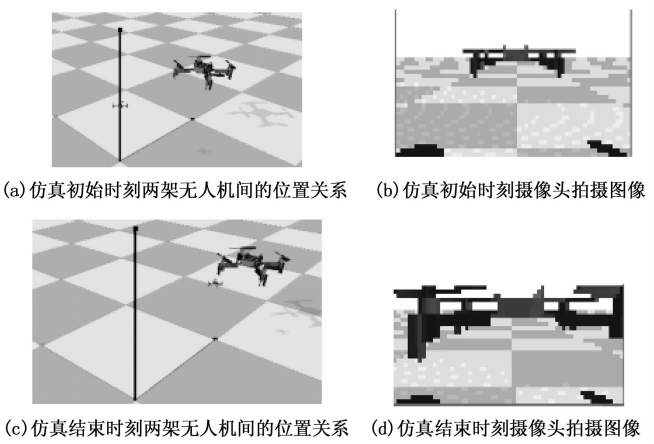
主动无人机摄像头拍摄的图像作为观测 O 输入 Actor 网络中，Actor 网络输出动作 A，A 为经 PID 控制后输出的 4 个电机转速，从而控制主动无人机自主跟踪被动无人机。跟踪条件为：

$$\begin{cases}x_{\text{track}}=1\ \text{m}\\ y_{\text{track}}=0.1\ \text{m}\\ z_{\text{track}}=0.1\ \text{m}\end{cases}$$

即主动无人机在世界坐标的 x 轴方向与被动无人机相对距离小于 1 m，在 y 轴和 z 轴方向距离小于 0.1 m 时，便视为跟踪成功，此时将获得数值为 10 的跟踪奖励 V_{track}。

3.2　空中目标跟踪实验结果分析

训练结束后利用训练过程中的最优模型进行仿真测试，仿真时间持续 6 s，测试情况如图 8 所示。图 8（a）为仿真初始时刻两架无人机间的位置关系，图 8（b）为初始时刻主动无人机摄像头所拍摄的图像，图 8（c）和图 8（d）分别为仿真结束时刻两架无人机间的位置关系与摄像头拍摄图像。由图 8 可以看出最终两架无人机保持着较近的跟踪距离[19-20]。



(a)仿真初始时刻两架无人机间的位置关系　(b)仿真初始时刻摄像头拍摄图像

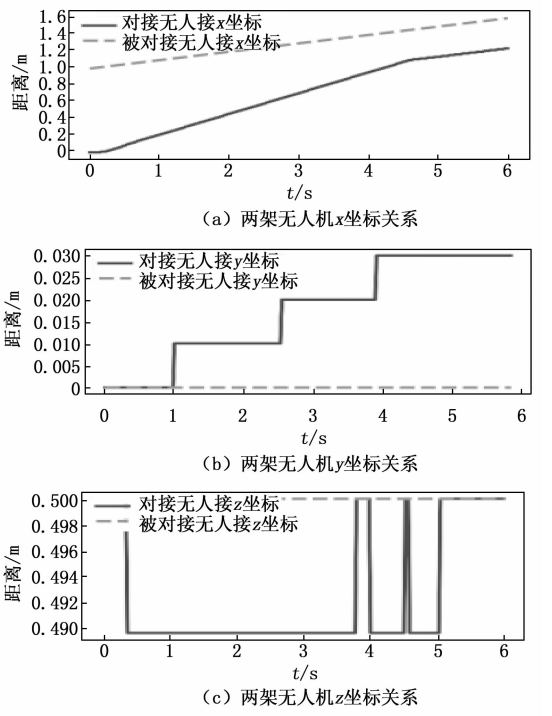
(c)仿真结束时刻两架无人机间的位置关系　(d)仿真结束时刻摄像头拍摄图像

图 8　仿真测试

空中目标跟踪仿真实验结果表明，本文采用 PPO 算法进行的空中目标跟踪能够达到跟踪要求，经过 6 s 的仿真测试最终沿世界坐标系 x 轴的跟踪距离最终保持在 0.5 m 左右，y 轴跟踪距离在 0.03 m 以内，z 轴跟踪距离在 0.01 m 以内。主动无人机与被动无人机间的位置关系与速度关系如图 9 和图 10 所示[21-22]。

从图 9 和图 10 可以看出被动无人机沿 x 轴正方向以 0.1 m/s 进行匀速直线运动，主动无人机沿 x 轴从静止开始加速到 0.25 m/s 匀速运动，当两架无人机间距离接近 0.5 m后进行减速，并保持与被动无人机相同速度进行飞行。同时主动无人机在 y 轴与 z 轴的位置与速度存在较小震荡，但最终基本稳定[23-24]。

主动无人机姿态角与角速度如图 11 和图 12 所示。由图 10（a）可以看出，初始阶段主动无人机需进行加速从而缩短与被动无人机间的距离，当速度达到 0.3 m/s 后进行减速，直至速度为 0.25 m/s 后进行匀速运动，当两架无人机距离达到 0.5 m 左右后减速至 0.1 m，并保持跟踪。因此主动无人机俯仰角与俯仰角速度需相应变化，由于本文的仿真环境中未考虑空气阻力，因此匀速运动状态姿态角均为 0°。但由图 11 和图 12 可知，姿态角在跟踪成功后存在一定波动，即不能保持与被动无人机的姿态角一致，因此在进行对接操作时还需考虑姿态角与姿态角速度的相对关系。

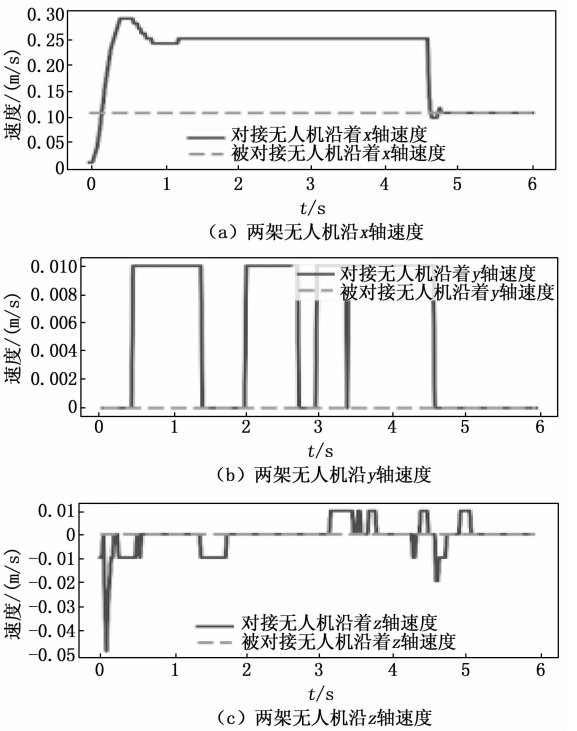


（a）两架无人机x坐标关系

（b）两架无人机y坐标关系

（c）两架无人机z坐标关系

图 9　两架无人机间的位置关系



（a）两架无人机沿x轴速度

（b）两架无人机沿y轴速度

（c）两架无人机沿z轴速度

图 10　两架无人机间的速度关系

仿真实验结果验证了采用深度强化学习的方法能够在仅输入主动无人机摄像头拍摄图像的情况下实现空中目标自主跟踪任务。该方法不需要通过 GPS 等传感器获得被动无人机的位置信息，也不需通过视觉算法对拍摄图像进行处理以解算被动无人机的位姿信息，通过端到端的方式即

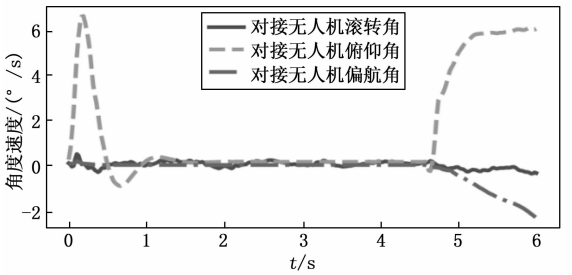


图 11 对接无人机姿态角变化

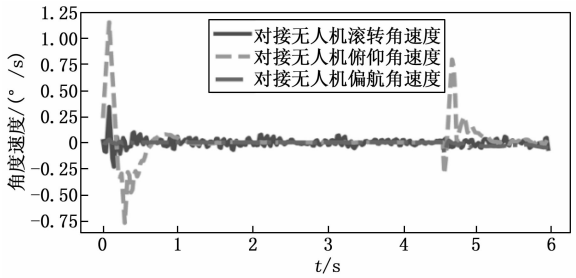


图 12 对接无人机角速度变化

可实现空中目标自主跟踪任务[25-26]。

4 结束语

本文针对空中目标自主跟踪问题，提出了一种基于深度强化学习的技术方法。主要贡献在于：提出了端到端的空中目标自主跟踪方法，通过将深度强化学习方法应用在空中目标自主跟踪领域，将无人机摄像头中拍摄图像作为输入，Actor 网络接受图像并输出 4 个电机的转速，从而控制无人机完成对空中目标的跟踪任务，无需获得目标位置信息，也不需额外设计图像处理算法，即可完成端到端的空中目标跟踪，提高了无人机的自主性与智能性。

本文仍有不足之处：实验中未考虑风阻力，也未考虑主动无人机与被动无人机间的相对姿态关系，仅适用于与相对姿态无关的空中目标跟踪任务，若需进行空中对接仍应进一步考虑两架无人机间的相对姿态与相对角速度关系。在未来的研究中将进一步考虑风阻力、两架无人机间的相对姿态与相对角速度关系。

参考文献：

[1] WANG X, KONG X, ZHI J, et al. Real-time drogue recognition and 3D locating for UAV autonomous aerial refueling based on monocular machine vision [J]. Chinese Journal of Aeronautics, 2015, 28 (6): 1667 - 1675.

[2] XU Y, DUAN H, LI C, et al. On-board visual navigation system for unmanned aerial vehicles autonomous aerial refueling [J]. Proceedings of the Institution of Mechanical Engineers, 2019, 233 (4): 1193 - 1203.

[3] CHEN S, DUAN H, DENG Y, et al. Drogue pose estimation for unmanned aerial vehicle autonomous aerial refueling system based on infrared vision sensor [J]. Optical Engineering, 2017, 56 (12): 124105 - 1 - 124105 - 12.

[4] 纪 超，王 庆. 基于双目视觉的自主空中加油算法研究与仿真 [J]. 系统仿真学报，2013 (6)：1327 - 1331.

[5] 单 尧，孙永荣，黄 斌，等. 自主空中加油飞行对接演示平台设计与实现 [J]. 电子测量技术，2016，39 (12)：176 - 179，188.

[6] 刘爱超，佘浩平，杨钦宁，等. 无人机空中对接中的视觉导航方法 [J]. 导航定位与授时，2019，6 (1)：28 - 34.

[7] 王宏伦，杜 熠，盖文东. 无人机自动空中加油精确对接控制 [J]. 北京航空航天大学学报，2011，37 (7)：822 - 826.

[8] 李大伟，王宏伦，盖文东. 基于 L1 自适应的自动空中加油对接段飞行控制技术 [J]. 控制理论与应用，2014 (6)：717 - 724.

[9] 黄永康，袁锁中，闫留浩. 基于直接升力的空中加油对接飞行控制 [J]. 兵工自动化，2021，40 (5)：62 - 67，93.

[10] 朱 虎，袁锁中，申 倩. 基于 L1 动态逆的自主空中加油对接控制 [J]. 兵工自动化，2018，37 (1)：19 - 23.

[11] 钱素娟，王水萍. 基于辅助视觉飞机空中加油对接优化过程仿真 [J]. 计算机仿真，2014，31 (8)：88 - 91，267.

[12] SUN S, YIN Y, WANG X, et al. Robust landmark detection and position measurement based on monocular vision for autonomous aerial refueling of UAVs [J]. IEEE Transactions on Cybernetics, 2018: 1 - 13.

[13] 王宏伦，刘一恒，苏子康. 无人机软管式自主空中加油精准对接控制 [J]. 电光与控制，2020，27 (9)：1 - 8.

[14] 张易明，艾剑良. 基于双目视觉的空中加油锥套定位与对接控制 [J]. 系统工程与电子技术，2021，43 (10)：2940 - 2953.

[15] 王浩龙. 飞行器自主对接的轨迹规划与近端策略优化控制方法 [D]. 北京：北京理工大学，2021.

[16] MELLINGER D, KUMAR V. Minimum snap trajectory generation and control for quadrotors [C] // IEEE International Conference on Robotics & Automation. IEEE, 2011: 2520 - 2525.

[17] COBBE K W, HILTON J, KLIMOV O, et al. Phasic policy gradient [C] //International Conference on Machine Learning, PMLR, 2021: 2020 - 2027.

[18] PANERATI J, ZHENG H, ZHOU S Q, et al. Learning to fly—a gym environment with pybullet physics for reinforcement learning of multi-agent quadcopter control [C] //2021 IEEE/RSJInternational Conference on Intelligent Robots and Systems (IROS), IEEE, 2021: 7512 - 7519.

[19] 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.

[20] RUBÍ B, MORCEGO B, PéREZ R. Deep reinforcement learning for quadrotor path following with adaptive velocity [J]. Autonomous Robots, 2021, 45 (1): 119 - 134.

[21] XU Y, LIU Z, WANG X. Monocular vision based autonomous landing of quadrotor through deep reinforcement learning [C] //37th Chinese Control Conference (CCC), IEEE, 2018: 10014 - 10019.

(下转第 102 页)