Journal of Systems Engineering and Electronics
Vol. 32, No. 6, December 2021, pp.1421 – 1438

UAV cooperative air combat maneuver decision based on multi-agent reinforcement learning

ZHANG Jiandong[1], YANG Qiming[1,*], SHI Guoqing[1], LU Yi[2], and WU Yong[1]

1. School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China;
2. Shenyang Aircraft Design Institute, Shenyang 110035, China

Abstract: In order to improve the autonomous ability of unmanned aerial vehicles (UAV) to implement air combat mission, many artificial intelligence-based autonomous air combat maneuver decision-making studies have been carried out, but these studies are often aimed at individual decision-making in 1v1 scenarios which rarely happen in actual air combat. Based on the research of the 1v1 autonomous air combat maneuver decision, this paper builds a multi-UAV cooperative air combat maneuver decision model based on multi-agent reinforcement learning. Firstly, a bidirectional recurrent neural network (BRNN) is used to achieve communication between UAV individuals, and the multi-UAV cooperative air combat maneuver decision model under the actor-critic architecture is established. Secondly, through combining with target allocation and air combat situation assessment，the tactical goal of the formation is merged with the reinforcement learning goal of every UAV, and a cooperative tactical maneuver policy is generated. The simulation results prove that the multi-UAV cooperative air combat maneuver decision model established in this paper can obtain the cooperative maneuver policy through reinforcement learning, the cooperative maneuver policy can guide UAVs to obtain the overall situational advantage and defeat the opponents under tactical cooperation.

Keywords: decision-making, air combat maneuver, cooperative air combat, reinforcement learning, recurrent neural network.

DOI: 10.23919/JSEE.2021.000121

1. Introduction

The world has been shocked by the advent of the unmanned aerial vehicle (UAV) because it has had a significant impact on war patterns in high-tech local wars in recent decades [1]. However, by reason of limitations of the communication technology, UAV cannot be used to perform air combat missions through ground-based remote control operations [2,3], so it is the future of UAV air combat to let UAVs make autonomous air combat maneuvering decisions based on the situation environment, and it is also an important development direction of UAV intelligence [4].

Autonomous air combat maneuver decision refers to the process of automatically generating flight control commands to gain the advantage during air combat confrontation based on mathematical optimization, artificial intelligence, and other methods. At present, the research methods of autonomous air combat maneuver decision can be mainly divided into three categories based on the game theory [5,6], the optimization method [7−9], and the artificial intelligence [10−15] method. The methods based on the game theory and optimization algorithms mostly divide the actions of the aircraft into a limited number of maneuver actions and then calculate the effect of each action on the situation to select the best maneuver action execution. The established model based on these methods can intuitively reflect the main factors of air combat confrontation. However, due to the limitation of real-time computing, the set of mobile actions is often simple and sparse, and it is impossible to achieve complex tactical actions in air combat. The methods based on artificial intelligence mainly include expert system method [10], neural network method [11], and reinforcement learning methods [12−17]. The core of the expert system method [10] is to summarize the pilot's flight experience into a rule library, and the flight control commands are generated by the rules in the rule library, while the neural network method stores the rules in the form of network parameters, so compared with table lookup traversal, the neural network has a faster response speed and stronger robustness, but the training of the neural network still requires a lot of rule data, and air combat is a highly complex game process, so it is difficult to build a

Manuscript received December 22, 2020.

*Corresponding author.

This work was supported by the Aeronautical Science Foundation of China (2017ZC53033) and the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University (CX2020156).

complete rule base. Reinforcement learning is a machine learning method for agents to learn action policy by interacting with the environment. The learning process does not require sample data, so it is an effective method to solve sequential decision problems that lack prior models [18]. Scholars have researched air combat maneuver decision-making based on reinforcement learning [12−18]. In [12], an air combat maneuver decision model is established by combining fuzzy logic and Q-learning, and in [13−17], an air combat maneuver decision model is constructed by using deep reinforcement learning algorithms.

The above-mentioned UAV air combat maneuvering decision-making researches are carried out in the context of 1v1 air combat. However, real air combat is usually a cooperative operation between multiple aircraft formations [19]. Multi-aircraft cooperative air combat is a closely coupled coupling process of three aspects [20]: air combat situation assessment, target allocation, and maneuvering decision. Compared with the maneuvering decision of the 1v1 confrontation, in addition to the increase in the number of aircraft, the tactical cooperation between the aircraft should be considered in multi-aircraft cooperative air combat, so the decision model is more complicated. The research on multi-aircraft cooperative air combat decision-making can be divided into centralized and distributed ones. The centralized method is to calculate the actions of all aircraft in the formation by a single center. The parameter scale of this type of model increases sharply with the increase in the number of aircraft, so the calculation complexity is very large. For example, the decision model established through the differential game method in [21,22], although the model can be realized in principle, it is very difficult to solve complex nonlinear differential game models. In [23], a training and evaluation model for mid-range cooperative air combat is established, and the correlation between mid-range air combat and cooperative decision-making is analyzed from the perspective of mid-range air combat tactics. However, there is still a problem of insufficient real-time performance in solving the model. The idea of the distributed method is that each aircraft in the formation calculates its maneuvers based on target allocation, thereby reducing the complexity of the model and achieving the coordination of formation tasks through target allocation. The current distributed methods mostly use target allocation to transform many-to-many cooperative air combat into multiple one-to-one confrontations [24−27]. This method cannot effectively play the multi-target attack capability and tactical coordination of formation operations. Therefore, the combat effectiveness of 1 + 1> 2 cannot be achieved.

Based on the research of 1v1 autonomous air combat maneuver decision and the distributed multi-agent reinforcement learning idea, a multi-UAV cooperative air combat maneuver decision model is established. In terms of model architecture, a bidirectional recurrent neural network is used to construct the UAV formation communication network, which connects the individual standalone actor-critic air combat maneuver decision models into a formation model. Through communication, the maneuver decisions made by each UAV in the model consider not only the state of their own but also the state of other teammates, thereby achieving collaboration on the organizational structure. In terms of tactical collaboration, a target assignment method is designed according to the characteristics of multi-target attacking, and each UAV's reinforcement learning reward value is calculated through combining the target assignment method and the air combat situation evaluation value, the individual's reinforcement learning process is guided by its reward value, making the combat goal of formation unified with the learning policy of every UAV. The simulation results prove that the multi-UAV cooperative air combat maneuver decision model established in this paper can independently learn to obtain cooperative air combat maneuver policy, and the policy makes UAVs achieve tactical cooperation in the air combat process to get the overall formation advantage and defeat the opponents.

The following part of the paper is arranged as follows. The research of the 1v1 maneuver decision is introduced in Section 2, and the multi-aircraft air combat maneuver decision model is introduced in Section 3. Section 4 introduces the training and testing of the model through simulation analysis. Finally, Section 5 concludes the full text.

2. Related work

In the previous research on air combat maneuver decision based on reinforcement learning, UAV's 1v1 air combat maneuver decision model was established based on deep deterministic policy gradient (DDPG) [16] and deep Q network (DQN) [17] respectively. The modeling framework is shown as Fig. 1.

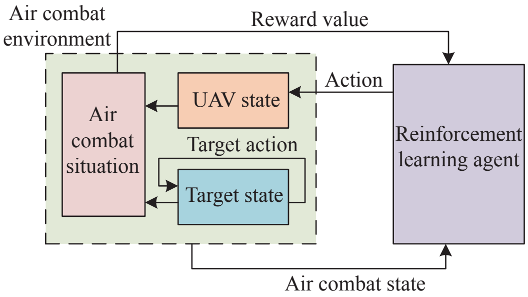

Fig. 1 UAV short-range air combat maneuver decision model framework based on reinforcement learning[17]

In the process of establishing the air combat maneuver decision model, firstly, a one-to-one air combat environment model was established. Secondly, the state space, action space and reward function of the reinforcement learning model were designed according to the air combat environment model. The reward of the reinforcement learning mode was calculated based on the air combat situation evaluation value, so as to guide the UAV to learn the maneuver policy that can obtain the advantage of air combat. In this paper, we will build a multi-to-multi cooperative air combat maneuver decision reinforcement learning model based on this modeling idea.

The training process of reinforcement learning models is prone to failure due to sparse rewards. In [17], we proposed a "basic-confrontation" training method, which effectively improved the training effect of reinforcement learning through a reasonable arrangement of the training process. In this paper, we will continue to use this training method in the training of the UAV cooperative air combat maneuver decision model.

3. Multi-UAV cooperative air combat maneuver decision model

In this section, firstly, a multi-to-multi air combat environment model is established to clarify the state space, action space and reward value of the decision model. Secondly, the proposed target assignment algorithm is introduced as a key step of maneuver policy coordination. Finally, the multi-UAV cooperative air combat maneuver decision model based on multi-agent reinforcement learning is introduced.

3.1 Air combat environment

3.1.1 Aircraft motion model

The motion model of the aircraft is the basis of the air combat model. The control commands for maneuver decisions are executed through the motion model to change the position and speed of the aircraft, thereby changing the air combat situation. The maneuver decision mainly considers the positional relationship and velocity vectors of the two sides in the 3D space, while the body attitude has little influence on the maneuver decision. Therefore, a three-degree-of-freedom particle model is used as the aircraft motion model.

In the ground coordinate system, the *ox* axis takes the east, the *oy* axis takes the north, and the *oz* axis takes the vertical direction. The motion model of the aircraft in the coordinate system is shown in

$$\begin{cases}\dot{x} = v\cos\gamma\sin\psi \\ \dot{y} = v\cos\gamma\cos\psi \\ \dot{z} = v\sin\gamma\end{cases} \tag{1}$$

where x, y, and z represent the position of the aircraft in the coordinate system. v represents speed, and $\dot{x}$, $\dot{y}$, and $\dot{z}$ represent values of speed v on three coordinate axes. The track angle γ represents the angle between the velocity vector and the horizontal plane o-x-y. The heading angle ψ represents the angle between the projection v' of the velocity vector on the o-x-y plane and the oy axis. In the same coordinate system, the dynamic model of the aircraft is shown in

$$\begin{cases}\dot{v} = g(n_x - \sin\gamma) \\ \dot{\gamma} = \dfrac{g}{v}(n_z\cos\mu - \cos\gamma) \\ \dot{\psi} = \dfrac{gn_z\sin\mu}{v\cos\gamma}\end{cases} \tag{2}$$

where g represents the acceleration of gravity.$[n_x,n_z,\mu]$ is a set of the control variables that control the aircraft to maneuver. n_x is the overload in the velocity direction, representing the thrust and deceleration of the aircraft. n_z represents the overload in the pitch direction, which is the normal overload. μ is the roll angle around the velocity vector. n_x controls the speed of the aircraft, while n_z and μ control the direction of the velocity vector, thereby controlling the aircraft to perform maneuvers. The parameters of the aircraft particle model are shown in Fig. 2.

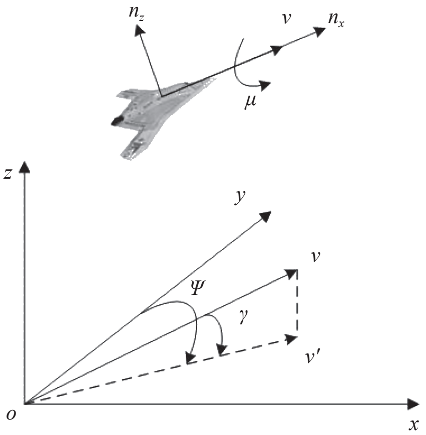

Fig. 2 Aircraft three-degree-of-freedom particle model

3.1.2 Situation evaluation model

(i) One-to-one scene

The purpose of the maneuver is to try to make the target into UAV's attack range while avoiding the UAV from entering the target's attack range, so that UAV can enter an advantageous position from any situation. In modern short-range air combat, air-to-air missiles can intercept and lock the target which is in the field of view of the seeker, and the missile can be launched after the target is intercepted. This paper sets the missile to have only the tail attack capability, and the missile interception area is shown in Fig. 3. In the figure, v_U and v_T represent the velocity of the UAV and the target, $\boldsymbol{D}$ is the distance vector, indicating the positional relationship between the UAV and the target, α_U and α_T represent the angle between the UAV velocity vector and the target velocity vector and $\boldsymbol{D}$ respectively.

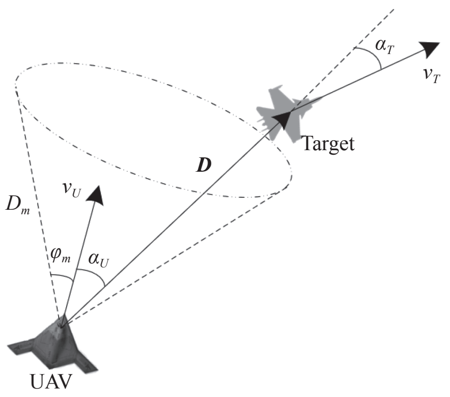

Fig. 3 One-to-one short-range air combat situation

Assuming that the maximum interception distance of the missile is D_m and the angle of view is φ_m, the interception area of the missile is a cone area Ω. The goal of UAV maneuvering in air combat is to make the target enter the UAV intercept area Ω_U while avoiding the UAV entering the target intercept area Ω_T.

Air combat situation evaluation is to quantitatively characterize the pros and cons of UAV in the current situation. According to the definition of the missile interception area, if the target is in the UAV's missile interception area, it means that UAV can launch weapons to attack the target, so UAV is in an advantageous situation. Define the advantage value when UAV intercepts the target:

$$\eta_U = \begin{cases} \text{Re}, & (x_T, y_T, z_T) \in \Omega_U;\ \alpha_T < \dfrac{\pi}{2} \\ 0, \text{otherwise} \end{cases}. \tag{3}$$

Similarly, the target can get the advantage value η_T when the UAV is intercepted by it. In air combat, the advantage value obtained by UAV based on interception opportunities is defined as

$$\eta_A = \eta_U - \eta_T. \tag{4}$$

Besides, in the close one-to-one air combat, due to the small field of view of the cannon and some missiles, the weapon launch conditions can generally be formed only in the case of tail-chasing, so the requirements for the angle relationship are more stringent, therefore, the advantage value calculated based on the angle parameters and distance parameters of both sides is defined as

$$\eta_B = \begin{cases} \dfrac{\pi - \alpha_U - \alpha_T}{\pi}, & D \leqslant D_m \\ \dfrac{\pi - \alpha_U - \alpha_T}{\pi} \mathrm{e}^{-\frac{(D-D_m)^2}{D_m^2}}, & D > D_m \end{cases}. \tag{5}$$

It can be seen from (5) that when the UAV is chasing the target, the advantage value will be 1, on the contrary, when the UAV is chased by the target, the advantage value will be −1. Also, when the distance between the two sides is greater than the longest interception distance of the missile, the advantage value decays exponentially. Based on (4) and (5), the evaluation function of the UAV air combat situation is

$$\eta = \eta_A + \eta_B. \tag{6}$$

(ii) Multi-to-multi scene

As shown in Fig. 4, in multi-aircraft air combat, set the number of UAVs to n, respectively recorded as UAV$_i$ (i=1, 2, $\cdots$, n), and the number of targets to m, respectively recorded as Target$_j$ (j= 1, 2, $\cdots$, m). The number of targets is set to be not greater than the number of UAVs, that is to say, $m \leqslant n$. The multi-aircraft air combat environment will be analyzed from three aspects of state space, action space, and reward value. In the follow-up content, we set a prerequisite, that is, each UAV is completely observable to the status of other individuals in the air combat environment. The focus of this paper is on the cooperative air combat maneuvering decision problem under the premise that the individual UAV status is fully observable.

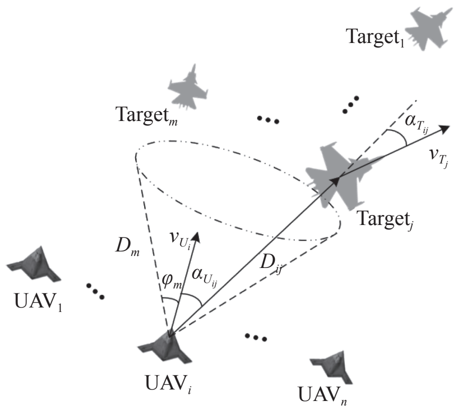

Fig. 4 Multi-UAV air combat state variables

3.1.3 State space

Compared with the one-to-one air combat maneuver decision model, due to the increasement of the number of UAVs and targets in multi-aircraft air combat, each UAV needs to consider the relative state with all other aircraft (targets and friends) when making maneuver decisions. According to [17], the relative situation of a UAV and target in air combat can be fully described by a 13-D vector space $\boldsymbol{s}$=[v_U, γ_U, ψ_U, v_T, γ_T, ψ_T, D, γ_D, ψ_D, α_U, α_T, z_U, z_T], where $D = \|\boldsymbol{D}\|$, γ_D represents the angle between $\boldsymbol{D}$ and the *o*-*x*-*y* plane, and ψ_D represents the angle between the projection vector of $\boldsymbol{D}$ on the *o*-*x*-*y* plane and the *oy* axis. Thus the relative state between UAV$_i$ and Target$_j$ is denoted as $\boldsymbol{s}_{ij} = [v_{U_i}, \gamma_{U_i}, \psi_{U_i}, v_{T_j}, \gamma_{T_j}, \psi_{T_j}, D_{ij}, \gamma_{D_{ij}}, \psi_{D_{ij}}, \alpha_{U_{ij}}, \alpha_{T_{ij}}, z_{U_i}, z_{T_j}]$, and the relative state between UAV$_i$ and any friend UAV$_k$ is denoted as $\boldsymbol{s}_{ik}$=[v_{U_i}, γ_{U_i}, ψ_{U_i}, v_{T_k}, γ_{T_k}, ψ_{T_k}, D_{ik}, $\gamma_{D_{ik}}$, $\psi_{D_{ik}}$, $\alpha_{U_{ik}}$, $\alpha_{T_{ik}}$, z_{U_i}, z_{T_k}], then the observation state of UAV$_i$ in multi-aircraft air combat is

$$\boldsymbol{S}_i = \left[\cup \boldsymbol{s}_{ij}\Big|_{j=1,2,\cdots,m}, \cup \boldsymbol{s}_{ik}\Big|_{k=1,2,\cdots,n(k\neq i)}\right]. \tag{7}$$

3.1.4 Action space

In the process of multi-aircraft air combat, each UAV makes its own maneuver decision according to its situation in the air combat environment. According to the aircraft dynamics model described in (2), UAV controls the flight through three variables n_x, n_z, and μ, so the action space of UAV$_i$ is $\boldsymbol{A}_i = [n_{xi}, n_{zi}, \mu_i]$.

3.1.5 Situation evaluation

In multi-aircraft cooperative air combat, the situation evaluation values η_A and η_B between each UAV and each target can be calculated according to (4) and (5) respectively, and the situation evaluation values between UAV$_i$ and Target$_j$ can be noted as $\eta_{A_{ij}}$ and $\eta_{A_{ij}}$. Besides, the influence of the relative state of the UAV$_i$ and its friend UAV$_k$ on its situation should also be considered. If the distance between the UAV$_i$ and UAV$_k$ is too close, it will increase the risk of collision, so the situation evaluation function of UAV$_i$ and UAV$_k$ is defined as

$$\eta_{C_{ik}} = \begin{cases} -P, \ D_{ik} < D_{\text{safe}} \\ 0, \ \text{otherwise} \end{cases} \tag{8}$$

where D_{safe} is the minimum safety distance and P is a large positive value.

3.2 Target assignment

In multi-aircraft cooperative air combat, from the overall perspective of air combat, UAV formation has the greatest advantage in air combat, which means that each target can be attacked by UAV weapons, but each UAV can only be maneuvered against one target at a time. Therefore, in multi-aircraft cooperative air combat, UAV must make target assignments while making maneuver decisions to achieve tactical coordination.

Target assignment methods can be divided into two types: centralized and distributed ones. The centralized method has high requirements on the computing and communication capabilities of the assignment center, but compared with the distribution method, the centralized method has stronger real-time performance and reliability and is more suitable for target assignment in air combat. This paper designs a target assignment algorithm for multi-UAV formation based on the Hungarian algorithm [28].

3.2.1 Target assignment model

In air combat, n UAVs fight against m targets, and $n \geqslant m$. Note that the target assignment matrix is $\boldsymbol{X} = [x_{ij}]$, when $x_{ij} = 1$, it means that Target$_j$ is assigned to UAV$_i$, and when $x_{ij} = 0$, it means that Target$_j$ is not assigned to UAV$_i$. In the process of multi-aircraft air combat, there may be situations that multiple targets are simultaneously in the attack area of an UAV. Therefore, the model sets that each UAV can simultaneously launch weapons against L targets in the attack zone, that is, $\sum_{j=1}^{m} x_{ij} \leqslant L$. In addition, each target should be assigned at least one UAV to attack, that is, $\sum_{i=1}^{n} x_{ij} \geqslant 1$, and all UAVs should be put into combat, that is, $\sum_{j=1}^{m} x_{ij} > 0$. Taking the maximization of the situation advantage of UAVs as the assignment goal, the target assignment model is established as follows:

$$\begin{aligned} &\max \quad \sum_{j=1}^{m}\sum_{i=1}^{n} \eta_{ij} \cdot x_{ij} \\ &\text{s.t.} \quad \begin{cases} \sum_{i=1}^{n} x_{ij} \geqslant 1 \\ 0 < \sum_{j=1}^{m} x_{ij} \leqslant L \\ x_{ij} \in \{0,1\} \end{cases} . \end{aligned} \tag{9}$$

3.2.2 Target assignment method

The purpose of UAV maneuvering in air combat is to let the target enter the attack zone to launch weapons. Therefore, in the process of target assignment, the targets in the attack zone should be assigned first, and then the targets outside the attack zone can be allocated. Therefore, the target assignment method is divided into the following two steps.

(i) Targets in the attack zone assigned first

Construct two $n \times m$-dimensional matrices $\boldsymbol{H}_A$ and $\boldsymbol{H}_B$ with $\eta_{A_{ij}}$ and $\eta_{B_{ij}}$ as elements, $\boldsymbol{H}_A = \left[\eta_{A_{ij}}\right]_{n\times m}$, $\boldsymbol{H}_B = \left[\eta_{B_{ij}}\right]_{n\times m}$. According to (3), if Target$_j$ is in the attack area of UAV$_i$, $\eta_{A_{ij}} = \text{Re}$, otherwise $\eta_{A_{ij}} \leqslant 0$. Therefore, let $\boldsymbol{H}_{A_1} = \left[\eta_{A_{ij}}\right]_{n\times m} - [\text{Re}]_{n\times m}$, then the targets numbered with column coordinates of all 0 elements in the matrix $\boldsymbol{H}_{A_1}$ are in the attack area of the UAVs numbered with the row coordinate, so the corresponding target should be assigned to the corresponding UAV, and set the corresponding x_{ij} = 1. If the number X of targets in the attack zone of UAV$_i$ exceeds L, $X > L$, then compare the element values of these X targets with UAV$_i$ in the $\boldsymbol{H}_B$ matrix and select the targets with the L largest values to be assigned to UAV$_i$.

(ii) Targets assigned outside the attack zone

For UAV$_i$, if a target within its attack zone has been assigned, it can no longer be assigned to a target outside the attack zone, and for multiple targets outside the attack zone, UAV cannot make maneuvers to make multiple targets enter the attack zone, therefore, when the targets are outside the attack zone, only one target can be assigned to the UAV. Therefore, after the assignment to the targets in the attack zone is completed, the remaining target assignment work is to assign a target to the unassigned UAV, and the assignment can be achieved by using the Hungarian algorithm. First, according to the current target assignment matrix $\boldsymbol{X}=[x_{ij}]_{n\times m}$, if x_{ij}= 1, all elements on the ith row and the jth column of $\boldsymbol{H}_B$ are deleted to obtain matrix $\boldsymbol{H}_{B_1}$. Based on $\boldsymbol{H}_{B_1}$, the Hungarian algorithm is used to calculate the target allocation. Since $n \geqslant m$ and L>0, if the number of rows in $\boldsymbol{H}_{B_1}$ is greater than the number of columns, the Hungarian algorithm is completed by using the complement method [28] to achieve the target assignment, and set the corresponding x_{ij} = 1.

After completing the above two steps, the target assignments are completed and the target assignment matrix $\boldsymbol{X}=[x_{ij}]_{n\times m}$ is obtained. The operation logic pseudo code of the target assignment method is shown in Algorithm 1.

Algorithm 1 Cooperative air combat target assignment algorithm

```
Calculate each $\eta_{A_{ij}}$ and $\eta_{B_{ij}}$ according to (4) and (5), respectively.
Initialize $\boldsymbol{H}_A=\left[\eta_{A_{ij}}\right]_{n\times m}$, $\boldsymbol{H}_B=\left[\eta_{B_{ij}}\right]_{n\times m}$, $\boldsymbol{X}=[x_{ij}]_{n\times m}=[0]_{n\times m}$.
Set $\boldsymbol{H}_{A_1}=\left[\eta_{A_{ij}}\right]_{n\times m}-[\mathrm{Re}]_{n\times m}$
for $i$ from 1 to $n$ do
    for $j$ from 1 to $m$ do
    if $\boldsymbol{H}_{A_1}(i,j)$==0
        set $x_{ij}$=1
for $i$ from 1 to $n$ do
    if $\sum_{j=1}^{m} x_{ij} > L$ do
        while $\sum_{j=1}^{m} x_{ij} > L$ do
            Select min $\boldsymbol{H}_B(i,j)$, set $x_{ij}$=1
for $x_{ij}$ in $\boldsymbol{X}$ do
    if $x_{ij}$ == 1 do
        Delete the $i$th row and the $j$th column of $\boldsymbol{H}_B$, get a new matrix $\boldsymbol{H}_{B_1}$.
Based on the matrix $\boldsymbol{H}_{B_1}$, using the Hungarian algorithm to complete the remaining target allocation.
Obtain the final target assignment matrix $\boldsymbol{X}$.
```

3.3 Maneuver decision

Multi-UAV cooperative air combat is a multi-agent system where each agent makes its own decision based on its state observation and cooperation with other agents.

In the multi-agent environment, the traditional reinforcement learning method can no longer be used directly. On the one hand, the policy of each agent changes with the training process, so from the perspective of any agent, the change in environmental state is not entirely caused by its own actions, that is, the environment becomes no longer stable. It is a serious challenge for the stability of each agent learning process. On the other hand, a completely discrete decision model cannot achieve information interaction between agents, so collaboration between agents cannot be achieved.

In the centralized decision-making model, as the number of agents increases, the parameter space of the model will increase exponentially. Compared with the centralized model, the distributed multi-agent system can effectively deal with the change of the number of agents and can organize the learning behavior of the individual agents into group collaboration through the coordination mechanism. Therefore, this paper creates a communication-based distributed multi-agent learning system to realize the maneuver decision of multi-UAV cooperative air combat.

3.3.1 Policy coordination mechanism

Multi-UAV cooperative air combat can be regarded as a competitive game between n UAVs and m targets. An air combat model is established based on the framework of a random game which can be represented by a tuple $(\boldsymbol{S},\{\boldsymbol{A}_i\}_{i=1}^{n},\{\boldsymbol{B}_i\}_{i=1}^{m},T,\{R_i\}_{i=1}^{n+m})$. $\boldsymbol{S}$ represents the state space of the current game, which can be shared by all agents. The action space of UAV$_i$ is defined as $\boldsymbol{A}_i$, and action space of Target$_i$ is defined as $\boldsymbol{B}_i$. $T:\boldsymbol{S}\times\boldsymbol{A}^n\times\boldsymbol{B}^m\rightarrow\boldsymbol{S}$ represents the deterministic transfer function of the environment, and $R_i:\boldsymbol{S}\times\boldsymbol{A}^n\times\boldsymbol{B}^m\rightarrow\mathbb{R}$ represents the reward value function of UAV$_i$. In the research of cooperative air combat, it is assumed that the performance of both sides are the same, so the aircraft in their respective formations have the same action space, that is, $\boldsymbol{A}_i=\boldsymbol{A}$ and $\boldsymbol{B}_i=\boldsymbol{B}$

for UAV$_i$ ($i \in [1,n]$) and Target$_j$ ($j \in [1,m]$), respectively.

The situation of UAV formation should be evaluated by the situation of all UAVs. The global reward value of the UAV formation is defined as the average value of each UAV reward value,

$$r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} R_i(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}). \tag{10}$$

For simplicity, the time subscript t of the global reward value $r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$ is omitted. $r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$ represents the reward value obtained by the UAV formation at time t when the environmental state is $\boldsymbol{s}$, the UAV formation takes action $\boldsymbol{a} \in \boldsymbol{A}^n$, and the target formation takes action $\boldsymbol{b} \in \boldsymbol{B}^m$. The goal of the UAV formation is to learn a policy to maximize the expected sum of the discount rewards, i.e., $\mathbb{E}\left[\sum_{k=0}^{+\infty} \lambda^k r_{t+k}\right]$, where $0 < \lambda \leqslant 1$ is the discount factor, indicating the uncertainty of future rewards. Contrary to UAV formations, the action policy of the target formation is to minimize the expected sum of the discount rewards of UAVs. In summary, the following minimax game can be obtained:

$$Q^*(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) = \\ r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) + \lambda \max_{\theta} \min_{\phi} Q^*(\boldsymbol{s}',\boldsymbol{a}_\theta(\boldsymbol{s}'),\boldsymbol{b}_\phi(\boldsymbol{s}')) \tag{11}$$

where $\boldsymbol{s}' \equiv \boldsymbol{s}^{t+1}$, which is determined by the state transition function $T(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$, representing the state at time t+1. $Q^*(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$ represents the optimal state-action value which follows the Bellman optimization equation. Suppose the UAV formation uses the parameterized deterministic policy $\boldsymbol{a}_\theta : \boldsymbol{S} \to \boldsymbol{A}^n$, and the target formation uses the parameterized deterministic policy $\boldsymbol{b}_\phi : \boldsymbol{S} \to \boldsymbol{B}^m$, where θ and ϕ are the parameters of the policy function. To simplify the problem, the policy of target is set to be fixed, that is, the target performs the same maneuver under the same state, and the effect of the target policy is not considered in the subsequent research, so the random game defined by (11) can be transformed into a Markov decision problem[29]:

$$Q^*(\boldsymbol{s},\boldsymbol{a}) = r(\boldsymbol{s},\boldsymbol{a}) + \lambda \max_{\theta} Q^*(\boldsymbol{s}',\boldsymbol{a}_\theta(\boldsymbol{s}')). \tag{12}$$

The global reward defined by (10) can reflect the overall situation of the UAV formation, but the global reward does not reflect the role of each UAV individual in coordination. The global coordination is driven by the goals of each individual. Therefore, the reward value of each UAV is defined as

$$r_i(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b}) = \sum_{j=1}^{m} x_{ij}\eta_{ij} + \sum_{k=0}^{n(k\neq i)} \eta_{C_{ik}} \tag{13}$$

which is used to characterize the reward of UAV$_i$ at the time t, under the conditions that environmental state is $\boldsymbol{s}$, UAV formation takes action $\boldsymbol{a} \in \boldsymbol{A}^n$, and target formation takes action $\boldsymbol{b} \in \boldsymbol{B}^m$. $\sum_{j=1}^{m} x_{ij}\eta_{ij}$ represents the situation evaluation value of UAV$_i$ relative to the targets assigned to it, and $\sum_{k=0}^{n(k\neq i)} \eta_{C_{ik}}$ represents a penalty term which is used to constrain the distance between UAV$_i$ and its teammates. Based on (13), for n UAV individuals, there are n Bellman equations as shown in (14), where the policy function $\boldsymbol{a}_\theta$ has the same parameters θ.

$$Q_i^*(\boldsymbol{s},\boldsymbol{a}) = r_i(\boldsymbol{s},\boldsymbol{a}) + \lambda \max_{\theta} Q_i^*(\boldsymbol{s}',\boldsymbol{a}_\theta(\boldsymbol{s}')) \tag{14}$$

In the training process of reinforcement learning, through the distribution of reward values, the behavior feedback of each UAV in target assignment, situation advantage, and collision avoidance are defined. After training, there is policy coordination that can be achieved which makes the behavior of each UAV reach a tacit agreement, and it is not necessary to carry out a centralized target assignment.

3.3.2 Policy learning mechanism

The premise of achieving cooperation is information interaction between individuals. Therefore, this paper builds a multi-UAV maneuver decision model based on a bidirectional recurrent neural network to ensure the information interaction between UAVs and achieve the coordination of the formation maneuvering.

As shown in Fig. 5, the single UAV air combat maneuvering decision model based on DDPG includes the Actor and Critic network modules. On this basis, a multi-UAV air combat maneuver decision model is constructed by connecting multiple single UAV models through the communication network, and the model is shown in Fig. 6. The multi-UAV air combat maneuver decision model is composed of the Actor network and the Critic network, and the Actor network and the Critic network are respectively formed by connecting the Actor and the Critic network of each UAV through bidirectional recurrent neural network (BRNN). In the model, the hidden layers in the policy network (Actor) and Q network (Critic) of the single UAV decision model are set as the recurrent unit of the BRNN, and then the BRNN is expanded according to the number of UAVs. The policy network inputs the current air combat state and outputs the action values of each UAV. Since BRNN can not only realize the communication between UAV individuals but also serve as a memory unit, UAV can save the individual action policy while exchanging state information with teammates.

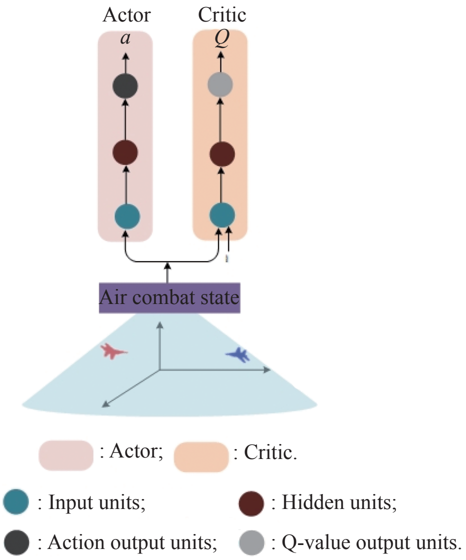

Fig. 5 Actor-Critic model framework for single UAV air combat maneuver decision

Since the model is constructed based on BRNN, the idea of network parameter learning is to expand the network into n (UAV number) sub-networks to calculate the reverse gradient, and then update the network parameters based on the time-based backpropagation algorithm. The gradient is propagated in the Q_i function and policy function of each UAV individual. During model learning, the reward value of each UAV affects the actions of each UAV, and then the resulting gradient is backpropagated and the model parameters are updated [29].

The objective function of the individual UAV$_i$ is defined as $J_i(\theta)=\mathbb{E}_{\boldsymbol{s}\sim\rho_{\boldsymbol{a}_\theta}^T}[r_i(\boldsymbol{s},\boldsymbol{a}_\theta(\boldsymbol{s}))]$, which represents the expected sum of the reward r_i, $\rho_{\boldsymbol{a}_\theta}^T$ represents the state distribution obtained by adopting the action $\boldsymbol{a}_\theta$ under the state transition function T. The state distribution is a generally stable distribution during the traversed Markov decision process, so the objective function of n UAVs can be denoted as $J(\theta)$.

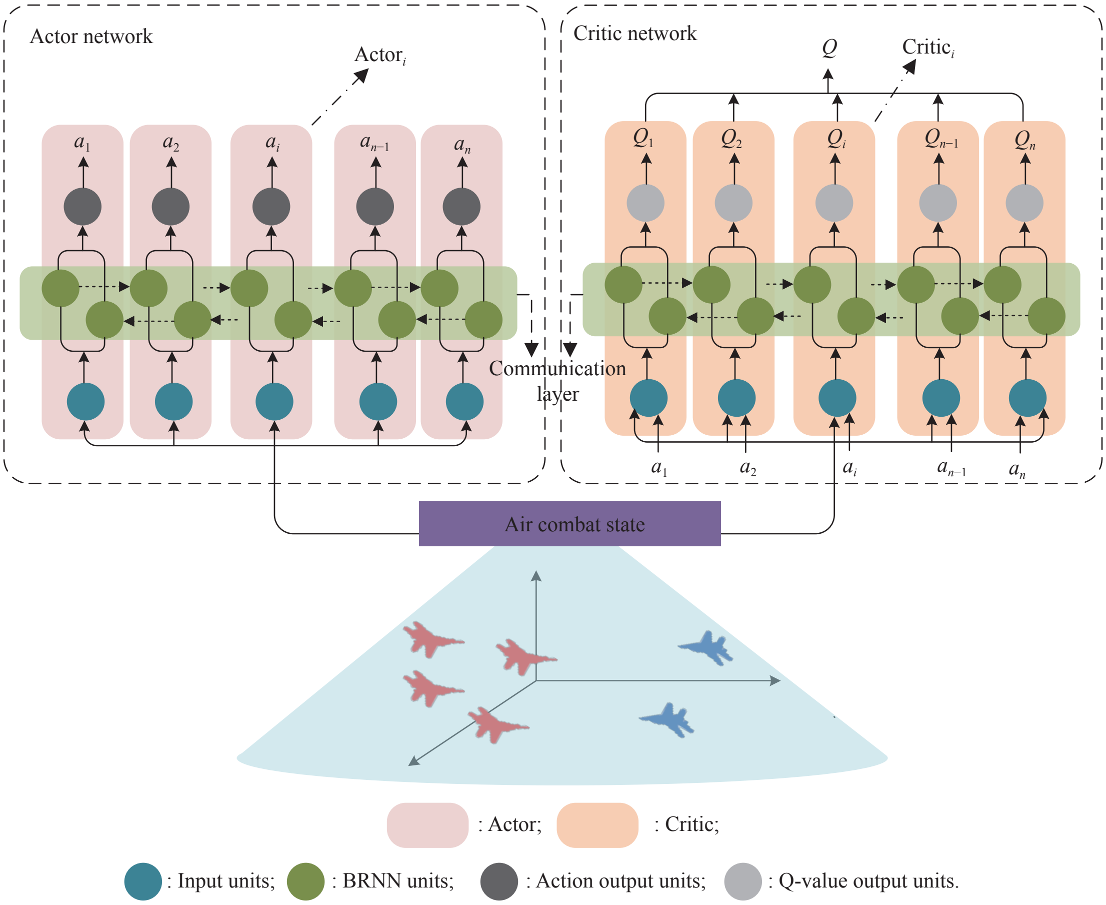

Fig. 6 Model structure of BRNN-based multi-UAV air combat maneuver decision

$$J(\theta)=\mathbb{E}_{\boldsymbol{s}\sim\rho_{\boldsymbol{a}_\theta}^T}\left[\sum_{i=1}^{n} r_i(\boldsymbol{s},\boldsymbol{a}_\theta(\boldsymbol{s}))\right] \tag{15}$$

The multiagent deterministic policy gradient theorem (MDPGT) is derived based on the deterministic policy gradient theory [30,31]. According to MDPGT, for the objective function $J(\theta)$ of n UAVs described in (15), the gradient of the policy network parameter θ is

$$\nabla_\theta J(\theta) = \mathbb{E}_{\boldsymbol{s} \sim \rho^T_{\boldsymbol{a}_\theta}} \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \nabla_\theta \boldsymbol{a}_{j,\theta}(\boldsymbol{s}) \cdot \nabla_{\boldsymbol{a}_j} Q_i^{\boldsymbol{a}_\theta}(\boldsymbol{s}, \boldsymbol{a}_\theta(\boldsymbol{s})) \right]. \quad (16)$$

The deterministic off-policy actor-critic algorithm can reduce the variance. The parameterized critic function $Q^\xi(\boldsymbol{s}, \boldsymbol{a})$ is used to estimate the state-action value function $Q_i^{\boldsymbol{a}_\theta}$ in (16). And the square sum loss function is used for training the critic. The gradient of the parameterized critic function $Q^\xi(\boldsymbol{s}, \boldsymbol{a})$ is shown in (17), where ξ is the parameter of the Q network.

$$\nabla_\xi L(\xi) = \\ \mathbb{E}_{\boldsymbol{s} \sim \rho^T_{\boldsymbol{a}_\theta}} \left[\sum_{i=1}^{n} \Big(r_i(\boldsymbol{s}, \boldsymbol{a}_\theta(\boldsymbol{s})) + \lambda Q_i^\xi(\boldsymbol{s}', \boldsymbol{a}_\theta(\boldsymbol{s}')) - \right. \\ \left. Q_i^\xi(\boldsymbol{s}, \boldsymbol{a}_\theta(\boldsymbol{s})) \Big) \cdot \nabla_{\partial\xi} Q_i^\xi(\boldsymbol{s}, \boldsymbol{a}_\theta(\boldsymbol{s})) \right] \quad (17)$$

Based on (16) and (17), the stochastic gradient descent method is used to optimize the actor and critic network. In the process of interactive learning, the network parameters are updated through the data obtained by trial and error to complete the optimization of the cooperative air combat policy.

3.3.3 Cooperative air combat maneuver decision model

According to the policy coordination mechanism and policy learning mechanism, we design the reinforcement learning process of the multi-UAV cooperative air combat maneuver decision model as follows.

First, initialize the model. Randomly initialize the online network parameters of Actor and Critic, and then assign the online network parameters to their corresponding target network parameters, namely $\theta' \leftarrow \theta$ and $\xi' \leftarrow \xi$, where θ' and ξ' are the parameters of Actor and Critic target network respectively. Initialize the experience replay space $\mathbb{R}$ to save the experience data obtained from the exploration interaction. Initialize a random process ε, which is used to explore the action value.

Second, determine the initial state of the training, that is, the relative situation at the beginning of the air combat. Set the initial position state and speed state of each aircraft in the UAV formation and target formation. According to the definition of state space, the initial state $\boldsymbol{s}^1$ of air combat is calculated.

Finally, repeat the multi-episode training according to the initial state, and perform the following operations in each episode air combat simulation. First, based on the target assignment algorithm described in Table 1, the target assignment matrix $\boldsymbol{X}^t$ is calculated according to the current situation $\boldsymbol{s}^t$. Then each UAV$_i$ generates an action value $\boldsymbol{a}_i^t = \boldsymbol{a}_{i,\theta}(\boldsymbol{s}^t) + \varepsilon_t$ based on the state $\boldsymbol{s}^t$ and the random process ε and executes it. At the same time, according to the predefined policy, each Target$_i$ in the target formation executes action $\boldsymbol{b}_j^t$. After the execution of all actions, the state shifts to $\boldsymbol{s}^{t+1}$, and the reward value $[r_i^t]_{i=1}^n$ can be calculated according to (13). The transfer process variables $\left\{\boldsymbol{s}^t, [\boldsymbol{a}_i^t, r_i^t]_{i=1}^n, \boldsymbol{s}^{t+1}\right\}$ are stored in $\mathbb{R}$ as a piece of empirical data. When learning, randomly sample a batch of M pieces of empirical data $\left\{\boldsymbol{s}_m^t, \left[\boldsymbol{a}_{m,i}^t, r_{m,i}^t\right]_{i=1}^n, \boldsymbol{s}_m^{t+1}\right\}_{m=1}^M$ from the experience pool $\mathbb{R}$, to calculate the target Q value of each UAV, that is, for each of the M pieces of data,

$$\hat{Q}_{m,i} = r_{m,i} + \lambda Q_{m,i}^{\xi'}\left(\boldsymbol{s}_m^{t+1}, \boldsymbol{a}_{\theta'}\left(\boldsymbol{s}_m^{t+1}\right)\right). \quad (18)$$

Then calculate the gradient estimate of Critic according to (17), namely

$$\Delta\xi = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} \Big[\Big(\hat{Q}_{m,i} - Q_{m,i}^\xi(\boldsymbol{s}_m^t, \boldsymbol{a}_\theta(\boldsymbol{s}_m^t)) \Big) \cdot \\ \nabla_\xi Q_{m,i}^\xi(\boldsymbol{s}_m^t, \boldsymbol{a}_\theta(\boldsymbol{s}_m^t)) \Big]. \quad (19)$$

Finally, calculate the Actor's gradient estimate according to (16), that is,

$$\Delta\theta = \\ \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\nabla_\theta \boldsymbol{a}_{j,\theta}(\boldsymbol{s}_m^t) \cdot \nabla_{\boldsymbol{a}_j} Q_{m,i}^\xi(\boldsymbol{s}_m^t, \boldsymbol{a}_\theta(\boldsymbol{s}_m^t)) \right]. \quad (20)$$

Based on the obtained gradient estimates $\Delta\xi$ and $\Delta\theta$, the optimizer is used to update the online network parameters of Actor and Critic. After the online network optimization is completed, the target network parameters are updated soft, that is,

$$\begin{cases} \xi' \leftarrow \kappa\xi + (1-\kappa)\xi' \\ \theta' \leftarrow \kappa\theta + (1-\kappa)\theta' \end{cases} \quad (21)$$

where $\kappa \in (0, 1)$.

In summary, the pseudo-code of the multi-UAV cooperative air combat maneuver decision algorithm is shown in Algorithm 2.

Algorithm 2 Multi-UAV cooperative air combat maneuver decision algorithm

```
Initialize the formation size of UAV and target with $n$ and $m$.
Initialize Actor online network and Critic online network with random parameters $\theta$ and $\xi$.
Initialize Actor target network and Critic target network with $\theta' \leftarrow \theta$ and $\xi' \leftarrow \xi$.
Initialize replay buffer $\mathbb{R}$.
Initialize a random process $\varepsilon$ for action exploration.
for episode = 1, $E$ do
    Initialize the initial state of UAVs and targets.
    Receive initial observation state $s^1$.
```

```
  for $t = 1, T$ do
      Execute Algorithm 1 to calculate the target allocation matrix $\boldsymbol{X}$.
      For each UAV$_i$, select and execute action $\boldsymbol{a}_i^t = \boldsymbol{a}_{i,\theta}(s^t) + \varepsilon_t$.
      For each Target$_j$, execute action $\boldsymbol{b}_i^t$ based on selected policy.
      Observe next state $\boldsymbol{s}^{t+1}$, and calculate the return value $[r_i^t]_{i=1}^n$ according to (13).
      Store transition $\left\{\boldsymbol{s}^t, [\boldsymbol{a}_i^t, r_i^t]_{i=1}^n, \boldsymbol{s}^{t+1}\right\}$ in $\mathbb{R}$.
      Sample a random minibatch of $M$ transitions $\left\{\boldsymbol{s}_m^t, \left[\boldsymbol{a}_{m,i}^t, r_{m,i}^t\right]_{i=1}^n, \boldsymbol{s}_m^{t+1}\right\}_{m=1}^M$ from $\mathbb{R}$.
      Calculate the target $Q$ value for each UAV in each transition.
      for $m = 1, M$ do
          $\hat{Q}_{m,i} = r_{m,i} + \lambda Q_{m,i}^{\xi'}(\boldsymbol{s}_m^{t+1}, \boldsymbol{a}_{\theta'}(\boldsymbol{s}_m^{t+1}))$
      end for
      Compute Critic gradient estimation $\Delta\xi$ according to (19).
      Compute Actor gradient estimation $\Delta\theta$ according to (20).
      Update the online networks based on optimizer using $\Delta\theta$ and $\Delta\xi$.
      Update the target networks according to (21).
  end for
end for
```

3.4 Target policy

In the training process of the multi-UAV cooperative air combat maneuver decision model, it is necessary to set the maneuver policy of the target formation to make the targets reflect the confrontation effect of the air combat simulation, so as to prove the autonomous learning ability of the multi-UAV cooperative air combat maneuver decision model and effectiveness of the learned air combat policy.

Because the size of the state space and action space of the multi-UAV cooperative air combat maneuver decision model is linearly increased compared to the single aircraft model, for this large-scale network model, if the policy is directly learned from the confrontation process, a large number of invalid samples will be generated, resulting in low efficiency of reinforcement learning, or even local optimization and learning failure. To solve this problem, this paper adopts the "basic-confrontation" training method proposed in [17].

3.4.1 Basic policy

During the basic training, the target formation maintains its initial motion law, such as performing uniform linear motion or circular motion, and does not change its motion according to changes of combat situation. And the training will be carried out at the beginning of the episode that the UAV formation is at advantage, balance, and disadvantage initial situation, to make the UAV familiar with the situation environment of air combat.

As shown in Fig. 7, when UAV is chasing the target, UAV is at an advantage. On the contrary, when UAV is chased by the target, UAV is at a disadvantage. When UAV and the target are heading towards each other, the two sides are in a balance. When the two sides depart from each other, it means that the two sides are going out of engagement. This state is not conducive to the learning of the maneuver policy. Therefore, the initial situation of departing is not adopted in the basic training process.

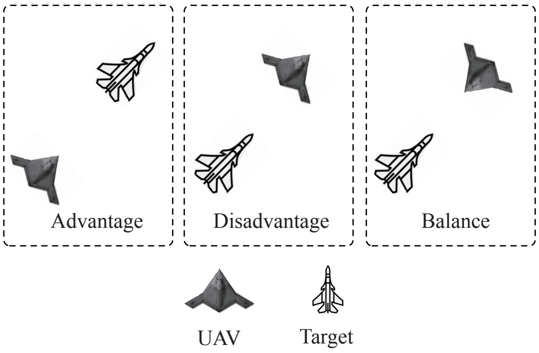

Fig. 7 Advantage, balance, and disadvantage situation of UAV

3.4.2 Confrontation policy

In confrontation training, target formation is required to have corresponding maneuver policy, which generates maneuvers for each target aircraft to fight against UAV formation. The combat policy of target formation includes target assignment and maneuver decision. We design the target assignment algorithm and the maneuver decision algorithm of target formation based on the idea of the greedy algorithm.

(i) Target assignment

Assuming that the target formation does not have a unified command, each aircraft in the formation preferentially selects the opponent aircraft closest to it as the attacking target during the combat process.

(ii) Maneuver decision

In the confrontation simulation, the target and UAV formations are both homogeneous formations, and the performance of the aircraft is the same. The target aircraft uses the same motion model as UAV. From (i) and (ii), it can be known that the control variables of the aircraft are $[n_x, n_z, \mu]$, and a group of control variables represents a maneuver. According to common air combat maneuver methods, NASA scholars have designed seven typical maneuvers [24], and any complex tactical maneuver can be composed of these basic maneuvers. In this paper, the maneuver library established in [17] is set as the target aircraft's action space. As shown in Fig. 8, the aircraft can maneuver forward, left, right, up, and down in five directions, each direction with constant speed, acceleration, and deceleration control. The maneuver library contains 15 actions.

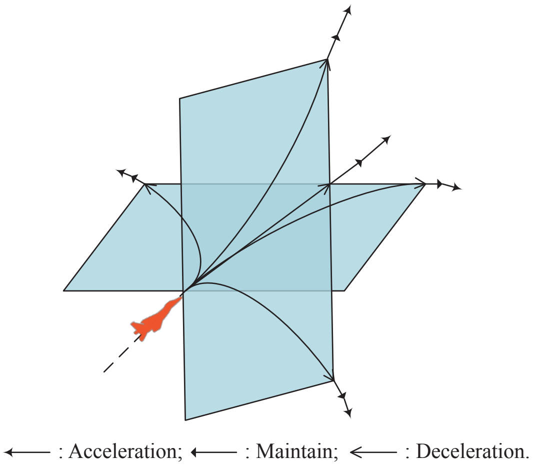

Fig. 8 Maneuver library of target

Based on the established maneuver library, the maneuver policy of the target in this paper adopts a maneuver decision method based on statistical principles [9].

The maneuver decision method based on statistical principles adopts four parameters of azimuth α, distance D, speed v, and altitude h to characterize the current air combat situation and define the membership functions η_α, η_D, η_v and η_h of each parameter separately. The membership functions not only enhance the robustness of situation description but also normalize the output to the interval [0,1]. When the membership functions are gradually approaching 1, the aircraft is at an advantage. On the contrary, when the membership function is approaching 0, the aircraft is at a disadvantage. Based on the membership functions, the process of selecting the optimal maneuver in the maneuver library is as follows.

First, based on the current air combat situation, the control values of each action in the maneuver library are sent to the motion model in turn, and maneuver trials are conducted one by one.

Second, through the previous step, all possible positions of the aircraft at the next moment are obtained, and the situation of each position can be solved to obtain a set of membership functions corresponding to all maneuvers.

Finally, calculate the mean and standard deviations of the four membership values corresponding to each maneuver. Then each maneuver corresponds to a binary group consisting of the mean and standard deviations, and select the largest expected element from all the binary groups, and take the corresponding maneuver as the forthcoming action. If the maximum number of means is greater than 1, the maneuver corresponding to the element with the smallest standard deviation among these elements is taken as the output of the upcoming action.

4. Simulation and analysis

4.1 Platform setting

4.1.1 Hardware

In this paper, the air combat environment model is established by using Python language, and the network model is built based on the Tensorflow module. The multi-UAV cooperative air combat maneuver decision model runs on one computer. The computer has an Intel(R) Core(TM) i7-8700k CPU and 16GB RAM. On this basis, a NVIDIA GeForce GTX 1 080 TI graphics card is also installed for Tensorflow acceleration.

4.1.2 Parameter setting

The air combat background of multi-UAV cooperative air combat is set as short-range air combat, and the parameters of the air combat environment model are set as follows. The farthest interception distance of the missile is $D_{\max}$=3 km, the angle of view is φ_m=π/4, the minimum safety distance between the two aircraft is D_{safe}=200 m, the advantage value when intercepting the target is Re=5, and the penalty value is P=10. In the aircraft motion model, set the maximum speed $v_{\max}$=400 m/s, the minimum speed $v_{\min}$=90 m/s. For the control space, set $n_x \in [-1,2]$, $n_z \in [0,8]$, $\mu \in [-\pi,\pi]$. According to the control space, the maneuver library of the target is shown in Table 1.

Table 1 Maneuver library of target

Number	Maneuver	Control value		
		n_x	n_z	μ
1	Forward maintain	0	1	0
2	Forward accelerate	2	1	0
3	Forward decelerate	−1	0	0
4	Left turn maintain	0	8	−arc cos (1/8)
5	Left turn accelerate	2	8	−arc cos (1/8)
6	Left turn decelerate	−1	8	−arc cos (1/8)
7	Right turn maintain	0	8	arc cos (1/8)
8	Right turn accelerate	2	8	arc cos (1/8)
9	Right turn decelerate	−1	8	arc cos (1/8)
10	Upward maintain	0	8	0
11	Upward accelerate	2	8	0
12	Upward decelerate	−1	8	0
13	Downward maintain	0	8	π
14	Downward accelerate	2	8	π
15	Downward decelerate	−1	8	π

The Actor network of the maneuver decision model is divided into three parts: the input layer, the hidden layer, and the output layer, where the input layer inputs the air combat state. The hidden layer is divided into two layers. The first layer is composed of 400 long short-term memory (LSTM) units in the forward direction and the reverse direction. This layer is expanded according to the number of UAVs according to the bidirectional recurrent neural network structure to form a communication layer. The second layer consists of 100 units, which is the tanh activation function, and the parameters are randomly initialized with uniform distribution $\left[-3\times10^{-4},3\times10^{-4}\right]$. The output layer outputs three control values, and the parameters are randomly initialized with uniform distribution $\left[-2\times10^{-5},2\times10^{-5}\right]$. Through linear adjustment, three output ranges are adjusted from [0,1] to [1,2], [0,8], and $[-\pi,\pi]$, respectively.

The Critic network is also divided into three parts: the input layer, the hidden layer, and the output layer. The input layer inputs the air combat state and the action value. The hidden layer is divided into two layers. The first layer is composed of 500 LSTM units in the forward direction and the reverse direction. This layer is expanded according to the number of UAVs according to the bidirectional recurrent neural network structure to form a communication layer. The second layer consists of 150 units, which is the tanh activation function, and the parameters are randomly initialized with uniform distribution $\left[-3\times10^{-4},3\times10^{-4}\right]$. The output layer outputs Q value, and the parameters are randomly initialized with uniform distribution $\left[-2\times10^{-4},2\times10^{-4}\right]$. Both the Actor and Cirtic models use the Adam optimizer, the learning rate of the Actor network is set to 0.001, and the learning rate of the Critic network is set to 0.000 1. The discount factor λ = 0.95, and the soft update factor of the target network κ= 0.005. The Ornstein-Uhlenbeck (OU) process is selected as the random process of action value exploration. The size of the experience replay space $\mathbb{R}$ is set to 10^6, and the size of the batch is set to 512.

4.2 Model training and testing

The large-scale recurrent neural network has many parameters and requires a long training period. In order to verify the effectiveness of the self-learning ability of the established model and the effectiveness of the learned policies, 2v1 and 2v2 air combat simulation training are carried out. In each scenario, basic training and confrontation training are carried out successively. Through simulation comparison, the advantages of the self-learned policy over the target policy are verified.

4.2.1 Simulations for 2v1 scene

(i) Basic training

In the basic training of two UAVs fighting against one target scenario, the target adopts uniform linear motion, and the training is carried out in sequence when the initial UAV situation is at an advantage, a balance, and a disadvantage. Through these three training items, the UAVs can be familiar with the air combat environment. There are 10^6 episodes per basic training, and an evaluation episode will be performed every 3 000 episodes. In the evaluation episode, the random process ε is not performed to add noise to the action value, and the online Actor network directly outputs the action value of each UAV. After performing evaluation episode, record the episode reward values to evaluate the previously learned maneuver policy.

In each training process, in order to make the UAVs fully familiar with the air combat environment, and improve the diversity of the samples, so that to prevent the network from overfitting, and making the learning policy can be more generalized, the initial states of UAVs and target in the training episode are randomly generated in a large range. While in order to ensure the uniformity of the evaluation of the maneuver policy, the constant initial situation is used in the evaluation episode. For example, for the first training, that is, when UAV is in an advantage initial position, the initial state of training episode and evaluating episode is shown in Table 2.

Table 2 Advantage initial state setting for 2v1 basic training

Initial state		x/m	y/m	z/m	v/(m/s)	γ/(°)	ψ/(°)
Training episode	UAV1	[−200, 200]	[−300, 300]	3 000	200	0	[−60, 60]
	UAV2	[2 500, 3 500]	[−500, 500]	3 500	200	0	[−60, 60]
	Target	[2 500, 3 500]	[2 500, 3 500]	[2 800, 3 800]	[150, 300]	0	[−60, 60]
Evaluation episode	UAV1	0	0	3 000	200	0	40
	UAV2	3 000	0	3 500	200	0	40
	Target	3 000	3 000	3 000	220	0	45

Fig. 9 shows the simulated maneuver trajectory of air combat based on the learned policy after the basic training of Item 1. It can be seen from the figure that UAV1 and UAV2 start chasing the target from the rear of both sides of the target, continuously adjust the course and speed, gradually reduce the distance from the target, so that two UAVs maintain the tail-chasing situation to the target, making the target crosswise surrounded from both sides, and the target always in two UAVs' interception area.

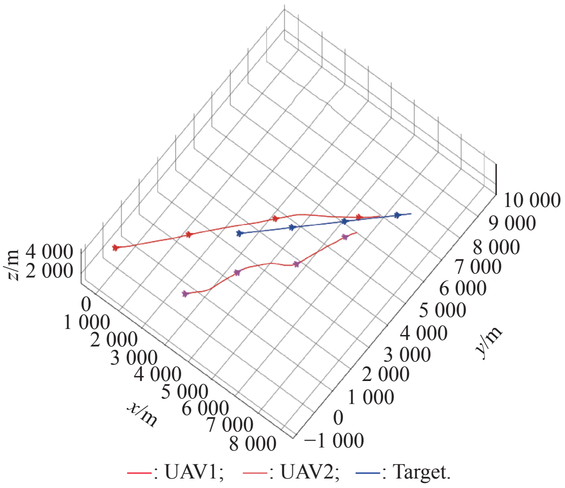

Fig. 9 2v1 maneuver trajectory after basic training of advantage initial state

Fig. 10 shows the simulated maneuver trajectory of air combat based on the learned policy after the training of Item 2. As can be seen from the figure, at the initial moment, both sides are in balance, UAV1 and UAV2 fly toward the target, and then UAV1 and UAV1 adjust their height while continuously approaching the target. After meeting with the target, UAVs turn to both sides of the target and begin to chase the target. Due to the limitation of the turning radius, UAV1 and UAV2 turn to the other side of the target's movement direction, and then continue to adjust the course and speed, gradually narrowing the distance to the target and achieving a tail-chasing situation that encircles the target from both left and right, this process realizes the cross attack tactics [19], that is, the target is always within the monitoring and attack range of two UAVs.

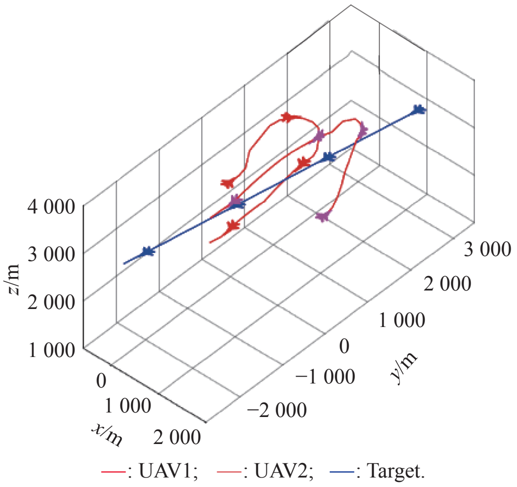

Fig. 10 2v1 maneuver trajectory after basic training of balance initial state

Fig. 11 shows the simulated maneuver trajectory of air combat based on the learned strategy after the training of Item 3. It can be seen from the figure that UAV1 and UAV2 are at a disadvantage situation relative to the target at the initial moment. In order to get rid of the disadvantageous situation of being chased by the target, UAV1 and UAV2 quickly change the flight direction, so that the target cannot form an interception condition, that is, the situation is changed to a balanced situation, and then the course, altitude, and speed are constantly adjusted, and the tail-chasing situation is finally achieved. The whole process described above achieves the conversion of disadvantage-balance-advantage, which proves that the established model can enable UAVs to learn the cooperative air combat maneuver policy to obtain advantages through maneuvering under any situation.

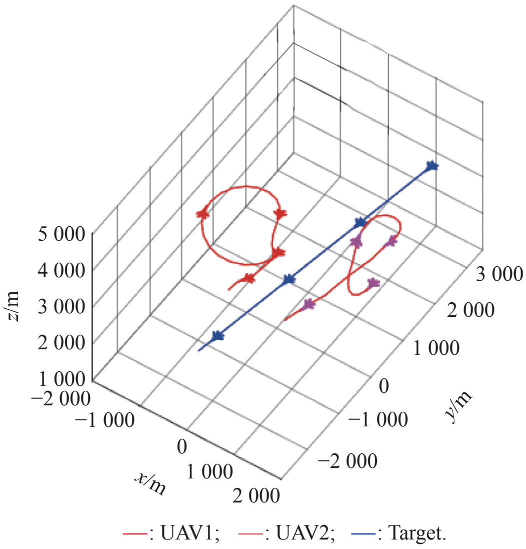

Fig. 11 2v1 maneuver trajectory after basic training of disadvantage initial state

(ii) Confrontation training

After completing the above basic training, UAVs with the basic policy continue to conduct confrontation training with the target with the maneuver policy introduced in Subsection 3.4.2. In order to ensure the diversity of air combat states and the generalization of the maneuver policy, the initial states of UAVs and target are randomly generated within a certain range during training episodes. The training effect will be explained by taking the balance initial state as an example. Table 3 shows the initial state of training in the case of the balance initial state.

Fig. 12 shows the maneuver trajectory of both sides in an episode after confrontation training of the balance initial state. The two sides start heading from the initial position. The target selects the nearest UAV1 as its attack target and flies towards it. UAV2 flies in formation on the right side of UAV1 and adjusts the course to reduce the distance to the target. UAV2 gradually realizes the tail-chasing to the target while turning to the left. On the contrary, the target intends to turn right into the tail of UAV1, but during the right turn, it is at a disadvantage because of the tail-chasing by UAV2. And UAV1 also adjusts the course to achieve the goal of chasing the target. In this case, the greedy algorithm executed by the target selects the optimal action to climb and accelerate to get rid of the attack, and UAV1 and UAV2 follow closely behind, keeping the advantage. Although in the end, the distance between UAVs and the target is widened, and the response of UAVs to the target's acceleration behavior is not good, from the overall situation, the learned policy can enable the two UAVs formation to obtain advantage during the battle against the target with the maneuver policy.

Table 3 Balance initial state setting for 2v1 confrontation training

Initial state		x/m	y/m	z/m	v/(m/s)	γ/(°)	ψ/(°)
Training episode	UAV1	[−200, 200]	[−300, 300]	3 000	200	0	[−60, 60]
	UAV2	[1 500, 2 500]	[−500, 500]	3 500	200	0	[−60, 60]
	Target	[1 000, 2 000]	[2 500, 3 500]	[2 800, 3 800]	[150, 300]	0	[120, 240]
Evaluation episode	UAV1	0	0	3 000	200	0	0
	UAV2	2 000	0	3 500	200	0	0
	Target	1 000	3 000	3 000	220	0	200

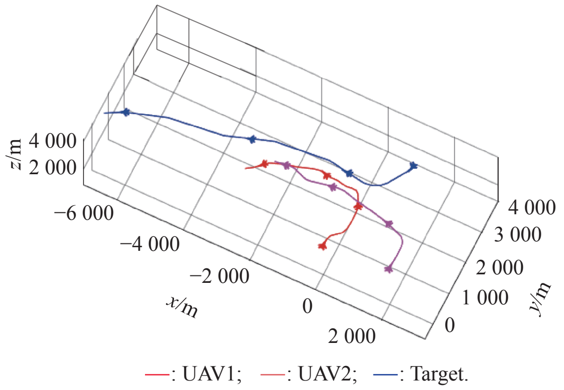

Fig. 12 2v1 maneuver trajectory after confrontation training of balance initial state

4.2.2 Simulations for 2v2 scene

(i) Basic training

In the basic training of two UAVs fighting against the two targets scenario, the targets adopt uniform linear motion, and the training is carried out in sequence when the initial UAV situation is at an advantage, a balance, and a disadvantage. Through these three training items, the UAVs can be familiar with the air combat environment. Due to the expansion of the air combat environment, the state space increases, so compared to the 2v1 scene, each training item increases to 3×10^6 episodes.

The following will take the balance initial state as an example to illustrate the training process and the training effect. The initial state of training episodes and evaluation episodes are shown in Table 4.

Table 4 Balance initial state setting for 2v2 basic training

Initial State		x/m	y/m	z/m	v/(m/s)	γ/(°)	ψ/(°)
Training episode	UAV1	[−200, 200]	[−300, 300]	3 000	200	0	[10, 70]
	UAV2	[2 800, 3 200]	[−300, 300]	3 200	200	0	[10, 70]
	Target1	[2 500, 3 500]	[2 500, 3 500]	[2 900, 3 100]	[180, 220]	0	[−165, −105]
	Target2	[5 500, 6 500]	[2 500, 3 500]	[2 900, 3 100]	[180, 220]	0	[−165, −105]
Evaluation episode	UAV1	0	0	3 000	200	0	40
	UAV2	3 000	0	3 200	200	0	40
	Target1	3 000	3 000	3 000	200	0	−135
	Target2	6 000	3 000	3 000	200	0	−135

Fig. 13 is the air combat simulation maneuver trajectory based on the learned policy after finishing the basic training in the balance initial state. It can be seen from the figure that, at the initial moment, UAV1 and UAV2 fly toward Target1 and Target2, respectively. According to the target assignment algorithm, UAV1 and UAV2 select

Target1 and Target2 as attack targets for maneuvering, respectively, and adjust the course and altitude when approaching their respective targets to avoid possible collisions in the intersection. During meeting with the target, UAV1 turns to the right and UAV2 turns to the left, realizing a cross cover. Instead of continuing to turn around to pursue their initial assigned targets, the two UAVs exchange their respective attack targets after turning, which reflects the tactical cooperation. It proves that after reinforcement learning, UAVs can learn the air combat maneuver policy to realize the tactical cooperation, and gain advantages in air combat, rather than decomposing the multi-aircraft air combat into multiple 1v1 confrontations.

(ii) Confrontation training

After completing the above basic training, UAVs with the basic policy continue to conduct confrontation training with targets with the maneuver policy introduced in Subsection 3.4.2. In order to ensure the diversity of air combat states and the generalization of the maneuver policy, the initial states of UAVs and targets are randomly generated within a certain range during training episodes. The training effect will be explained by taking the balance initial state as an example. Table 5 shows the initial state of training in the case of the balance initial state. In order to improve the training speed, compared with the basic training, the value range of the initial position in the confrontation training is narrowed.

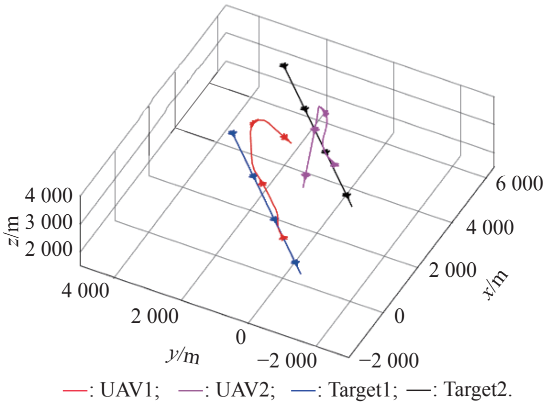

Fig. 13 2v2 maneuver trajectory after basic training of balance initial state

Table 5 Balance initial state setting for 2v2 confrontation training

Initial state		x/m	y/m	z/m	v/(m/s)	γ/(°)	ψ/(°)
Training episode	UAV1	[−200, 200]	[−200, 200]	3 000	200	0	[20, 60]
	UAV2	[2 800, 3 200]	[−200, 200]	3 200	200	0	[20, 60]
	Target1	[2 500, 3 500]	[2 800, 3 200]	[2 900, 3 100]	[180, 220]	0	[−155, −115]
	Target2	[5 500, 6 500]	[2 800, 3 200]	[2 900, 3 100]	[180, 220]	0	[−125, −115]
Evaluation episode	UAV1	0	0	3 000	200	0	40
	UAV2	3 000	0	3 200	200	0	40
	Target1	3 000	3 000	3 000	200	0	−135
	Target2	6 000	3 000	3 000	200	0	−135

Fig. 14 shows the maneuver trajectory of both sides in the evaluation episode after the confrontation training of the balance initial state. The two sides start heading from the initial position. Target1 and Target2 choose the closest UAV2 as their attack target and fly towards it. UAV1 flies in a formation on the left side of UAV2 and adjusts the course to reduce the distance to the target. In the process of UAV2 meeting Target1 and turning left, UAV1 adjusts the course to the right, gradually faces the tail of Target1, and covers UAV2 from the side and the rear. At the same time, UAV1 changes the attack target from Target2 to Target1, and gradually achieves the advantage to Target1. On the other hand, UAV2 adjusts the course and speed after meeting Target1 to prevent Target2 from entering the tail, and finally realize the tail-chasing situation to Target2. Throughout the process, UAVs have achieved tactical cooperation such as covering and alternating attack targets.

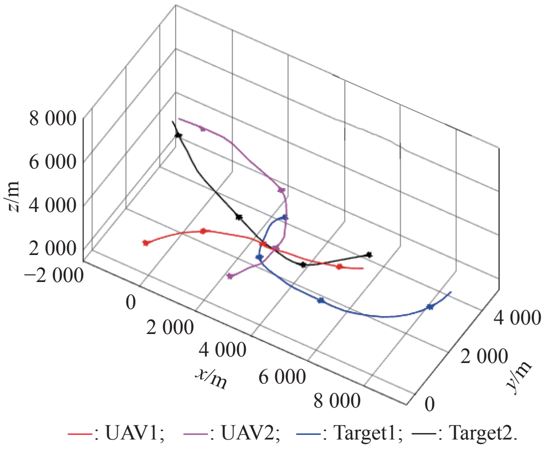

Fig. 14 2v2 maneuver trajectory after confrontation training of balance initial state

Based on the above simulations, it is proved that the policy obtained by the self-learning of the decision model can enable the multi-UAV formation to obtain advantages during the combat against targets with the maneuver policy under the condition of equal or superior strength, to achieve coordinative air combat and win the victory.

4.2.3 Target assignment performance

In order to study the training effect of the target assignment algorithm proposed in this paper on the maneuver decision model, the target assignment performance test simulation is carried out.

The reinforcement learning model using the target assignment algorithm proposed in this paper is marked as the maneuver decision Model 1, while the reinforcement learning model which uses the target assignment algorithm of the confrontation policy is marked as the maneuver decision Model 2. Model 1 and Model 2 are trained 3×10^6 episodes with the confrontation policy respectively under the balance initial state. The training episode initial state settings are shown in Table 6.

Table 6 Balance initial state setting for target assignment performance testing

Initial state		x/m	y/m	z/m	v/(m/s)	γ/(°)	ψ/(°)
Training episode	UAV1	[−200, 200]	[−200, 200]	3000	200	0	[20, 60]
	UAV2	[2800, 3200]	[−200, 200]	3200	200	0	[20, 60]
	Target1	[2500, 3500]	[2800, 3200]	[2900, 3100]	[180, 220]	0	[−155, −115]
	Target2	[5500, 6500]	[2800, 3200]	[2900, 3100]	[180, 220]	0	[−125, −115]
Evaluation episode	UAV1	[−100, 100]	[−100, 100]	3000	200	0	[35, 45]
	UAV2	[2900, 3100]	[−100, 100]	3200	200	0	[35, 45]
	Target1	[2500, 3500]	[2800, 3200]	[2900, 3100]	[180, 220]	0	[−155, −115]
	Target2	[5500, 6500]	[2800, 3200]	[2900, 3100]	[180, 220]	0	[−125, −115]

After completing the reinforcement learning training, Model 1 and Model 2 are evaluated with the confrontation policy for 500 times respectively. The initial state of the confrontation test is randomly selected within a certain area, as evaluation episode setting shown in Table 6. The results of the confrontation test are expressed in terms of winning percentage, and the winning conditions are shown in Table 7.

Table 7 Conditions of air combat result

Result	Condition
Win	All targets are shot down
Draw	Number of remaining targets and drones are equal
Lose	All UAVs are shot down

The test results of Model 1 and Model 2 after training are shown in Fig. 15. It can be seen from Fig. 15 that the winning percentage of Model 1 is much higher than that of Model 2 under the conditions that the network structure, training process, and target policy are the same. This fully demonstrates that the target assignment algorithm proposed in this paper can effectively guide the individual reinforcement learning process to form a formation coordination policy, and give full play to the situational advantage in coordinated air combat.

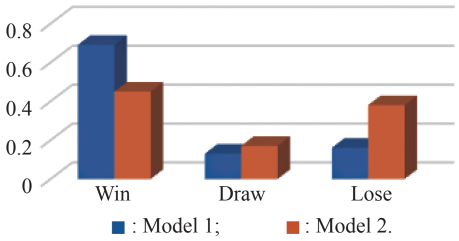

Fig. 15 Result of the confrontation test

5. Conclusions

In this paper, a multi-UAV cooperative air combat maneuver decision model based on reinforcement learning and recurrent neural network is built. First of all, based on the single-UAV air combat environment model, an environment model of multi-UAV air combat is established, and the state space, action space, and situation evaluation model for individual maneuver decision in multi-aircraft cooperative air combat are designed. At the same time, according to the characteristics of multi-target attacking, a multi-aircraft cooperative air combat target assignment method is designed based on the Hungarian algorithm, and based on the air combat situation evaluation value, the reward calculation method of individual UAV is designed. Then, based on the single-UAV Actor-Critic air combat maneuver decision-reinforcement learning model, a bidirectional recurrent neural network is used as the communication network between UAV individuals, and the individual UAVs are connected to form a collaborative decision-making network of the formation. The multi-UAV cooperative air combat maneuver decision model realizes the unity of UAV individual behavior learning and the combat objectives of the formation.

However, due to the constraints of training time and equipment resources, this paper does not carry out some more detailed simulation analysis. For example, there is no simulation verification in a larger-scale force scenario. In addition, in this paper, a 3-DOF aircraft motion model is used to establish the air combat environment model. In subsequent research, a 6-DOF motion model can be used to improve model accuracy. At the same time, a detailed sensor model can be added and the attack zone model can be refined to build an air combat environment with incompletely observable target information. Carry out air combat decision-making research in the context of being closer to real air combat.

References

[1] ZHOU K, WEI R, XU Z, et al. An air combat decision learning system based on a brain-like cognitive mechanism. Cognitive Computation, 2020, 12(1): 128–139.

[2] YANG Q M, ZHANG J D, SHI G Q. Modeling of UAV path planning based on IMM under POMDP framework. Journal of Systems Engineering and Electronics, 2019, 30(3): 545–554.

[3] MCGREW J S, HOW J P, WILLIAMS B, et al. Air-combat strategy using approximate dynamic programming. Journal of Guidance, Control, and Dynamics, 2010, 33(5): 1641–1654.

[4] ZHOU K, WEI R X, XU Z F, et al. A brain like air combat learning system inspired by human learning mechanism. Proc. of IEEE/CSAA Guidance, Navigation and Control Conference, 2018: 286–293.

[5] XU G, WEI S, ZHANG H. Application of situation function in air combat differential games. Proc. of the 36th Chinese Control Conference, 2017: 5865–5870.

[6] PARK H, LEE B Y, TAHK M J, et al. Differential game based air combat maneuver generation using scoring function matrix. International Journal of Aeronautical & Space Sciences, 2015, 17(2): 204–213.

[7] SMITH R E, DIKE B A, MEHRA R K, et al. Classifier systems in combat: two-sided learning of maneuvers for advanced fighter aircraft. Computer Methods in Applied Mechanics & Engineering, 2000, 186(2): 421–437.

[8] HANG C Q, DONG K S, HUANG H Q, et al. Autonomous air combat maneuver decision using Bayesian inference and moving horizon optimization. Journal of Systems Engineering and Electronics, 2018, 29(1): 86–97.

[9] GUO H F, HOU M Y, ZHANG Q J, et al. UCAV robust maneuver decision based on statistics principle. Acta Armamentarii, 2017, 38(1): 160–167. (in Chinese)

[10] FU L, XIE H. An UAV air-combat decision expert system based on receding horizon control. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(11): 1994–1999. (in Chinese)

[11] ROGER W S, ALAN E B. Neural network models of air combat maneuvering. Las Cruces, U.S.: New Mexico State University, 1992.

[12] DING L J, YANG Q M. Research on air combat maneuver decision of UAVs based on reinforcement learning. Avionics Technology, 2018, 49(2): 29–35. (in Chinese)

[13] LIU P, MA Y. A deep reinforcement learning based intelligent decision method for UCAV air combat. Proc. of Asian Simulation Conference, 2017: 274–286.

[14] ZUO J L, YANG R N, ZHANG Y, et al. Intelligent decision-making in air combat maneuvering based on heuristic reinforcement learning. Acta Aeronautica et Astronautica Sinica, 2017, 38(10): 217–230. (in Chinese)

[15] ZHANG X B, LIU G Q, YANG C J, et al. Research on air confrontation maneuver decision-making method based on reinforcement learning. Electronics, 2018, 7(11): 279.

[16] YANG Q M, ZHU Y, ZHANG J D, et al. UAV air combat autonomous maneuver decision based on DDPG algorithm. Proc. of the IEEE 15th International Conference on Control and Automation, 2019: 37−42.

[17] YANG Q M, ZHANG J D, SHI G Q, et al. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning. IEEE Access, 2020, 8: 363–378.

[18] WAN K F, GAO X G. Robust motion control for UAV in dynamic uncertain environments using deep reinforcement learning. Remote Sensing, 2020, 12: 640.

[19] ROBERT L S. Fighter combat—tactics and maneuvering. Maryland: Naval Institute Press, 1985.

[20] XI Z F, XU A, KOU Y X, et al. Decision process of multi-aircraft cooperative air combat maneuver. Systems Engineering and Electronics, 2020, 42(2): 381–389. (in Chinese)

[21] LI J X, TONG M A, JIN D K. Bargaining differential game theory and application to multiple-airplane combat analysis. Systems Engineering-Theory & Practice, 1997, 6(6): 68–72. (in Chinese)

[22] WANG Y N, JIANG Y X. An intelligent differential game on air combat decision. Flight Dynamics, 2003, 21(1): 66–70. (in Chinese)

[23] ZUO J L, ZHANG Y, YANG R N, et al. Reconstruction and evaluation of medium-rang cooperation air combat decision-making process with two phase clustering. Systems Engineering and Electronics, 2020, 42(1): 108–117. (in Chinese)

[24] XIE R Z, LI J Y, LUO D L. Research on maneuvering decisions for multi-UAVs Air combat. Proc. of the 11th IEEE International Conference on Control & Automation, 2014: 767–772.

[25] LUO D L, SHEN C L, WANG B, et al. Air combat decision-making for cooperative multiple target attack using heuristic adaptive genetic algorithm. Proc. of the International Conference on Machine Learning and Cybernetics, 2005: 473–478.

[26] SU M C, LAI S C, LIN S C, et al. A new approach to multi-aircraft air combat assignments. Swarm and Evolutionary Computation, 2012, 6: 39–46.

[27] WANG Y, ZHANG W, LI Y. An efficient clonal selection algorithm to solve dynamic weapon-target assignment game model in UAV cooperative aerial combat. Proc. of the 35th Chinese Control Conference, 2016: 9578–9581.

[28] TAL S, STEVE R, DAVE G. Assigning micro UAVs to task tours in an urban terrain. IEEE Trans. on Control Systems Technology, 2007, 15(4): 601–612.

[29] . PENG P, WEN Y, YANG Y, et al. Multiagent bidirectionally-coordinated nets: emergence of human-level coordination in learning to play star craft combat games. arXiv preprint arXiv: 1703.10069v4, 2017.

[30] SILVER D, LEVER G, HEESS N, et. al. Deterministic policy gradient algorithms. Proc. of the 31st International Conference on Machine Learning, 2014: 605–619.

[31] SUTTON R, MCALLESTER D, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation. Proc. of the 13th Annual Neural Information Processing Systems Conference, 1999: 1057–1063.

Biographies

ZHANG Jiandong was born in 1974. He is an associate professor at the Department of System and Control Engineering in Northwestern Polytechnical University, China. He received both his M.S. and Ph.D. degrees in system engineering from the same university. His research interests include modeling simulation and effectiveness evaluation of complex systems, development and design of integrated avionics system, and system measurement & test technologies.
E-mail: jdzhang@nwpu.edu.cn

YANG Qiming was born in 1988. He received his master degree from Northwestern Polytechnical University (NPU), Xi’an, China in 2013. He was awarded with a Ph.D. degree in electronic science and technology in 2020. He is an assistant researcher of the NPU. His main research interests are artificial intelligence and its application on control and decision of UAV.
E-mail: yangqm@nwpu.edu.cn

SHI Guoqing was born in 1974. He is an associate professor at the Department of System and Control Engineering in Northwestern Polytechnical University, China. He received his M.S. and Ph.D. degrees in system engineering from the same university. His research interests include integrated avionics system measurement & test technologies, development and design of embedded real-time systems, modeling simulation and effectiveness evaluation of complex systems, etc.
E-mail: shiguoqing@nwpu.edu.cn

LU Yi was born in 1975. He graduated from Nanjing University of Aeronautics and Astronautics in 1998, majoring in aircraft guidance control and simulation. He is currently the deputy chief designer of Shenyang Aircraft Design Institute, and mainly engaged in fighter avionics system design work.
E-mail: yiluemail@126.com

WU Yong was born in 1964. He is a professor at the Department of System and Control Engineering in Northwestern Polytechnical University, China. He received his M.S. degree in system fire control from the same university in 1988. His research interests include integrated avionics system measurement & test technologies, development and design of embedded real-time systems, modeling simulation and effectiveness evaluation of complex systems, etc.
E-mail: yongwu@nwpu.edu.cn