文章编号：2097-1974(2025)06-0001-09 DOI：10.7654/j.issn.2097-1974.20250601

基于改进型TD3强化学习的高速飞行器姿态控制

王伟丽，黄万伟，刘晓东，路坤锋，贾晨辉
（北京航天自动控制研究所，宇航智能控制技术全国重点实验室，北京，100854）

摘要：针对高速飞行器再入段面临的强非线性、高不确定性以及参数快时变等挑战，结合航天器智能化发展需求，提出了一种改进型的双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）端到端智能姿态控制方法。为解决TD3算法在姿态控制学习过程中存在训练不稳定、收敛困难的问题，在其马尔可夫决策过程中，设计了混合奖励机制，融合连续跟踪误差惩罚和稀疏任务完成奖励，协同引导智能体收敛；在其训练过程中，引入基于现代控制理论的先验知识约束，提出了基于行为克隆的Actor网络优化更新策略，以平衡专家经验模仿与累计回报最大化目标。仿真结果表明，在14种参数偏差组合的工况下，所提方法能够精确跟踪三通道姿态指令。

关键词：高速飞行器；姿态控制；深度强化学习；行为克隆；强适应控制
中图分类号：V448.2 **文献标识码**：A

Attitude Control of High-speed Vehicles Based on Improved TD3 Reinforcement Learning

WANG Weili, HUANG Wanwei, LIU Xiaodong, LU Kunfeng, JIA Chenhui
(National Key Laboratory of Science and Technology on Aerospace Intelligent Control, Beijing Aerospace Automatic Control Institute, Beijing, 100854)

Abstract: To address the challenges of strong nonlinearity, high uncertainty, and rapid time-varying parameters during the reentry phase of high-speed vehicles, this study proposes an end-to-end intelligent attitude control method based on an improved Twin Delayed Deep Deterministic Policy Gradient algorithm, aligned with the demands of intelligent spacecraft development. To overcome the issues of training instability and convergence difficulties in TD3-based attitude control learning, two key innovations are introduced: a hybrid reward mechanism combining continuous tracking error penalties and sparse task-completion rewards is designed within the Markov Decision Process framework to synergistically guide agent convergence. Prior knowledge constraints derived from modern control theory are incorporated into the training process, proposing a behavior cloning-based optimization strategy for the Actor network to balance expert experience imitation and cumulative reward maximization. Simulation results show that the proposed method can accurately track the three-channel attitude commands under 14 combinations of parameter deviations.

Keywords: high-speed vehicles; attitude control; deep reinforcement learning; behavior cloning; strongly adaptive control

0 引 言

高速飞行器凭借其全空域机动、宽速域巡航和强突防能力的优势，已成为现代远程精确打击体系的核心装备[1]。然而，其飞行包线内表现出的强非线性动力学特性、多通道耦合效应以及由气动热或结构形变引发的快时变参数，使得传统基于精确数学模型的控制器设计方法面临理论局限[2]。尤其在再入段高动态环境下，飞行器同时承受极端气动载荷、复杂干扰和模型不确定性，基于先验知识的经典控制方法（如增益调度PID、鲁棒自适应控制）进而设计可靠的姿态控制器变得困难。

为应对这些挑战，并响应航天器智能化发展趋势[3]，智能飞行控制（Intelligent Flight Control，IFC）技术应运而生。其中，数据驱动方法因其对模

收稿日期：2025-07-05；修回日期：2025-09-15
基金项目：国家自然科学基金(U24B20159)

型依赖程度低的优势，正逐渐成为IFC领域的研究重点。深度强化学习（Deep Reinforcement Learning，DRL）因其特有的“环境交互-自主优化”机制，展现出解决复杂控制问题的独特潜力[4-6]。目前DRL在高速飞行器控制中的应用主要呈现3个研究方向[7]：控制参数自适应整定[8-9]、不确定性补偿控制[10]以及端到端自主控制[11]，形成了该领域新的技术突破点。

基于DRL的端到端控制架构通过直接从原始传感器数据学习控制策略以生成执行器命令，无需人工特征提取或控制律设计，实现高速飞行器主控制律的高自主设计，有效降低对其地面设计模型的依赖程度。然而，DRL在高速飞行器再入段三通道姿态控制中的应用研究仍处于探索阶段，其核心挑战源于两大特性：一是再入过程中气动参数与惯性参数的快时变特性导致系统动力学环境剧烈变化；二是三通道间的强耦合效应使得控制策略的训练难以稳定收敛。针对这些问题，现有研究主要从两个方向突破：在算法层面，学者们通过优化网络架构设计[12]等方式和改进经验回放机制[13]来提升收敛性能；在训练策略层面，结合行为克隆（Behavior Cloning，BC）技术以修正策略偏差[14]。

综上所述，本研究针对高速飞行器再入段端到端姿态控制问题，创新性提出了基于知识引导的双延迟深度确定性策略梯度（Knowledge-Guided Twin Delayed Deep Deterministic Policy Gradient，KG-TD3）算法，该算法通过融合现代控制理论中的先验知识与双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）算法的数据驱动特性，构建了新的混合驱动控制架构。这种知识嵌入式的DRL范式不仅能够解决纯数据驱动方法在复杂动态环境中的训练不稳定问题，同时保留了端到端控制的自适应优势，为高速飞行器的智能控制提供了新的技术途径。

1 模型建立

本文研究对象为升力式面对称无动力高速飞行器，其采用BTT控制模式，控制物理量为攻角α、侧滑角β和速度倾侧角γ_v。典型的无动力高速飞行器有美国的HTV-2、SR-72等，如图1所示。

根据升力式面对称高速飞行器自身及其再入段飞行环境特点，提出如下可行性假设条件：忽略地球自转的影响，此时发射惯性坐标系与地面坐标系始终重合，而且不需考虑离心惯性力和哥氏惯性力的作用；将地球视为均质圆球，忽略地球扁率以及切向引力加速度的影响；惯量积$J_{x_1y_1}$为小量，且忽略不计；将飞行器视为刚体，即不考虑弹性影响。

HTV-2

SR-72

图1 典型的无动力高速飞行器

Fig.1 Typical unpowered high-speed vehicle

参考文献［15］建立了飞行器六自由度数学模型，并写为如式（1）所示的仿射非线性形式，以便于控制系统设计。

$$\begin{cases}\dot{\boldsymbol{x}}_{\Omega}=\boldsymbol{F}_{\Omega}+\boldsymbol{G}_{\Omega}\boldsymbol{x}_{\omega}+\boldsymbol{d}_{\Omega}\\ \dot{\boldsymbol{x}}_{\omega}=\boldsymbol{F}_{\omega}+\boldsymbol{G}_{\omega}\boldsymbol{u}+\boldsymbol{d}_{\omega}\end{cases}\tag{1}$$

式中 $\boldsymbol{x}_{\Omega}$为姿态环状态量，即攻角α、侧滑角β、速度倾侧角γ_v；$\boldsymbol{x}_{\omega}$为姿态角速度环状态量，即滚转角速度ω_x、偏航角速度ω_y、俯仰角速度ω_z；$\boldsymbol{u}$表示控制输入，即滚转舵偏角δ_x、偏航舵偏角δ_y、俯仰舵偏角δ_z；$\boldsymbol{F}_{\Omega}$，$\boldsymbol{F}_{\omega}$表示受控对象的集中动力学；$\boldsymbol{G}_{\Omega}$，$\boldsymbol{G}_{\omega}$为控制信号的增益，描述了控制信号对系统动力学的影响；$\boldsymbol{d}_{\Omega}=\begin{bmatrix}d_{\alpha} & d_{\beta} & d_{\gamma_v}\end{bmatrix}^{\mathrm{T}}$，$\boldsymbol{d}_{\omega}=\begin{bmatrix}d_{\omega_x} & d_{\omega_y} & d_{\omega_z}\end{bmatrix}^{\mathrm{T}}$为外界干扰。

在实际工程中，由于高速飞行器飞行过程中存在参数摄动大、外界干扰严重等问题，则将飞行器模型写为如式（2）所示的仿射非线性系统：

$$\begin{cases}\dot{\boldsymbol{x}}_{\Omega}=\bar{\boldsymbol{F}}_{\Omega}+\bar{\boldsymbol{G}}_{\Omega}\boldsymbol{x}_{\omega}+\varsigma_{\Omega}\\ \dot{\boldsymbol{x}}_{\omega}=\bar{\boldsymbol{F}}_{\omega}+\bar{\boldsymbol{G}}_{\omega}\boldsymbol{u}+\varsigma_{\omega}\end{cases}\tag{2}$$

式中 $\boldsymbol{F}_{\Omega}=\bar{\boldsymbol{F}}_{\Omega}+\Delta\boldsymbol{F}_{\Omega}$，$\boldsymbol{F}_{\omega}=\bar{\boldsymbol{F}}_{\omega}+\Delta\boldsymbol{F}_{\omega}$，$\boldsymbol{G}_{\Omega}=\bar{\boldsymbol{G}}_{\Omega}+\Delta\boldsymbol{G}_{\Omega}$，

$\boldsymbol{G}_\omega=\bar{\boldsymbol{G}}_\omega+\Delta\boldsymbol{G}_\omega$。$\bar{\boldsymbol{F}}_\Omega$，$\bar{\boldsymbol{G}}_\Omega$是姿态环标称状态下的已知模型，$\bar{\boldsymbol{F}}_\omega$，$\bar{\boldsymbol{G}}_\omega$是姿态角速度环标称状态下的已知模型，具体如式（3）和式（4）所示。$\Delta\boldsymbol{F}_\Omega$，$\Delta\boldsymbol{G}_\Omega$，$\Delta\boldsymbol{F}_\omega$，$\Delta\boldsymbol{G}_\omega$为未知动态。$\varsigma_\Omega$和$\varsigma_\omega$表示模型未知动态、参数摄动、外界干扰等系统未知项，$\varsigma_\Omega=\Delta\boldsymbol{F}_\Omega+\Delta\boldsymbol{G}_\Omega\boldsymbol{x}_\omega+\boldsymbol{d}_\Omega$，$\varsigma_\omega=\Delta\boldsymbol{F}_\omega+\Delta\boldsymbol{G}_\omega\boldsymbol{u}+\boldsymbol{d}_\omega$，该项的存在是导致系统性能下降的主要原因，需采用自适应强抗扰的控制器应对，本文采用的是KG-TD3智能控制器。

$$\begin{cases}\dot{\boldsymbol{x}}_\Omega=\begin{bmatrix}\dot\alpha & \dot\beta & \dot\gamma_v\end{bmatrix}^{\mathrm{T}}\\ \boldsymbol{x}_\omega=\begin{bmatrix}\omega_x & \omega_y & \omega_z\end{bmatrix}^{\mathrm{T}}\\ \bar{\boldsymbol{F}}_\Omega=\begin{bmatrix}-\dfrac{L-mg\cos\theta\cos\gamma_v}{mV\cos\beta}\\ \dfrac{Z+mg\cos\theta\sin\gamma_v}{mV}\\ \dfrac{\left[L\left(\tan\beta+\tan\theta\sin\gamma_v\right)+Z\tan\theta\cos\gamma_v-mg\cos\theta\tan\beta\cos\gamma_v\right]}{mV}\end{bmatrix}\\ \bar{\boldsymbol{G}}_\Omega=\begin{bmatrix}-\cos\alpha\tan\beta & \sin\alpha\tan\beta & 1\\ \sin\alpha & \cos\alpha & 0\\ \cos\alpha\sec\beta & -\sin\alpha\sec\beta & 0\end{bmatrix}\end{cases}\tag{3}$$

$$\begin{cases}\dot{\boldsymbol{x}}_\omega=\begin{bmatrix}\dot\omega_x & \dot\omega_y & \dot\omega_z\end{bmatrix}^{\mathrm{T}}\\ \boldsymbol{u}=\begin{bmatrix}\delta_x & \delta_y & \delta_z\end{bmatrix}^{\mathrm{T}}\\ \bar{\boldsymbol{F}}_\omega=\begin{bmatrix}\dfrac{J_y-J_z}{J_x}\omega_y\omega_z+\dfrac{\left(C_{mx}^{\alpha}\alpha+C_{mx}^{\beta}\beta\right)qSl}{J_x}\\ \dfrac{J_z-J_x}{J_y}\omega_x\omega_z+\dfrac{C_{my}^{\beta}\beta qSl}{J_y}\\ \dfrac{J_x-J_y}{J_z}\omega_x\omega_y+\dfrac{C_{mz}^{\alpha}\alpha qSl}{J_z}\end{bmatrix}\\ \bar{\boldsymbol{G}}_\omega=qSl\begin{bmatrix}\dfrac{1}{J_x} & 0 & 0\\ 0 & \dfrac{1}{J_y} & 0\\ 0 & 0 & \dfrac{1}{J_z}\end{bmatrix}\begin{bmatrix}C_{mx}^{\delta_x} & C_{mx}^{\delta_y} & C_{mx}^{\delta_z}\\ C_{my}^{\delta_x} & C_{my}^{\delta_y} & C_{my}^{\delta_z}\\ C_{mz}^{\delta_x} & C_{mz}^{\delta_y} & C_{mz}^{\delta_z}\end{bmatrix}\end{cases}\tag{4}$$

式中　m，V，θ分别为飞行器的质量、速度和弹道倾角；L，Z分别为气动升力和气动侧向力；g为重力加速度；M_{x1}，M_{y1}和M_{z1}分别为气动滚转力矩、偏航力矩和俯仰力矩；J_{x1}，J_{y1}和J_{z1}为飞行器的主转动惯量；q，S，l分别为动压、气动参考面积和参考长度；$C_{mx}^{(\cdot)}$，$C_{my}^{(\cdot)}$，$C_{mz}^{(\cdot)}$分别为相对于$(\cdot)$的滚动、偏航和俯仰力矩系数。

至此，面向控制的高速飞行器三通道姿态运动数学模型构建完成。接下来，将根据该模型研究知识与数据融合的智能姿态控制方法。

2　知识与数据融合的智能控制器设计

在深度强化学习中，TD3算法虽然在连续控制任务中表现出色，但仍面临探索效率低、训练初期不稳定以及局部最优陷阱等问题。为此，本文引入基于知识的控制器约束，即动态面控制器（Dynamic Surface Control，DSC）（本文将其定义为“专家控制器”），并结合行为克隆方法，提出了基于知识引导的TD3（KG-TD3）算法。其中，“知识”与“数据”均依据文献［16］界定。

2.1　基于知识的控制器设计

动态面控制器依赖于精确、解析的飞行器数学模型，故依据文献［16］可称为基于知识的控制器。

对于非线性系统，定义跟踪误差$s_{\boldsymbol{\Omega}}$和其微分：

$$\begin{cases}\boldsymbol{s}_\Omega=\boldsymbol{x}_\Omega-\boldsymbol{x}_{\Omega d}\\ \dot{\boldsymbol{s}}_\Omega=\bar{\boldsymbol{F}}_\Omega+\bar{\boldsymbol{G}}_\Omega\boldsymbol{x}_\omega+\varsigma_\Omega-\dot{\boldsymbol{x}}_{\Omega d}\end{cases}\tag{5}$$

式中　$\boldsymbol{x}_\Omega$为系统实际状态；$\boldsymbol{x}_{\Omega d}$为系统制导指令状态。

取虚拟控制输入$\boldsymbol{x}_v$：

$$\boldsymbol{x}_v=-\bar{\boldsymbol{G}}_\Omega^{-1}\left(\bar{\boldsymbol{F}}_\Omega+\boldsymbol{W}_\Omega\boldsymbol{s}_\Omega-\dot{\boldsymbol{x}}_{\Omega d}+\varsigma_\Omega\right)\tag{6}$$

经一阶滤波器，得到：

$$\boldsymbol{\tau}\dot{\boldsymbol{x}}_\tau+\boldsymbol{x}_\tau=\boldsymbol{x}_v\tag{7}$$

式中　$\boldsymbol{\tau}$为待设计的滤波器系数；$\boldsymbol{x}_\tau$为经过滤波器后的控制输入。

定义跟踪误差$\boldsymbol{s}_\omega$并对其微分：

$$\begin{cases}\boldsymbol{s}_\omega=\boldsymbol{x}_\omega-\boldsymbol{x}_\tau\\ \dot{\boldsymbol{s}}_\omega=\bar{\boldsymbol{F}}_\omega+\bar{\boldsymbol{G}}_\omega\boldsymbol{u}+\varsigma_\omega-\dot{\boldsymbol{x}}_\tau\end{cases}\tag{8}$$

从而设计最终的控制器$\boldsymbol{u}$：

$$\boldsymbol{u}=-\bar{\boldsymbol{G}}_\omega^{-1}\left[\bar{\boldsymbol{F}}_\omega+\boldsymbol{W}_\omega\boldsymbol{s}_\omega-\dot{\boldsymbol{x}}_\tau+\varsigma_\omega\right]\tag{9}$$

综上，动态面姿态控制律：

$$\begin{cases}\boldsymbol{s}_\Omega=\boldsymbol{x}_\Omega-\boldsymbol{x}_{\Omega d}\\ \boldsymbol{x}_v=-\bar{\boldsymbol{G}}_\Omega^{-1}\left(\bar{\boldsymbol{F}}_\Omega+\boldsymbol{W}_\Omega\boldsymbol{s}_\Omega-\dot{\boldsymbol{x}}_{\Omega d}+\varsigma_\Omega\right)\\ \boldsymbol{\tau}\dot{\boldsymbol{x}}_\tau+\boldsymbol{x}_\tau=\boldsymbol{x}_v\\ \boldsymbol{s}_\omega=\boldsymbol{x}_\omega-\boldsymbol{x}_\tau\\ \boldsymbol{u}=-\bar{\boldsymbol{G}}_\omega^{-1}\left(\bar{\boldsymbol{F}}_\omega+\boldsymbol{W}_\omega\boldsymbol{s}_\omega-\dot{\boldsymbol{x}}_\tau+\varsigma_\omega\right)\end{cases}\tag{10}$$

式中　$\boldsymbol{W}_\Omega$和$\boldsymbol{W}_\omega$均为正定矩阵。

DSC控制器所得到的$\boldsymbol{u}$，即滚转舵偏角δ_x、偏航舵偏角δ_y、俯仰舵偏角δ_z，将其视为“专家动作”，引导后续设计的智能控制器训练环节。

2.2　基于知识引导的TD3智能控制器设计

2.2.1　马尔可夫决策过程模型设计

在训练智能控制器前，需要建立高速飞行器再入段飞行的马尔可夫决策过程（Markov Decision Process，MDP），以创建环境与智能体之间的联系。

a）状态空间。

状态空间包括智能体可以从环境中收集到的有用信息。在本研究中，我们优先考虑高速飞行器的可观测状态，形成如式（11）所示的状态空间。

$$\boldsymbol{s}_t=\left[\boldsymbol{\rho}_{\Omega}\boldsymbol{e}_{\Omega},\boldsymbol{\rho}_{\omega}\boldsymbol{e}_{\omega}\right]^{\mathrm{T}}=\left[e_{\alpha},e_{\beta},e_{\gamma},e_{\omega_x},e_{\omega_y},e_{\omega_z}\right]^{\mathrm{T}} \tag{11}$$

式中 $\boldsymbol{e}_{\Omega}$表示当前时刻姿态角跟踪误差；$\boldsymbol{e}_{\omega}$表示当前时刻姿态角速度跟踪误差；$\boldsymbol{\rho}_{\Omega}$和$\boldsymbol{\rho}_{\omega}$为归一化正定对角矩阵，用于保证状态量的尺度大小基本相同。

b）动作空间。

本研究为高速飞行器的三通道姿态控制，故智能体直接学习控制指令，设计如式（12）所示动作空间。

$$\boldsymbol{a}_t=\left[\delta_x,\delta_y,\delta_z\right]^{\mathrm{T}} \tag{12}$$

式中 δ_x为滚转舵偏角；δ_y为偏航舵偏角；δ_z为俯仰舵偏角。

此外，考虑气动舵作动范围的物理限制，舵偏角的幅值应满足给定的约束范围。

$$-35^{\circ}\leqslant\delta_i<35^{\circ},i=x,y,z \tag{13}$$

c）奖励函数。

奖励函数被设计为连续和稀疏奖惩的混合函数，其组成部分可表示如下：

1）姿态角和姿态角速度误差惩罚。

$$P_1=K_1\left(\left|e_{\alpha}\right|+\left|e_{\beta}\right|+\left|e_{\gamma}\right|\right)+K_2\left(\left|e_{\omega_x}\right|+\left|e_{\omega_y}\right|+\left|e_{\omega_z}\right|\right) \tag{14}$$

2）姿态角误差奖励。

$$R_1=K_3\mathrm{e}^{-\eta_1\left(\left|e_{\alpha}\right|+\left|e_{\beta}\right|+\left|e_{\gamma}\right|\right)} \tag{15}$$

3）动作及动作变化率抖动惩罚。

首先，为抑制过大动作，对动作的绝对值之和进行惩罚；其次，若当前时刻的动作为$\boldsymbol{\delta}_{x_c}$，$\boldsymbol{\delta}_{y_c}$，$\boldsymbol{\delta}_{z_c}$，上一时刻动作为$\boldsymbol{\delta}_{x_l}$，$\boldsymbol{\delta}_{y_l}$，$\boldsymbol{\delta}_{z_l}$，动作变化率可依次写为$\boldsymbol{\Delta\delta}_x=\boldsymbol{\delta}_{x_c}-\boldsymbol{\delta}_{x_l}$，$\boldsymbol{\Delta\delta}_y=\boldsymbol{\delta}_{y_c}-\boldsymbol{\delta}_{y_l}$，$\boldsymbol{\Delta\delta}_z=\boldsymbol{\delta}_{z_c}-\boldsymbol{\delta}_{z_l}$，惩罚相邻时间步动作的变化幅度，以抑制高频抖动。

$$P_2=K_4\mathrm{e}^{-\boldsymbol{\eta}_2\left(\left|\boldsymbol{\delta}_x\right|+\left|\boldsymbol{\delta}_y\right|+\left|\boldsymbol{\delta}_z\right|\right)}+K_5\mathrm{e}^{-\boldsymbol{\eta}_3\left(\left|\boldsymbol{\Delta\delta}_x\right|+\left|\boldsymbol{\Delta\delta}_y\right|+\left|\boldsymbol{\Delta\delta}_z\right|\right)} \tag{16}$$

4）动作安全性惩罚。

若动作超出阈值$\boldsymbol{\delta}_{\mathrm{M}}$，则对超限部分进行二次惩罚。

$$P_3=K_6\sum_i\left(\left|\boldsymbol{\delta}_i\right|-\boldsymbol{\delta}_{\mathrm{M}}\right)^2,\text{if}\,\forall\left|\boldsymbol{\delta}_i\right|>\boldsymbol{\delta}_{\mathrm{M}}(i=x,y,z) \tag{17}$$

5）高精度跟踪奖励。

$$R_2=K_7,\text{if}\sum\left(\left|e_{\alpha}\right|+\left|e_{\beta}\right|+\left|e_{\gamma}\right|\right)<M \tag{18}$$

最终得到混合奖励函数：

$$R=-\sum_{j=1}^{3}P_j+\sum_{r=1}^{2}R_r \tag{19}$$

式中 $K_l\in\mathbb{R}^+$，$\forall l\in\{1,2,\cdots,7\}$；$\boldsymbol{\eta}_1$，$\boldsymbol{\eta}_2$，$\boldsymbol{\eta}_3\in\mathbb{R}^+$均为奖惩系数；$M\in\mathbb{R}^+$为设计要求所提的姿态角值。

2.2.2 智能控制器训练

a）Actor网络更新策略的改进。

为解决TD3算法在复杂任务中Actor网络可能面临因探索不足或训练初期Critic不准确，从而学到次优策略，且若单纯模仿“专家控制器”又无法超越“专家”水平，为解决该问题，本文提出了一种基于行为克隆的约束优化方法，从而对Actor网络的更新策略进行改进。

该方法的核心思想是将“专家控制器”的先验知识以软约束的形式融入DRL框架，具体来讲，在Actor网络策略优化目标中引入“专家动作”的行为克隆损失作为正则项，构建如式（20）所示的复合目标函数，以确保Actor输出的动作不会偏离专家动作太远，同时最大化Critic评估的q值，达到平衡“模仿专家”和“最大化累计回报”两个目标。

$$L(\boldsymbol{\theta})=\underbrace{-\mathbb{E}_{\boldsymbol{s}_j\sim\mathcal{D}}\left[q\left(\boldsymbol{s}_j,\hat{\boldsymbol{a}}_j;\boldsymbol{\omega}_{1,o}\right)\right]}_{\text{强化学习目标项}}+\boldsymbol{\lambda}\underbrace{\mathbb{E}_{\boldsymbol{s}_j\sim\mathcal{D}}\left[\left\|\hat{\boldsymbol{a}}_j-\boldsymbol{a}_{j\_k}\right\|^2\right]}_{\text{行为克隆约束项}} \tag{20}$$

式中 $L(\boldsymbol{\theta})$为损失函数，优化的目标即为最小化损失函数；$\mathbb{E}_{\boldsymbol{s}_j\sim\mathcal{D}}\left[q\left(\boldsymbol{s}_j,\hat{\boldsymbol{a}}_j;\boldsymbol{\omega}_{1,o}\right)\right]$表示当前策略$\boldsymbol{\omega}_{1,o}$在状态$\boldsymbol{s}_j$下生成的动作$\hat{\boldsymbol{a}}_j$，由Critic网络评估得到$q$值后，在数据分布$\mathcal{D}$（从经验回放缓冲区$\mathfrak{R}$中采样得到）上的平均$q$值；$\mathbb{E}_{\boldsymbol{s}_j\sim\mathcal{D}}\left[\left\|\hat{\boldsymbol{a}}_j-\boldsymbol{a}_{j_k}\right\|^2\right]$表示当前策略$\boldsymbol{\omega}_{1,o}$在状态$\boldsymbol{s}_j$下生成的动作$\hat{\boldsymbol{a}}_j$与当前时刻的“专家动作”$\boldsymbol{a}_{j_k}$的均方误差在数据分布$\mathcal{D}$上的期望；$\boldsymbol{\lambda}$为行为克隆的权重因子。

b）KG-TD3算法训练框架。

KG-TD3算法训练与TD3有两点不同：一是提出了一种基于行为克隆的约束优化方法，利用式（20）进行Actor网络更新策略进行改进；二是为将智能控制器与“专家控制器”在时间尺度上对齐，需额外存储“专家动作”，本研究选取对经验回放缓冲区进行改进。具体来讲，在智能体与环境交互过程后，将智能体轨迹整理为$\left(\boldsymbol{s}_t,\ \boldsymbol{a}_t,\ r_t,\ \boldsymbol{s}_{t+1},\ \boldsymbol{a}_{t_k}\right)$的五元组，即$t$时刻的状态$\boldsymbol{s}_t$、智能体动作$\boldsymbol{a}_t$、奖励$r_t$以及$t+1$时刻的状态$\boldsymbol{s}_{t+1}$、“专家控制器”动作$\boldsymbol{a}_{t_k}$。而后通过经验回放缓存区采用时，即可匹配同一时间的“专家动作”。

除上述两点外的KG-TD3训练环节与TD3别无二致，此处不再赘述，可参考文献［12］。KG-TD3完整的算法伪代码如下所示。

随机初始化策略网络$\boldsymbol{\mu}(\boldsymbol{s};\boldsymbol{\theta})$和价值网络$q_1(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\omega}_1)$、$q_2(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\omega}_2)$

初始化目标策略网络参数$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$和目标价值网络$\boldsymbol{\omega}_1^- \leftarrow \boldsymbol{\omega}_1$、$\boldsymbol{\omega}_2^- \leftarrow \boldsymbol{\omega}_2$

初始化经验回放缓冲区$\Re$

for episode = 1 **to** M **do**

　初始化噪声$\boldsymbol{\varepsilon}$用于噪声探索

　策略网络接收初始观察状态$\boldsymbol{s}$

　for t = 1 **to** T **do**

　　根据当前策略和探索噪声选择动作$\boldsymbol{a}_t = \boldsymbol{\mu}(s_t;\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$

　　执行动作$\boldsymbol{a}_t$，获取奖励r_t，观测新状态$\boldsymbol{s}_{t+1}$

　　经验存储：在$\Re$中存储$(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1},\boldsymbol{a}_{t_k})$

　　经验回放：从$\Re$中随机取出N个$(\boldsymbol{s}_j,\boldsymbol{a}_j,r_j,\boldsymbol{s}_{j+1},\boldsymbol{a}_{j_k})$

　　$\hat{\boldsymbol{a}}_{j+1}^- = \boldsymbol{\mu}(s_{j+1};\boldsymbol{\theta}_o^-) + \boldsymbol{\xi}$

　　两个目标价值网络预测：$\hat{q}_{i,j+1}^- = q(\boldsymbol{s}_{j+1},\hat{\boldsymbol{a}}_{j+1}^-;\boldsymbol{\omega}_{i,o}^-)(i=1,2)$

　　$\hat{y}_j = r_j + \boldsymbol{\gamma}\cdot\min\{\hat{q}_{1,j+1}^-,\hat{q}_{2,j+1}^-\}$

　　两个价值网络预测：$\hat{q}_{i,j} = q(\boldsymbol{s}_j,\boldsymbol{a}_j;\boldsymbol{\omega}_{i,o})$

　　最小化损失函数：$L(\boldsymbol{\omega}_i) = \frac{1}{N}\left[(\hat{q}_{1,j}-\hat{y}_j)^2 + (\hat{q}_{2,j}-\hat{y}_j)^2\right]$

　　更新价值网络$\boldsymbol{\omega}_{i,e} \leftarrow \boldsymbol{\omega}_{i,o}$

　　if t mod k **then**

　　$\hat{\boldsymbol{a}}_j = \boldsymbol{\mu}(\boldsymbol{s}_j;\boldsymbol{\theta}_o)$

　　最小化损失函数$L(\boldsymbol{\theta})$，更新策略网络$\boldsymbol{\theta}_e \leftarrow \boldsymbol{\theta}_o$

$$L(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{s}_j\sim\mathcal{D}}\left[q(\boldsymbol{s}_j,\hat{\boldsymbol{a}}_j;\boldsymbol{\omega}_{1,o})\right] + \boldsymbol{\lambda}\mathbb{E}_{\boldsymbol{s}_j\sim\mathcal{D}}\left[\left\|\hat{\boldsymbol{a}}_j - \boldsymbol{a}_{j\_k}\right\|^2\right]$$

　　软更新目标策略网络$\boldsymbol{\theta}_e^- \leftarrow \boldsymbol{\tau}\boldsymbol{\theta}_o + (1-\boldsymbol{\tau})\boldsymbol{\theta}_o$

　　软更新目标价值网络$\boldsymbol{\omega}_{i,e}^- \leftarrow \boldsymbol{\tau}\boldsymbol{\omega}_{i,o} + (1-\boldsymbol{\tau})\boldsymbol{\omega}_{i,o}$

　end if

end for

end for

2.2.3　智能控制器部署

当Actor和Critic网络收敛，则智能体训练完成。如图2所示，训练好的Actor网络将作为神经网络控制策略在线实施，并以端到端的方式生成高速飞行器三通道姿态控制指令。具体来讲，训练好的Actor网络接收到归一化后的姿态角跟踪误差和姿态角速度跟踪误差，经过全连接层后，输出三轴舵偏角。图2中的$\boldsymbol{x}_{\Omega d}$为系统姿态角制导指令，$\boldsymbol{x}_{\Omega}$为实际姿态角状态量，即攻角$\alpha$、侧滑角$\beta$、速度倾侧角$\gamma_v$。

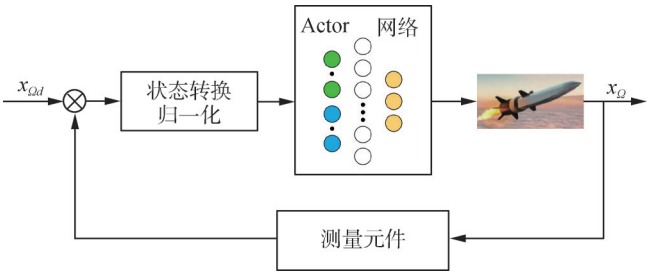


图2　神经网络模型的在线部署

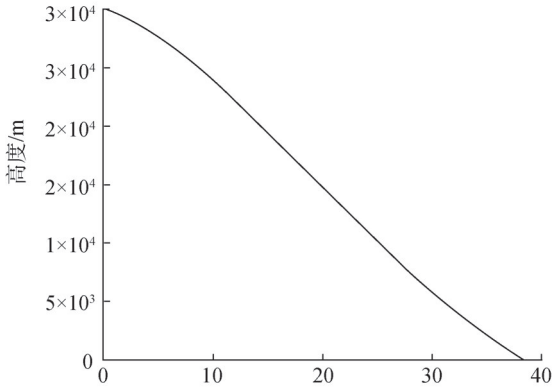
Fig.2　Online deployment of neural network model

3　仿真试验及其结果分析

3.1　试验环境设置

通过仿真软件Visual Studio Code，编程语言Python对提出的KG-TD3模型进行训练和测试。Visual Studio Code版本为1.102.3，Python版本为3.12.3。用于试验的硬件平台配置如下：操作系统为Win11，CPU为Intel Core i5，内存为16 GB。

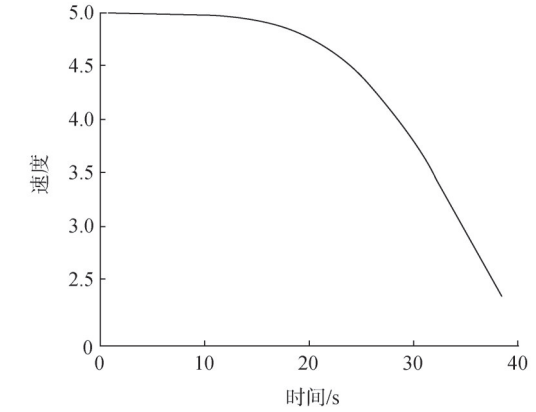
本研究选取高速飞行器的再入段，其飞行持续时间为38.4 s，初始高度为30 km，初始速度约为Ma=5，动压范围从27.4 kPa到588.8 kPa。图3为标准轨迹的典型参数，从图3中可以看出，该飞行阶段的高度、速度和动压呈现出显著而快速的变化，同时伴随着模型参数的变化。飞行器机体参数如表1所示，气动参数为参考文献［17］中提供的公开数据。



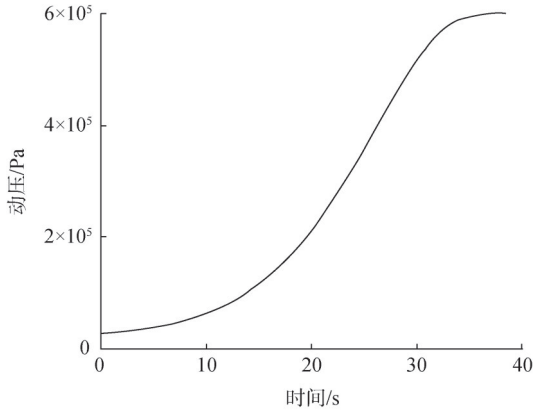
a）高度

图3　标准轨迹的典型参数

Fig.3　Typical parameters of standard trajectories



b）速度



c）动压

续图3

表1 飞行器机体参数

Tab.1 Aircraft body parameters

参数	数值
m/kg	1 000
l/m	0.7
J_y/(kg·m^{-2})	3 000
S/m^2	0.45
J_x/(kg·m^{-2})	200
J_z/(kg·m^{-2})	2 800

使用KG-TD3作为高速飞行器再入段三通道智能姿态控制器，其网络结构如表2所示，智能体训练过程中的超参数如表3所示。

表2 KG-TD3算法的网络结构

Tab.2 Network structure of the KG-TD3 algorithm

网络名称	层类型	神经元个数	激活函数
Actor	输入层	6	None
	全连接层	128	Relu
	输出层	3	Tanh
Critic	输入层	9	None
	全连接层	512	Relu
	全连接层	256	Relu
	输出层	1	Linear

表3 KG-TD3算法训练过程超参数

Tab.3 KG-TD3 algorithm training process hyper-parameters

超参数	数值
探索噪声$\boldsymbol{\varepsilon}$	$\mathcal{N}(0, 0.1^2)$
目标策略平滑噪声$\boldsymbol{\xi}$	$\mathcal{CN}(0, 0.2^2, -0.5, -0.5)$
延迟更新策略k	2
目标网络更新率$\boldsymbol{\tau}$	0.005
批量大小	128
Actor学习率	$1 \times e^{-4}$
Critic学习率	$1 \times e^{-4}$
行为克隆权重$\boldsymbol{\lambda}$	0.1

3.2 适应能力分析

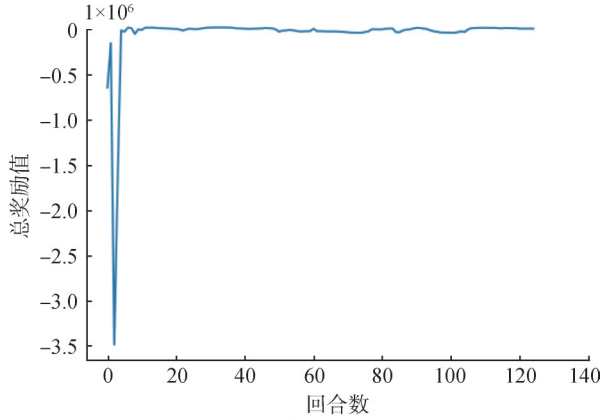
为了验证本文提出的KG-TD3姿态控制器的适应能力，我们考虑了三个通道中的初始姿态偏差$\Delta\boldsymbol{\alpha}_0$、$\Delta\boldsymbol{\beta}_0$、$\Delta\boldsymbol{\gamma}_{v0}$；质量、气动参考面积、参考长度和大气密度偏差$\Delta m$、$\Delta S$、$\Delta l$、$\Delta\boldsymbol{\rho}$；三通道中的惯性力矩偏差$\Delta J_x$、$\Delta J_y$、$\Delta J_z$；气动力系数偏差$\Delta C_{\mathrm{L}}$、$\Delta C_{\mathrm{Z}}$；俯仰力矩系数偏差$\Delta C_{mz}^{\alpha}$、$\Delta C_{mz}^{\beta}$。表4为具体偏差值。由于风干扰不改变系统本身动力学模型结构，可通过添加扰动观测器等行为进行干扰估计并补偿。本研究重点针对需依赖控制器本身适应性的参数摄动，故适应能力分析只涉及参数摄动。

表4 参数偏差范围

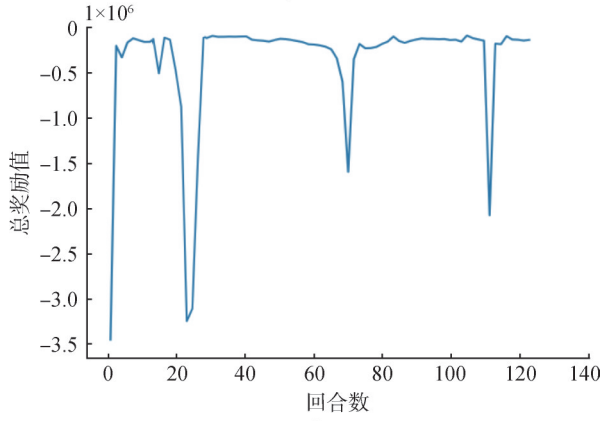
Tab.4 Margins of error for parameters

参数	偏差范围
$\Delta\boldsymbol{\alpha}_0$/(°)	1
$\Delta\boldsymbol{\gamma}_{v0}$/(°)	3
ΔS	10%
$\Delta\boldsymbol{\rho}$	20%
ΔJ_y	15%
ΔC_{L}	20%
ΔC_{mz}^{α}	20%
$\Delta\boldsymbol{\beta}_0$/(°)	1
Δm	10%
Δl	10%
ΔJ_x	15%
ΔJ_z	15%
ΔC_{Z}	20%
ΔC_{mz}^{β}	20%

图4展示了所提KG-TD3算法和TD3算法在训练过程中每回合智能体的动作奖励回报，反映了Actor网络的学习情况。KG-TD3算法在训练10轮左右奖励值趋近稳定，且无高位振荡的状态，反映了其训练过程稳定，收敛快速。相比之下，TD3算法在训练过程中奖励值波动较大，且数值略低于KG-TD3算法。



a）KG-TD3



b）TD3

图4　训练过程中的奖励值

Fig.4　Reward progression during training

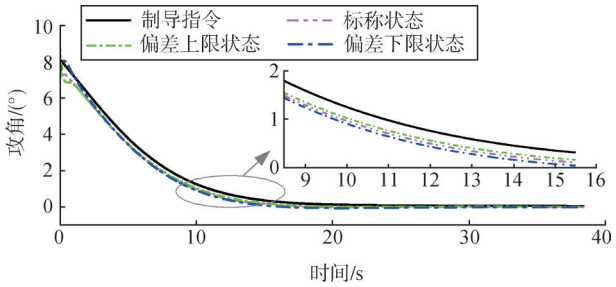
与TD3算法相比，本文提出的KG-TD3算法主要改进了奖励机制以及Actor网络优化更新策略。为了评估这一改进对姿态角跟踪精度的影响，我们将其与TD3进行对比分析，两种智能控制器的姿态角跟踪最大误差（MAX）如表5所示。其中，用于分析误差的飞行段为1~38.4 s。通过表5分析可知，KG-TD3相比TD3算法跟踪精度有较为明显的提升。

表5　姿态角跟踪误差对比

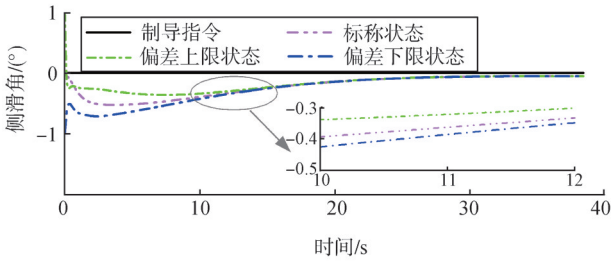
Tab.5　Comparison of attitude angle tracking errors

最大误差	TD3	KG-TD3
MAX_α/(°)	2.403	0.420
MAX_β/(°)	0.836	0.528
MAX_γ_V/(°)	34.563	0.725

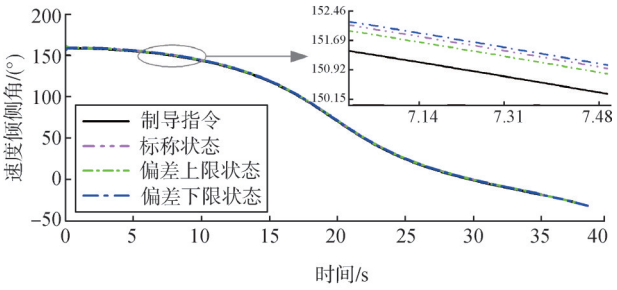
为验证所提控制方法的自适应能力，在三种典型工况下进行了仿真试验：标称状态、偏差上限状态和偏差下限状态。KG-TD3智能控制器的姿态角跟踪性能以及根据控制器得出的等效舵偏曲线如图5至图7所示。仿真试验结果分析表明，在14组涵盖上、下边界的参数摄动工况下，系统稳态控制偏差均小于1°。



a）攻角



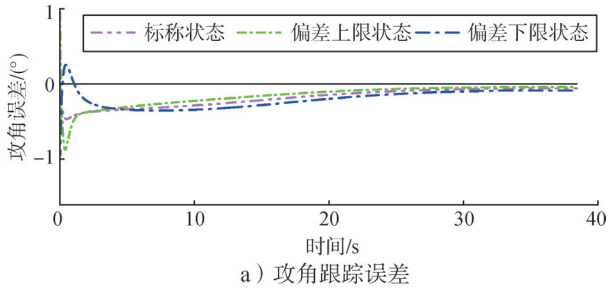
b）侧滑角



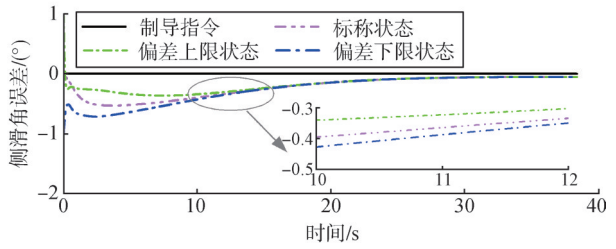
c）速度倾侧角

图5　姿态角跟踪效果

Fig.5　Attitude angle tracking effect



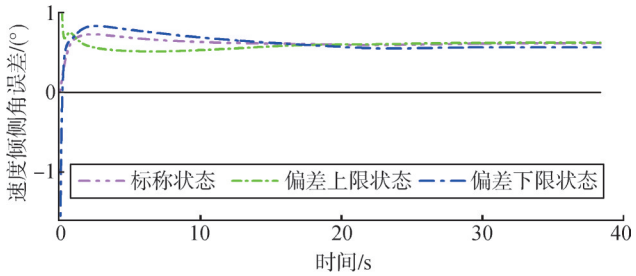
a）攻角跟踪误差



b）侧滑角跟踪误差

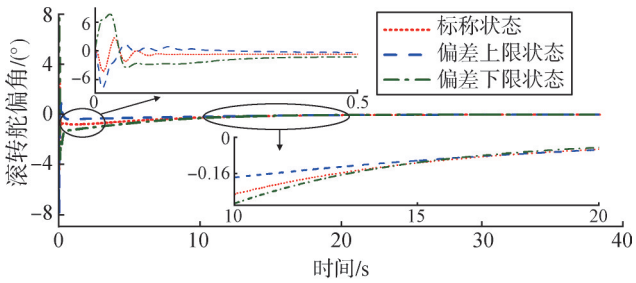
图6　姿态角跟踪误差

Fig.6　Attitude angle tracking error

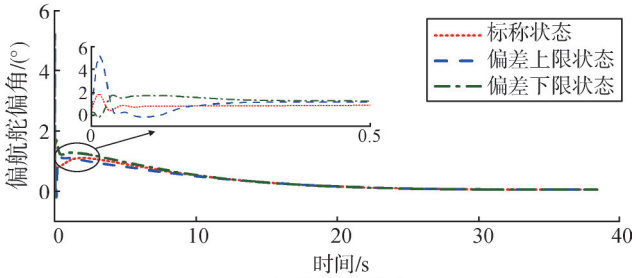


c）速度倾侧角跟踪误差

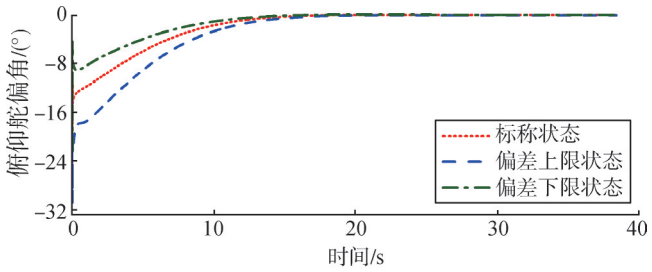
续图6



a）滚转舵偏角



b）偏航舵偏角



c）俯仰舵偏角

图7 基于KG-TD3方法的三轴舵偏

Fig.7 Three-axis rudder deflection of the KG-TD3 method

4 结束语

本研究针对高速飞行器再入段强非线性、高不确定性和参数快时变等复杂控制问题，提出了一种基于改进型TD3算法的端到端智能姿态控制方法。通过融合混合奖励机制和基于行为克隆的先验知识约束，有效解决了传统深度强化学习在姿态控制中训练不稳定、收敛困难的问题。

参 考 文 献

[1] 刘双喜, 刘世俊, 李勇, 等. 国外高超声速飞行器及防御体系发展现状[J]. 空天防御, 2023, 6(3): 39-51.
LIU Shuangxi, LIU Shijun, LI Yong, et al. Current developments in foreign hypersonic vehicles and defense systems[J]. Air and Space Defense, 2023, 6(3): 39-51.

[2] 樊轶, 秦昌茂, 董添, 等. 基于MIMO-ESO的高速飞行器自抗扰控制[J]. 导弹与航天运载技术(中英文), 2024(1): 64-70.
FAN Yi, QIN Changmao, DONG Tian, et al. ADRC attitude controller design for hypersonic vehicle based on MIMO-ESO[J]. Missiles and Space Vehicles, 2024(1): 64-70.

[3] 包为民. 航天智能控制技术让运载火箭“会学习”[J]. 航空学报, 2021, 42(11): 8-17.
BAO Weimin. Space intelligent control technology enables launch vehicle to “self-learning” [J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(11): 8-17.

[4] ZHANG Z Y, MO Z B, CHEN Y T, et al. Reinforcement learning behavioral control for nonlinear autonomous system[J]. IEEE-CAA Journal of Automatica Sinica, 2022, 9(9): 1561-1573.

[5] LUO B, SUN J Y, TANG R, et al. Reinforcement learning-based 3D trajectory tracking control of hypersonic gliding vehicles with time-varying uncertainties[J]. IEEE Transactions on Automation Science and Engineering, 2025(22): 8187-8199.

[6] LIU C, DONG C Y, ZHOU Z J, et al. Barrier Lyapunov function based reinforcement learning control for air-breathing hypersonic vehicle with variable geometry inlet[J]. Aerospace Science and Technology, 2020(96): 105537.

[7] LU K F, WANG W L, LIU X D, et al. Research progress and prospect of high-speed vehicle control technology based on reinforcement learning[J]. Advances in Astronautics, 2025, 8(2): 201-209.

[8] WANG G, AN H, WANG Y, et al. Intelligent control of air-breathing hypersonic vehicles subject to path and angle-of-attack constraints[J]. Acta Astronautica, 2022(198): 606-616.

[9] GAO Q, LI X, JI Y, et al. Research on active disturbance rejection control of hypersonic vehicle based on Q-learning[J]. Control Engineering of China, 2024, 31(4): 577-582.

[10] LI X, JI Y H, SONG Y, et al. Modified deep deterministic policy gradient based on active disturbance rejection control for hypersonic vehicles[J]. Neural Computing and Applications, 2024, 36(8): 4071-4081.

[11] 路坤锋, 贾晨辉, 黄旭, 等. 面向变构型飞行器的强化学习位置姿态一体化控制方法[J]. 宇航学报, 2024, 45(7): 1100-1110.

LU Kunfeng, JIA Chenhui, HUANG Xu, et al. Reinforcement learning-based integrated position and attitude control method towards morphing flight vehicles[J]. Journal of Astronautics, 2024, 45(7): 1100-1110.

[12] 姜凌峰, 李新凯, 张海, 等. 基于改进TD3算法的无人机动态环境无地图导航[J]. 航空学报, 2025, 46(8): 298-313.
JIANG Lingfeng, LI Xinkai, ZHANG Hai, et al. Mapless navigation of UAVs in dynamic environments based on an improved TD3 algorithm[J]. Acta aeronautica et Astronautica Sinica, 2025, 46(8): 298-313.

[13] 彭博, 王晓波, 魏祥麟, 等. 基于SPER-TD3的无人机编队三维航迹规划[J]. 计算机系统应用, 2025, 34(2): 61-73.
PENG Bo, WANG Xiaobo, WEI Xianglin, et al. 3D trajectory planning for unmanned aerial vehicle formation based on SPER-TD3[J]. Computer Systems & Applications, 2025, 34(2): 61-73.

[14] 闫雷鸣, 刘健, 朱永昕. DPC-DQRL: 动态行为克隆约束的离线-在线双Q值强化学习[J]. 计算机应用研究, 2025, 42(4): 1003-1010.
YAN Leiming, LIU Jian, ZHU Yongxin. DPC-DQRL: offline to online double Q value reinforcement learning with dynamic behavior cloning constraints[J]. Application Research of Computers, 2025, 42(4): 1003-1010.

[15] 刘晓东, 黄万伟, 禹春梅. 含扩张状态观测器的高超声速飞行器动态面姿态控制[J]. 宇航学报, 2015, 36(8): 916-922.
LIU Xiaodong, HUANG Wanwei, YU Chunmei. Dynamic surface attitude control for hypersonic vehicle containing extended state observer[J]. Journal of Astronautics, 2015, 36(8): 916-922.

[16] 黄旭, 柳嘉润, 张远, 等. 知识与数据混合驱动的高速飞行控制方法综述[J]. 宇航学报, 2023, 44(8): 1113-1126.
HUANG Xu, LIU Jiarun, ZHANG Yuan, et al. Review on knowledge-based and data-driver cooperating control methods of high-speed vehicle[J]. Journal of Astronautics, 2023, 44(8): 1113-1126.

[17] LIU X, HUANG W, DU L. An integrated guidance and control approach in three-dimensional space for hypersonic missile constrained by impact angles[J]. ISA Transactions, 2017(66): 164-175.

作 者 简 介

王伟丽（1997—），女，博士研究生，主要研究方向为飞行器智能控制、自适应控制等。

黄万伟（1970—），男，博士，研究员，主要研究方向为飞行器制导与控制、智能控制、自适应控制等。

刘晓东（1987—），男，博士，研究员，主要研究方向为飞行器制导与控制、智能控制、自适应控制等。

路坤锋（1983—），男，博士，研究员，主要研究方向为飞行器制导与控制、智能控制、自适应控制等。

贾晨辉（1985—），男，博士，高级工程师，主要研究方向为飞行器制导与控制、智能控制等。