航空科学技术
Aeronautical Science & Technology
Nov. 25 2023 Vol. 34 No.11 103-110

基于深度迁移强化学习的无人机投放自主引导机动控制算法

张堃[1,2*]，李珂[1]，邹杰[2]，栗鸣[3]，李阳[4]
1.西北工业大学，陕西 西安 710072
2.洛阳电光设备研究所 空基信息感知与融合全国重点实验室，河南 洛阳 471000
3.西安机电信息技术研究所，陕西 西安 710065
4.航空工业沈阳飞机设计研究所，辽宁 沈阳 110035

摘 要：针对无人机精确投放引导问题，本文提出基于深度迁移强化学习的无人机投放自主引导机动控制算法，分别建立基于马尔可夫决策过程的引导机动决策模型、引导机动评估模型等，并设计基于迁移学习和课程学习的引导机动策略训练方法，拟合基于深度学习的引导机动策略和评估网络，最后开展仿真训练和验证试验。仿真结果表明，该算法实现了无人机在任意姿态和位置条件下，能够自主规避区域威胁并自主引导至目标投放点，成功完成投放瞄准任务，有效地提升了无人机投放引导机动控制的自主性。

关键词：投放引导；机动控制；深度迁移强化学习；投放瞄准；马尔可夫决策过程

中图分类号：V249.4 **文献标识码：**A DOI：10.19452/j.issn1007-5453.2023.11.014

随着无人机技术和计算机技术的迅猛发展，无人机的性能得到了快速的提升，它的功能也不断得到完善，其被广泛应用到区域搜索、目标监视/跟踪、精确投放等各种任务场景中，无人机的智能化成为当前及未来很长一段时间的研究热点[1]。对于无人机在实际应用场景中存在的问题，各国的无人机专家学者专注于预防人为损失、提升平台自主飞行能力及减少人为干预的次数[2-3]。针对无人机投放引导过程中的机动控制问题，有专家学者提出无人机投放自主引导机动控制算法，以引导无人机规避飞行过程中存在的雷达探测等威胁，完成对投放目标点的瞄准。

针对无人机在精确投放任务中的自主引导问题，相关领域专家学者提出了航路规划算法和轨迹跟踪控制技术相结合的算法模型。一般使用直觉模糊博弈[4-5]、遗传算法[6]、动态贝叶斯网络[7]、影响图[8]、滚动时域[9-10]，以及近似动态规划[11]等方法，实现固定区域内的航路规划。但上述方法都存在一些局限性，如直觉模糊博弈、影响图等都要求对自主引导问题的模型构建清晰而完整，这使得构建自主引导问题模型的过程十分复杂；动态贝叶斯网络对未知环境的适应能力差，要求对问题有全面的了解；近似动态规划则要求有清晰的状态转移模型；对于在线问题的解决，采用遗传算法等优化类方法，效率往往并不高。在执行阶段，为使无人机具有相应的机动，还需要设计轨迹追踪控制器。这些因素共同降低了无人机航路引导的自主性，增加了未来无人机智能化作战需求的困难。

随着人工智能技术的飞速发展，各种人工智能算法被应用于解决无人机投放引导问题。其中，因为深度强化学习方法[12]具备端对端特性，在解决无人机投放自主引导机动控制问题时具有一定的优势；同时由于无人机投放引导问题的复杂性，引入迁移学习方法[13]，将领域知识融入模型中，将复杂问题拆解为若干子问题。因此，本文基于深度迁移强化，提出无人机投放自主引导机动控制算法。首先，建立基于马尔可夫决策过程的无人机投放引导机动决策模

收稿日期：2023-05-27；退修日期：2023-08-29；录用日期：2023-10-09
基金项目：航空科学基金（20200051053001）；陕西省重点基础研发计划（2022GXLH-02-09），中央高校基本科研业务费（D5000230311，G2023KY0601）
引用格式：Zhang Kun，Li Ke，Zou Jie，et al. Autonomous guidance maneuvering control algorithm for UAV dropping based on deep transfer reinforcement learning[J].Aeronautical Science & Technology，2023，34(11)：103-110.张堃，李珂，邹杰，等.基于深度迁移强化学习的无人机投放自主引导机动控制算法[J].航空科学技术，2023，34(11)：103-110.

型，并设计基于回报重塑的无人机投放引导机动决策评价模型；在此基础上，构建基于强化学习的无人机投放自主引导机动控制策略学习方法，拟合基于深度神经网络的无人机投放自主引导机动控制策略网络和评估网络。其次，建立基于迁移学习和课程学习的无人机投放引导机动策略学习机制。最后，仿真实现无人机投放自主引导飞行，验证本文所提算法的有效性。

1 基于马尔可夫决策过程的无人机投放引导机动决策模型

1.1 马尔可夫决策过程

马尔可夫决策过程是离散事件动态系统中一个重要的状态分析工具[14]，其特征在于决策者在一个特定的时间尺度上，通过对带有马尔可夫特性的随机动态系统进行周期或连续的观测，并按一定的顺序做出相应的决策。马尔可夫决策过程可通过五元组$\langle T,S,A(s),P(\cdot|s,a),R(s,a)\rangle$来描述。

马尔可夫决策过程的执行过程如图1所示，s_0为系统的初始状态，决策者选取动作执行a_0，系统按照转移概率$P(\cdot|s_0,a_0)$向下一个状态s_1转移，如此迭代循环。

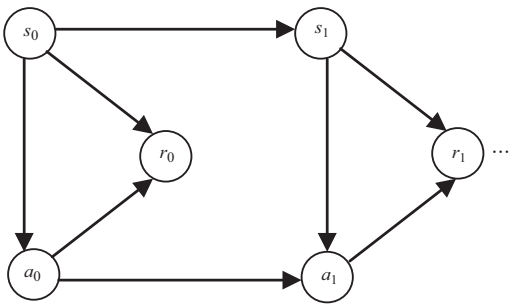

图1 马尔可夫决策过程执行过程

Fig.1 The execution processes of Markov decision processes model

在决策过程中，决策者可获得$(r_0,r_1,\cdots)$的即时回报。此过程中，决策者会受外部收益的激励，在决策中不断地调整自身决策策略，以使自身收益最大化。决策者所采取的策略确定为$a=\pi(s)$，马尔可夫理论效用函数（在系统状态$s\in S$下，利用决策者所采取的策略π所能够得到的期望回报）定义为$v(s,\pi)$，因此，如果此刻的策略是最优策略，应该满足式（1）

$$v(s)=\sup_{\pi} v(s,\pi),s\in S \tag{1}$$

如式（2）所示，针对无人机航路自动导向机动控制问题的特征，构建效用函数无限阶段折扣模型

$$v(s,\pi)=\sum_{t=0}^{\infty}\gamma^t \mathrm{E}_{\pi}^{s}\left[R(s_t,a_t)\right],s\in S \tag{2}$$

式中，$\gamma\in[0,1]$为未来报酬折扣因子，$R(s,a)$为回报函数。

1.2 无人机投放自主引导问题

针对无人机投放自主引导问题，基于三自由度运动方程构造无人机运动模型，通过对无人机的方位过载进行控制，可以在任务范围内进行动态规避，同时向目标点自主引导。图2所示为无人机投放自主引导任务示意图。

设无人机速度矢量为$\boldsymbol{V}_{\mathrm{UAV}}$，无人机方位为$\psi_{\mathrm{UAV}}$，任务区域内第$i$个威胁的位置为$\boldsymbol{X}_{\mathrm{thr}}^{i}$，其影响范围为半径$R_{\mathrm{thr}}^{i}$的圆形区域，目标分布在以$\boldsymbol{X}_{\mathrm{tgt}}^{pr}$为中心、$R_{\mathrm{tgt}}^{pr}$为半径的圆形区域内。无人机的引导目标为：在规避任务区域内所有威胁的前提下，飞入目标点所在区域并完成对目标点的瞄准。

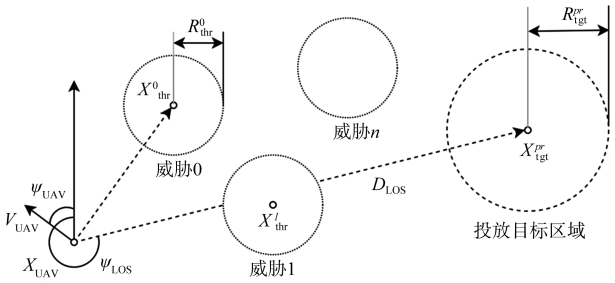

图2 自主引导任务示意图

Fig.2 The schematic diagram of autonomous guidance mission

1.3 无人机投放自主引导状态空间/动作空间

针对无人机投放自主引导问题，基于马尔可夫决策过程的定义，设计无人机投放自主引导状态空间和动作空间。

1.3.1 状态空间

图3所示为无人机投放自主引导威胁感知示意图。在无人机投放自主引导过程中，根据无人机对周围环境威胁实时感知信息，建立如下所示的状态空间

$$S=\left\{v_{\mathrm{UAV}},H_{\mathrm{UAV}},d_{\mathrm{thr}}^{f},\delta_{\psi_{\mathrm{thr}}}^{f},d_{\mathrm{thr}}^{l},\delta_{\psi_{\mathrm{thr}}}^{l},d_{\mathrm{thr}}^{r},\delta_{\psi_{\mathrm{thr}}}^{r},d_{\mathrm{LOS}}^{xz},\psi_{\mathrm{LOS}}^{f},A_{\mathrm{Bomb}}\right\} \tag{3}$$

式中，v_{UAV}为无人机速度；H_{UAV}为无人机高度；d_{thr}^{f}为无人机正前方威胁距离；$\delta_{\psi_{\mathrm{thr}}}^{f}$为无人机正前方威胁相对方位；$d_{\mathrm{thr}}^{l}$为无人机左前方威胁距离；$\delta_{\psi_{\mathrm{thr}}}^{l}$为无人机左前方威胁相对方位；$d_{\mathrm{thr}}^{r}$为无人机右前方威胁距离；$\delta_{\psi_{\mathrm{thr}}}^{r}$为无人机右前方威胁相对方位；$d_{\mathrm{LOS}}^{xz}$为目标点相对无人机的水平距离；$\psi_{\mathrm{LOS}}^{f}$为目标点相对无人机的方位；$A_{\mathrm{Bomb}}$为当前态势下无人机投放物水平射程。

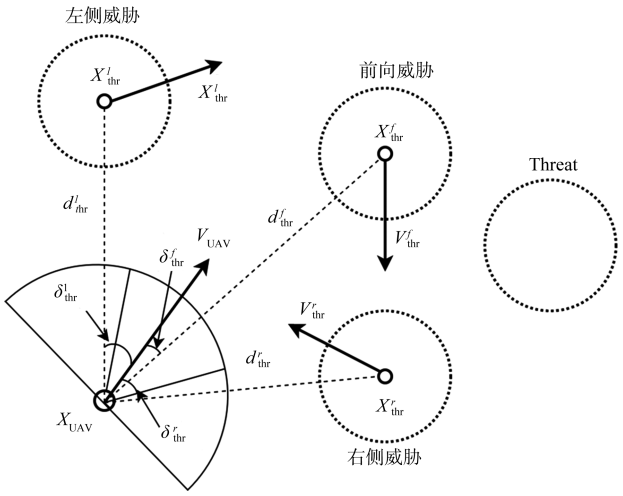

图3 自主引导威胁感知示意图

Fig.3 The schematic diagram of threat perception for autonomous guidance

1.3.2 动作空间

针对无人机投放自主引导问题，建立了如下所示的动作空间

$$A=\{N_{\mathrm{T}}\} \tag{4}$$

式中，N_{T}为无人机的转向过载；T表示转向。

1.4 无人机投放自主引导机动决策评价模型

针对无人机投放自主引导任务，基于回报重塑方法和航空火力控制理论，迁移专家经验辅助解决火控问题，构建无人机投放自主引导机动决策评价模型，如式(5)所示

$$R(s,a)=\gamma\Phi(s')-\Phi(s) \tag{5}$$

式中，$R(s,a)$为回报函数，γ为折扣参数，$\Phi(s)$为状态势函数。$\Phi(s)$定义如式(6)所示

$$\Phi(s)=U_{\mathrm{att}}(s)-U_{\mathrm{rep}}(s) \tag{6}$$

式中，$U_{\mathrm{att}}(s)$为目标点吸引势函数；$U_{\mathrm{rep}}(s)$为威胁排斥势函数。式(7)所示为$U_{\mathrm{att}}(s)$的定义

$$U_{\mathrm{att}}(s)=\frac{1}{2}\cdot k_{\mathrm{att}}\cdot\left(\frac{d_{\mathrm{LOS}}^{\max}-d_{\mathrm{LOS}}^{xz}}{d_{\mathrm{LOS}}^{\max}}\right)^2 \tag{7}$$

式中，k_{att}为吸引势权重因子，$d_{\mathrm{LOS}}^{\max}$为目标点相对无人机最大水平距离。$U_{\mathrm{rep}}(s)$定义如式(8)所示

$$U_{\mathrm{rep}}(s)=u(s_{\mathrm{f}})+u(s_{\mathrm{l}})+u(s_{\mathrm{r}}) \tag{8}$$

式中，$u(\cdot)$为无人机威胁影响势函数；s_{f}、s_{l}和s_{r}分别为无人机正前方、左前方和右前方威胁状态。$u(\cdot)$定义如式(9)所示

$$u(\cdot)=\frac{1}{2}\cdot k_{\mathrm{rep}}\cdot\left(\frac{R_{\mathrm{thr}}^{\max}-d_{\mathrm{thr}}^{i}}{R_{\mathrm{thr}}^{\max}-R_{\mathrm{thr}}^{i}}\right)^2 \tag{9}$$

式中，k_{rep}为威胁排斥势权重因子；$R_{\mathrm{thr}}^{\max}$为威胁感知最远距离；d_{thr}^{i}为当前感知威胁的水平距离。

2 基于深度迁移强化学习的无人机投放自主引导机动控制算法

2.1 无人机投放自主引导机动决策框架

基于Actor-Critic架构的深度确定性策略梯度方法[15]是一种无模型且异策略的深度强化学习方法。该方法能够很好地处理连续性控制问题，图4所示为深度确定性策略梯度(DDPG)方法组织结构图。

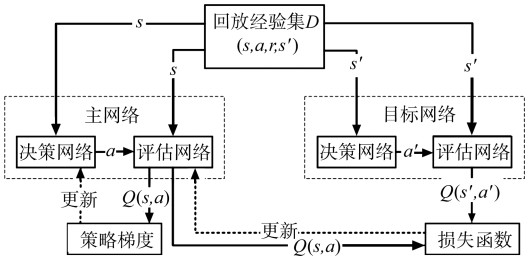

图4 DDPG方法组织结构图

Fig.4 The schematic diagram of DDPG method

该算法主要由决策网络$\mu(s;\theta^{\mu})$、评估网络$Q(s,a;\theta^{Q})$、目标决策网络$\mu'(s;\theta^{\mu'})$和目标评估网络$Q'(s,a;\theta^{Q'})$共4个网络与回放经验集共D5部分组成，在学习过程中，通过专家经验收集历史数据建立经验库，并使用强化学习算法对经验进行学习和优化。在开始阶段，通过结合加入噪声的当前环境状态，行为网络选择执行对应的动作，接着将此刻的系统状态、决策者的行动动作、决策者获得的回报收益以及之后的系统状态数据储存在经验存储区中，之后，行为网络从回放经验集中随机少量地抽取部分样本，然后使用梯度下降法等优化算法来更新行为网络与评判网络的参数，最后平滑更新目标网络参数。

2.2 基于深度神经网络的无人机投放自主引导机动控制策略模型

Actor-Critic的深度强化学习结构如图5所示。在强化学习训练时，动态演化环境的作用是产生系统状态$s\in S$，决策网络以此为基础，生成动作$a\in A(s)$，在整个训练中，采用TD-error[16]优化评估网络参数，决策网络参数优化则是通过在动态演化环境中进行迭代，依据$\max Q(s,a)$原则获取最优策略。

基于深度神经网络设计无人机航路自动引导机动控制决策算法中的决策网络和评估网络，从而更好地模拟无人

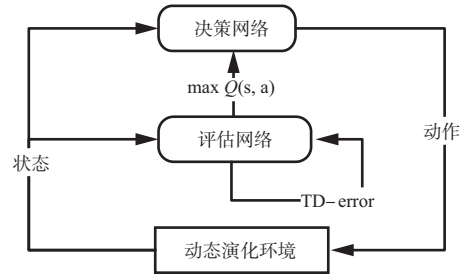

图5 Actor-Critic深度强化学习结构图

Fig.5 The schematic diagram of Actor–Critic deep reinforcement learning

机的飞行状态和“端到端”的无人机机动决策。

2.2.1 决策网络

决策网络$\mu(s;\theta^{\mu})$主要是基于此刻的系统状态来进行实时的判断并做出决策，它的网络输入为此刻的系统状态$s\in S$，而网络输出则是系统根据此刻状态而应该采取的行动动作$a\in A(s)$。按照上文中对无人机运动状态空间的定义，用$\dim(S)$表示网络输入神经元数量，$\dim(A)$表示网络的输出神经元数量，图6所示为决策网络组织结构图。

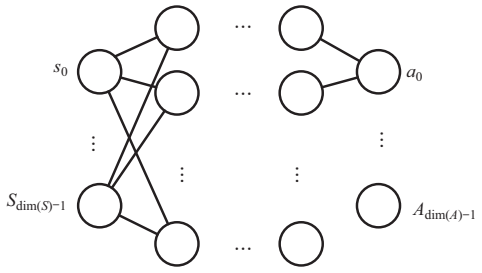

图6 决策网络组织结构图

Fig.6 The schematic diagram of decision network

根据决策网络的定义，决策网络输入层由11个单元组成，与状态空间的维度相同；隐藏层全部是全连接的线性层，分别由20、40、40和40个修正线性单元组成；输出层也是全连接的线性层，具有一个单元，与动作空间维度相同。

2.2.2 评估网络

评估网络的功能是对此刻决策的行动动作$a\in A(s)$的最优程度进行评估，它的网络输入与输出分别定义为$[s,a]$和$Q(s,a)$。图7所示为评估网络组织结构图。

根据评估网络的定义，评估网络输入层由12个单元组成，与状态空间和动作空间的维度相同；隐藏层全部是全连接的线性层，分别由20、40、40和40个修正线性单元组成；

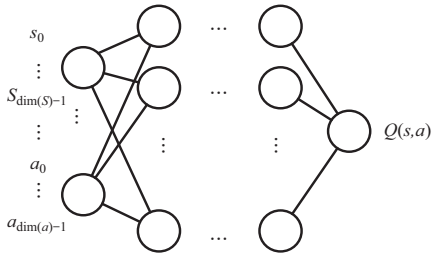

图7 评估网络组织结构图

Fig.7 The schematic diagram of critic network

输出层也是全连接的线性层，具有一个单元，输出状态和动作对应的Q值。

根据前文所定义的状态空间与动作空间，在将状态$s\in S$和动作$a\in A(s)$归一化之后，将其输入网络。在DDPG中，目标决策网络$\mu'(s;\theta^{\mu'})$与目标评估网络$Q'(s,a;\theta^{Q'})$的结构与$\mu(s;\theta^{\mu})$和$Q(s,a;\theta^{Q})$相同。

2.2.3 回放经验集

回放经验集D记录了算法与环境交互产生的历史数据，从D中重新随机抽样，打破序列相关性并重复利用历史经验，生成决策网络和评估网络的训练样本集，完成决策网络和评估网络的训练。训练样本与当前状态$s\in S$、下一时刻状态$s'\in S$、动作$a\in A(s)$和回报$r=R(s,a)$相关。

2.3 基于DDPG的无人机投放自主引导机动控制算法

在式(2)基础上，通过分析马尔可夫决策过程理论效用函数，得到了相应的描述状态—动作评价函数，如式(10)所示

$$Q(s,a)=_{\pi}\left[v(s,\pi)\right] \tag{10}$$

式(10)为状态动作值函数，因此，最优决策可以定义为

$$a_t=\arg\max_{a}Q(s_t,a) \tag{11}$$

式(11)表示在系统状态为$s_t\in S$时，最优决策为a_t。因此，可通过求解$Q(s,a)$的方法来得到最优策略。根据式(2)及式(10)，可得到Q-Learning方法迭代公式，如式(12)所示

$$Q(s,a)=Q(s,a)+\alpha\left[r+\gamma\max_{a}Q(s',a)-Q(s,a)\right] \tag{12}$$

式中，$s\in S$为系统当前状态；$a\in A(s)$为当前决策结果；$r=R(s,a)$为当前回报；$s'\in S$为系统下一时刻状态。在此基础上，得到$Q(s,a;\theta^{Q})$网络训练损失函数，如式(13)所示

$$L(\theta^{Q})=\left[r+\gamma\max_{a'}Q'(s,a;\theta^{Q'})-Q(s,a;\theta^{Q})\right]^2 \tag{13}$$

进而可得到$Q(s,a;\theta^{Q})$网络的损失函数的梯度，如式

(14)所示

$$\nabla_{\theta^Q}L(\theta^Q)=\mathop{}_{s,a,r,s'}\left[\begin{matrix}\left(r+\gamma\max\limits_{a'}Q'(s,a;\theta^{Q'})-Q(s,a;\theta^Q)\right)\\ \nabla_{\theta^Q}Q(s,a;\theta^Q)\end{matrix}\right] \tag{14}$$

在实际训练中，可以根据式(13)不断优化改变$Q(s,a;\theta^Q)$的网络参数θ^Q。

Policy Gradient算法[17]作为一种以策略为导向的强化学习方法，与值函数方法相比，具有可以直接求解最优策略的优势，而DDPG的决策网络正是源自此算法。根据DPG定理，直接获得决策网络$\mu(s;\theta^\mu)$的优化目标函数$v(s,\mu)$的梯度方程，如式(15)所示

$$\nabla_{\theta^\mu}\left[v(s,\mu)\right]=\mathop{}_{s,a,r,s'}\left[\nabla_aQ(s,a;\theta^Q)\nabla_{\theta^\mu}\mu(s;\theta^\mu)\right] \tag{15}$$

在训练过程中，通过式(15)优化决策网络$\mu(s;\theta^\mu)$的参数θ^μ。由于$\nabla_aQ(s,a;\theta^Q)$为常量，因此，在实际训练中算法对参数θ^μ的优化如式(16)所示

$$\min_{\theta^\mu}\left\{-Q(s,a;\theta^Q)\right\} \tag{16}$$

另外，DDPG还定义了用于存放先前数据的回放经验集D，通过使用D中的历史数据，训练决策网络和评价网络，经验集D的元素定义如式(17)所示

$$D=\left\{[s,a,r,s']\right\} \tag{17}$$

式中，$s\in S$为系统当前状态；$a\in A(s)$为当前决策结果；$r=R(s,a)$为当前回报；$s'\in S$为系统下一时刻状态。

对于目标网络$\mu'(s;\theta^{\mu'})$和$Q'(s,a;\theta^{Q'})$的参数，本文采用平滑更新的方式进行更新，如式(18)所示

$$\begin{cases}\theta^{Q'}=\tau\theta^Q+(1-\tau)\theta^{Q'}\\ \theta^{\mu'}=\tau\theta^\mu+(1-\tau)\theta^{\mu'}\end{cases} \tag{18}$$

式中，$\tau\in(0,1)$为目标网络更新参数。

在训练过程中，因为确定性策略的动作探索性不强，所以采取了将噪声附加在决策网络输出上来处理该问题，如式(19)所示

$$a_t=\mu(s_t;\theta^\mu)+N(t) \tag{19}$$

式中，$N(t)$为Ornstein-Uhlenbeck过程[18]。

在上文所述的基础上，本文给出的一种基于DDPG的无人机航路自主引导机动控制决策算法的训练流程如下：(1) 预置回放经验集D；(2) 预置决策网络$\mu(s;\theta^\mu)$和目标决策网络$\mu'(s;\theta^{\mu'})$，评价网络$Q(s,a;\theta^Q)$和目标评价网络$Q'(s,a;\theta^{Q'})$；(3) 预置Ornstein-Uhlenbeck的过程$\mathcal{N}(t)$，采集无人机飞行模拟环境系统的初始状态s_0；(4) 基于$a_t=\mu(s_t;\theta^\mu)+N(t)$产生行动动作；(5) 在无人机飞行模拟环境系统中执行行动动作a_t，得到反馈回报收益r_t；(6) 获取后一时间段无人机飞行模拟环境系统的状态s_{t+1}，并在D中记录当前数据(s_t,a_t,r_t,s_{t+1})；(7) 根据式(14)，更新参数θ^Q；(8)根据式(16)，更新参数θ^μ；(9)根据式(18)，更新目标网络参数$\theta^{Q'}$和$\theta^{\mu'}$；(10)重复第(4)~(9)步至$t=T$；(11)重复第(3)~(10)步M次至训练结束。

按照上述流程进行训练，当训练结束后，就能够获得对应的最优决策网络$\mu(s;\theta^\mu)$，此流程中，决策结果可直接被用作决策网络的输出，式(20)所示为动作生成公式

$$a=\mu(s;\theta^\mu),s\in S \tag{20}$$

3 仿真验证与分析

给定无人机飞行试验的空域范围为100km×100km的正方形区域，对模型进行训练的周期数为$M=1000$，一个循环周期内最大决策时刻数目$T=500$。通过建立随机的无人机初始状态，设置不同的目标点位置和无人机初始航向，实现无人机自主引导和瞄准。

图8~图11所示为部分仿真试验的可视化结果。图中横轴East表示正东方向，纵轴North表示正北方向。红色实线为无人机飞行轨迹，红色虚线为瞄准线，红色实心点为无人机起点，红色“X”为无人机终点，蓝色“+”为目标点，绿色点画线为威胁影响范围，绿色虚线为威胁截止区域，绿色“X”为威胁位置。

仿真试验过程中，目标距离初始生成无人机约为80km，无人机最大过载为5，任务区域内包含三个威胁。无人机在任意位置、姿态下，能够规避任务区域内威胁，快速抵达投放目标点附近，并完成瞄准。

仿真试验过程中，将决策网络$\mu(s;\theta^\mu)$与目标决策网络$\mu'(s;\theta^{\mu'})$、评价网络$Q(s,a;\theta^Q)$与目标评价网络$Q'(s,a;\theta^{Q'})$作为整体进行训练。输入飞机初始状态$s\in S$到决策网络，得到输出，继续将输出输入评价网络，得到评估结果，根据评估结果与预期目标计算损失函数$Q(s,a;\theta^Q)$，更新参数θ^μ，优化网络。

从图中可看出，无人机在飞行过程中，面对不同位置的敌机威胁，从起始位置到结束位置约80km，通过控制无人机转向过载实现威胁规避，并向目标点飞行；到达目标点附近后，控制无人机转向过载，能够消除无人机瞄准偏差，完成对目标点的瞄准。

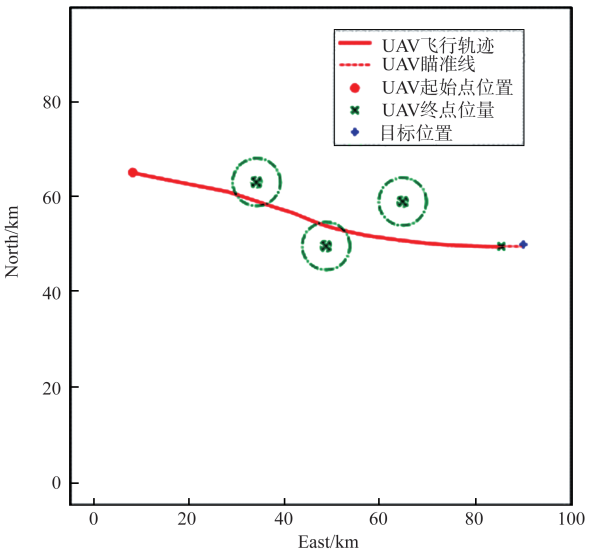

图8 无人机投放自主引导试验1结果图

Fig.8 The visualization of autonomous guidance of UAV dropping experiment 1

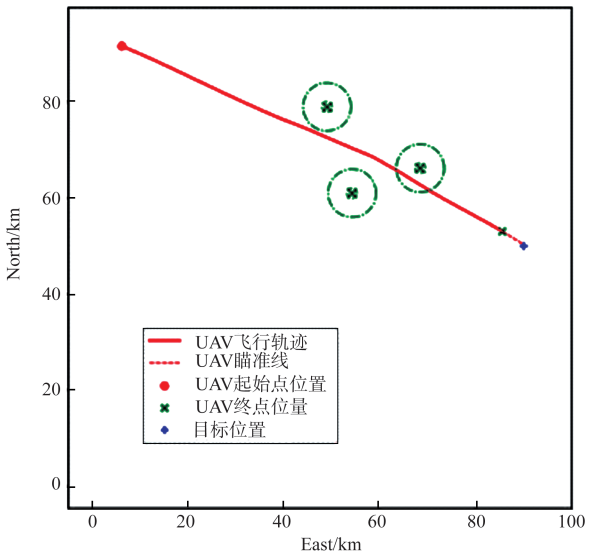

图9 无人机投放自主引导试验2结果图

Fig.9 The visualization of autonomous guidance of UAV dropping experiment 2

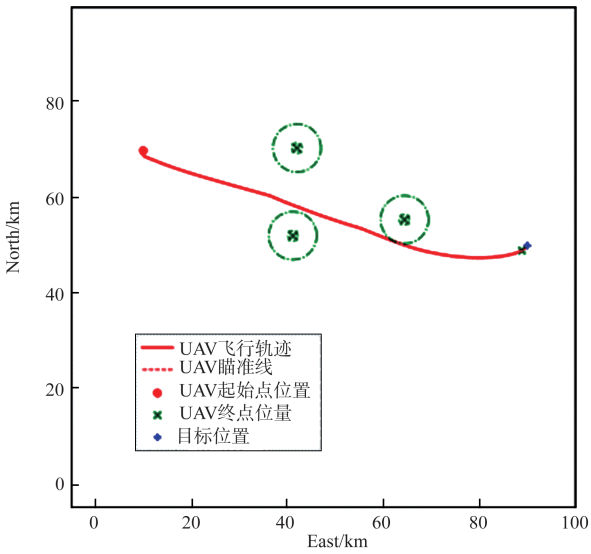

图10 无人机投放自主引导试验3结果图

Fig.10 The visualization of autonomous guidance of UAV dropping experiment 3

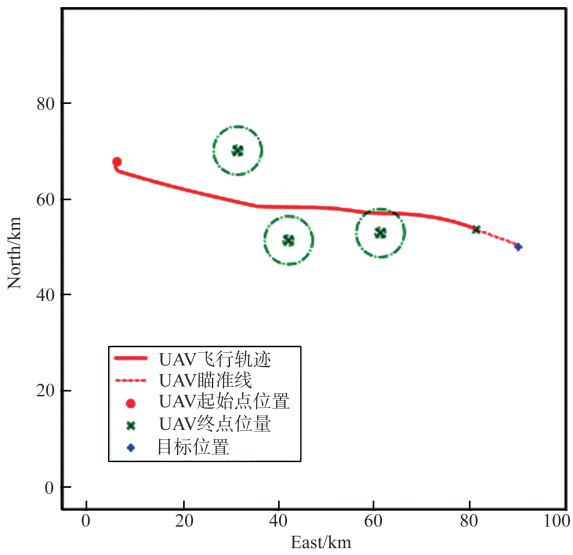

图11 无人机投放自主引导试验4结果图

Fig.11 The visualization of autonomous guidance of UAV dropping experiment 4

4 结论

本文针对无人机投放自主引导机动控制问题，提出了基于深度迁移强化学习的无人机投放自主引导机动控制算法，提炼了无人机投放自主引导机动控制问题，采用马尔可夫决策过程构建了无人机投放引导机动决策模型，设计了无人机投放引导状态空间、动作空间和改进的回报函数模型，实现了无人机投放自主引导仿真环境，开展了无人机投放自主引导机动控制算法的仿真训练，并进行了大量仿真验证。结果表明了无人机投放自主引导机动控制算法的有效性，证明了本文所提算法能够有效提高无人机执行投放引导任务的自主性。

参考文献

[1] 尹欣繁，章贵川，彭先敏，等. 军用无人机技术智能化发展及应用[J]. 国防科技，2018，39(5):30-34.

Yin Xinfan, Zhang Guichuan, Peng Xianmin, et al. Intelligent

development and application of military UAV technology[J]. National Defense Science & Technology, 2018, 39(5): 30-34. (in Chinese)

[2] 黄长强. 未来空战过程智能化关键技术研究[J]. 航空兵器，2019，26(1):11-19.
Huang Changqiang. Research on key technology of future air combat process intelligentization[J]. Aero Weaponry, 2019, 26 (1):11-19. (in Chinese)

[3] 周思羽，吴文海，张楠，等. 自主空战机动决策方法综述[J]. 航空计算技术，2012(1):27-31.
Zhou Siyu, Wu Wenhai, Zhang Nan, et al. Overview of autonomous air combat maneuver decision[J]. Aeronautical Computing Technique, 2012(1):27-31. (in Chinese)

[4] 李世豪，丁勇，高振龙. 基于直觉模糊博弈的无人机空战机动决策[J]. 系统工程与电子技术，2019，41(5):1063-1070.
Li Shihao, Ding Yong, Gao Zhenlong. UAV air combat maneuvering decision based on intuitionistic fuzzy game theory [J]. Systems Engineering and Electronics, 2019, 41(5): 1063-1070. (in Chinese)

[5] 李世豪. 复杂空战环境下基于博弈模型的无人机机动决策方法研究[D]. 南京：南京航空航天大学，2019.
Li Shihao. Research on UAV maneuvering decision method based on game theory in complex air combat[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2019. (in Chinese)

[6] 邓可，彭宣淇，周德云. 基于矩阵对策与遗传算法的无人机空战决策[J]. 火力与指挥控制，2019，44(12):61-66+71.
Deng Ke, Peng Xuanqi, Zhou Deyun. Study on air combat decision method of UAV based on matrix game and genetic algorithm[J]. Fire Control & Command Control, 2019, 44(12): 61-66+71. (in Chinese)

[7] 孟光磊，罗元强，梁宵，等. 基于动态贝叶斯网络的空战决策方法[J]. 指挥控制与仿真，2017，39(3):49-54.
Meng Guanglei, Luo Yuanqiang, Liang Xiao, et al. Air combat decision-making method based on dynamic Bayesian network [J]. Command Control & Simulation, 2017, 39(3): 49-54. (in Chinese)

[8] 董彦非，申洋，张恒喜. 空战机动决策中的影响图方法[J]. 电光与控制，2001(1):49-53.
Dong Yanfei, Shen Yang, Zhang Hengxi. Influence diagram used in air combat maneuvering decision[J]. Electronics Optics & Control, 2001(1):49-53. (in Chinese)

[9] 傅莉，谢福怀，孟光磊，等. 基于滚动时域的无人机空战决策专家系统[J]. 北京航空航天大学学报，2015，41(11):1994-1999.
Fu Li, Xie Fuhuai, Meng Guanglei, et al. An UAV air-combat decision expert system based on receding horizon control[J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(11):1994-1999. (in Chinese)

[10] 付昭旺，李战武，强晓明，等. 基于滚动时域控制的战斗机空战机动决策[J]. 电光与控制，2013，20(3):20-24.
Fu Zhaowang, Li Zhanwu, Qiang Xiaoming, et al. Tactical decision-making method based on receding horizon control for air combat[J]. Electronics Optics & Control, 2013, 20(3): 20-24. (in Chinese)

[11] 黄长强，赵克新，韩邦杰，等. 一种近似动态规划的无人机机动决策方法[J]. 电子与信息学报，2018，40(10):166-171.
Huang Changqiang, Zhao Kexin, Han Bangjie, et al. Maneuvering decision-making method of UAV based on approximate dynamic programming[J]. Journal of Electronics & Information Technology, 2018, 40(10):166-171. (in Chinese)

[12] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. 2nd Edition. Cambridge: MIT Press, 2018.

[13] Zhu Zhuangdi, Lin Kaixiang, Zhou Jiayou, et al. Transfer learning in deep reinforcement learning: A survey[J/OL]. (2023-07-04). https://ieeexplore.ieee.org/document/10172347.

[14] Song H, Liu C C, Lawarree J, et al. Optimal electricity supply bidding by Markov decision process[J]. IEEE Transactions on Power Systems, 2000, 15(2):618-624.

[15] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. Computer Science, 2015, 8(6): A187.

[16] Tesauro G. Temporal difference learning and TD-Gammon[J]. Communications of the Acm, 1995, 38(3):58-68.

[17] Peters J, Schaal S. Reinforcement learning of motor skills with policy gradients[J]. Neural Networks, 2008, 21(4):682-697.

[18] Barndorff-Nielsen O E, Shephard N. Non-Gaussian Ornstein Uhlenbeck-based models and some of their uses in financial economics[J]. Journal of the Royal Statistical Society, 2001, 63 (2):167-241.

Autonomous Guidance Maneuvering Control Algorithm for UAV Dropping Based on Deep Transfer Reinforcement Learning

Zhang Kun[1,2], Li Ke[1], Zou Jie[2], Li Ming[3], Li Yang[4]

1. Northwestern Polytechnical University, Xi'an 710072, China

2. National Key Laboratory of Space Based Information Perception and Fusion, Luoyang Institute of Electro-Optical Equipment, Luoyang 471000, China

3. Xi'an Institute of Electromechanical Information Technology, Xi'an 710065, China

4. AVIC Shenyang Aircraft Design and Research Institute, Shenyang 110035, China

Abstract: Aiming at the problem of long-distance guidance for precise dropping of UAVs, this paper proposed the autonomous guidance maneuvering control algorithm for UAV dropping based on deep transfer reinforcement learning. This paper established the guidance maneuvering decision-making model for UAV dropping based on Markov decision processes. Specifically, it constructed an improved reward model for evaluating the action obtained by the algorithm we proposed based on traditional aviation fire control theory. And it presented the training process of the autonomous guidance maneuvering policy for UAV dropping based on transfer learning and curriculum learning. It constructed the autonomous guidance maneuvering policy and evaluation networks based on deep learning. Finally, the simulation results show that this algorithm achieves autonomous flight of UAV in any initial position/attitude to avoid threats in the mission area and towards the target point, ultimately performing the targeting of the dropping, and effectively improving the autonomy of maneuvering control during the UAV dropping guidance process.

Key Words: dropping guidance; maneuvering control; deep transfer reinforcement learning; dropping aiming; Markov decision processes

Received: 2023-05-27; **Revised:** 2023-08-29; **Accepted:** 2023-10-09

Foundation item: Aeronautical Science Foundation of China(20200051053001); Key Research and Development Program of Shaanxi(2022GXLH-02-09); Fundamental Research Funds for the Central Universities(D5000230311,G2023KY0601)