SCIENCE CHINA
Information Sciences

• LETTER •

February 2026, Vol. 69, Iss. 2, 129207:1–129207:2
https://doi.org/10.1007/s11432-024-4728-1

Counterfactual baseline-based MAPPO for asymmetric UAV swarm confrontation game

Ershen WANG[1,2], Zeqi TONG[1], Chen HONG[3,4*], Xiaotong WU[1], Mingming XIAO[5], Chang LIU[4] & Jihao CHEN[1]

[1]*School of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, China*
[2]*Key Laboratory of Technology and Equipment of Tianjin Urban Air Transportation System, Civil Aviation University of China, Tianjin 300300, China*
[3]*Multi-Agent Systems Research Centre, Beijing Union University, Beijing 100101, China*
[4]*College of Robotics, Beijing Union University, Beijing 100101, China*
[5]*College of Smart City, Beijing Union University, Beijing 100101, China*

Received 24 December 2024/Revised 6 July 2025/Accepted 8 September 2025/Published online 16 January 2026

Citation Wang E S, Tong Z Q, Hong C, et al. Counterfactual baseline-based MAPPO for asymmetric UAV swarm confrontation game. Sci China Inf Sci, 2026, 69(2): 129207, https://doi.org/10.1007/s11432-024-4728-1

With the rapid development of unmanned aerial vehicle (UAV) manufacturing technology and the increasing complexity of the UAV working environment, UAV swarms are bound to thrive in future military and civilian tasks. Effective internal cooperation across UAVs is crucial for accomplishing the mission successfully in the UAV swarm confrontation game [1]. How to make optimal decision in asymmetric UAV swarm confrontation tasks is still an open issue [2,3]. In this study, we propose a novel multi-agent deep reinforcement learning method named counterfactual baseline-based MAPPO (CB-MAPPO) to tackle the decision-making issues of asymmetric UAV swarm confrontation missions. Our contributions are summarized as follows.

• A MADRL-based multi-UAV confrontation game model is established on a 3D-based JSBSim flight simulation platform.

• We propose a novel mult-agent deep reinforcement learning method named CB-MAPPO, which introduces a counterfactual baseline mechanism within the MAPPO framework, enabling more accurate credit assignment for each agent in multi-agent environments with partial observability and asymmetric team structures. Unlike standard MAPPO, which uses a shared baseline, CB-MAPPO instead computes baselines for each agent, thereby reducing the variance of the policy gradient estimates and improving learning efficiency.

• The experimental results show that the Nash equilibrium can be well achieved. By employing CB-MAPPO, the number of destroyed enemies is increased by 37.5%, and the survival rate is increased by 40%.

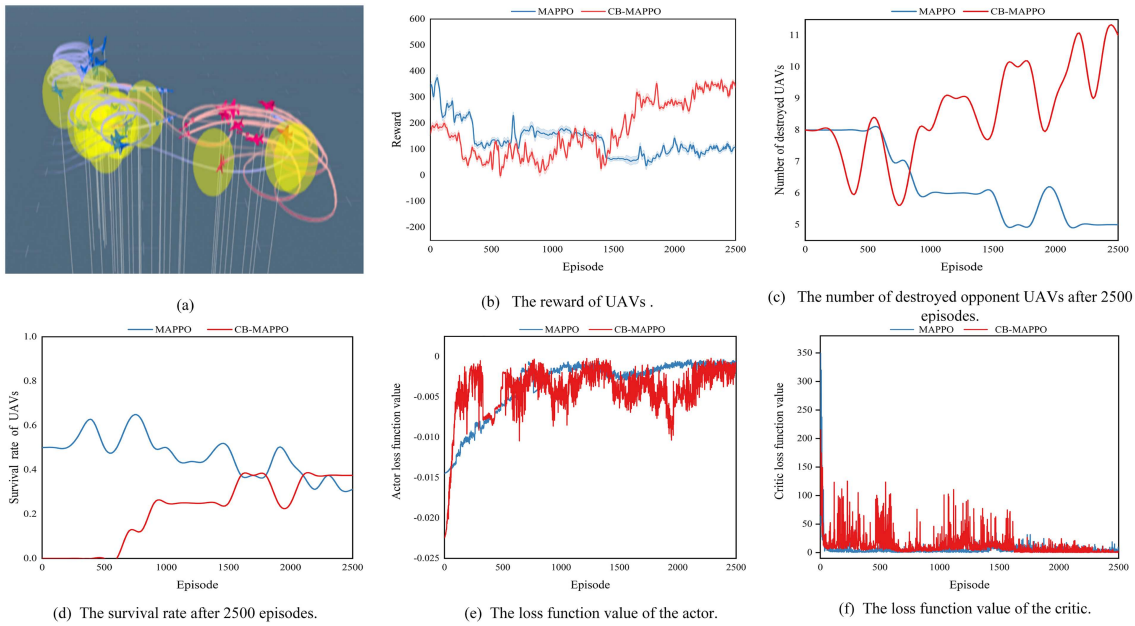
Problem formulation. Due to the incomplete information, the UAV swarm game can be formulated as a partially observable Markov game, which is represented by an eight-tuple: $\langle N, S, O, A, P, Z, R, \gamma\rangle$, where N is the number of total UAVs, S is the set of interaction states between UAVs and the environment, $s \in S$; O denotes the joint observation space of all UAVs, $O = O_1 \times O_2 \times \cdots \times O_N$, where O^i is the observation space of the i-th UAV. At each time step, each UAV can only obtain its own local observation $o^i \in O^i$. A is the set of UAV actions, and $A = \{A_1 \times A_2 \times \cdots \times A_N\}$. The state transition probability P is indeed related to the joint actions of all UAVs. Specifically, $P(s'|s,a)$ denotes the probability that the environment transitions to the next state s' given the current state s and the joint action a taken by all UAVs. The observation function Z is defined as $Z = P(o|s)$, which represents the probability distribution over the joint observation o that all UAVs receive, given the current environment state s. In other words, Z describes how the state determine the observations. R is the immediate reward function of the UAV, which is obtained when all UAVs transition to the state s' after executing the joint action; γ is the discount factor, which represents the trade-off between the immediate reward and the future reward of the UAV.

To establish the UAV swarm confrontation game environment, both red and blue UAVs adopt a 6-DOF dynamic model, and the dynamics of each UAV can be described as follows:

$$\begin{cases} \dot{\phi} = r_\phi + (r_\varphi \cos\phi + r_\theta \sin\phi)\tan\theta, \\ \dot{\theta} = r_\theta \cos\phi - r_\varphi \sin\phi, \\ \dot{\varphi} = (r_\varphi \cos\phi + r_\theta \sin\phi)/\cos\theta, \\ \dot{v}_x = v_x - \dfrac{F}{m}(\cos\varphi\sin\theta\cos\phi + \sin\varphi\sin\phi)\mathrm{d}t, \\ \dot{v}_y = v_y - \dfrac{F}{m}(\sin\varphi\sin\theta\cos\phi - \cos\varphi\sin\phi)\mathrm{d}t, \\ \dot{v}_z = v_z + \left(g - \dfrac{F}{m}\cos\phi\cos\theta\right)\mathrm{d}t, \\ -0.2 \text{ rad/s} < r_\phi < 0.2 \text{ rad/s}, \\ -0.2 \text{ rad/s} < r_\varphi < 0.2 \text{ rad/s}, \\ -0.1 \text{ rad/s} < r_\theta < 0.1 \text{ rad/s}, \end{cases} \tag{1}$$

where ϕ represents the roll angle, r_ϕ is the angular velocity in roll; θ denotes the pitch angle, r_θ is the pitch angle velocity; φ

* Corresponding author (email: xxthongchen@buu.edu.cn)



(a) (b) The reward of UAVs . (c) The number of destroyed opponent UAVs after 2500 episodes.

(d) The survival rate after 2500 episodes. (e) The loss function value of the actor. (f) The loss function value of the critic.

Figure 1 (Color online) (a) The middle scene of UAV swarm confrontation; (b)–(f) are the comparisons between CB-MAPPO and MAPPO under different evaluation metrics.

represents the yaw angle, r_φ is the yaw angular velocity; F represents the driving force, m is UAV's mass, g is the gravitational acceleration, and dt denotes the differential variable of time t.

Proposed solution. To improve the capability of UAVs in swarm confrontation scenarios, we incorporate MAPPO with counterfactual baseline and propose a method named CB-MAPPO. The state-action value function Q_w, the state value function $V_{\boldsymbol{\pi}}$, the advantage function $A_{\boldsymbol{\pi}}$, and the expected discount reward function $\eta(\boldsymbol{\pi})$ are respectively computed by [4]

$$\begin{aligned} &Q_w(s_t, \boldsymbol{a}_t) = \mathbb{E}_{s_{t+1}, \boldsymbol{a}_{t+1}}[R_t \mid s_t, \boldsymbol{a}_t], \\ &V_{\boldsymbol{\pi}}(s_t) = \mathbb{E}_{\boldsymbol{a}_t, s_{t+1}, \ldots}[R_t \mid s_t], \\ &A_{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) = Q_w(s_t, \boldsymbol{a}_t) - V_{\boldsymbol{\pi}}(s_t), \\ &\eta(\boldsymbol{\pi}) = \mathbb{E}_{s_0 \sim p_0(s_0)}\{V_{\boldsymbol{\pi}}(s_0)\}, \end{aligned} \tag{2}$$

where $\boldsymbol{\pi}$ is the joint policy, s_t is the state at time step t, $\boldsymbol{a}_t$ is UAV's action at time step t, R_t is the reward obtained by the UAV at time step t, and $p_0(s_0)$ is the probability distribution of the initial state s_0.

To solve the credit assignment problem, we use counterfactual baseline, where a centralized critic network is adopted to estimate the state-action value function, which can be denoted as

$$\hat{Q}^i(s_t, \boldsymbol{a}_t) = Q_{\overline{w}^i}(s_t, \boldsymbol{a}_t) + \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{\mathrm{T}}\delta_T, \tag{3}$$

where the temporal difference error $\delta_t = r_t + \lambda Q_{w^i}(s_{t+1}, a_{t+1}) - Q_{w^i}(s_t, a_t)$, and $Q_{\overline{w}^i}(s_t, \boldsymbol{a}_t)$ is the target state-value function of UAV i.

Simulation. The middle scene of two UAV swarms confrontation is displayed in Figure 1(a), and Figure 1(b) shows that the reward of UAVs is higher than that of blue UAVs when the number of episodes is larger than 1500, indicating CB-MAPPO outperforms MAPPO even though its UAV swarm size is only half of the latter. Figure 1(c) shows the number of destroyed opponent UAVs under the two methods. One can see that at the initial stage of learning, the number of destroyed UAVs under MAPPO is 8 per episode, while fluctuating near lower values under CB-MAPPO. As the number of episodes increases, although the number of destroyed blue UAVs with respect to CB-MAPPO fluctuates, it generally rises and reaches a peak value of 11, which demonstrates that 11 blue UAVs adopting MAPPO can be killed by 8 red UAVs using CB-MAPPO.

As shown in Figure 1(d), the survival rate of CB-MAPPO is higher than that of MAPPO after 1500 episodes. We can see that the loss function has reached convergence (Figures 1(e) and (f)), indicating that the UAV countermeasure mission is stabilized.

Conclusion. To summarize, we have proposed a multi-agent deep reinforcement learning method named CB-MAPPO to address the decision-making issues of asymmetric UAV swarm confrontation tasks. The results indicate that CB-MAPPO can converge to the Nash equilibrium and obtain the best performance. By employing CB-MAPPO, the number of destroyed opponent UAVs has increased by 37.5%, and the survival rate can be increased by 40%.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2018AAA0100804), Basic Science Center Program of National Natural Science Foundation of China (Grant No. 62388101), National Natural Science Foundation of China (Grant No. 62173237), Aeronautical Science Foundation of China (Grant No. 20240055054001), Open Fund of Key Laboratory of Technology and Equipment of Tianjin Urban Air Transportation System (Grant No. TJKL-UAM-202305), Joint Fund of Ministry of Natural Resources Key Laboratory of Spatiotemporal Perception and Intelligent Processing (Grant No. 232203), and Applied Basic Research Programs of Liaoning Province (Grant No. 2025JH2/101300011).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chen J, Li T, Zhang Y, et al. Global-and-local attention-based reinforcement learning for cooperative behaviour control of multiple UAVs. IEEE Trans Veh Technol, 2024, 73: 4194–4206
2 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. 2017. ArXiv:1707.06347
3 de Marco A, D'Onza P M, Manfredi S. A deep reinforcement learning control approach for high-performance aircraft. Nonlinear Dyn, 2023, 111: 17037–17077
4 Guo D, Tang L, Zhang X, et al. Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning. IEEE Trans Veh Technol, 2020, 69: 13124–13138