•智能交叉科学与工程•
DOI:10.12454/j.jsuese.202400419

本刊网刊

基于时间尺度分离理论的空战深度强化学习分层算法

谭　泰[1],江泰民[1],黎博文[1],李　杰[1],李　辉[1,2*],化晨昊[1]

(1.四川大学 计算机学院,四川 成都 610065;2.四川大学 视觉合成图形图像技术国家级重点实验室,四川 成都 610065)

摘　要:六自由度无人机空战是一个具有复杂多维状态、耦合连续动作和高度非线性动力学的挑战性场景。深度强化学习不需要标签数据,仅通过与环境交互优化策略,在自主空战机动决策中的应用受到广泛关注。然而,高维度的状态和动作空间导致端到端训练难以学习到有效策略、收敛缓慢且泛化性差;奖励函数的设计多依赖人工经验,获得好的奖励并不等同于学习到好的策略。针对这些问题,本文提出了一种基于时分框架的两阶段时间尺度状态分离近端策略优化(two stage time-scale states separation proximal policy optimization,TTS–PPO)算法。因飞控参数对不同状态量控制效果的时间尺度差异,该算法将空战机动划分为短周期转动运动和长周期轨迹运动两部分,短周期部分采用比例–积分–微分(PID)算法完成飞控参数实时输出,长周期部分通过近端策略优化(PPO)算法对短周期 PID 控制接口进行训练,使两类运动的动作空间解耦,从而使无人机更容易学到有效策略;同时,将环境状态量分离得到长短周期状态量,降低状态空间的维度从而加快收敛并提高模型的泛化性。此外,本文在训练过程中对长周期决策的 PPO 网络采取两阶段训练方式:第一阶段设计单步奖励并采用较低的决策频率,使无人机训练过程快速度过冷启动时期;第二阶段只保留终局奖励并采用更高的决策频率,避免陷入追求高奖励而损失性能的误区。实验结果表明:使用该框架的算法能够收敛到更高的奖励值;引入长短周期状态量能提升约 67% 的收敛速度,且在不同空战场景中的泛化性更强;TTS–PPO 算法增加了第二阶段训练,性能进一步提升,仅以直线飞行的敌机作为对手训练后就能击败专家无人机。

关键词:时间尺度分离;比例–积分–微分;近端策略优化;两阶段训练;两阶段时间尺度状态分离近端策略优化
中图分类号:TP181;V249.1　　**文献标志码:**A　　**文章编号:**2096-3246(2026)02-0069-15

制空权在现代战争中具有重要地位,随着智能化技术的迅猛发展,无人机在军事领域扮演的角色越来越多,如进行战场监视及毁伤评估、完成枯燥危险的长航时反恐任务、作为忠诚僚机协同空战[1],其多功能性和在极端环境中执行任务的能力受到各国广泛关注。

在自主空战机动决策领域中的无人机研究不断涌现,文献[2]将现有方法分为基于博弈论的方法、基于优化理论的方法和基于数据驱动的方法 3 大类。基于博弈论的方法主要包括微分博弈[3]、影响图[4]等,这些方法计算量巨大,求解困难;基于优化理论的方法主要包括遗传算法[5]、粒子群优化[6]等,这些方法实时性差,难以适应复杂多变的实际空战环境;基于数据驱动的方法主要包括专家系统[7]、神经网络[8]等,专家系统受限于飞行员认知且不具备学习能力,神经网络需要大量有标签的数据且可解释性差。

近年来,随着深度强化学习(DRL)在 Atari[9]、围棋[10]和 Dota2[11]等即时策略游戏中的巨大成功,越来越多的研究者将其应用到空战自主决策任务中。基于 DRL 的智能空战研究依据无人机的动作空间可分为离散和连续两类。Zhang 等[12]针对超视距空战,基于包含 9 种固定动作的动作集,提出一种融合专家经验的启发式 Q–网络方法;Yang 等[13]针对近距空战,在美国国家航空航天局的 7 种基本动作[14]基础上扩展了机动库,提出一个基于强化学习的无人机自主机动决策模型。文献[12–13]都设计了包含离散动作的机动库,这

收稿日期:2024 – 06 – 02　　**修回日期:**2024 – 12 – 16　　**网络出版日期:**2025 – 03 – 27
基金项目:国家自然科学基金–联合基金项目(U20A20161)
作者简介:谭　泰(1995—),男,硕士生. 研究方向:深度强化学习. E-mail: tantai@stu.scu.edu.cn
* **通信作者:**李　辉,教授, E-mail: lihuib@scu.edu.cn

种通过人类先验知识将动作离散化的做法虽能大大降低探索空间的复杂程度，但限制了算法对复杂情境的适应能力，不能满足高精度操作的需求。尽管使用离散动作能够极大程度简化问题，但使用连续控制量时能完成的机动动作更丰富，能完成更复杂的任务，空战仿真更贴合实战[2]。Li 等[15]针对协同作战，在连续动作空间下构建了深度确定性策略梯度的基本结构，将动作与油门、攻击角和飞机路径倾角 3 个连续参数关联。文献[13–15]的研究对象都是三自由度飞机，而真正的无人战斗机空战通常会使用六自由度飞机模型，因其能够全面、准确地描述和控制飞机在三维空间中的所有运动，适应复杂的空战环境和动态变化。使用六自由度模型的问题是在连续状态空间和动作空间使用端到端的训练方式，探索空间太大，收敛难度急剧提升。分层强化学习通过分而治之的方法缓解了这一问题，Pope 等[16]将分层架构与最大熵强化学习相结合，先训练几种基础的无人机策略，然后配合高层的策略选择器进行决策，在 Alpha Dog Fight 比赛中击败了 F-16 教官 Banger。Li 等[17]通过构建六自由度模型并结合基于粒子群优化算法径向基函数（PSO-RBF）的敌机操作预测方法和改进的深度确定性策略梯度（DDPG），在模拟和决策层次上实现了无人机自主空战机动决策的优化。Chai 等[18]提出了一种用于六自由度无人机空战的分层深度强化学习框架，通过将决策过程分为外环和内环两部分来解决复杂的空战问题。文献[16–18]虽然都将分层思想用于六自由度无人机空战来解决连续动作空间下的探索困难问题，但方法和标准各不相同，且进行训练时，上下层输入的状态量仍是全部的状态量，上下层的决策频率也未作差异化处理，没有达到实质上的分层。此外，奖励函数设计对于策略学习至关重要，目前仍没有统一设计规律可遵循[2]，以往的大多数研究都依赖人工经验设计奖励函数，通过实验来调整奖励设计，这种受限于人类知识的奖励设计无法完全准确地表征环境的奖励反馈。文献[19–20]使用瞄准敌机、被敌机瞄准等关键空战事件进行奖励塑造来替代以往研究中的单步奖励，一定程度上缓解了人类经验设计奖励函数对模型性能的影响。

针对上述问题，提出一种两阶段时间尺度状态分离近端策略优化（two stage time-scale states separation proximal policy optimization，TTS–PPO）算法。首先，基于操作杆对空战状态量控制效果的时间尺度差异提出一套时分空战框架，将空战机动决策划分为短周期和长周期两部分，前者利用比例–微分–积分（PID）算法实现操作杆的实时控制，后者通过近端策略优化（PPO）算法对短周期 PID 控制接口进行训练。其次，将环境状态量进行时间尺度分离，设计对应的长短周期状态量和相对态势转化模块，降低强化学习的状态空间维度，从而提升模型的收敛速度和泛化性。最后，考虑到奖励函数的局限性和分层后的长短周期决策频率差异，对 PPO 网络进行两阶段训练，第一阶段将空战任务分为追击和打击两种情况来设计相应的单步奖励，并采用较低的决策频率，以加快模型的收敛；第二阶段的训练去掉单步奖励，仅保留终局的稀疏奖励，减小单步奖励对无人机的影响，并提高决策频率，尽可能提升其作战性能。仿真实验中选用开源的 F-16 无人机模型和 JSBSim[21] 开源动力学模型搭建了六自由度 1 对 1 近距空战场景，针对算法的 3 处改进进行了消融实验验证其有效性。结果发现，训练时仅以直线飞行的敌机作为对手，得到的无人机就能与专家无人机对抗，具备近距空战的能力。

1 相关理论

1.1 时间尺度分离

控制效果是指被控制的状态量 j 对应控制输入 δ_j 的敏感程度，即状态量 j 随控制输入 δ_j 的变化率，用 $\partial j/\partial\delta_j$ 表示。基于不同状态量的控制效果不同，将不同状态量按照控制效果的优劣在时间尺度上进行划分，控制效果好意味着状态量变化较快，控制效果差意味着状态量变化较慢[22–23]。

无人机空战中涉及复杂多维的连续状态量，六自由度无人机的 4 个操作杆对这些状态量的控制效果在时间尺度上存在差异。当有多个状态量需要控制时，可以使用时间尺度分离方法，根据控制效果将其在时间尺度下进行划分。

图 1 为状态量的时间尺度分离。将状态量 $\boldsymbol{S}$ 划分成 5 个不同时间尺度的状态量 $\boldsymbol{S}_1$、$\boldsymbol{S}_2$、$\boldsymbol{S}_3$、$\boldsymbol{S}_4$、$\boldsymbol{S}_5$，下标值越大控制有效性越高，此时就可以根据时间尺度对状态量进行分层，将快速变化的 $\boldsymbol{S}_4$、$\boldsymbol{S}_5$ 划分为短周期状态量 $\boldsymbol{S}_{\text{fast}}$，将变化较慢的 $\boldsymbol{S}_1$、$\boldsymbol{S}_2$、$\boldsymbol{S}_3$ 划分为长周期状态量 $\boldsymbol{S}_{\text{slow}}$。

针对长、短周期状态量在时间尺度上的差异，可采用分层控制策略。首先，利用 $\boldsymbol{S}_{\text{fast}}$ 实现对系统状态的快速调节与稳定控制；在此基础上，引入 $\boldsymbol{S}_{\text{slow}}$ 对系统整体行为进行趋势性引导。具体而言，长周期策略根据长短周期接口状态量输出其期望值 $\boldsymbol{S}_{4,\text{desire}}$，短周期策略接收该期望值并结合当前短周期状态量，计算并输出最终的操作杆控制量，从而协同完成系统的整体控制。

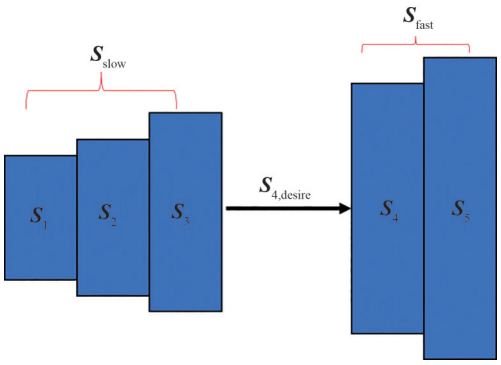


图1　状态量的时间尺度分离

Fig. 1　Time-cale separation of state variables

1.2　PID 算法

PID 算法是一种经典的控制算法，PID 控制器[24]是工业控制中最常用的控制器之一，常用于实时控制系统。假设系统在当前时刻 t 的目标值为 $r(t)$，实际输出为 $y(t)$，误差为 $e(t)=r(t)-y(t)$，则 PID 控制器的输出 $u(t)$ 可以表示为：

$$u(t)=K_{\mathrm{p}}e(t)+K_{\mathrm{i}}\int_0^t e(\tau)\mathrm{d}\tau+K_{\mathrm{d}}\frac{\mathrm{d}e(t)}{\mathrm{d}t} \tag{1}$$

式中：

1）$K_{\mathrm{p}}e(t)$ 为比例控制部分；K_{p} 为比例增益参数，表示误差信号对控制输出的影响程度。比例控制部分使系统的响应速度与误差成正比，可以快速响应系统变化。

2）$K_{\mathrm{i}}\int_0^t e(\tau)\mathrm{d}\tau$ 为积分控制部分；K_{i} 为积分增益参数，表示误差积分对控制输出的影响程度；τ 为积分里的占位时间变量。积分控制部分用于消除系统的稳态误差，保证系统达到期望值时误差为 0。

3）$K_{\mathrm{d}}\frac{\mathrm{d}e(t)}{\mathrm{d}t}$ 为微分控制部分；K_{d} 为微分增益参数，表示误差变化率对控制输出的影响程度。微分控制部分用于抑制系统的振荡和超调，使系统更加稳定。

综合比例、积分和微分 3 个部分，PID 控制器能够实现对系统的精确控制，快速响应系统变化，并消除稳态误差和振荡现象。

1.3　PPO 算法

强化学习通过无人机与环境交互得到最佳策略。无人机通过与环境交互得到状态量 o_k，然后根据策略得到相应的动作 a_k，下标 k 表示不同的时间步。环境接收该动作后，发生状态转移，并返回用于评估该动作效果的奖励 r_{k+1}。强化学习的目标是最大化折扣奖励和，又称为回报 G_k，其计算方式如下：

$$G_k=r_{k+1}+\gamma r_{k+2}+\gamma^2 r_{k+3}+\cdots \tag{2}$$

式中，γ 为折扣因子，$\gamma\in[0,1]$，表示后续奖励对当前的影响随时间步的衰减程度。

PPO 是一种基于策略梯度的强化学习算法，采用 Actor−Critic 架构，由策略网络（Actor）和价值网络（Critic）两部分组成，二者均使用神经网络进行参数化表示[25]。Actor 网络表示策略 $\pi_\Theta(a|o)$，用于根据当前状态 o 输出动作 a 的概率分布；Critic 网络表示状态值函数 $V(o;\Phi)$，用于近似在策略 π_Θ 下从状态 o 出发所能获得的期望回报；Θ 和 Φ 分别为 Actor 和 Critic 网络的参数。

智能体无人机与环境进行交互，采集“状态–动作–奖励–下一状态”四元组 $(o_k|a_k|r_k|o_{k+1})$，并将其存入经验缓冲区。在每个回合结束后，根据采样得到的奖励序列计算回报 G_k，也存入缓冲区。优势函数 A_k 采用广义优势估计（GAE）方法计算，其形式如下：

$$A_k=\sum_{n=0}^{T-k-1}(\gamma\lambda)^n\delta_{k+n} \tag{3}$$

式中，δ_k 为时序差分（TD）误差，$\delta_k=r_k+\gamma V(o_{k+1};\Phi)-V(o_k;\Phi)$，$\lambda$ 为 GAE 的平滑参数，T 为当前回合的终止时间步，n 为从当前时刻开始向后看的步数索引。优势函数用于衡量在给定状态下所选动作相对于当前策略平均行为的优劣程度。

PPO 算法利用缓冲区中的采样数据对 Actor 和 Critic 网络进行联合优化。Critic 网络通过最小化预测状态值与回报之间的均方误差进行训练，其损失函数 $L_{\mathrm{Critic}}(\Phi)$ 定义如下：

$$L_{\mathrm{Critic}}(\Phi)=\mathbb{E}[V(o_k;\Phi)-G_k]^2 \tag{4}$$

式中，$\mathbb{E}[\cdot]$ 为计算期望的函数。

Actor 网络的优化目标为最大化裁剪策略目标函数 $L_{\mathrm{Actor}}(\Theta)$，其表达式如下：

$$L_{\mathrm{Actor}}(\Theta)=\mathbb{E}[\min(r_k(\Theta)A_k,\mathrm{clip}[r_k(\Theta),1-\epsilon,1+\epsilon]A_k)] \tag{5}$$

式中：$r_k(\Theta)$ 为新策略与采样时旧策略之间的概率比，$r_k(\Theta)=\pi_\Theta(a_k|o_k)/\pi_{\Theta_{\mathrm{old}}}(a_k|o_k)$，下标 Θ_{old} 为采样时 Actor 网络的参数；clip[·]为裁剪函数；ϵ 为裁剪系数，用于限制策略更新幅度，从而提高训练过程的稳定性。

通过最小化 Critic 网络的损失函数并最大化 Actor 网络的目标函数，PPO 算法在保证策略更新稳定性的同时逐步提升策略性能。

2　两阶段时间尺度状态分离 PPO 算法

2.1　空战建模

六自由度无人机的运动分为平动和转动。在平动中，无人机改变的状态量包括其三维坐标 (p_x,p_y,p_z)、在 3 个轴方向的速度矢量 $\boldsymbol{v}=(v_x,v_y,v_z)$、速度大小 V。在转动中，无人机改变的状态量包括滚转角 ϕ、俯仰角 θ、偏航角 ψ、滚转角的角速度 p、俯仰角的角速度 q、偏航

角的角速度 r_ω、迎角 α（$\boldsymbol{v}$ 在无人机几何对称平面 $o_bx_bz_b$ 内的投影与机体轴 o_bx_b 间的夹角）、侧滑角 β（$\boldsymbol{v}$ 与无人机几何对称平面 $o_bx_bz_b$ 间的夹角）。

无人机的运动受重力、空气动力和发动机功率的综合影响。在发动机功率方面，发动机的油门控制指令 δ_T 决定了在机体坐标系 x 轴的加速度。无人机使用 3 个控制舵对动力学项进行控制，通过指令 δ_a、δ_e、δ_r 对副翼、升降舵和方向舵的偏转角分别进行调整，用于控制无人机的横滚、俯仰和偏航。无人机的状态由 12 个非线性运动方程控制，动力学部分可简化为：

$$\dot{\xi}=f(\xi,u) \tag{6}$$

式中：$\dot{\xi}$ 为无人机的状态向量 ξ 对时间的导数，$\xi=(V,A,\alpha,\beta,\phi,\theta,p,q,r_\omega)$，其中 A 为无人机高度；u 为无人机的输入指令，$u=\{\delta_T,\delta_a,\delta_e,\delta_r\}$；$f(\cdot)$ 为无人机的非线性动力学方程。

在 1 对 1 近距离空战中，每个无人机的目标都是击落对手且不被对手击落，是一种零和博弈。图 2 为红方无人机的可攻击范围。将双方无人机的有效打击距离设定为 500～3 000 m，最大有效攻击角（ATA）角度 ω_0 设置为 2°，即只要双方距离为 500～3 000 m，一方的 ATA 角度 ω 小于 2°就能对另一方进行打击。

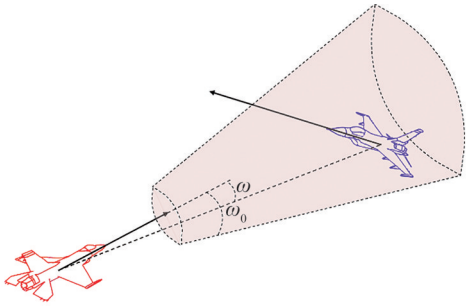


图 2　红方无人机的可攻击范围

Fig. 2　Attack range of red team's drone

每次打击的伤害 D_{Hit} 与双方的距离 d_r（单位为 m）的关系由分段函数表示如下：

$$D_{Hit}=\begin{cases}0, d_r>3\,000;\\ \dfrac{3\,000-d_r}{2\,500}, 500\le d_r\le 3\,000;\\ 0, d_r<500\end{cases} \tag{7}$$

2.2　时分空战框架

在六自由度无人机空战场景中，由于状态空间的高维性和非线性，以及动作空间的连续性，使用端到端训练方式存在难以学习到有效策略、收敛缓慢和泛化性差的问题。

为解决上述问题，在六自由度空战场景中引入时间尺度分离概念，将空战机动划分成短周期运动和长周期运动。短周期运动主要是转动运动；长周期运动是力的再平衡运动过程，表现为无人机质心的平移或轨迹运动[26]。据此，将空战任务分为短周期的转动速度变化决策和长周期的平移、轨迹运动决策两层。

之前的研究虽然对空战场景进行了分层，但在低层和高层的训练中仍使用全状态量，缺乏实质性的高、低层分离。本文在时间尺度分离的框架上，进一步将六自由度空战场景中涉及的状态量分为短周期和长周期两种，并对两种状态量进行分组决策，以实现状态空间的降维和决策机动的解耦合。其中，短周期状态量包括滚转角的角速度、俯仰角的角速度、偏航角的角速度、俯仰角、滚转角、偏航角和速度，长周期状态量包括位置等，实现了时间尺度上的状态量分离。

图 3 为时分空战框架。图 3 中，$\boldsymbol{S}_{slow}$ 输入长周期策略后输出期望的状态量 $\boldsymbol{S}_{4,desire}=(\phi_c,\theta_c,V_c,\beta_c)$，$\phi_c$、$\theta_c$、$V_c$、$\beta_c$ 分别为滚转角、俯仰角、速度和侧滑角的期望值，将 $\boldsymbol{S}_{fast}$ 和 $(\phi_c,\theta_c,V_c,\beta_c)$ 一起作为短周期决策的输入后得到六自由度无人机 4 个操作杆的控制指令

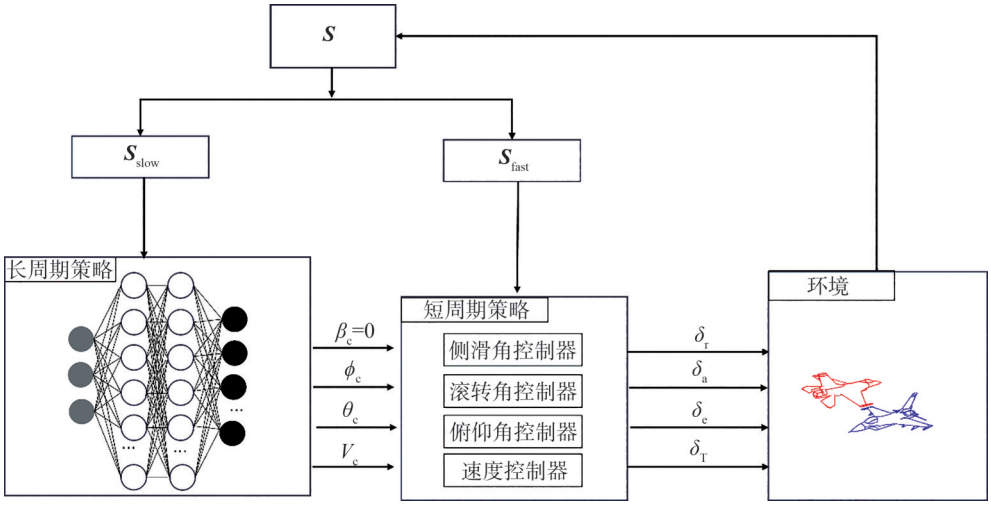


图 3　时分空战框架

Fig. 3　Time-division air combat framework

$\{\delta_{\mathrm{T}},\delta_{\mathrm{a}},\delta_{\mathrm{e}},\delta_{\mathrm{r}}\}$,通过长短周期决策共同完成空战决策。

2.2.1　短周期PID控制

在空战任务中,短周期PID控制扮演着关键角色,能够快速响应和调整飞行器的姿态和速度,对保持空中机动性和战斗效率至关重要。

使用PID算法对无人机的俯仰角、滚转角、侧滑角和速度进行控制[26]。滚转角控制器主要负责调节无人机绕纵轴的滚转角度,使其可以进行横向运动;俯仰角控制器调节无人机绕横轴的俯仰角度,控制无人机的纵向运动;侧滑角控制器调节无人机的侧滑角度,保证其横向稳定性。此外,还使用了速度油门控制,配合俯仰角控制器实现对无人机的纵向控制。

1)滚转角控制器

无人机在水平面上转弯时,需要通过调整其滚转角ϕ来实现。因此,滚转角控制器需要具备快速响应和最小超调能力。该控制器接收滚转角的期望值ϕ_{c}作为输入,并生成相应的指令δ_{a},使无人机逐步达到ϕ_{c}。滚转角控制器的内、外回路分别采用滚转角的角速度p和滚转角ϕ作为反馈。整个控制过程的数学表达式如下:

$$\delta_{\mathrm{a}}=-[k_{\phi}(\phi_{\mathrm{c}}-\phi)]-k_{p}p \tag{8}$$

式中,k_{ϕ}为滚转角误差增益,k_{p}为滚转角的角速度阻尼增益。

2)俯仰角控制器

在控制无人机的纵向飞行时,即改变其高度或速度时,需要通过俯仰角控制器来调整其俯仰角。该控制器接收俯仰角的期望值θ_{c}作为输入,并生成相应的指令δ_{e},使无人机逐步达到θ_{c}。俯仰角控制器的内、外回路分别采用俯仰角的角速度q和俯仰角θ作为反馈。整个控制过程的数学表达式如下:

$$\delta_{\mathrm{e}}=-[k_{\theta}(\theta_{\mathrm{c}}-\theta)]-k_{q}q \tag{9}$$

式中,k_{θ}为俯仰角误差增益,k_{q}为俯仰角的角速度阻尼增益。

3)侧滑角控制器

由于转弯过程中无人机会产生侧滑角β,需要通过操纵方向舵来消除侧滑角,以维持无人机的稳定性。该控制器接收侧滑角的期望值β_{c}作为输入,输出相应的指令δ_{r}。侧滑角控制器的内、外回路分别采用偏航角的角速度r_{ω}和侧滑角β作为反馈。考虑到侧滑角控制存在稳态误差问题,采用比例积分控制方法。整个控制过程的数学表达式如下:

$$\delta_{\mathrm{r}}=\left(k_{\beta}+k_{\beta\mathrm{i}}\frac{1}{s}\right)(\beta_{\mathrm{c}}-\beta)-k_{r,\omega}r_{\omega} \tag{10}$$

式中,k_{β}为侧滑角误差比例增益,$k_{\beta\mathrm{i}}$为侧滑角误差积分增益,$k_{r,\omega}$为偏航角的角速度阻尼增益,$1/s$为积分的拉普拉斯表示形式。

4)油门速度控制器

调节油门控制指令可以改变无人机的动力输出,从而调整无人机的速度,通过与俯仰角控制系统配合使用,实现对无人机速度的控制。将油门速度控制器接收期望的速度值V_{c}作为输入,并输出相应的指令δ_{T},调整无人机的速度以达到预期目标。该控制系统的反馈量为速度大小V,采用了比例积分控制以解决稳态误差问题。整个控制过程的数学表示达式如下:

$$\delta_{\mathrm{T}}=\left(k_{V}+k_{V\mathrm{i}}\frac{1}{s}\right)(V_{\mathrm{c}}-V) \tag{11}$$

式中,k_{V}为速度误差比例增益,$k_{V\mathrm{i}}$为速度误差积分增益。

2.2.2　长周期PPO决策

无人机空战场景可以被建模成一个对称零和博弈的马尔可夫过程$\boldsymbol{M}_{\mathrm{p}}=(\boldsymbol{S}_{\mathrm{MP}},\overline{\boldsymbol{S}}_{\mathrm{MP}},\boldsymbol{A}_{\mathrm{MP}},\overline{\boldsymbol{A}}_{\mathrm{MP}},R_{\mathrm{MP}},\overline{R}_{\mathrm{MP}},T_{\mathrm{MP}})$[27],其中,$(\boldsymbol{S}_{\mathrm{MP}},\boldsymbol{A}_{\mathrm{MP}},R_{\mathrm{MP}})$和$(\overline{\boldsymbol{S}}_{\mathrm{MP}},\overline{\boldsymbol{A}}_{\mathrm{MP}},\overline{R}_{\mathrm{MP}})$分别代表了博弈双方的状态空间、动作空间和奖励,$T_{\mathrm{MP}}$为状态转移概率函数。

1)状态空间

为了增强训练无人机的能力并提高其泛化性能,无人机获取的信息$\boldsymbol{S}_{k}$包括我方位置$\boldsymbol{p}_{\mathrm{m}}$、我方速度$\boldsymbol{v}_{\mathrm{m}}$、我方姿态角(包括迎角$\alpha_{\mathrm{m}}$和侧滑角$\beta_{\mathrm{m}}$)、我方姿态角速度(包括机体轴系下的滚转角、俯仰角、偏航角的角速度p_{m}、q_{m}和$r_{\omega\mathrm{m}}$)、敌方位置$\boldsymbol{p}_{\mathrm{e}}$和敌方速度$\boldsymbol{v}_{\mathrm{e}}$。然而,由于状态量较多,如果直接将所有状态量用于强化学习,将导致状态空间非常庞大,降低训练效率和网络最终的泛化性能。因此,需要对状态量进行适当处理。

根据第2.2.1节的分析,短周期决策部分已经完成了短周期状态量的控制,因此长周期决策部分只需关注双方相对态势,包括相对位置$\boldsymbol{p}_{\mathrm{r}}=(p_{\mathrm{rx}},p_{\mathrm{ry}},p_{\mathrm{rz}})$、相对速度$\boldsymbol{v}_{\mathrm{r}}=(v_{\mathrm{rx}},v_{\mathrm{ry}},v_{\mathrm{rz}})$,以及我方此时的速度$\boldsymbol{v}_{\mathrm{m}}=(v_{\mathrm{mx}},v_{\mathrm{my}},v_{\mathrm{mz}})$。与使用所有状态量进行训练相比,本文仅使用9个状态量进行训练,大大减小了状态空间的规模,降低了训练难度。

为了进一步增强训练网络的泛化性能,还需对状态量进行适当处理,调整原始的坐标系,将其绕z轴旋转,将x轴旋转到$\boldsymbol{v}$在xoy平面的投影位置,从而得到新的坐标系。然后,在新的坐标系下表示上述矢量信息。整个转换过程使用的转换矩阵$\boldsymbol{M}$为:

$$\boldsymbol{M}=\begin{bmatrix}\dfrac{v_y}{l} & -\dfrac{v_x}{l} & 0\\ \dfrac{v_x}{l} & \dfrac{v_y}{l} & 0\\ 0 & 0 & 1\end{bmatrix} \tag{12}$$

式中，l为$\boldsymbol{v}$在xoy平面投影的模长（长度），$l=\sqrt{v_x^2+v_y^2}$。

通过式（13）进行转换后可以得到新的状态量：

$$\begin{cases}\boldsymbol{p}'_{\mathrm{r}}=\boldsymbol{M}\cdot\boldsymbol{p}_{\mathrm{r}},\\ \boldsymbol{v}'_{\mathrm{r}}=\boldsymbol{M}\cdot\boldsymbol{v}_{\mathrm{r}},\\ \boldsymbol{v}'_{\mathrm{m}}=\boldsymbol{M}\cdot\boldsymbol{v}_{\mathrm{m}}\end{cases}\tag{13}$$

式中，$\boldsymbol{p}'_{\mathrm{r}}$、$\boldsymbol{v}'_{\mathrm{r}}$、$\boldsymbol{v}'_{\mathrm{m}}$为转换后的相对位置、相对速度和我方速度。

综上所述，最终的状态量为$(p'_{\mathrm{rx}},p'_{\mathrm{ry}},p'_{\mathrm{rz}},v'_{\mathrm{rx}},v'_{\mathrm{ry}},v'_{\mathrm{rz}},v'_x,v'_y,v'_z)$。

2）动作空间

如图3所示，短周期策略和长周期决策的接口是ϕ_{c}、θ_{c}、V_{c}及β_{c}。侧滑角的目的是保持无人机稳定，因此β_{c}始终设置为0，不再由长周期决策控制。

长周期策略并不直接输出短周期策略中4个操作杆的控制量，而是输出长短周期接口状态量的期望值，包括ϕ_{c}、θ_{c}和V_{c}，输出动作可表示为action$(\phi_{\mathrm{c}},\theta_{\mathrm{c}},V_{\mathrm{c}})$。短周期决策的侧滑角期望值始终为0，其与另外3个期望值一起被传递给短周期策略，短周期策略随后输出真实的动作值。

由于底层PID控制器具有一定的延迟，需要一定的时间来达到相应的控制量，因此在实验中，将高层策略网络的更新频率设置为0.5 Hz，而底层PID控制器的更新频率为10.0 Hz。

与常规使用离散底层控制动作的方法相比，本文采用连续动作空间，将俯仰角范围设置为(–90°, 90°)，滚转角范围设置为(–180°, 180°)，速度（单位为m/s）范围设置为(0, 400)。并且，所选择的动作非常接近最底层的动作，使得训练的网络能够更有效地学习到更多、更好的策略。

3）奖励函数

为了避免仅在终局奖励情况下获得稀疏奖励，提高训练效率，将无人机的打击任务划分为两个阶段：追击和打击阶段。在追击阶段，主要目标是使敌机进入我机的攻击范围；而在打击阶段，当敌机进入我机的攻击范围时，我机会调整角度对敌方进行打击。因此，奖励设计也分别针对这两个阶段，引导我机靠近敌机和对敌机进行打击。

a. 相对位置奖励。为了激励我方无人机在敌方无人机后方并使自身速度方向朝向敌方，同时抑制敌方在我后方形成相似态势，构建相对位置奖励。其计算公式如下：

$$\begin{cases}R_1=k_1\cos\theta_1,\\ R_2=k_2\cos\theta_2,\\ R_{\mathrm{closure1}}=R_1+R_2\end{cases}\tag{14}$$

式中：R_1为我方在敌方后方的奖励；R_2为敌方在我方后方的惩罚；R_{closure1}为综合相对位置奖励；k_1和k_2为超参数，$|k_1|+|k_2|=1$，$k_1>0$，$k_2<0$，可以通过调节k_1和k_2的大小来让无人机的策略偏向攻击或防守；θ_1和θ_2分别为我方速度与我方、敌方连线的夹角，敌方速度与敌方、我方连线的夹角。

b. 距离奖励。我方无人机在追击时靠近敌机获得奖励R_3，我方无人机在被追击时距离变近获得惩罚R_4，R_{closure2}为综合距离奖励。其计算公式如下：

$$\begin{cases}R_3=\ln\left(\dfrac{1-1/\mathrm{e}}{5\,000}d_{\mathrm{r}}+\dfrac{1}{\mathrm{e}}\right),\\ R_4=-\ln\left(\dfrac{1-1/\mathrm{e}}{5\,000}d_{\mathrm{r}}+\dfrac{1}{\mathrm{e}}\right),\\ R_{\mathrm{closure2}}=R_3+R_4\end{cases}\tag{15}$$

c. 打击奖励。当敌方无人机在我方攻击范围或我方出现在敌方攻击范围（双方距离为500 m≤d_{r}≤3 000 m、ATA角度$\omega\le2°$或敌方攻击角$\omega^{\mathrm{enemy}}\le2°$）时，设置打击奖励。若在攻击距离范围内攻击敌方，$d_{\mathrm{r}}$越小奖励越大，$\omega$越小奖励越大；而被敌方攻击时，$d_{\mathrm{r}}$越小惩罚越大，$\omega^{\mathrm{enemy}}$越小惩罚越大。打击奖励的计算公式如下：

$$\begin{cases}R_{\mathrm{Hit}}=k_1\dfrac{3\,000-d_{\mathrm{r}}}{2\,500}+k_2\dfrac{\omega_0-\omega}{\omega_0},500\le d_{\mathrm{r}}\le3\,000;\\ R_{\mathrm{BeHit}}=k_1\dfrac{3\,000-d_{\mathrm{r}}}{2\,500}+k_2\dfrac{\omega_0-\omega^{\mathrm{enemy}}}{\omega_0},500\le d_{\mathrm{r}}\le3\,000;\\ R_{\mathrm{hit}}=k_3R_{\mathrm{Hit}}-k_4R_{\mathrm{BeHit}}\end{cases}\tag{16}$$

式中：R_{Hit}为我方攻击奖励；R_{BeHit}为我方被攻击惩罚；R_{hit}为综合打击奖励；$|k_3|+|k_4|=1$，$k_3>0$，$k_4<0$，可以通过调节k_3和k_4的值来让无人机的策略偏向攻击或偏向防守，进而有目的地对我方无人机策略进行训练。

d. 高度引导奖励。当无人机高度A超过8 500 m或低于2 500 m，即为接近高度限制时给予惩罚R_{deck}。其计算公式如下：

$$R_{\mathrm{deck}}=\begin{cases}\ln\left(\dfrac{1-1/\mathrm{e}}{1\,500}(A-1\,000)+\dfrac{1}{\mathrm{e}}\right),A\in(1\,000,2\,500);\\ \ln\left(\dfrac{1-1/\mathrm{e}}{1\,500}(10\,000-A)+\dfrac{1}{\mathrm{e}}\right),A\in(8\,500,10\,000)\end{cases}\tag{17}$$

e. 终局奖励。当训练回合因以下任一条件而终止时，将触发终局奖励机制：①高度失控；②超出有效距离；③任务超时；④胜利；⑤失败。如果无人机高度小于1 000 m或大于10 000 m，给予惩罚r_1；如果无人机间的距离大于设置的最大距离$d_{\max}$，给予惩罚r_2；如果无人机生存时间t_{a}超过最大对局时长$t_{\mathrm{a,end}}$而对局未结束，给予奖励r_3；如果击败敌方无人机则获得奖励r_4；如果被对手击败获得惩罚r_5。在设定中，为了让我方无

人机学到更好的攻击策略，更注重打击对手，将超参数 $(r_1, r_2, r_3, r_4, r_5)$ 设为 $(-100, -60, 20, 1\,000, -100)$。终局奖励 R_{other} 的计算方式如下：

$$R_{\text{other}} = \begin{cases} r_1, A \notin [1\,000, 10\,000]; \\ r_2, d_{\text{r}} > d_{\max}; \\ r_3, t_{\text{a}} > t_{\text{a,end}}; \\ r_4, \text{胜利}; \\ r_5, \text{失败} \end{cases} \tag{18}$$

2.3　两阶段训练的 TTS–PPO 算法

采取两阶段训练对长周期策略的 PPO 网络进行训练。长周期策略的两阶段训练如图 4 所示。

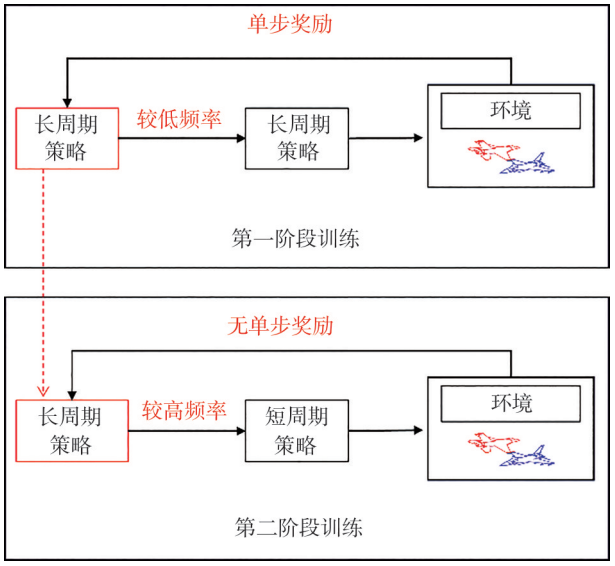


图 4　长周期策略的两阶段训练

Fig. 4　Long-cycle policy two-stage training

图 4 上方的框图代表第一阶段训练：在设计奖励时，根据两机间的距离是否在可打击范围，又将空战任务分为追击和打击阶段，各自设计相应的单步奖励；由于短周期控制具有一定的延时性，长周期策略使用较低的决策频率 0.5 Hz，即每 2 s 决策一次。通过这两种方式使无人机训练过程能快速度过冷启动时期，以加快模型收敛。

将第一阶段训练收敛得到的模型作为第二阶段训练的初始模型，图 4 下方的框图代表第二阶段训练。为了进一步提高无人机的目标决策能力，去掉单步奖励，仅保留终局奖励以解决奖励重塑对无人机训练目标带来的影响，并将决策频率升高为 2 Hz，即每 0.5 s 决策一次，减小单步奖励对无人机的影响，以提高长周期策略的性能上限，尽可能提升无人机的作战性能。

针对六自由空战任务，在深度强化学习中引入时分空战框架，将环境状态量进行时间尺度分离，并在训练过程中使用两阶段训练，最后得到本文提出的 TTS–PPO 算法，算法的流程如下。

算法 TTS–PPO

1. 利用短周期状态量完成对滚转角、侧滑角、俯仰角和速度油门的控制。
2. 长周期策略第一阶段训练。
3. 初始化 Actor 和 Critic 网络参数。
4. 对于每个训练回合（episode，记为 e，e=1,2,3…）。
5. 初始化缓冲区 Buffer。
6. 使用当前策略与环境交互，将采集的信息存储到 Buffer 中。
7. 更新 Actor 和 Critic 网络参数。
8. 存储收敛的策略。
9. 长周期策略第二阶段训练。
10. 使用第一阶段训练的网络作为第二阶段的初始值，并调整奖励函数以及提高决策频率。
11. 对于每个训练回合 e=1,2,3…。
12. 初始化缓冲区 Buffer。
13. 使用当前策略与环境交互，将采集的信息存储到 Buffer 中。
14. 更新 Actor 和 Critic 网络参数。
15. 存储收敛的策略。
16. 结束训练。

3　实验结果及分析

3.1　短周期 PID 控制实验

3.1.1　实验设置

实验选用开源 F-16 无人机模型作为研究对象，并利用开源动力学模型 JSBSim 进行仿真。在短周期决策部分的实验中，采用 PID 算法对无人机的俯仰角、滚转角、侧滑角和速度进行控制。将无人机的初始状态设定为俯仰角、侧滑角、滚转角均为 0°，速度为 125 m/s。为了探究不同控制目标下系统的性能，分别设置 3 种实验条件：在俯仰角控制实验中，目标俯仰角分别设置为 10°、20°和 30°；在滚转角控制实验中，目标滚转角分别设置为 20°、30°和 40°，同时要求侧滑角保持为 0°；在速度控制实验中，目标速度分别设定为 100、200 和 300 m/s。通过以上实验设计，能够在不同目标参数下评估系统的控制性能及响应特性。

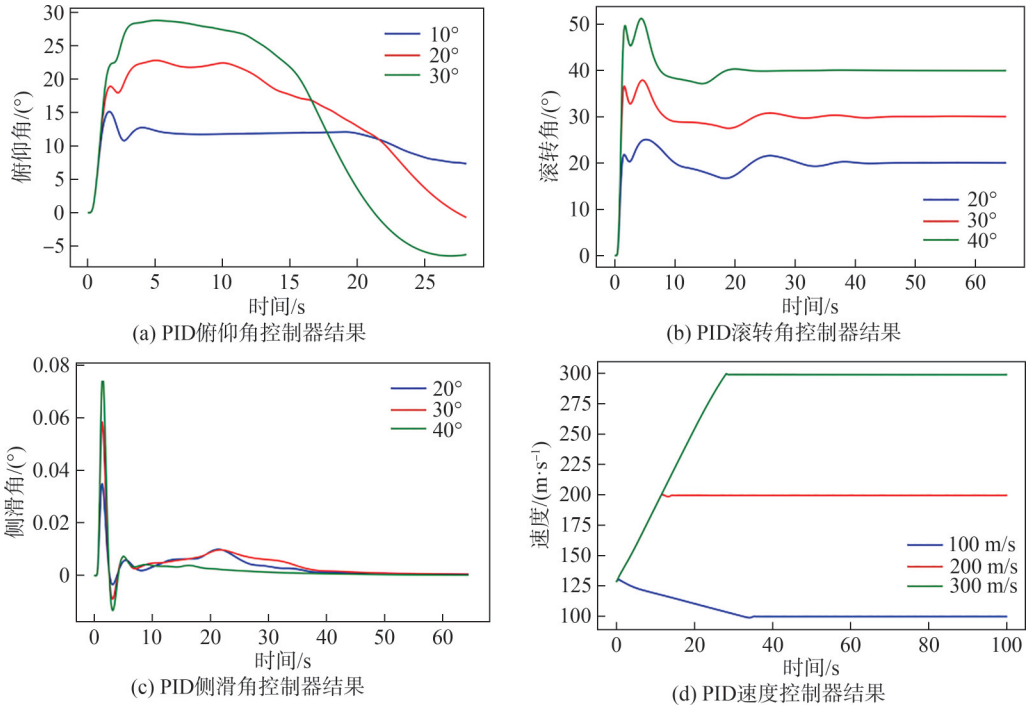
3.1.2　实验结果

图 5 为 PID 控制效果。图 5(a) 为俯仰角控制器结果。该控制器反应非常灵敏，通常在 4 s 左右即可接近目标俯仰角。然而，由于仅控制了俯仰角而未考虑无人机的速度变化等因素，该俯仰角并不能长时间保持稳定。因此，俯仰角控制需要与油门速度控制相结合，才能对无人机进行有效的纵向控制。

图5(b)、(c)分别为滚转角和侧滑角控制器结果。由于无人机在横向转弯时会产生侧滑,因此滚转角和侧滑角控制需要同时进行。实验结果表明,滚转角控制器通常在2 s左右即可达到目标滚转角附近,然后经过反复微调达到稳定值。在滚转角控制过程中,当滚转角变化较快时,侧滑角也会相应变化较快。但是,在侧滑角控制下,侧滑角始终能够保持在0°附近,表现出较好的控制效果。

图5(d)为速度控制器结果。速度控制是整个控制过程中的重要组成部分,它与俯仰角控制结合使用才能有效控制无人机的纵向运动。实验结果表明,无人机通常在20～30 s达到目标速度,并保持稳定。

短周期控制部分的表现对长周期决策至关重要,实验结果表明,使用PID算法的短周期控制部分能够较快地响应长周期决策,展现出了良好的控制效果和可行性。



(a) PID俯仰角控制器结果　(b) PID滚转角控制器结果

(c) PID侧滑角控制器结果　(d) PID速度控制器结果

图5　PID控制效果

Fig. 5　PID control performance

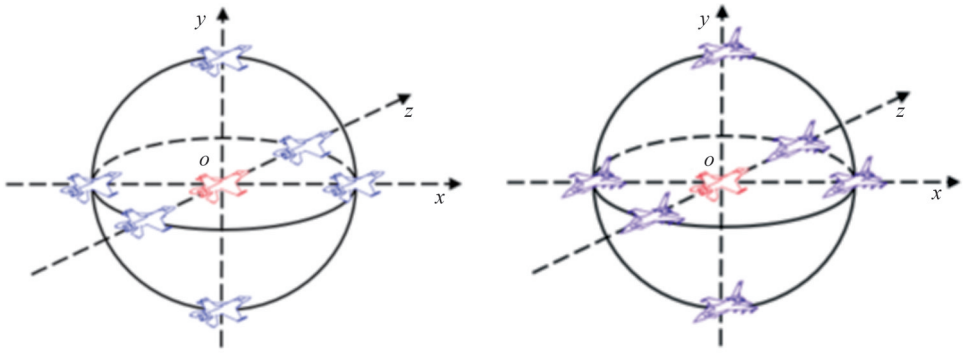
3.2　长周期空战策略实验

3.2.1　实验设置

1)实验场景

用PPO算法对长周期决策网络进行训练时,为了在数量尽可能少的空战初始场景中包含所有的相对态势,使无人机与环境交互的过程中能收集到各类不同场景下的训练数据,设计了球形相对态势场景,从中选取了最具代表性的12类场景作为训练的初始化环境。

图6为敌我双方初始位置关系。我方机体初始位置固定,而敌方机体则分布在以我方机体为球心的球面上,且处于前、后、左、右4个方位及上、下方向的不同高度位置。为考察无人机应对复杂态势的能力,在每个位置选取敌我双方初始朝向夹角为0°和



(a) 敌我双方初始朝向夹角0°　(b) 敌我双方初始朝向夹角180°

图6　敌我双方初始位置关系

Fig. 6　Initial positional relationship between enemy and us

180°两个端点值。12类场景有效覆盖了六自由度空战决策问题中三维空间的典型相对态势，并通过在训练初期强调相对态势的两个极端状态，使无人机能够在动态追击和规避过程中自然经历180°到0°的连续转变过程，从而在实际训练中并不仅局限于端点值，有助于无人机在中间态势下的策略泛化与灵活应对，最终提高模型在复杂多变的实际空战中的决策稳定性和作战性能。

2）实验设计

为验证本文提出的时分空战框架、时间尺度状态分离和两阶段训练对六自由度空战无人机训练的作用，以直线飞行的无人机作为对手分别训练了4个无人机，使用的算法为PPO、全状态PPO（FS–PPO）、时间尺度状态分离PPO（TS–PPO）、两阶段时间尺度状态分离PPO（TTS–PPO）。其中：PPO为近端策略优化算法，作为对照的基础算法；FS–PPO在PPO的基础上引入了时分空战框架；TS–PPO在FS–PPO的基础上将环境状态量进行了时间尺度分离和相对态势转化；TTS–PPO则在TS–PPO训练收敛的结果上做了舍弃单步奖励的第二阶段训练，并将决策频率从0.5 Hz提高至2.0 Hz。为便于论述，分别将4个无人机命名为PPO、FS–PPO、TS–PPO、TTS–PPO无人机。

除上述改进外，使用4种算法训练的无人机在空战初始场景、敌机策略、网络结构与超参数等方面均完全相同。表1为神经网络参数。

表1　神经网络参数

Tab. 1　Neural network parameters

参数名	取值
并行线程数	10
GAE超参数	0.95
每轮迭代次数	5
缓冲区大小	51 200
批量大小	5 120
折扣因子	0.99
动作者网络神经元	[64,256,256,64]
评价者网络神经元	[64,256,64]
学习率	0.000 1
激活函数	Tanh
优化器	Adam

注：中括号内数据表示每层神经网络的神经元数目。

3.2.2　训练过程对比

图7为PPO、FS–PPO和TS–PPO无人机的奖励变化曲线。从奖励值的变化和收敛结果来看：PPO无人机的平均奖励最终收敛到负数，表明其在训练场景中未完成打击任务，没有学习到有效的空战策略；FS–PPO无人机在约1 200次迭代训练后，平均奖励最后收敛到2 300左右，表明其完成了击败敌机的目标，验证了时分空战框架在无人机习得有效策略方面的作用。从奖励值的收敛结果和收敛速度来看：FS–PPO和TS–PPO无人机最终的平均奖励都收敛到了2 000以上，说明二者都完成了击败敌机的任务；TS–PPO无人机的最终平均奖励在2 200左右，略低于FS–PPO无人机，更重要是，其在约400次迭代训练后便收敛到了最优值，相较于FS–PPO无人机迭代次数减少67%，验证了长短周期状态量分离能使高维状态空间降维，提升了无人机的训练速度。

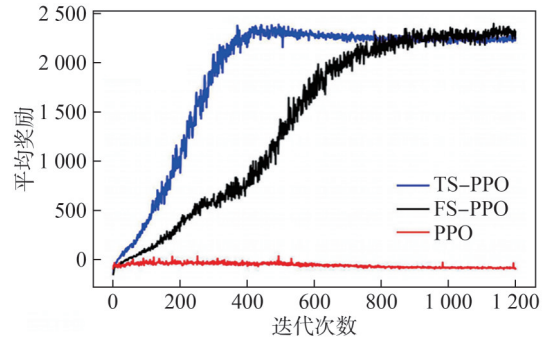


图7　PPO、FS–PPO和TS–PPO无人机的奖励变化曲线

Fig. 7　Reward variation curves of PPO, FS–PPO, and TS–PPO UAV

图8为TTS–PPO无人机的奖励曲线。TTS–PPO无人机在TS–PPO无人机的基础上，获得的奖励值进一步提升并在500次迭代训练后收敛，说明TS–PPO无人机在分阶段单步奖励和终局奖励的引导下虽然奖励值已经收敛，但是其性能仍有上升空间。这是因为凭借人工经验设计的单步稠密奖励虽能在训练初期帮助无人机更快探索达到好的终局情况，但由于人类知识的局限性和单步奖励因环境复杂而表征失真，无人机会为了获得更多单步奖励而损失一部分决策效率。实验结果证明，去掉单步奖励和提高长周期决策频率，无人机的决策能力将进一步提升。

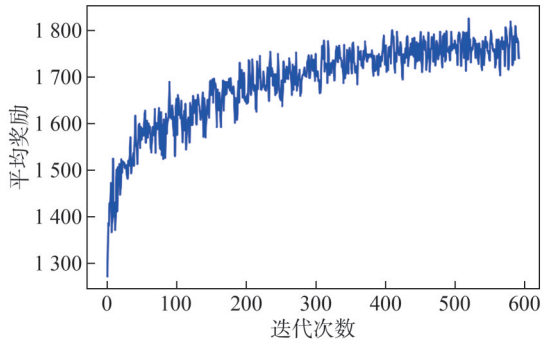


图8　TTS–PPO无人机奖励变化曲线

Fig. 8　Reward variation curve of TTS–PPO UAV

3.2.3 空战行为分析

为了更好地理解 FS–PPO、TS–PPO、TTS–PPO 无人机在训练环境中学习到的空战策略，分别在 5 种初始态势下，对 3 个无人机进行测试，从各自的运动轨迹来进行空战行为分析。图 9 为训练过程的轨迹记录。

图 9(a)为我方优势、敌方劣势情况下 3 个无人机的轨迹。TS–PPO 无人机先通过降低高度获取较大速度来靠近敌机，到达敌机正后方，调整距离和角度完成对敌机的打击；FS–PPO 和 TTS–PPO 无人机也通过相同方法靠近敌机，但在敌机下方完成打击过程，而非到敌机正后方才进行打击。图 9(b)为我方劣势、敌方优势情况下 3 个无人机的轨迹。3 个无人机都是通过提升高度来脱离敌机攻击范围，然后调整高度和角度绕到敌机后方完成对敌机的打击。图 9(c)为均势情况下 3 个无人机的轨迹。FS–PPO 无人机通过转弯调整运动方向，高度基本保持不变，到达敌机后方完成打击；TS–PPO 和 TTS–PPO 同时调整高度和运动方向到达敌机后方击落敌机。图 9(d)、(e)分别为我方高度占优和敌方高度占优，即不同初始高度态势下 3 个无人机的轨迹。FS–PPO 无人机先将高度调整到与敌方一致，然后再运动到敌机正后方完成打击；TS–PPO 和 TTS–PPO 无人机先调整高度，运动到高度低于敌机获取更大速度，然后追击敌机，在追击过程中完成打击。

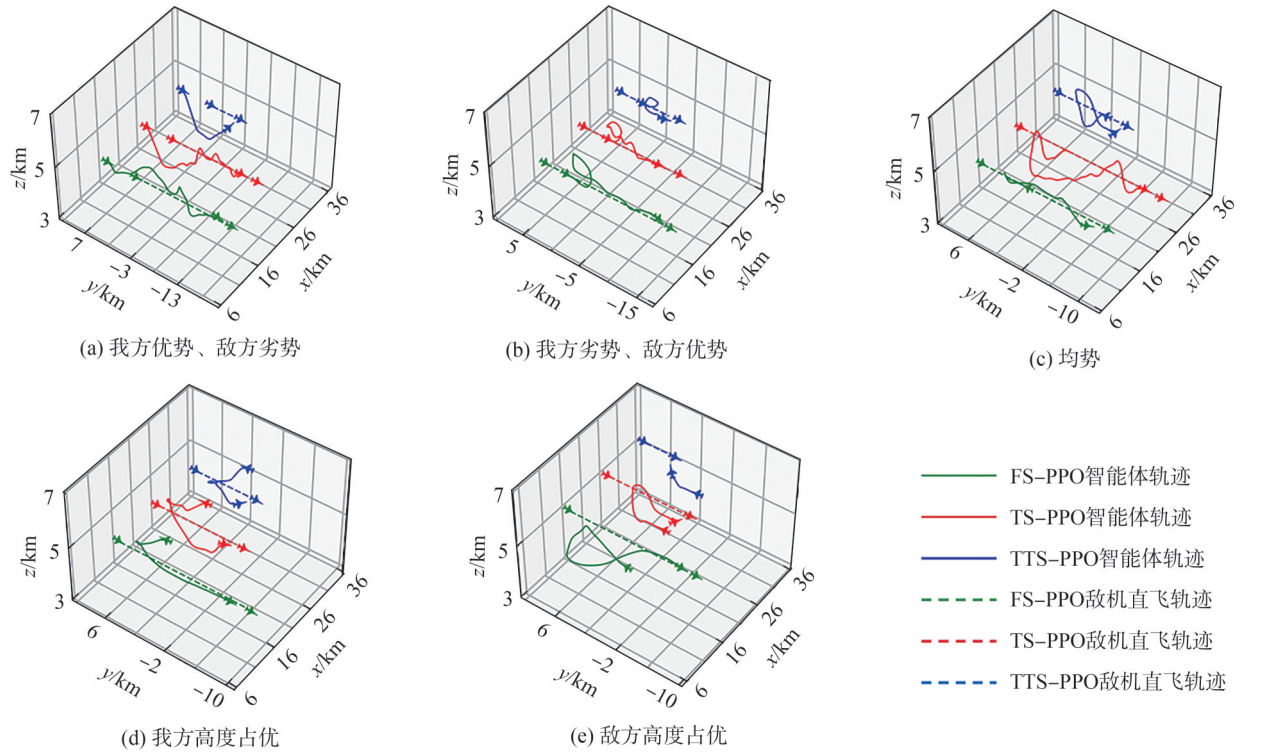


图 9 训练过程的轨迹记录

Fig. 9 Trajectory records of the training process

3 个无人机在躲避时都会通过提高自身高度进行躲避，追击时通过降低高度获取更大速度进行追击，其中 TS–PPO、TTS–PPO 无人机的高度调整更激进。与 FS–PPO 和 TS–PPO 无人机相比，TTS–PPO 无人机的策略更加有效，打击策略更多，可以在多个方位完成打击，且完成任务的运动轨迹明显更短，即使用的决策步更少。

3.2.4 两两对抗结果

训练结束后，将成功学习到有效空战策略的 FS–PPO、TS–PPO、TTS–PPO 无人机两两一组，分别进行 100 局实战对抗，通过各个无人机的胜利、失败和平局的次数来对比其空战能力，图 10 为两两对抗结果记录，其中胜利和失败的结果以该组的第一个无人机为本机视角。为了避免其他因素的干扰，在空战初始环

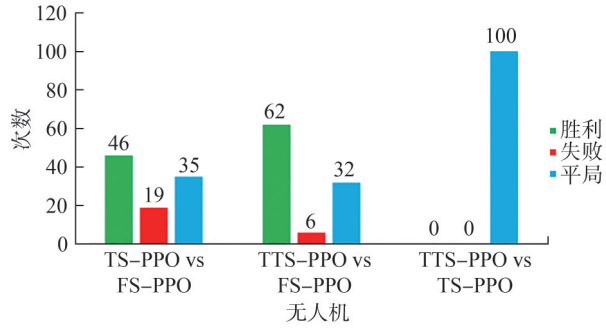


图 10 两两对抗的结果记录

Fig. 10 Records of pairwise confrontation results

境中，双方无人机互为均势，都直接面朝对方即两机初始朝向夹角为180°，且初始速度、高度及各类基本态势皆完全相同。由图10可见：在TS–PPO与FS–PPO无人机的对战中，TS–PPO无人机胜利次数为46，FS–PPO无人机的胜利次数为19，剩余35局为平局，表明TS–PPO无人机在实际空战中的作战性能优于FS–PPO无人机，进一步验证了长短周期状态分离不仅能够加速算法收敛，还能使训练的无人机在实际空战场景中面对新对手时，有更强的机动性和更好的空战策略，提高了模型的泛化性。在TTS–PPO与FS–PPO无人机的对战中，TTS–PPO无人机胜利次数为62，FS–PPO无人机的胜利次数仅为6，剩余32局为平局。在与FS–PPO无人机的对抗中，TTS–PPO无人机的胜率比TS–PPO无人机有了明显的提升，表明TTS–PPO无人机在实际空战中的作战性能更优于TS–PPO无人机，进一步证明了第二阶段舍弃单步奖励的训练不仅提升了无人机获得的奖励值，还使其学习到了更优秀的空战策略。在TTS–PPO与TS–PPO的100局对抗中，最终结果都是平局。

为了更好地分析二者的空战表现，需要对抗过程中更详细的态势信息，图11为TS–PPO和TTS–PPO对抗过程记录，展示了对抗过程中双方的距离，以及各自的攻击角变化和血量变化。

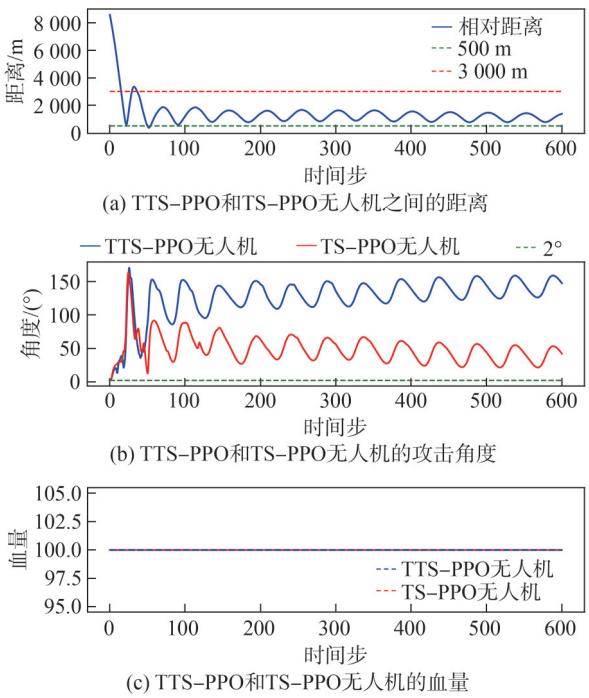


图11　TS–PPO和TTS–PPO对抗过程记录

Fig. 11　Process records of the confrontation between TS–PPO and TTS–PPO

图11(a)为双方距离。两机间的距离在15 s左右就减小到最大攻击距离3 000 m以内，随后在可攻击范围500～3 000 m不断波动。图11(b)为双方攻击角度。在整个对抗过程中，双方的攻击角度都在不断波动，但在达到可攻击距离范围后，都未曾小于可攻击的最大值，因而有了图11(c)所示的双方血量，即双方一直保持满血到对局结束。进一步分析图11(b)可知，虽然双方无人机的攻击角都未达到可攻击的范围，但在整个对抗周期里，TTS–PPO无人机的攻击角比TS–PPO无人机更小，证明了TTS–PPO无人机在机动性能和决策效率上优于TS–PPO无人机，即在空战对抗中的表现更佳。

3.2.5　TTS–PPO与专家无人机对抗

在前面的实验中，已经对TTS–PPO训练的无人机进行了比较全面的评估。为进一步检验该无人机的实战能力，本文设计并引入一套规则型专家无人机作为性能参照。该专家无人机通过明确的控制准则来优化自身的朝向、速度与高度，以在近距离空战中获得优势态势，具体策略如下。

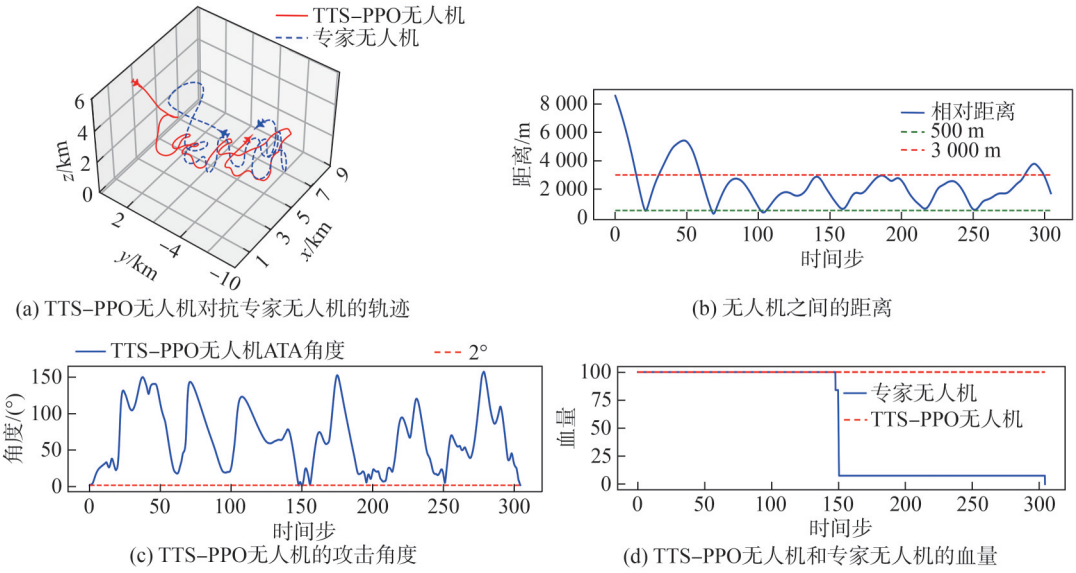
1)攻击与朝向控制：无人机在敌机距本机500～3 000 m且攻击角小于2°时有效打击，并在更近距离内给予更高打击伤害。为实现这一目标，无人机根据实时计算的朝向矢量与相对位置矢量精准调整转向，在恰当的攻击角下高效打击。

2)速度控制策略：在较远距离下，无人机维持高于敌机的速度以迅速接近目标；随战斗距离缩短，逐步降低速度并逼近敌机速度水平。在满足有效攻击条件的范围内，无人机将速度控制在与敌机相当水平，以确保较佳的命中率与机动性。

3)高度控制策略：无人机通常选择略高于敌机的飞行高度，以获得潜在的重力势能储备。必要时，可迅速将重力势能转化为动能，提升瞬时机动能力，从而在进攻与防御中更具优势。

图12为TTS–PPO无人机与专家无人机对抗过程记录。初始时，双方处于均势，距离8 500 m。0～147.7 s，双方距离先不断减小，然后双方ATA角度和距离不断变化，双方优劣势不断转换。在147.7～150.5 s，双方距离在1 500～2 000 m，TTS–PPO无人机ATA角度都达到攻击要求，对专家无人机进行连续打击，专家无人机血量从100.0减小到7.5。然后，双方再次进入缠斗阶段，最终在304.4 s时专家无人机再次出现在TTS–PPO无人机攻击范围内，专家无人机被击落。整个过程中，专家无人机未能对TTS–PPO无人机进行打击。

通过分析与专家无人机的对抗过程，发现TTS–PPO无人机能够在与对手缠斗中寻找攻击机会，且具有较强的躲避能力。



(a) TTS–PPO无人机对抗专家无人机的轨迹　(b) 无人机之间的距离

(c) TTS–PPO无人机的攻击角度　(d) TTS–PPO无人机和专家无人机的血量

图12　TTS–PPO与专家无人机对抗过程记录

Fig. 12　Process records of the confrontation between TTS–PPO and the Expert UAV

4　结　论

针对六自由度空战场景中高维空间导致的DRL算法学习策略困难、收敛缓慢和模型泛化性差以及奖励函数设计依赖人类经验的局限性等问题，本文提出了一种基于时分空战框架的TTS–PPO算法。该框架基于时间尺度分离理论将空战决策分为长短周期决策，而后设计了相应的长短周期状态量，并在训练过程中采取两阶段训练策略。通过多个实验验证了框架对无人机习得有效策略的作用，时间尺度状态分离对收敛速度和模型泛化性的提升，以及两阶段训练中后期去除单步奖励和提高决策频率对无人机作战性能的增强。上述工作为六自由度无人机空战的机动决策任务提供了理论支撑和实验验证，也为存在时间尺度差异状态量的同类型强化学习任务提供了一种新的解决思路。

参考文献：

[1] Yu Huangchao,Niu Yifeng,Wang Xiangke.Stages of development of unmanned aerial vehicles[J].National Defense Technology,2021,42(3):18–24.[喻煌超,牛轶峰,王祥科.无人机系统发展阶段和智能化趋势[J].国防科技,2021,42(3):18–24.]

[2] Chen Hao,Huang Jian,Liu Quan,et al.Review and prospects of autonomous air combat maneuver decisions[J].Control Theory & Applications,2023,40(12):2104–2129.[陈浩,黄健,刘权,等.自主空战机动决策技术研究进展与展望[J].控制理论与应用,2023,40(12):2104–2129.]

[3] Horie K,Conway B A.Optimal fighter pursuit-evasion maneuvers found *via* two-sided optimization[J].Journal of Guidance,Control,and Dynamics,2006,29(1):105–112.

[4] Zhou Siyu,Wu Wenhai, Kong Fan'e,et al.Improved multi-stage influence diagram maneuvering decision method based on stochastic decision criterions[J].Transactions of Beijing Institute of Technology,2013,33(3):296–301.[周思羽,吴文海,孔繁峨,等.基于随机决策准则的改进多级影响图机动决策方法[J].北京理工大学学报,2013,33(3):296–301.]

[5] Smith R E,Dike B A,Mehra R K,et al.Classifier systems in combat:Two-sided learning of maneuvers for advanced fighter aircraft[J].Computer Methods in Applied Mechanics and Engineering,2000,186(2/3/4):421–437.

[6] Chen Xia,Liu Min,Hu Yongxin.Study on UAV offensive/defensive game strategy based on uncertain information[J].Acta Armamentarii,2012,33(12):1510–1515.[陈侠,刘敏,胡永新.基于不确定信息的无人机攻防博弈策略研究[J].兵工学报,2012,33(12):1510–1515.]

[7] Wang Xuan,Wang Weijia,Song Kepu,et al.UAV air combat decision based on evolutionary expert system tree[J].Ordnance Industry Automation,2019,38(1):42–47.[王炫,王维嘉,宋科璞,等.基于进化式专家系统树的无人机空战决策技术[J].兵工自动化,2019,38(1):42–47.]

[8] Zhang Hongpeng,Huang Changqiang,Xuan Yongbo,et al.Maneuver decision of autonomous air combat of unmanned combat aerial vehicle based on deep neural network[J].Acta Armamentarii,2020,41(8):1613–1622.[张宏鹏,黄长强,轩永波,等.基于深度神经网络的无人作战飞机自主空战机动决策[J].兵工学报,2020,41(8):1613–1622.]

[9] Mnih V,Kavukcuoglu K,Silver D,et al.Human-level control through deep reinforcement learning[J].Nature,2015,518(7540):529–533.

[10] Silver D,Huang A,Maddison C J,et al.Mastering the game of Go with deep neural networks and tree search[J]. Nature,2016,529(7587):484–489.

[11] Berner C, Brockman G, Chan B, et al. Dota 2 with large scale deep reinforcement learning[EB/OL].(2019 – 12 – 13)[2024–06–02].https://arxiv.org/abs/1912.06680v1.

[12] Zhang Xianbing,Liu Guoqing,Yang Chaojie, et al.Research on air confrontation maneuver decision-making method based on reinforcement learning[J].Electronics, 2018,7(11):279.

[13] Yang Qiming,Zhang Jiandong,Shi Guoqing,et al.Maneuver decision of UAV in short-range air combat based on deep reinforcement learning[J].IEEE Access,2020,8:363–378.

[14] Austin F,Carbone G,Falco M,et al.Automated maneuvering decisions for air-to-air combat[C]//Proceedings of the Guidance, Navigation and Control Conference. Monterey: AIAA,1987:659–669.

[15] Li Yue,Han Wei,Wang Yongqing.Deep reinforcement learning with application to air confrontation intelligent decision-making of manned/unmanned aerial vehicle cooperative system[J].IEEE Access,2020,8:67887–67898.

[16] Pope A P,Ide J S,Mićović D,et al.Hierarchical reinforcement learning for air-to-air combat[C]//Proceedings of the 2021 International Conference on Unmanned Aircraft Systems(ICUAS).Athens:IEEE,2021:275–284.

[17] Li Yongfeng, Lyu Yongxi, Shi Jingping, et al. Autonomous maneuver decision of air combat based on simulated operation command and FRV-DDPG algorithm[J].Aerospace, 2022,9(11):658.

[18] Chai Jiajun,Chen Wenzhang,Zhu Yuanheng,et al.A hierarchical deep reinforcement learning framework for 6-DOF UCAV air-to-air combat[J].IEEE Transactions on Systems, Man,and Cybernetics:Systems,2023,53(9):5417–5429.

[19] Piao Haiyin,Sun Zhixiao,Meng Guanglei,et al.Beyond-visual-range air combat tactics auto-generation by reinforcement learning[C]//Proceedings of the 2020 International Joint Conference on Neural Networks(IJCNN). Glasgow:IEEE,2020:1–8.

[20] Sun Zhixiao,Piao Haiyin,Yang Zhen,et al.Multi-agent hierarchical policy gradient for Air Combat Tactics emergence via self-play[J]. Engineering Applications of Artificial Intelligence,2021,98:104112.

[21] Berndt J. JSBSim: An open source flight dynamics model in C++[C]//Proceedings of the AIAA Modeling and Simulation Technologies Conference and Exhibit. Providence: AIAA,2004:AIAA2004–4923.

[22] Buffington J M,Adams R J,Banda S S.Robust,nonlinear, high angle-of-attack control design for a supermaneuverable vehicle[C]//Proceedings of the Guidance, Navigation and Control Conference.Monterey:AIAA,1993:690–700.

[23] Reiner J,Balas G J,Garrard W L.Flight control design using robust dynamic inversion and time-scale separation[J]. Automatica,1996,32(11):1493–1504.

[24] Li Yun,Ang K H,Chong G C Y.PID control system analysis and design[J]. IEEE Control Systems Magazine, 2006, 26(1):32–41.

[25] Schulman J,Wolski F,Dhariwal P,et al.Proximal policy optimization algorithms[EB/OL].(2017–08–28)[2024–06–02].https://arxiv.org/abs/1707.06347v2.

[26] Rauschelbach P. Aircraft automatic flight control system: US3848833[P].1974–11–19.

[27] Garcia F, Rachelson E.Markov Decision processes in artificial intelligence Markov decision processes[M]. Hoboken:Wiley,2013:1–38.

Hierarchical Deep Reinforcement Learning Algorithm for Air Combat Based on Time Scale Separation Theory

TAN Tai[1], *JIANG Taimin*[1], *LI Bowen*[1], *LI Jie*[1], *LI Hui*[1,2*], *HUA Chenhao*[1]

(1.College of Computer Science, Sichuan University, Chengdu 610065, China;
2.National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China)

Abstract:

Objective Six-degree-of-freedom (6-DoF) unmanned aerial vehicle (UAV) air combat scenarios present substantial challenges for strategy learning when reinforcement learning methods are applied. These challenges stem from high-dimensional state spaces, continuously coupled action domains, and strongly nonlinear flight dynamics. Conventional end-to-end deep reinforcement learning (DRL) approaches struggle to achieve rapid convergence, to identify effective maneuver strategies, and to generalize learned policies beyond narrowly constrained conditions. In addition, reward functions often rely on handcrafted rules derived from human expertise, which do not ensure that higher reward values correspond to genuinely effective combat strategies. This study addresses these limitations by introducing a hierarchical framework based on time scale separation theory. The proposed framework employs a two-stage training procedure that accounts for differences in how flight parameters influence state variables across multiple time scales, improving learning efficiency, enhancing strategy quality, and increasing generalization capability in com-

plex and diverse combat environments.

Methods A novel algorithm, termed TTS–PPO, was developed. TTS–PPO stood for a Two-Stage Training framework leveraging Time-Scale separation within Proximal Policy Optimization. The method focused on partitioning the 6 DoF UAV air combat decision-making process into short-cycle and long-cycle segments, which reflected differences in how control inputs influenced state variables across distinct time scales. A time-division framework was established. The short–cycle component addressed rapid rotational and attitude adjustments. Instead of allowing the DRL procedure to directly manage these fine-grained actions, a Proportional-Integral-Derivative (PID) controller was employed to output real-time joystick commands. This configuration allowed classical low-level stability and attitude control to be handled independently, which reduced the complexity encountered by the DRL policy at the higher strategic level. With low-level stability assured, the DRL agent focused on tactical and strategic decision-making. The long-cycle component used Proximal Policy Optimization to manage trajectory planning and tactical maneuvers. The long-cycle PPO agent effectively decoupled strategic decision-making from low-level actuation tasks by issuing high-level commands to guide the PID-driven short-cycle layer. This hierarchical decomposition allowed learning to proceed more efficiently. The long-cycle agent encountered a reduced problem space and concentrated on discovering effective combat strategies without being burdened by the complexities of rapid stabilization maneuvers. Time scale separation was further implemented within the state space. Environmental states were divided into long-cycle and short-cycle groups. The long-cycle states captured slowly evolving features such as relative positions, energy conditions, and global situational parameters, whereas the short-cycle states encompassed rapidly changing variables such as angular rates and orientation deviations. Aligning state variables with their corresponding time scales accelerated learning and improved policy robustness. A relative situation transformation module was introduced to refine and compress the state representation, which ensured that the agent received relevant information at appropriate decision intervals and minimized computational complexity at each step. A two-stage training strategy was employed. In the first stage, single-step rewards designed for specific subtasks, such as pursuit or strike, were introduced with a lower decision frequency to assist the agent during the initial “cold start” period. This incremental guidance supported the stabilization of fundamental behavioral patterns and facilitated the acquisition of essential tactical principles. During this phase, the agent overcame early-stage instability, which resulted in more reliable initial policies. In the second stage of training, single-step rewards were removed, and only sparse terminal rewards were retained. In the second stage of training, single-step rewards were removed, and only sparse terminal rewards were retained, while the decision frequency was increased. In the absence of frequent intermediate rewards, the policy emphasized long-term outcomes rather than short-term objectives. The higher decision frequency enabled more refined tactical adjustments and encouraged the emergence of maneuvers that improved overall performance. The gradual transition from a guided, intermediate-reward scenario to a sparse-reward, high-frequency regime allowed the policy to progress from basic stability toward advanced strategic competence. A simulation environment was constructed using an open-source F-16 UAV model coupled with the JSBSim flight dynamics engine to evaluate the effectiveness of the proposed hierarchical DRL algorithm founded on time scale separation theory. This configuration provided realistic 6 DoF conditions and supported one-on-one close-range air combat simulations. Ablation experiments were conducted to assess the contribution of individual components within the TTS−PPO framework. One configuration trained the agent against a non-maneuvering linear opponent, which served as a controlled baseline for determining whether the learned policy can scale from simple engagements to more complex combat scenarios.

Results and Discussions The results demonstrated that the TTS−PPO approach, which incorporated hierarchical decomposition and time scale separation, achieved faster convergence and improved final performance metrics compared to baseline end-to-end DRL methods that lacked time scale separation or a two-stage training procedure. Assigning state variables to short-cycle and long-cycle categories, together with hierarchical action decomposition, significantly reduced overall problem complexity. Training convergence speed improved by approximately 67%, which reduced computational costs and enabled more frequent iterative policy refinements. With enhanced efficiency, the DRL agent discovered more stable and effective combat strategies within fewer training episodes. Generalization performance was evaluated by testing agents trained under different variants of the approach across various initial conditions, velocities, and adversary tactics. Comparisons were conducted among three agent types: an agent trained with PPO on full-state inputs without time scale division (FS−PPO), an agent using time scale-separated states with a single-stage training approach (TS−PPO), and the two-stage time scale-separated TTS−PPO. The agent trained with TTS−PPO outperformed both FS−PPO and TS−PPO agents in pairwise confrontations, which indicated that combining time scale separation with two-stage training not only enhanced learning speed but also enabled the agent to acquire more generalizable combat principles rather than narrowly optimizing for a specific scenario. Further validation involved testing the TTS−PPO-trained agent against rule-based expert opponents. The policy derived from TTS−PPO successfully defeated these expert systems. Even

when training was conducted exclusively against a simple linear adversary, the learned policy surpassed expert-level strategies, which confirmed that hierarchical time scale separation and the two-stage training design facilitated the development of adaptable policies with robust tactical proficiency. The ability to transfer from minimal training complexity to outperforming expert opponents highlighted the scalability and versatility of the learned strategies.

Conclusions Accordingly, the hierarchical DRL algorithm, grounded in time scale separation theory and employing a two-stage training strategy, addressed significant challenges associated with applying DRL to 6-DoF UAV air combat tasks. The method substantially improved training efficiency and enhanced both the robustness and generalization capability of the resulting policies by decomposing decision-making into short-cycle and long-cycle phases, introducing a PID-controlled low-level stabilization layer, and separating state variables based on their respective time scales. The hierarchical framework enabled the agent to focus on strategic maneuvers at the long-cycle level, while the short-cycle PID layer managed rapid stabilization tasks. Time scale-aware state representations and a staged training procedure guided the policy from basic stability to advanced tactical competence. The observed increases in convergence speed and the ability to manage a range of adversarial conditions highlight the value of applying time scale separation principles in challenging reinforcement learning domains. The TTS–PPO framework can serve as a reference for addressing other complex reinforcement learning problems characterized by distinct time scale dynamics, fostering more efficient, generalizable, and strategically effective decision-making in advanced autonomous systems.

Key words: time-scale separation; PID; PPO; two-stage training; TTS–PPO

（编辑　李轶楠）

引用格式: Tan Tai,Jiang Taiming,Li Bowen,et al.Hierarchical deep reinforcement learning algorithm for air combat based on time scale separation theory[J].Advanced Engineering Sciences,2026,58(2):69–83.[谭泰,江泰民,黎博文,等.基于时间尺度分离理论的空战深度强化学习分层算法[J].工程科学与技术,2026,58(2):69–83.]