Computer Science and Application 计算机科学与应用, 2026, 16(1), 102-114
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161009

Hans汉斯

基于分层策略与世界模型的多智能体深度确定性策略梯度算法

张华东[1]，王友鑫[1]，王于婷[1]，徐衍亮[2]，侯恩广[1*]

[1]山东交通学院轨道交通学院，山东 济南
[2]山东大学电气工程学院，山东 济南

收稿日期：2025年12月5日；录用日期：2026年1月5日；发布日期：2026年1月14日

摘 要

针对三维环境中多无人机路径规划面临着样本效率低、长时程决策困难和鲁棒性不足等挑战，本文提出一种基于分层策略与世界模型增强的多智能体深度确定性策略梯度算法框架(HWC-MADDPG)。首先，引入对比学习机制，从高维观测中提取时序一致性的鲁棒状态表征，增强了状态表征的区分度；其次，设计多智能体层次化策略网络架构，通过高层策略网络规划宏观意图，低层策略网络执行具体动作的方式，将路径规划任务分解，提升决策能力；最后，集成共享的世界模型，通过其内在的前瞻性推演生成想象奖励，优化Critic网络的价值评估，提升了决策前瞻性和收敛速度。实验结果表明，本文提出的算法在学习速度、策略稳定性和飞行安全性上均优于传统的多智能体深度确定性策略梯度算法(MADDPG)。该研究为解决三维环境下的多智能体路径规划问题提供了一种更高效的解决方案，具有一定的理论价值与应用前景。

关键词

无人机，多智能体深度确定性策略梯度算法(MADDPG)，层次化策略，世界模型，对比学习

Multi-Agent Deep Deterministic Policy Gradient Algorithm Based on Hierarchical Strategies and World Models

Huadong Zhang[1], Youxin Wang[1], Yuting Wang[1], Yanliang Xu[2], Enguang Hou[1*]

[1]School of Rail Transportation, Shandong Jiaotong University, Jinan Shandong
[2]School of Electrical Engineering, Shandong University, Jinan Shandong

Received: December 5, 2025; accepted: January 5, 2026; published: January 14, 2026

*通讯作者。

文章引用: 张华东, 王友鑫, 王于婷, 徐衍亮, 侯恩广. 基于分层策略与世界模型的多智能体深度确定性策略梯度算法[J]. 计算机科学与应用, 2026, 16(1): 102-114. DOI: 10.12677/csa.2026.161009

Abstract

Addressing challenges in multi-UAV path planning within 3D environments—such as low sample efficiency, difficulties in long-term decision-making, and insufficient robustness—this paper proposes a hierarchical strategy and world model-enhanced multi-agent deep deterministic policy gradient algorithm framework (HWC-MADDPG). First, a contrastive learning mechanism is introduced to extract temporally consistent robust state representations from high-dimensional observations, enhancing the discriminative power of state representations. Second, a hierarchical multi-agent policy network architecture is designed. By decomposing the path planning task—where the high-level policy network formulates macro-intentions and the low-level policy network executes specific actions—decision-making capabilities are enhanced. Finally, an integrated shared world model generates imagined rewards through its inherent forward-looking inference, optimizing the value assessment of the Critic network and improving decision foresight and convergence speed. Experimental results demonstrate that the proposed algorithm outperforms the traditional Multi-Agent Deep Deterministic Policy Gradient (MADDPG) in learning speed, policy stability, and flight safety. This research offers a more efficient solution for multi-agent path planning in 3D environments, holding significant theoretical value and practical application potential.

Keywords

UAV, Multi-Agent Deep Deterministic Policy Gradient (MADDPG), Hierarchical Strategy, World Model, Contrastive Learning

1. 引言

近年来，随着人工智能、芯片等技术的发展，无人机因其低成本和高机动性的特点，广泛应用在侦察监视[1]、协同打击[2]、搜索救援[3]和货物运输[4]等军事与民用领域，并且展现出了出色的能力。但是，随着任务复杂度和环境不确定性的增加，单无人机在面对复杂环境和突发情况时，往往难以满足任务要求。研究者们将注意力集中在多无人机系统上，多无人机系统因其任务执行效率高、系统鲁棒性强等优势已经成为当前研究热点[5]。然而如果要充分发挥多无人机系统的效能，核心在于路径规划问题。针对这一关键问题，国内外学者提出了多种解决方案。

现有的路径规划主要分为传统算法和智能算法两大类。传统的基于图搜索的算法[6]虽然在静态环境中可以生成最短路径，但是基于图的搜索方法并不适合智能体之间的协调。启发式算法[7]运行速度快，但是计算量大，在面对环境的变化时往往缺乏适应性。群智能体算法[8]虽然在避障和编队控制方面取得了一定效果，但是这种方法依赖局部规则容易陷入局部最优，不能保证全局最优。基于博弈论的方法[9]将无人机任务分配与规划等问题视为一种博弈问题，从而为其提供了新的解决方法，但是这些公式预先定义了环境信息，缺乏面对未知环境的能力。遗传算法[10]虽然在多无人机环境中发挥优势，但是其迭代次数多，计算开销大。

强化学习[11]方法不需要先验知识，根据环境迭代反馈来优化决策，可以捕捉复杂环境下的不确定性。唐峯竹等[12]在深度 Q 网络(DQN)算法的基础上提出了一种多无人机分布式动态任务分配方法，对任务

进行动态分配，提高了系统的任务完成度。周彬等[13]在 Q-Learning 算法上根据接收信号的强度作为回报值，提出了基于导向强化 Q 学习的无人机路径规划。任君凯等[14]将世界模型与强化学习算法相结合用于机器人运动控制，降低了对真实环境交互的依赖性。

随着神经网络的发展，将深度学习和强化学习相结合的深度强化学习(DRL)为多无人机智能控制提供了新的解决思路[15]。Zeng Y 等[16]利用多步学习技术结合双深度 Q 网络(DDQN)算法提出了基于直接强化学习的无人机导航算法。张天浩等[17]将人工势场法与 MADDPG 算法融合，解决了 MADDPG 算法早期盲目探索、收敛性差的问题。王娜[18]利用混合主动行为选择机制评估策略，设计了一种深度学习强化方法，实现了无人机的航迹控制和任务规划。Yan 等[19]利用分层强化学习的方法将总任务分解为子任务，结合 MAXQ 算法降低任务复杂性，提高学习速度。

虽然上述研究都已经取得了一定进展，但是三维环境下的多无人机路径规划面临着决策复杂性提高、状态空间维度上涨和长期规划能力受限等瓶颈。本文聚焦于多无人机路径规划，为解决三维环境下多无人机自主决策目标并且规划安全高效的飞行路径的问题，提出了一种基于分层策略和世界模型增强的多智能体深度确定性策略梯度算法(HWC-MADDPG)。该算法首先引用对比学习机制，从高维观测中提取出具有区分度的状态信息；其次，设计了一种层次化策略架构，将决策过程分解为高层宏观意图选择与低层动作执行，从而实现决策的有效分解；最后，构建了一个共享的世界模型，通过预测智能体未来状态和奖励，克服长期规划能力不足的瓶颈，增强决策前瞻性。

2. 问题建模

在三维环境中，n 架无人机从不同位置出发，通过自主规划到达目标点，同时保持自身安全，这要求无人机不仅要规划自身路径，还要进行隐式交互。

我们将多无人机路径规划建模为去中心化的部分可观测马尔可夫决策过程(Dec-POMDP)。Dec-POMDP 可以用元组 $\langle U,S,A,P,\Omega,O,\gamma\rangle$ 表示，其中 $U=\{1,2,\cdots,n\}$ 为智能体集合，n 为智能体数量；S 表示状态空间；A 表示动作空间；P 为状态转移函数；R 表示奖励函数；Ω表示观测空间；γ 为折扣率。

2.1. 状态空间

全局状态 $s_t \in S$ 描述了在 t 时刻环境中的所有的信息，包括所有智能体的位置与速度、目标位置与完成情况。

2.2. 动作空间

每一个智能体 i 的动作空间 A_i 是一个三维连续动作集合 $A_i=(x_i,y_i,z_i)$。

2.3. 观测空间

由于问题的部分可观测性，在每个时刻 t，智能体 i 只能获得一个局部的、以自我为中心的观测 $o_i \in \Omega_i$，观测空间 $o_i=(p_i,v_i,p_{ti},p_{oi})$。其中 $p_i=[x_i,y_i,z_i]$ 表示智能体自身的位置，$v_i=[v_{xi},v_{yi},v_{zi}]$ 表示智能体当前的速度信息，$p_{ti}=[x_{ti},y_{ti},z_{ti}]$ 表示智能体距最近未完成目标点的位置，$p_{oi}=[x_{oi},y_{oi},z_{oi}]$ 表示与其他智能体群的平均相对位置。

2.4. 奖励函数

智能体在执行动作后会受到奖励，为了正确引导智能体的动作，该奖励函数被精心设计以鼓励智能体高效安全完成任务，总奖励函数如公式(1)所示。

$$R=R_{\text{task}}+R_{\text{coop}}+R_{\text{dist}}+R_{\text{coll}}+R_{\text{time}} \tag{1}$$

其中 R_{task} 为任务奖励， R_{coop} 为协作奖励， R_{dist} 为距离惩罚， R_{coll} 为碰撞惩罚， R_{time} 为时间惩罚。

首先给予完成任务的智能体一个任务奖励,如公式(2)所示。

$$R_{\text{task}} = \sum_i r_{i,\text{task}} \tag{2}$$

奖励值设置如公式(3)所示。

$$r_{i,\text{task}} = \begin{cases} 500, \left\| p_i - p_{ti} \right\|_2 < 30 \\ 0，\text{其他} \end{cases} \tag{3}$$

当有任务被完成时，给予所有无人机一个协作奖励，如公式(4)所示。

$$R_{\text{coop}} = \sum_i r_{i,\text{coop}} \tag{4}$$

奖励值设置为公式(5)。

$$r_{i,\text{coop}} = \begin{cases} 100, \left\| p_{i,j} - p_{ti,j} \right\|_2 < 30 \\ 0，\text{其他} \end{cases} \tag{5}$$

该式表示只要有任务被完成时，智能体就会获得一个协作奖励。

为引导智能体飞向目标位置，我们设置一个距离惩罚，如公式(6)所示。

$$R_{\text{dist}} = \alpha_{\text{dist}} \left\| p_i - p_{ti} \right\|_2 \tag{6}$$

以鼓励智能体接近最近的任务，无人机之间的碰撞惩罚如公式(7)所示。

$$R_{\text{coll}} = \sum_i r_{\text{coll}} \tag{7}$$

奖励值设置如公式(8)所示。

$$r_{\text{coll}} = \begin{cases} -50，\ d_{ij} < 10 \\ 0，\text{其他} \end{cases} \tag{8}$$

其中 d_{ij} 为无人机之间的距离，时间惩罚如公式(9)所示。

$$R_{\text{time}} = \sum_i r_{i,\text{time}} \tag{9}$$

其中 $r_{i,\text{time}} = -1$，鼓励尽快完成任务。

在 Dec-POMDP 过程中，智能体的联合动作 A 与环境交互后从状态 S 转移到 S'，在新状态下智能体做出动作并且与环境重新交互。智能体将上述过程不断更新迭代，最终目的是最大化奖励的同时学会最优策略，实现高效的路径规划。

3. 层次化策略和世界模型的 MADDPG 算法

本节阐述了多智能体深度学习框架，该框架在标准 MADDPG 集中训练、分布执行式基础上对 Actor 和 Critic 进行了增强，集成了用于提升状态区分度的对比学习机制、Actor 的层次化策略以及想象增强的世界模型，提升了多无人机路径规划的效率和鲁棒性。

3.1. 整体框架

算法的整体框架如图 1 所示。在训练阶段，全局信息 (s,a,r,s') 存入经验回放区。一个增强的 Critic 网络利用对比编码器的状态表征信息和世界模型的想象推演指导层次化 Actor 的更新。同时，对比编码

器和世界模型利用经验回放区的数据进行训练。目标网络通过软更新的方式进行同步，确保训练稳定。在执行阶段，每个智能体接受对比编码器的状态表征信息，由高层次策略网络输出宏观意图选择，低层次策略网络执行具体动作。层次化 Actor 网络集成了对比编码器和层次化策略以实现状态到动作的映射。

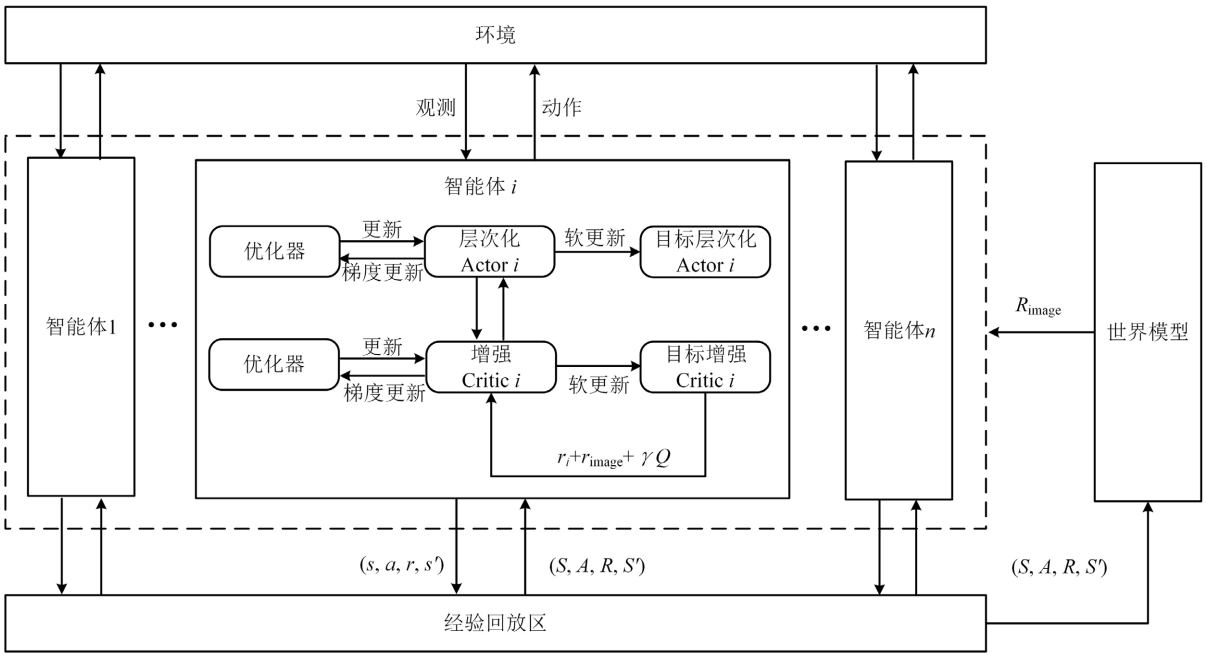


Figure 1. HWC-MADDPG algorithm framework
图 1. HWC-MADDPG 算法框架

3.2. 对比学习

在三维环境下的多无人机路径规划中，每架无人机都能够获得以自身为中心的高维局部观测(包含自身状态、最近目标点位置、其他无人机相对位置等)，这将导致原始的观测空间维度较高且包含冗余信息，直接使用原始观测作为输入，策略网络难以获得对决策有效的信息特征，导致策略早期盲目的探索严重。

针对无人机的观测数据随时间变化但是其拓扑结构相对稳定的特点，我们引入对比编码器机制，采用时序一致的采样策略，通过最大化相似样本对和最小化不相关样本对以提升状态区分度。

对比编码器 $E_{\varnothing}$ 采用三层 MLP 架构，将原始高维观测信息映射到一个潜在空间。我们从经验回放区为每一个观测 o_t 采样一个正样本 o_t^+ 和负样本 o_t^-，考虑到无人机飞行轨迹的连续性，其中正样本 o_t^+ 为同一轮次中相邻时间步内的随机采样，以捕捉时间维度的相关性；负样本 o_t^- 为不同轮次中的随机采样，以确保样本的上下文的无关性。通过最小化损失函数来优化对比编码器，如公式(10)所示。

$$L_{\text{contrastive}} = -\mathbb{E}\left[\log \frac{\exp\left(\text{sim}\left(z, z^+\right)/\tau\right)}{\exp\left(\text{sim}\left(z, z^+\right)/\tau\right) + \sum_{k=1}^{K}\exp\left(\text{sim}\left(z, z_k^-\right)/\tau\right)}\right] \tag{10}$$

其中 $z = E_{\varnothing}$ 为锚点样本，z^+ 和 z^- 分别为正、负样本，$\text{sim}\left(z, z^+\right)$ 为正负样本的余弦相似度，τ 为超参数，用于调节相似度分布。对比编码器被嵌入到 Actor 和 Critic 网络，促使智能体学习到有利于决策、鲁棒的状态表征，为决策和评估提供高质量的状态表征，有效缓解信息冗余和前期策略学习困难等问题。

3.3. 层次化策略

由于目标点选择的离散性、长时间特点和无人机动作的连续性、实时性特点在时间尺度和动作空间上存在天然矛盾。无人机同时感知自身状态和环境状态等信息，状态空间维度上涨，而动作空间的连续性进一步加剧了决策难度。传统的单网络仅学习状态到动作的映射关系难以同时兼顾宏观调度和微观控制，容易出现梯度消失和收敛缓慢等问题。

为了解决路径规划行为的复杂性，我们采用层次化策略网络作为 Actor 的核心，将决策拆分为高层和低层两部分，构建“意图–动作”双层决策框架，通过决策粒度拆分化解这一矛盾。

高层策略采用两层 MLP 网络，通过接受对比编码器 $E_{\varnothing}$ 编码后的状态表示，选择一个离散高层次宏观意图 k_t，该决策以一定的频率 *H_freq* 运行，以适应任务分配的长周期特点。高层策略如公式(11)所示。

$$k_t = \arg\max_{k\in\{1,2,\cdots,K\}} \pi_{\text{high}}\left(k\middle|z_t\right) \tag{11}$$

其中 $\pi_{\text{high}}\left(k\middle|z_t\right)$ 为高层次策略网络在给定状态表征 z_t 的前提下，选中宏观意图 k 的概率。低层策略由四个专门化的子网络构成，每个子网络由三层 MLP 网络构成，一旦高层次策略选择了意图，对应的低层次子网激活并输出具体的动作 a_t，如公式(12)所示。

$$a_t = \pi_{\text{low},k_t}\left(z_t\right) \tag{12}$$

这种针对性的层次化设计实现了目标点选择与动作执行的解耦，有效降低了无人机学习策略的难度，使得无人机既能对宏观局面有着良好的判断，又能灵活应对局部情况。

3.4. 世界模型

在多无人机快速飞行的场景中，往往伴随着碰撞风险并且试错成本极高。传统的强化学习算法依赖环境交互的数据，样本利用率较低并且智能体的决策缺乏对未来状态的预判能力。为了提升决策的前瞻性并且加速收敛，本文引入世界模型，该模型通过监督学习的方式从经验回放区中采样来预测环境的变化。

世界模型由动力学网络和奖励网络组成，通过接受所有智能体的状态和动作，输出下一个状态和奖励的预测，如公式(13)所示。

$$\left(\hat{s}_{t+1},\hat{r}_t\right) = W_\theta\left(s_t,a_t\right) \tag{13}$$

其中 s_t 和 a_t 为当前时刻的状态和动作，$\hat{s}_{t+1}$ 为下一时刻的预测状态，$\hat{r}_t$ 为执行动作后的预测奖励。动力学网络和奖励网络分别为三层 MLP 和两层 MLP 网络。世界模型从当前状态 s 出发，使用目标策略在网络中模拟未来 *H_imagine* 步的想象推演，得到一系列未来轨迹，提前评估策略可能导致的风险或收益。累计折扣奖励如公式(14)所示。

$$R_{\text{imagine},i} = \sum_{h=0}^{H-1}\gamma^h \hat{r}_{i,t+h} \tag{14}$$

其中 γ 为折扣因子，$\hat{r}_{i,t+h}$ 为智能体 i 第 $t+h$ 步的想象奖励，通过引入想象奖励 R_{imagine}，将其作为内在奖励与外部环境奖励结合，Critic 网络能更准确的评估当前动作，增强 Critic 网络的价值估计准确性，从而引导 Actor 网络生成具有前瞻性的飞行策略，减少对真实环境交互的依赖，提高了飞行安全性。

世界模型的损失函数用预测状态和奖励相对真实状态和奖励的均方误差来表示，如公式(15)所示。

$$L_{\text{world}} = \mathbb{E}_{s_t,a_t,s_{t+1},r_t\sim\mathcal{D}}\left[\left\|s_{t+1}-\hat{s}_{t+1}\right\|^2 + \left\|r_t-\hat{r}\right\|^2\right] \tag{15}$$

其中 $\left\|s_{t+1}-\hat{s}_{t+1}\right\|^2$ 为预测状态与真实状态的均方误差，$\left\|r_t-\hat{r}_t\right\|^2$ 为预测奖励与真实奖励的均方误差。

3.5. Actor-Critic 框架

每个智能体的策略网络都采用层次化策略，接收对比编码器处理的状态特征，高层策略选择技能，低层策略生成具体动作，如图 2 所示。

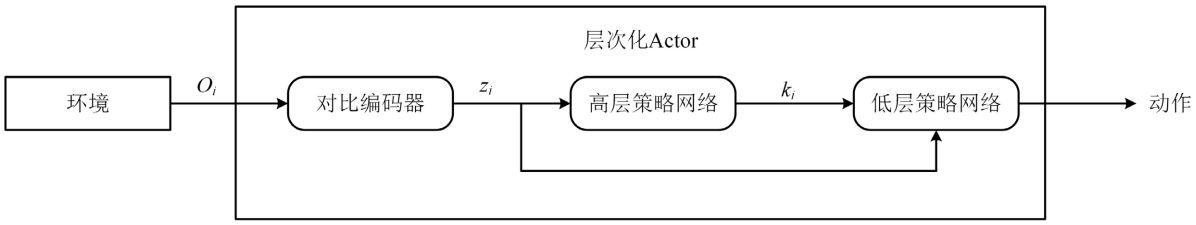


Figure 2. Hierarchical Actor network architecture
图 2. 层次化 Actor 网络结构

Actor 网络 π_i 采用确定性梯度策略，通过最大化 Critic 网络的期望进行更新，其策略梯度可表示为公式(16)。

$$\nabla_{\theta_i} J\left(\pi_i\right) \approx \mathbb{E}_{s,\mathrm{a}\sim\mathcal{D}}\left[\nabla_{\theta_i}\pi_i\left(o_i\right)\nabla_{a_i} Q_i^{\pi}\left(s,a_1,\ldots,a_N\right)\Big|_{a_i=\pi_i\left(o_i\right)}\right] \tag{16}$$

其中 θ_i 是 Actor 网络 π_i 的可学习参数，Q_i^{π} 为 Critic 网络。

Critic 网络结构如图 3 所示。

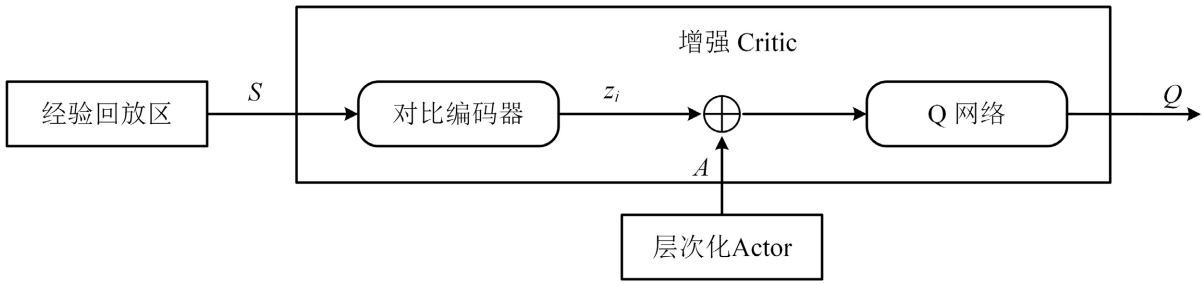


Figure 3. Enhanced Critic network architecture
图 3. 增强 Critic 网络结构

Critic 网络 Q_i 通过最小化时序差分(TD)误差损失函数进行更新，如公式(17)所示。

$$L\left(Q_i\right)=\mathbb{E}_{s,\mathrm{a},r,s'\sim\mathcal{D}}\left[\left(y_i^{\text{enhanced}}-Q_i\left(s,a_1,\cdots,a_N\right)\right)^2\right] \tag{17}$$

其中 $Q_i\left(s,a_1,\cdots,a_N\right)$ 为当前 Critic 网络对状态动作的价值估计，y_i^{enhanced} 为增强的 TD 目标，Critic 网络除了接受智能体的观测和动作外，还利用世界模型生成的想象奖励构建一个增强的 TD 目标，其计算方式如公式(18)所示。

$$y_i^{\text{enhanced}}=\left(r_i+\alpha R_{\text{imagine},i}\right)+\gamma\left(1-d\right)Q_i'\left(s',\pi_i'\left(o_i'\right),\cdots,\pi_{N_d}'\left(o_{N_d}'\right)\right) \tag{18}$$

其中 Q_i' 为目标 Critic 网络。

3.6. 训练流程

该框架通过对比编码器、世界模型与策略网络协同优化，总损失函数如公式(19)所示。

$$L_{\text{total}} = \lambda_{\text{r}} L_{\text{rl}} + \lambda_{\text{c}} L_{\text{constrastive}} + \lambda_{\text{w}} L_{\text{world}} \tag{19}$$

其中 L_{rl} 为所有智能体 Actor 和 Critic 网络损失之和， $L_{\text{constrastive}}$ 为对比学习损失， L_{world} 为世界模型损失， λ_{rl} 、 λ_{c} 和 λ_{w} 为对应的权重系数。通过构建一个多任务学习目标，促使智能体更好地应对动态三维环境。

根据上述描述，算法伪代码如表 1 所示。

Table 1. HWC-MADDPG algorithm flowchart
表 1. HWC-MADDPG 算法流程

Algorithm：HWC-MADDPG
初始化：
for 智能体 i to N do
初始化层次化 Actor 网络 π_i 和增强 Critic 网络 Q_i
初始化目标层次化 Actor 网络 π_i' 和目标增强 Critic 网络 Q_i'
end for
初始化世界模型 M_Ψ
初始化对比编码器 $E_\varnothing$
初始化空的经验回放区 D
训练循环：for Episode = 1 to M do
初始化全局状态 $s0$
初始化智能体当前宏观意图
for t = 0 to T − 1 do
for 智能体 i = 1 to N do
获得局部观测 $o_{i,t}$
使用对比编码器提取特征
if t % H_freque = 0 then
高层次策略选择意图 $k_{i,t}$
end if
低层次策略执行动作 $a_{i,t}$
end for
在环境中执行联合动作 a_t ，接受奖励 r_t 并更新下一状态 s_{t+1}
将元组 $\langle s_t, a_t, r_t, s_{t+1} \rangle$ 存储到经验回放区 D
if $\lvert D \rvert$ > B then
从经验回放区 D 采样随机数据
更新世界模型 M_Ψ (公式 15)
计算想象奖励 R_{imagine} (公式 14)
for 智能体 i to N do
更新对比编码器 $E_\varnothing$ (公式 10)

续表

更新 Critic 网络 Q_i (公式 17)
更新 Actor 网络 π_i (公式 16)
软更新目标网络 π_i' 和 Q_i'
end for
end if
end for
end for

4. 实验和分析

为了验证算法的有效性，本节构建了三维环境下的多无人机场景进行实验评估。

4.1. 实验设置

仿真实验在 500 m*500 m*500 m 的三维虚拟场景下进行。任务点数量为 4，无人机数量为 5，无人机初始化位置在以原点为中心半径为 150 m 的水平圆周上，高度随机分布在 50 m~100 m 范围内。任务点数量为 4，在高度 30 m~150 m 范围内随机分布，任务完成半径为 30 m，无人机安全距离为 10 m。在达到最大时间步数 $T = 400$ 或者任务完成后，训练结束。

4.2. 训练设置

本文所有优化器均采用 Adam 优化器，HWC-MADDPG 网络结构和超参数分别如表 2 和表 3 所示。

Table 2. HWC-MADDPG network component architecture
表 2. HWC-MADDPG 网络组件结构

网络组件	结构描述
对比编码器	[256, 256, 128]
高层策略网络	[64, 4]
低层策略网络	[128, 64, 3]
Critic 网络(编码器)	[256, 128]
Critic 网络(Q 网络)	[256, 128, 1]
世界模型(动力学)	[256, 256, 60]
世界模型(奖励)	[128, 5]

Table 3. HWC-MADDPG parameter values
表 3. HWC-MADDPG 参数值

参数	值
Actor 学习率	1*10-4
Critic 学习率	3*10-4
世界模型学习率	3*10-4
折扣因子	0.95

续表

软更新系数	0.01
经验回放区大小	100,000
批次大小	256
高层决策频率	10
想象步数	5
对比学习特征维度	128

4.3. 基线算法

为了对比验证所提方法的有效性，选取三类代表性算法与 MADDPG 算法进行对比，分别为多智能体深度确定性策略梯度算法(MADDPG)、分层演员－评论家算法(HAC)、多智能体近端策略优化算法(MAPPO)。

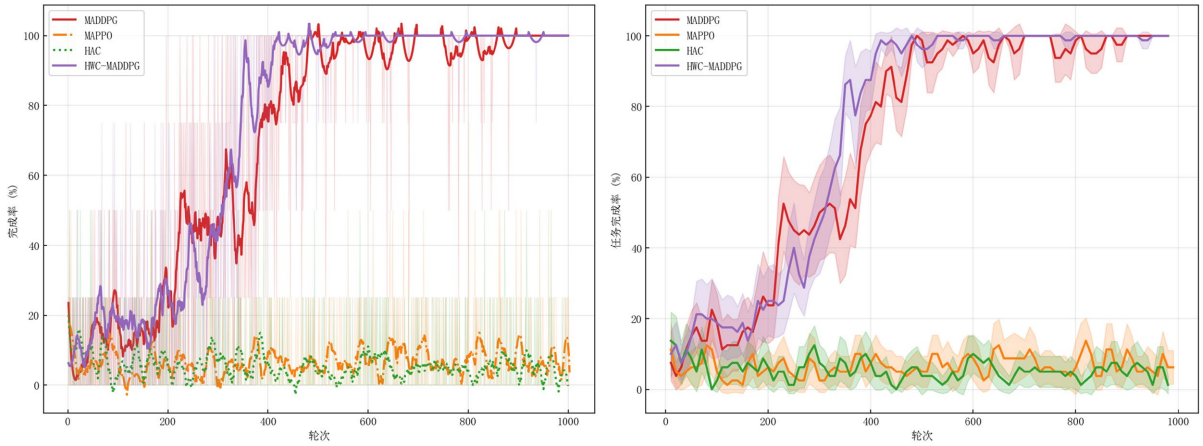
MADDPG 算法作为多智能体算法中表现先进的算法之一，直观体现出 HWC-MADDPG 算法相较于原始 MADDPG 算法的改进效果。HAC 算法作为层次化强化学习算法，其采用目标条件化层次结构，与 HWC-MADDPG 算法中的层次化决策进行对比。MAPPO 算法作为 on-policy 算法与 HWC-MADDPG 算法的 off-policy 进行对比。

4.4. 结果分析

从图 4 可以看出，随着训练轮次的增加，HWC-MADDPG 算法和 MADDPG 算法的奖励值逐渐增加。在训练到达 500 轮次左右时，HWC-MADDPG 算法的奖励值趋于平缓并总体到达收敛。HWC-MADDPG 算法的收敛速度和稳定性高于基线 MADDPG 算法，相较之下，MAPPO 算法和 HAC 算法的奖励始终处于一个较低范围，表明他们并未学习到有效的策略，陷入了局部最优。

Figure 4. Reward function curve
图 4. 奖励函数曲线

四种算法的任务完成率情况如图 5 所示，从图 5(a)中可以直观的看出 HWC-MADDPG 算法在 350 轮次展现出了较高的学习速度，并且早于其他算法到达了接近 100%的任务完成率。MADDPG 算法到达 100%任务完成率后的轮次明显晚于 HWC-MADDPG 算法，并且其稳定性低于 HWC-MADDPG 算法，而其他两种算法的任务完成率始终在较低水平徘徊，证明他们难以到达目标点。95%置信区间的任务完成率曲线如图 5(b)所示，置信区间的宽度反映了学习过程中的性能的波动程度。从图中可以看出，在初始阶段，所有算法都表现出较高的不确定性，但是随着轮次的增加 HWC-MADDPG 算法波动程度小于 MADDPG 算法，到达收敛后的 HWC-MADDPG 算法的性能稳定性也更高。这表明 HWC-MADDPG 算法具有更稳健的学习梯度。



(a) 任务完成率曲线　　(b) 95%置信区间任务完成率曲线

Figure 5. Task completion rate curve
图 5. 任务完成率曲线

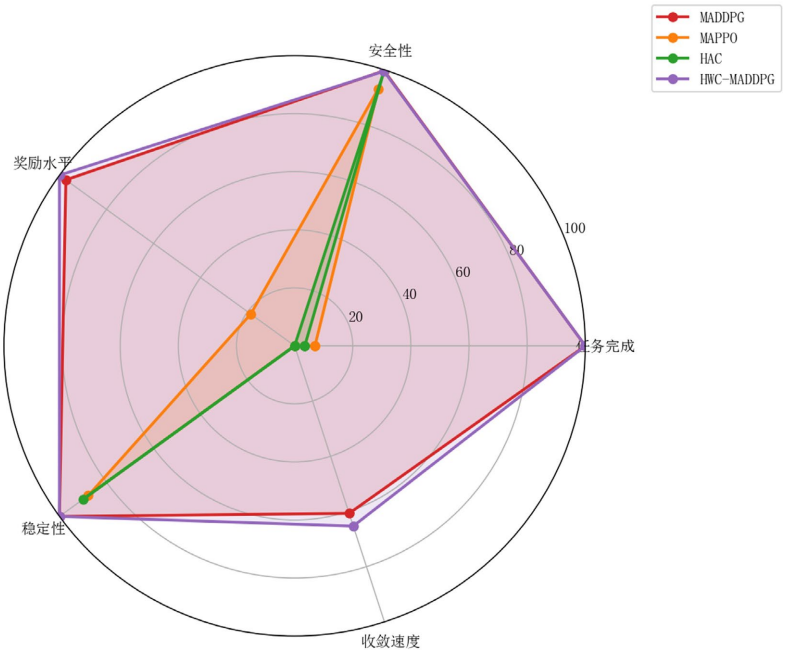


Figure 6. Comparison radar chart
图 6. 对比雷达图

我们引入雷达图从安全性、任务完成、收敛速度、稳定性、奖励水平五个维度进行了对话的对比，雷达图的覆盖面积直观的代表了算法的性能表现。如图 6 所示，HWC-MADDPG 算法在五个维度都高于其他算法，尤其在奖励水平、稳定性和收敛速度方面显著优于 MADDPG 算法。

为了更准确量化各算法的性能差异，我们统计了 HWC-MADDPG 算法与各基线算法在训练后期的多项指标，对比结果如表 4 所示。从最终完成率来看，本文提出的 HWC-MADDPG 算法相比 MADDPG 算法不仅均值更高，其标准差显著降低，表明 HWC-MADDPG 算法具有更强的鲁棒性，展现出稳定的高质量策略。相比之下 MAPPO 和 HAC 算法均为达到有效水平。HWC-MADDPG 算法在碰撞率较低的情况下，其仅用 350 个轮次就可以达到 85%的任务完成率，而 MADDPG 算法则需要 425 个轮次，在保证安全性的同时有效减少了多无人机的无效探索。

Table 4. Comparison of performance metrics for various algorithms
表 4. 各算法综合性能指标对比

算法	任务完成率(%)	碰撞率(%)	平均奖励	最高任务完成率(%)	收敛轮次(85%)
MADDPG	98.9 ± 6.8	0.5 ± 1.7	−4342	100.0	425
MAPPO	7.4 ± 13.2	7.0 ± 0.5	−56528	75.0	N/A
HAC	4.6 ± 10.9	0.5 ± 0.7	−73197	50.0	N/A
HWC-MADDPG	99.9 ± 1.8	0.4 ± 1.5	−1896	100.0	350

5. 结论与展望

本文针对三维环境中多无人机路径规划面临的样本利用率低、策略前瞻性不足和鲁棒性差等挑战，提出了 HWC-MADDPG 算法框架。通过自监督的方式训练对比学习编码器，提取观测信息中具有区分度的状态表征，提升策略的泛化能力。将每个智能体的 Actor 网络设置为层次化策略架构，通过高层选择宏观意图，低层执行具体动作实现了策略的细粒度化拆分。利用世界模型的内在推演生成想象奖励，对 Critic 网络中的 TD 目标进行增强，提升了决策前瞻性。仿真结果表明，HWC-MADDPG 算法相较其他主流基线算法在任务完成率、学习速度、安全性等方面展现出优越的性能。研究结果表明该方法为多智能体路径规划提供了一个更高效、更鲁棒的解决方案。

当前研究是在特定的静态环境下进行的，其在复杂场景下的适用性有待研究，未来可以进一步研究在受限环境下的多智能体协同优化，增加多智能体的数量，面向大规模真实场景下的多无人机协同优化。

基金项目

山东省重点研发计划项目(2023CXGC010202)、德州市科技型中小企业创新能力提升工程项目(2025DZTSGC013)。

参考文献

[1] Mayand, V.C., Nugraha, Y.E. and Alkaff, A. (2024) Three-Dimensional Coordination Control of Multi-UAV for Partially Observable Multi-Target Tracking. *Journal of Robotics and Control* (*JRC*), **5**, 1227-1240.

[2] Hu, R., Li, Y., Xu, C. and Li, Y. (2024) Analysis of Model and Simulation for UAVs Equipment Swarm Attack-Defense Tactics Based on Non-Static Bayesian Architecture. 2024 *International Conference on Electronics and Devices, Computational Science* (*ICEDCS*), Marseille, 23-25 September 2024, 706-712. https://doi.org/10.1109/icedcs64328.2024.00133

[3] Yanmaz, E., Balanji, H.M. and Güven, İ. (2024) Dynamic Multi-UAV Path Planning for Multi-Target Search and Connectivity. *IEEE Transactions on Vehicular Technology*, **73**, 10516-10528. https://doi.org/10.1109/tvt.2024.3363840

[4] 陈群, 李超. 城市物流末端卡车-无人机协同运输研究综述[J]. 长沙理工大学学报(自然科学版), 2025, 22(4): 104-115.

[5] 宁聪, 范菁, 孙书魁. 多无人机协同规划研究综述[J]. 计算机工程与应用, 2025, 61(1): 42-58.

[6] Kelner, J.M., Burzynski, W. and Stecz, W. (2024) Modeling UAV Swarm Flight Trajectories Using Rapidly-Exploring Random Tree Algorithm. *Journal of King Saud University-Computer and Information Sciences*, **36**, Article 101909. https://doi.org/10.1016/j.jksuci.2023.101909

[7] 曹晓意, 罗煦琼, 李景, 等. 改进人工势场法下的多无人机编队路径规划方法[J]. 计算机应用, 2025, 45(S1): 183-187.

[8] Elmokadem, T. and Savkin, A. (2021) Computationally-Efficient Distributed Algorithms of Navigation of Teams of Autonomous UAVs for 3D Coverage and Flocking. *Drones*, **5**, Article 124. https://doi.org/10.3390/drones5040124

[9] Zhang, R., Lu, R., Cheng, X., Wang, N. and Yang, L. (2021) A UAV-Enabled Data Dissemination Protocol with Proactive Caching and File Sharing in V2X Networks. *IEEE Transactions on Communications*, **69**, 3930-3942. https://doi.org/10.1109/tcomm.2021.3064569

[10] Hou, K., Yang, Y., Yang, X. and Lai, J. (2021) Distributed Cooperative Search Algorithm with Task Assignment and Receding Horizon Predictive Control for Multiple Unmanned Aerial Vehicles. *IEEE Access*, **9**, 6122-6136. https://doi.org/10.1109/access.2020.3048974

[11] 杨浅舒, 阮迪望, 吴先宇, 等. 多智能体强化学习在飞行器协同控制中的研究进展[J]. 战术导弹技术, 2025(4): 90-106.

[12] 唐峯竹, 唐欣, 李春海, 等. 基于深度强化学习的多无人机任务动态分配[J]. 广西师范大学学报(自然科学版), 2021, 39(6): 63-71.

[13] 周彬, 郭艳, 李宁, 等. 基于导向强化 Q 学习的无人机路径规划[J]. 航空学报, 2021, 42(9): 506-513.

[14] 任君凯, 张洪川, 瞿宇珂, 等. 基于世界模型强化学习的机器人运动控制方法[J/OL]. 机器人, 1-15. https://doi.org/10.13973/j.cnki.robot.250061, 2025-10-12.

[15] 李波, 黄晶益, 万开方, 等. 基于深度强化学习的无人机系统应用研究综述[J]. 战术导弹技术, 2023(1): 58-68.

[16] Zeng, Y., Xu, X., Jin, S. and Zhang, R. (2021) Simultaneous Navigation and Radio Mapping for Cellular-Connected UAV with Deep Reinforcement Learning. *IEEE Transactions on Wireless Communications*, **20**, 4205-4220. https://doi.org/10.1109/twc.2021.3056573

[17] 张天浩, 池晴佳, 林永水, 等. 基于人工势场法改进 MADDPG 算法的 AUV 协同应召搜潜航路规划研究[J/OL]. 中国舰船研究, 1-12. https://doi.org/10.19693/j.issn.1673-3185.04229, 2025-10-12.

[18] 王娜, 马利民, 姜云春, 等. 基于多 Agent 深度强化学习的无人机协作规划方法[J]. 计算机应用与软件, 2024, 41(9): 83-89+96.

[19] Yan, Y., Wang, H. and Chen, X. (2020) Collaborative Path Planning Based on MAXQ Hierarchical Reinforcement Learning for Manned/Unmanned Aerial Vehicles. 2020 39*th Chinese Control Conference* (*CCC*), Shenyang, 27-29 July 2020, 4837-4842. https://doi.org/10.23919/ccc50068.2020.9188401