IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 4, 2019, accepted December 19, 2019, date of publication December 23, 2019, date of current version January 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961426

Maneuver Decision of UAV in Short-Range Air Combat Based on Deep Reinforcement Learning

QIMING YANG[1], JIANDONG ZHANG[1], GUOQING SHI[1], JINWEN HU[2], AND YONG WU[1]
[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China
[2]School of Automation, Northwestern Polytechnical University, Xi'an 710129, China

Corresponding author: Jiandong Zhang (jdzhang@nwpu.edu.cn)

This work was supported in part by the Aeronautical Science Foundation of China under Grant 2017ZC53033, in part by the National Natural Science Foundation of China under Grant 61603303 and Grant 61803309, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JQ6070, and in part by the China Postdoctoral Science Foundation under Grant 2018M633574.

ABSTRACT With the development of artificial intelligence and integrated sensor technologies, unmanned aerial vehicles (UAVs) are more and more applied in the air combats. A bottleneck that constrains the capability of UAVs against manned vehicles is the autonomous maneuver decision, which is a very challenging problem in the short-range air combat undergoing highly dynamic and uncertain maneuvers of enemies. In this paper, an autonomous maneuver decision model is proposed for the UAV short-range air combat based on reinforcement learning, which mainly includes the aircraft motion model, one-to-one short-range air combat evaluation model and the maneuver decision model based on deep Q network (DQN). However, such model includes a high dimensional state and action space which requires huge computation load for DQN training using traditional methods. Then, a phased training method, called ''basic-confrontation'', which is based on the idea that human beings gradually learn from simple to complex is proposed to help reduce the training time while getting suboptimal but efficient results. Finally, one-to-one short-range air combats are simulated under different target maneuver policies. Simulation results show that the proposed maneuver decision model and training method can help the UAV achieve autonomous decision in the air combats and obtain an effective decision policy to defeat the opponent.

INDEX TERMS Deep reinforcement learning, maneuver decision, independent decision, deep Q network, network training.

NOMENCLATURE

$\boldsymbol{p}_U$	position vector of UAV, U indicate UAV
$\dot{\boldsymbol{p}}_\mathrm{T}$	velocity vector of Target,T indicate Target
α_U	angle between $\dot{\boldsymbol{p}}_\mathrm{U}$ and $(\boldsymbol{p}_\mathrm{T}-\boldsymbol{p}_\mathrm{U})$
α_T	angle between $\dot{\boldsymbol{p}}_\mathrm{T}$ and $(\boldsymbol{p}_\mathrm{T}-\boldsymbol{p}_\mathrm{U})$
n_x	overload in the velocity direction
n_z	normal overload
μ	roll angle around the velocity vector
η_A	situation advantage
η_R	distance advantage
v	speed
ψ_T	heading angle of target
γ_U	track angle of UAV
γ	reward discount factor
D	distance
θ	parameters of Q network
s_t	state at time t
r_t	reward at time t
a_t	action at time t
$\boldsymbol{a}_i$	control values of ith action of maneuver library

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

I. INTRODUCTION

Compared with manned aircraft, military UAVs have attracted much attention for their low cost, long flight time and fearless sacrifice. With the development of sensor technology, computer technology and communication technology, the performance of military UAVs is evolved significantly, and the range of executable tasks continues to expand [1]. Although the military UAV can perform reconnaissance and ground attack missions [2], most of the mission functions are inseparable from human intervention, and ultimately the decision is made by the control personnel in ground station. This ground-based remote operation mode

is mainly dependent on the data link which is vulnerable to weather and electromagnetic interference. So the traditional ground-based remote operation is difficult to command UAV to conduct air combat due to the difficulty of adapting to fast and varied air combat scenarios [3].Therefore, to let UAV automatically make control decisions according to the situation faced, and to realize UAV independent air combat is a major research direction of UAV intelligence.

The autonomous maneuver decision in short-range air combat is the most challenging application direction, because the two sides in the short-range air combat perform the most violent maneuvers, making the situation change very quickly. Autonomous maneuver decision requires automatically generating flight control commands under various air combat situations based on technical methods such as mathematical optimization and artificial intelligence.

The methods of autonomous maneuver decision can be roughly divided into three categories: game theory [4]–[7], optimization method [8]–[14] and artificial intelligence [15]–[20]. Representative methods based on game theory include differential games [4], [5] and influence diagrams [7]. The model established by this kind of method can directly reflect the various factors of air combat confrontation, but it is difficult to solve the model with complex decision set due to the limitation of real-time calculation [7].

The maneuver decision based on optimization theory includes genetic algorithm [9], Bayesian inference [10], statistical theory [14], etc. This kind of method is to transform the maneuvering decision problem into the optimization problem. By solving the optimization model, the maneuver policy can be obtained. However, many of these optimization algorithms have poor real-time performance on large-scale problems and cannot implement online decision making for air combat. Therefore, they can only be used for offline air combat tactics optimization research [9].

Artificial intelligence-based methods mainly include expert system method [16], neural network method [17] and reinforcement learning method [18]–[21]. The core of expert system method is to refine the flight experience into a rule base and generate flight control commands through rules. The problem with this method is that the rule base is too complex to build, and the policy in rule base is simple and fixed, so it is easy to be cracked by the opponent. The core of the neural network method is to use neural networks to store maneuver rules. The neural network maneuver decision is robust and can be continuously learned from a large number of air combat samples. However, the production of air combat samples also requires a lot of artificial work to complete, so this method faces the problem of insufficient learning samples. Compared with expert system method and neural network method, reinforcement learning does not require labeled learning samples. The agent updates the action policy by interacting with the external environment autonomously [22]. This method better realizes the combination of online real-time decision-making and self-learning. It is an effective method to solve the problem of sequential decision-making without a priori model.

Many scholars have carried out research on maneuver decision-making based on the idea of reinforcement learning. In [3], the approximate dynamic programming method is used to construct the maneuver decision model, and the flight experiment is carried out. It is verified that the UAV autonomous maneuver decision can be realized based on the idea of reinforcement learning. However, in the paper, the UAV is assumed to move in a 2D plane, and the actual situation of the aircraft moving in 3D space in air combat is not considered. In [18], the fuzzy logic method is used to divide the state space of the air combat environment, and the linear approximation method is used to approximate the Q value. However, with the continuous subdivision of the state space, the dramatic increase in the number of states leads to a decline in the efficiency of reinforcement learning, and learning tends to fall into the dimension explosion and leads to failure. In [19], [20], the deep reinforcement learning algorithm is used to construct the maneuver decision model of air combat, but the speed is not set as the decision variable in the model, and the speeds of both sides are set to constant values, which is inconsistent with the actuality of air combat. In [21], the DQN algorithm is used to construct the maneuver decision model of over-the-horizon air combat. The speed control is considered in the model. However, as in [3], the model still assumes that both sides move in the same 2D plane, without considering the influence of altitude changes on air combat. None of these models have fully and realistically reflected the air combat model, especially the characteristics of short-range air combat.

In this paper, based on reinforcement learning, the UAV short-range air combat autonomous maneuver decision modeling is carried out. Firstly, a second-order aircraft motion model and one-to-one short-range air combat model in 3D space are established in succession. And the evaluation model of air combat advantage is proposed by combining two factors of situation and distance. The model can quantitatively reflect the advantages of the aircraft in any situation of short-range air combat. Secondly, the maneuver decision model is constructed under the framework of DQN, and a new maneuver library is designed. The 7 classic maneuver actions are extended to 15, which improves the action space of decision. At the same time, based on the advantage evaluation function, the reward function is designed to comprehensively reflect the changing of the maneuver to the situation. Finally, for the problem that the maneuver decision model with high-dimensional state space (13-dimensional) and action space (15-dimensional) is difficult to be trained to converge traditionally, a training method called ''basic-confrontation'' is proposed, which effectively improves the training efficiency. Through a large number of machine-machine confrontation and man-machine confrontation simulation experiments under the initial states of advantages, disadvantages and balance, the maneuver decision model established in this paper is proved to be able

to learn the maneuver policy autonomously and therefore gain advantages in air combat. Compared with the previous researches of maneuver decision with reinforcement learning in which the simplifications of the air combat state space and action space was made due to the difficulty of convergence of high-dimensional space models, the proposed method can make the model more close to the actual motion state space of air combat, and can learn effective air combat maneuver decision policy, and further demonstrate the effectiveness of using reinforcement learning to solve the air combat maneuver decision problem.

The following part of the paper is arranged as follows. In section 2, the short-range air combat maneuver decision model is established. And the training method of DQN 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.

II. SHORT-RANGE AIR COMBAT MANEUVER DECISION MODEL

A. AIRCRAFT MOTION MODEL

The aircraft's motion model is the basis of the air combat model. The research focus of this paper is maneuvering decision-making, which mainly considers the positional relationship and velocity vector of the two sides in the three-dimensional space. Therefore, this paper uses a three-degree-of-freedom particle model as the motion model of the aircraft. The angle of attack and the side slip angle are ignored, assuming that the velocity direction coincides with the body axis.

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 as shown in (1).

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

In the same coordinate system, the dynamic model of the aircraft is shown in (2).

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

In (1) and (2), x, y, and z represent the position coordinates of the aircraft in the coordinate system, v represents the speed, and $\dot{x}$, $\dot{y}$, and $\dot{z}$ represent the values of the velocity v on the 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. g represents the acceleration of gravity. The position vector is recorded as $\boldsymbol{p} = [x, y, z]$, and $\dot{\boldsymbol{p}} = [\dot{x}, \dot{y}, \dot{z}]$.

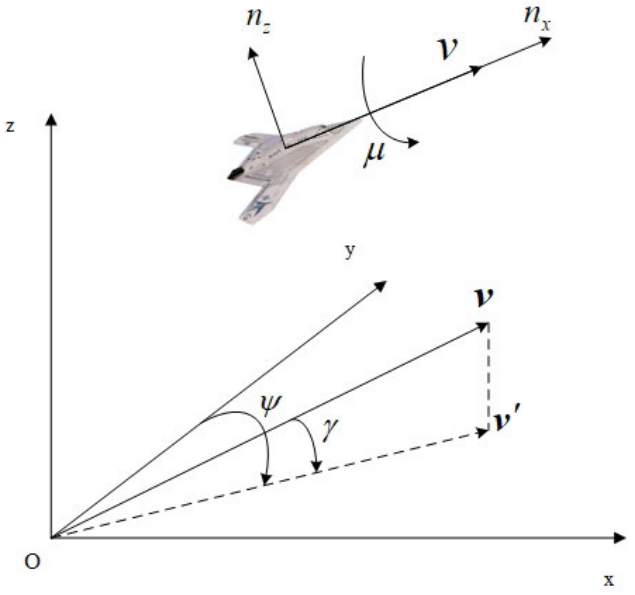


FIGURE 1. Aircraft three-degree-of-freedom particle model.

$[n_x, n_z, \mu]$ is a set of control variables that control the maneuvering of the aircraft, set $\boldsymbol{\Lambda}$ as the control space of the UAV. n_x is the overload in the velocity direction, which represents the thrust and deceleration of the aircraft. n_z represents the overload in the pitch direction, that is, the normal overload. μ is the roll angle around the velocity vector. The speed of the aircraft is controlled by n_x, and the direction of the velocity vector is controlled by n_z and μ, thereby controlling the aircraft to perform maneuvers. The parameters of the aircraft particle model are shown as in Figure 1.

B. ONE-TO-ONE SHORT-RANGE AIR COMBAT EVALUATION MODEL

Short-range air combat is also known as dog fighting.The goal in this air combat is to perform maneuvering to let the aircraft chase the tail of the target aircraft while avoiding letting the target enter its own tail.

In the one-to-one air combat situation as shown in Figure 2, from the idea of attack, the UAV should try to fly toward the target, chasing the target, that is, let the projection of the velocity vector $\dot{\boldsymbol{p}}_{\mathrm{U}}$ on $(\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}})$ be the largest, and from the defense idea, the UAV should avoid the target flying towards itself, and the projection of the target velocity vector $\dot{\boldsymbol{p}}_{\mathrm{T}}$ on $(\boldsymbol{p}_{\mathrm{U}} - \boldsymbol{p}_{\mathrm{T}})$ should be minimized. Then the situation advantage of UAV in air combat can be defined as

$$\eta_{\mathrm{A}} = \frac{\dot{\boldsymbol{p}}_{\mathrm{U}}\,(\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}})}{\|\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}}\|} - \frac{\dot{\boldsymbol{p}}_{\mathrm{T}}\,(\boldsymbol{p}_{\mathrm{U}} - \boldsymbol{p}_{\mathrm{T}})}{\|\boldsymbol{p}_{\mathrm{U}} - \boldsymbol{p}_{\mathrm{T}}\|}. \tag{3}$$

The larger η_{A} is, the larger the UAV has the advantage of the situation. On the contrary, when η_{A} is smaller, the target has a larger situation advantage, and the UAV is at a disadvantage.

In addition to the situation, distance is also a key factor in close air combat. The weapon used in close air combat has a generally limited attack range. It cannot attack the target beyond the range. Within the range, the closer distance,

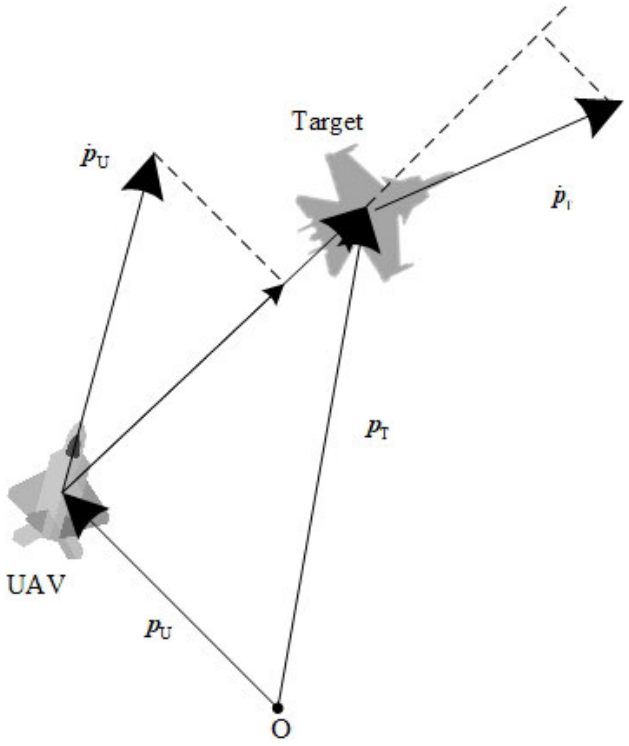


FIGURE 2. One-to-one short-range air combat situation.

the greater the probability of destroying the target. In addition, in order to avoid collision or accidental injury to the target debris, the distance between the two sides cannot be too close and there is a minimum safe distance. Let the maximum attack distance of the weapon be $D_{\max}$ and the minimum safe distance be $D_{\min}$, and then the distance advantage of the UAV in the air combat can be defined as

$$\eta_{\mathrm{D}} = \beta_1 \frac{D_{\max} - \left\|\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}}\right\|}{D_{\max} - D_{\min}} \left(1 - e^{-(\|\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}}\| - D_{\min})^{\beta_2}}\right), \quad (4)$$

where β_1 and β_2 are adjustment coefficients. When the distance between the UAV and the target is less than $D_{\max}$, η_{D} gradually increases as the distance decreases. When the distance is less than $D_{\min}$ or greater than $D_{\max}$, η_{D} decreases as the distance decreases or increases.

Combining situation advantage and distance advantage, UAV's advantage evaluation function in air combat can be defined as

$$\eta = \omega_1 \eta_{\mathrm{A}} + \omega_2 \eta_{\mathrm{D}}, \quad (5)$$

where ω_1 and ω_2 are weight coefficients. In summary, the UAV short-range air combat maneuver decision can be regarded as an optimization problem that solves the action in the control space $\boldsymbol{\Lambda}$ to maximize the advantage evaluation function η.

For the optimization problem $\max\eta(n_x, n_z, \mu)$, $[n_x, n_z, \mu] \in \boldsymbol{\Lambda}$, the objective function is a complex high-order nonlinear function, and it is difficult to obtain the extreme points through the derivative function, so it is difficult to obtain the analytical optimal solution. However, in the case where the air combat situation is determined, the advantage function can be calculated as the evaluation value of the current action to evaluate the current action. Therefore, we use the method of reinforcement learning to learn maneuver policy from the large number of feedback evaluation values.

The advantage function allows the UAV to understand its pros and cons in the current situation. In addition, it has to find a set of variables to describe the current air combat situation to let the UAV understand the current relative situation. The air combat state at any time can be completely determined by the information contained in the UAV position vector $\boldsymbol{p}_{\mathrm{U}}$, the UAV velocity vector $\dot{\boldsymbol{p}}_{\mathrm{U}}$, the target position vector $\boldsymbol{p}_{\mathrm{T}}$, and the target velocity vector $\dot{\boldsymbol{p}}_{\mathrm{T}}$ in the same coordinate system. However, the range of values of the three-dimensional coordinates is too large, so the coordinate values are not suitable as state inputs of reinforcement learning directly. Therefore, the absolute vector information of the UAV and the target should be transformed into the relative relationship between the two sides, and the angle should be used to represent the vector information. This will not only reduce the dimension of the state space, but also facilitate the normalization of the value range of the state information, thereby improving the efficiency of reinforcement learning. The one-to-one short-range air combat state space can be composed of the following five aspects:

1. UAV velocity information, including speed v_U, track angle γ_U and heading angle ψ_U.

2. Target velocity information, including speed v_T, track angle γ_T and heading angle ψ_T.

3. The relative positional relationship between the UAV and the target is characterized by the distance vector $(\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}})$. The coordinate information of the distance vector is converted into the form of the modulus and angle of the vector. The modulus of distance vector is $D = \left\|\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}}\right\|$. γ_D represents the angle between $(\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}})$ and the o-x-y plane, and ψ_D represents the angle between the projection vector of $(\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}})$ on the o-x-y plane and the oy axis. The relative positional relationship between the UAV and the target is represented by D, γ_D, and ψ_D.

4. The relative motion relationship between the UAV and the target, including the angle α_U which is between the UAV velocity vector $\dot{\boldsymbol{p}}_{\mathrm{U}}$ and the distance vector $(\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}})$ and the angle α_T which is between the target velocity vector $\dot{\boldsymbol{p}}_{\mathrm{T}}$ and the distance vector $(\boldsymbol{p}_{\mathrm{T}} - \boldsymbol{p}_{\mathrm{U}})$.

5. The height of the UAV z_U and the height of the target z_T.

Based on the above variables, the one-to-one air combat situation at any time can be fully characterized.

C. MANEUVER DECISION MODELING BY DEEP Q NETWORK

1) ARCHITECTURE OF MODEL

Reinforcement learning is a method for Agent to optimize its action policy in an unknown environment [22]. The Markov Decision Process (MDP) is usually used as the theoretical

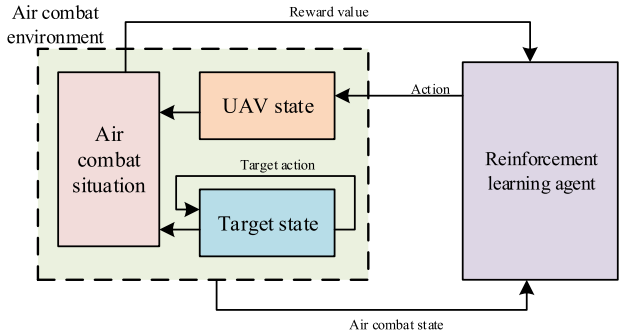


FIGURE 3. UAV short-range air combat maneuver decision model framework based on reinforcement learning.

framework for reinforcement learning. For model-free reinforcement learning problems, the Markov decision process is described by a 4-tuple (S, A, R, γ). Where S represents the state space, A represents the action space, R represents the reward function, and γ represents the discount factor. The interaction process of reinforcement learning is as follows. At time t, the Agent applies an action a_t to the environment. After the action a_t is executed, the state is transferred from s_t to s_{t+1} at the next moment, and the agent obtains the reward value r_{t+1} from the environment.

The state is evaluated by a state value function or a state-action value function. According to the Bellman formula, the state-action value function can be shown as $Q_\pi(s,a) = E_\pi[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) | s_t = s, a_t = a]$. It can be seen from state-action value function that the calculation of the state value is the accumulation of the immediate reward value and the subsequent state value, so the evaluation of the state has taken into account the subsequent state, so the action selected according to the state value has farsightedness, that is, the obtained policy is long-term optimal in theory. The purpose of reinforcement learning is to solve out an optimal policy $\pi_*(a|s) = \arg\max_{a \in A} Q_*(s,a)$ in the policy space π, which can make the long-term cumulative reward largest.

According to the interactive process, the reinforcement learning model framework of the UAV short-range air combat maneuver decision is shown in figure 3. The state of the UAV and the state of the target are integrated and calculated to form a state description of the air combat environment, which is output to the agent. The air combat environment model calculates the UAV's advantage evaluation value based on the current situation of both sides, and outputs it as a reward value to the reinforcement learning agent. In the interaction process, the agent outputs the action value to the air combat environment model to change the state of the UAV, and the target changes the state according to its own maneuver policy. The reinforcement learning agent continuously interacts with the air combat environment to obtain air combat states, action values, and reward values as transitions. Based on the transitions, the maneuver policy of UAV is dynamically updated, so that the output action value tends to be optimal, thus realizing the self-learning of the UAV air combat maneuver policy.

TABLE 1. The state space for the DQN model.

State	definition	State	definition
s_1	$\frac{v_U}{v_{\max}}a - b$	s_8	$2b\frac{\gamma_D}{\pi}$
s_2	$\frac{\gamma_U}{2\pi}a - b$	s_9	$\frac{\psi_D}{2\pi}$
s_3	$\frac{\psi_U}{2\pi}a - b$	s_{10}	$\frac{\alpha_U}{2\pi}a - b$
s_4	$\frac{v_U - v_T}{v_{\max} - v_{\min}}a - b$	s_{11}	$\frac{\alpha_T}{2\pi}a - b$
s_5	$\frac{\gamma_T}{2\pi}a - b$	s_{12}	$\frac{z_U}{z_{\max}}a - b$
s_6	$\frac{\psi_T}{2\pi}a - b$	s_{13}	$\frac{z_U - z_T}{z_{\max} - z_{\min}}a - b$
s_7	$\frac{D}{D_{\text{threshold}}}a - b$		

Facing a high-dimensional continuous state space such as an air combat environment, the DQN algorithm [23] is selected as the algorithm framework of reinforcement learning. The core of the DQN algorithm is to use the deep neural network to approximate the value function. At the same time, based on the Q learning algorithm, the TD error is used to continuously adjust the parameters θ of the neural network, so that the state value of the network output is constantly approaching the true value, $Q(s,a) \approx Q(s,a|\theta)$. Based on the short-range air combat environment model in section 2, the maneuver decision model is constructed under DQN framework.

2) STATE SPACE

The state space of the maneuver decision model is used to describe the air combat situation which is divided into five aspects as section 2.B. Therefore, the state space consists of the following 13 variables, v_U, γ_U, ψ_U, $v_U - v_T$, γ_T, ψ_T, D, γ_D, ψ_D, α_U, α_T, z_U, $z_U - z_T$. In order to unify the range of each state variable and improve the efficiency of network learning, each state variable is normalized into a range as shown in Table 1.

$v_{\max}$ and $v_{\min}$ represent the maximum and minimum speeds of the aircraft motion model, respectively. $z_{\max}$ and $z_{\min}$ represent the ceiling and minimum safe height of the aircraft, respectively. $D_{\text{threshold}}$ is the distance threshold, taking the starting distance of short-range air combat. a and b are two positive numbers and satisfy $a = 2b$. State space is defined as a vector $\boldsymbol{S} = [s_1, s_2, \cdots, s_{13}]$.

3) ACTION SPACE

The action space for the DQN model is the UAV's maneuver library. The establishment of the maneuver library can draw on the tactical actions of fighter pilots during air combat. Pilots can derive many tactical actions such as barrel rolling, cobra maneuvering, high yo-yo, and low yo-yo based on various factors such as aircraft performance, physical endurance and battlefield situation. However, these complex maneuvers are ultimately derived from basic maneuvering actions, so as long as the UAV's maneuvering library contains these basic maneuvers, it can meet the requirements of simulation research.

According to the common air combat maneuver, NASA scholars designed 7 basic maneuvers [24], which are uniform

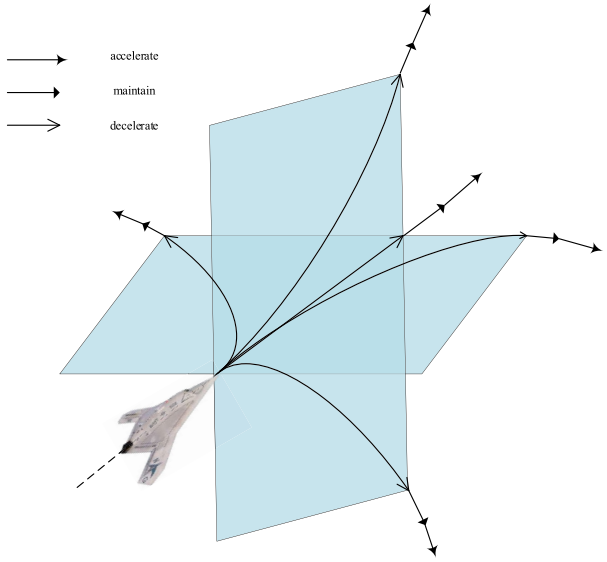


FIGURE 4. UAV maneuver library.

linear flight, accelerated straight flight, deceleration straight flight, left turn flight, right turn flight, upward flight, and downward flight. These maneuvers enable UAV movement in three-dimensional space, but the increase or decrease of speed can only be achieved when flying in a straight line, which makes it difficult to control speed when performing other maneuvers. Therefore, based on the above basic maneuver, the maneuver library can be expanded. As shown in Figure 4, the UAV can be maneuvered in the forward, left, right, up, and down directions. Maintenance, acceleration, and deceleration control are provided in each direction. So the maneuver library can be arbitrarily expanded to perform the actions required by different handling precisions. Each action $\boldsymbol{a}_i$ in the maneuver library corresponds to a set of control values $[n_x, n_z, \mu]$. Thus, action space $\boldsymbol{A}$ consists of a discrete set of action values, which is a subset of the control space, $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n] \subseteq \boldsymbol{\Lambda}$.

4) REWARD FUNCTION

The reward value is an immediate assessment of the agent's actions by the environment. The reward value in this paper is calculated based on the advantage evaluation function of the UAV air combat situation and is used as an immediate evaluation of the maneuver decision. At the same time, the reward value should reflect the penalty for the action beyond the flight range during the simulation. The limits of the flight range include the limitation of the flight altitude. Define the penalty function as

$$\eta_p = \begin{cases} P, & \text{if } (z_U < z_{\min}) \text{ or } (z_U > z_{\max}); \\ 0, & \text{otherwise}. \end{cases} \tag{6}$$

Based on the situation assessment function η and the penalty function η_P, the reward function of the reinforcement learning algorithm is $r = \eta + \eta_P$.

In summary, the model of UAV short-range air combat maneuver decision based on DQN algorithm is shown in Figure 5. The running process of the model is as follows. The current air combat state is s_t, the online Q network outputs the action value $a_t \in \boldsymbol{A}$ to the air combat environment based on the ϵ-greedy method. The UAV flies according to the action value a_t, and the target flies according to the preset policy, then the status is updated to s_{t+1}, and the return value r_t is obtained. Store (s_t, a_t, r_t, s_{t+1}) as a transition sample in experience replay memory to complete an interaction. In the learning process, the minibatch transitions are extracted from the experience replay memory based on the prioritized experience replay policy, and the parameters θ of the online network is updated according to the gradient descent principle by using the TD error, and the parameters of the online network θ is periodically assigned to the parameter of the target network θ'. Continue to learn and update until the TD error approaches 0, and finally come up with a short-range air combat maneuver policy.

III. TRAINING METHOD OF THE DQN MODEL

Since the state space of the air combat model is very large, if the unenlightened UAV is directly confronted with the target which has smart maneuver policy, the result will definitely be very bad, and a large number of invalid transition samples will be generated. This will lead to extremely low efficiency of reinforcement learning, even learning failure due to sparse sample. In response to this problem, a training method named basic-confrontation training is designed based on the process of human learning which is gradually transitioning from simple cognition to complex knowledge.

Based on this idea, the method divides the training process of the maneuver decision DQN model into two parts. First, let the target fly with simple basic action, such as uniform linear motion and horizontal spiral motion in different initial states such as advantage, disadvantage, and balance, make the UAV familiar with the air combat situation and learn the basic maneuver policy, which is called basic training. The basic training items are carried out according to the target's maneuver strategy and the initial situation of air combat from simple to complex. The follow-up training is carried out directly on the previously trained network, thus achieving a gradual superposition of learning effects. Second, after the UAV learned the basic flight strategy, carry out the confrontation training in different initial states in which the target has smart maneuver policy, let UAV learn the maneuver policy under the confrontation condition to defeat the target in the air combat.

A. TARGET POLICY

When conducting confrontation training, the target's maneuver policy adopts a robust maneuver decision algorithm based on statistics principle [14]. This algorithm has strong robustness, and the simulation proves that the algorithm is better than the traditional min-max method [25]. The framework of the algorithm is to test all the actions in the maneuver library

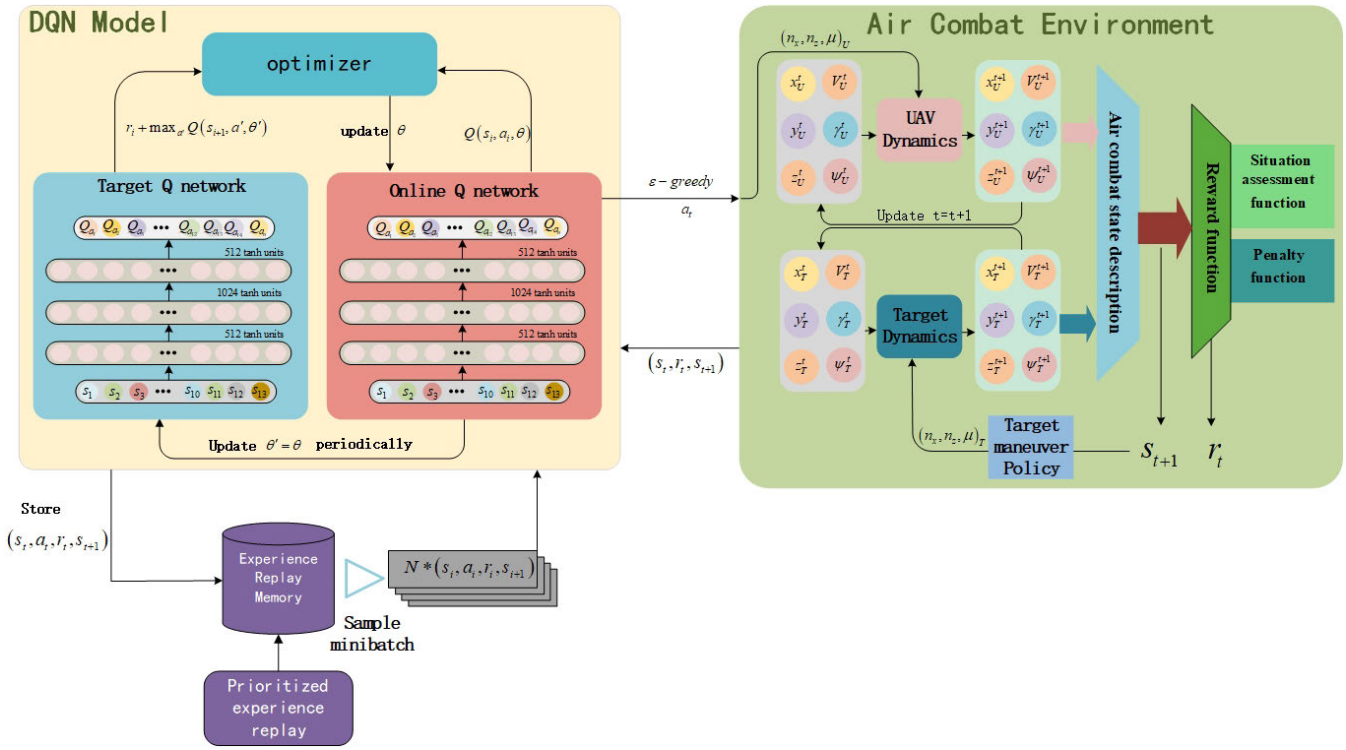


FIGURE 5. Model of UAV short-range air combat maneuver decision based on DQN algorithm.

under the current situation, obtain the value of membership functions after the execution of each action, and select the action that makes the statistical information of the membership functions the best as the next maneuver action. The following is a brief introduction to this algorithm.

The current air combat situation is characterized by four parameters: azimuth α, distance D, speed v and altitude h, and the membership functions of each parameter are defined separately to enhance the robustness of the situation description. The membership function of the azimuth parameter is

$$f_\alpha = \frac{\alpha_U \alpha_T}{\pi^2}. \tag{7}$$

The membership function of the distance parameter is

$$f_D = \begin{cases} 1, & \text{if } \Delta D \le 0; \\ e^{-\frac{\Delta D^2}{2\sigma^2}}, & \text{otherwise } \Delta D > 0, \end{cases} \tag{8}$$

where σ is the standard deviation of the weapon attack distance, and $\Delta D = D - D_{\max}$. The membership function of the speed parameter is

$$f_v = \frac{v_U}{v_*} e^{-\frac{2|v_U - v_*|}{v_*}}, \tag{9}$$

where v_* represents the optimal speed of the target attack UAV, set $\Delta v = v_{\max} - v_T$, and the value is set as

$$v_* = \begin{cases} v_T + \Delta v \left(1 - e^{\frac{\Delta D}{D_{\max}}}\right), & \text{if } \Delta D > 0; \\ v_T, & \text{otherwise } \Delta D \le 0. \end{cases} \tag{10}$$

The membership function of the height parameter is defined as

$$f_h = \begin{cases} 1, & \text{if } h_s \le \Delta z \le h_s + \sigma_h; \\ e^{-\frac{(\Delta z - h_s)^2}{2\sigma_h^2}}, & \text{if } \Delta z < h_s; \\ e^{-\frac{(\Delta z - h_s - \sigma_h)^2}{2\sigma_h^2}}, & \text{otherwise } \Delta z > h_s + \sigma_h, \end{cases} \tag{11}$$

where h_s represents the optimal attack height difference of the target to the UAV and σ_h is the standard deviation of the optimal attack height.

When the membership functions of the above four parameters are gradually approaching 1, the target is in an advantage, and when approaching 0, the target is at a disadvantage. The steps of the algorithm are as follows:

1. At time t, based on the current state of the target and the UAV, control commands for all actions in the action library are sent to the motion model for heuristic maneuvering.

2. Step 1 is performed to obtain all possible positions of the target at time t + 1, and the situation of each position is solved to obtain a set

$$F_i^{t+\Delta t} = \left\{ f_\alpha^{i,t+\Delta t}(\alpha), f_D^{i,t+\Delta t}(D), f_v^{i,t+\Delta t}(v), f_h^{i,t+\Delta t}(\Delta z) \right\} \tag{12}$$

where i represents the number of the action in the maneuver library, and the set of membership functions of the

parameters corresponding to all maneuvers is $F^{t+\Delta t} = \left\{F_1^{t+\Delta t}, F_2^{t+\Delta t}, \cdots F_n^{t+\Delta t}\right\}$.

3. Calculate the mean $m_i^{t+\Delta t}$ and the standard deviation $s_i^{t+\Delta t}$ of $F_i^{t+\Delta t}$.

$$m_i^{t+\Delta t} = E\left[F_i^{t+\Delta t}\right]. \tag{13}$$

$$s_i^{t+\Delta t} = \sqrt{\begin{pmatrix} \left(f_\alpha^{i,t+\Delta t}(\alpha) - m_i^{t+\Delta t}\right)^2 \\ + \left(f_D^{i,t+\Delta t}(R) - m_i^{t+\Delta t}\right)^2 \\ + \left(f_h^{i,t+\Delta t}(\Delta z\alpha) - m_i^{t+\Delta t}\right)^2 \\ + \left(f_v^{i,t+\Delta t}(v) - m_i^{t+\Delta t}\right)^2 \end{pmatrix}}. \tag{14}$$

Get a binary array $MS_i^{t+\Delta t} = \left(m_i^{t+\Delta t}, s_i^{t+\Delta t}\right)$, and build a set $MQ^{t+\Delta t} = \left(MS_i^{t+\Delta t}\right)$ for i $= 1, 2, \ldots,$ n. Select the element with the largest mean in $MQ^{t+\Delta t}$, and use the corresponding maneuver as the action to be executed for the target. If the elements with the largest mean are more than 1, output the maneuver corresponding to the element with the smallest standard deviation among these elements.

4.Execute the action, update the time, and return to step 1.

B. TRAINING EVALUATION

In confrontation training, the target makes maneuver decisions according to the above algorithm. A training process consists of multiple episodes, and each episode represents an air combat process consisting of multiple steps, each step representing a decision cycle. In order to evaluate the training performance of the maneuver decision model, three indicators are defined, including the advantage steps rate, the average advantage reward and the maximum episode value. When the reward value is not less than $0.8*\max(\eta)$, the UAV is considered to be in the advantage state, and the advantage steps rate is the ratio of the number of steps in the advantage state to the total number of execution steps in this episode. The average advantage reward is the average of the reward values of the advantage state in episode. The maximum episode value is the sum of all the reward values in this episode. if the UAV flies out of the height limit described in equation (6), causing the episode to be interrupted, the maximum episode value is set to 0.

In order to reflect the effect of the agent learning, the evaluation episode is set to be executed periodically during a training process. In the evaluation episode, the ε-greedy algorithm is not executed, and the online Q network directly outputs the action value with the largest Q value. Advantage steps rate, the average advantage reward and the maximum episode value in the episode are recorded to evaluate the previously learned maneuver policy.

In the next section, we will discuss the training process in detail through simulation experiments. The DQN model training process is shown as following algorithm.

Algorithm 1 DQN Model Training Process

```
Initialize online network Q with random parameters θ
Initialize target network Q' with random parameters
θ' ← θ
Initialize replay buffer R
Set the target maneuver policy (basic/ confrontation)
for episode = 1, M do
    Initialize the initial state of air combat
    Receive initial observation state s_1
    If episode % evaluation frequency = 0
     Perform evaluation episode
    for t = 1, T do
     With probability ϵ select a random action a_t
     Otherwise select a_t = max_a Q(s_t, a; θ)
     UAV executes action a_t, and target executes action
     according to its policy
     Receive reward r_t and observe new state s_{t+1}
     Store transition (s_t, a_t, r_t, s_{t+1}) in R
     Sample a random minibatch of N transition
     (s_i, a_i, r_i, s_{i+1}) from R
     Set y_i = r_i + γ max_{a'} Q'(s_{i+1}, a'; θ')
     Perform a gradient descent step on (y_i − Q(s_i, a_i; θ))^2
      with respect to the network parameters θ
     Every C steps reset θ' = θ
    end for
end for
```

IV. SIMULATION AND ANALYSIS

A. PLATFORM SETTING

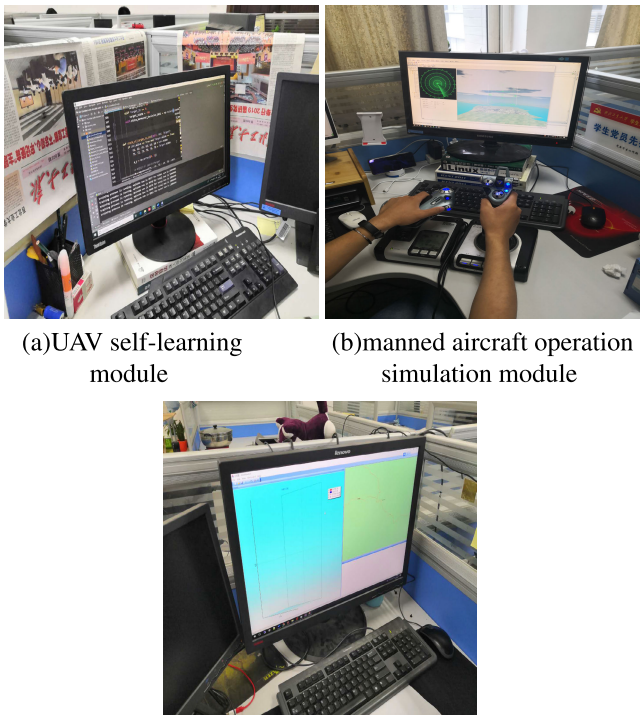
In this paper, the short-range air combat environment model is established by using Python language, and the DQN network model is built based on TensorFlow module.

1) HARDWARE

Based on the UAV autonomous maneuvering decision model, the man-machine air combat confrontation system is developed. As shown in Figure 6, the man-machine air combat confrontation system consists of three modules: the UAV self-learning module, the manned aircraft operation simulation module and the air combat environment module. The three modules reside on three computers, and the computers are connected by Ethernet to exchange information. Each computer has an Intel(R) Core(TM) i7-8700k CPU and 16GB RAM. The UAV self-learning module computer is also equipped with a NVIDIA GeForce GTX 1080 TI graphics card for Tensorflow acceleration.

2) PARAMETERS SETTING

The parameters of the short-range air combat environment model are set as follows. The farthest attack distance is $D_{\max} = 3$km, minimum distance between the two aircrafts $D_{\min} = 200$m, the maximum return value is adjusted to 5, punish value $P = -10$, and the maximum speed $v_{\max} = 400$ m/s, minimum speed $v_{\min} = 90$ m/s, ceiling height $z_{\max} = 12000$m, minimum height $z_{\min} = 1000$m, distance threshold



(a)UAV self-learning module

(b)manned aircraft operation simulation module

(c)air combat environment module

FIGURE 6. The man-machine air combat confrontation system.

$D_{\text{Threshold}} = 10000$ m, $a = 10$, $b = 5$. For control space, set $n_x \in [-1, 2]$, $n_z \in [0, 8]$, $\mu \in [-\pi, \pi]$. Extend 7 basic maneuvers to 15 to achieve direction and speed control. The action space contains 15 basic maneuver actions. The control values of the basic actions in the maneuver library are shown in Table 2.

The parameters in the DQN model are set as follows. According to the definition of the state space and the definition of the maneuver library, it is clear that the DQN has 13 input states and 15 output Q values. An online Q network and a target Q network are constructed using a fully connected network. Networks have 3 hidden layers with 512,1024 and 512 units respectively. The output layer has none activation function, and the remaining layers are all tanh layers. The learning rate is 0.001, and the discount factor is $\gamma = 0.9$. The target network is updated every 2,500 steps. The weights of each layer of the network are initialized by the variance scaling initializer, and the biases of the fully connected network are initialized by the zeros initializer. The size of Minibatch is set to 512, and the size of replay buffer is set to 10^5.

In the short-range air combat simulation process, the decision period T is set to 1s, and an episode contains 30 decision steps. In an episode simulation, if the UAV flies out of the boundary shown by equation (6), this episode will end.

Next, the basic training and confrontation training experiments of the maneuver decision model are carried out in turn. After the UAV learns a certain maneuver policy, the man-machine confrontation training is implemented.

TABLE 2. Maneuver library.

No.	Maneuver	control values		
		n_x	n_z	μ
a_1	forward maintain	0	1	0
a_2	forward accelerate	2	1	0
a_3	forward decelerate	-1	0	0
a_4	left turn maintain	0	8	-acos(1/8)
a_5	left turn accelerate	2	8	-acos(1/8)
a_6	left turn decelerate	-1	8	-acos(1/8)
a_7	right turn maintain	0	8	acos(1/8)
a_8	right turn accelerate	2	8	acos(1/8)
a_9	right turn decelerate	-1	8	acos(1/8)
a_{10}	upward maintain	0	8	0
a_{11}	upward accelerate	2	8	0
a_{12}	upward decelerate	-1	8	0
a_{13}	downward maintain	0	8	π
a_{14}	downward accelerate	2	8	π
a_{15}	downward decelerate	-1	8	π

TABLE 3. Basic training items.

Item	Target maneuver	Initial situation of UAV
1	Uniform linear flight	advantage
2	Uniform linear flight	balance
3	Uniform linear flight	disadvantage
4	horizontal spiral maneuver	advantage
5	horizontal spiral maneuver	balance
6	horizontal spiral maneuver	disadvantage

B. MODEL TRAINING AND TESTING

1) BASIC TRAINING

In the basic training, the target performs uniform linear motion and horizontal spiral maneuver respectively. The initial situation of UAV is in advantage, balance and disadvantage, respectively, making UAV fully familiar with the situation of air combat. The training items are shown in Table 3.

The advantage in Table 3 means that the UAV pursues the target from behind. The balance refers to the UAV and the target heading toward each other. The disadvantage is that the target pursues the UAV from behind. Training is carried out item by item according to the serial number in Table 3. Each item is trained with 10^6 episodes, and an evaluation episode is performed every 3000 episodes during training.

In each training process, in order to make the UAV fully familiar with the air combat environment and improve the diversity of transitions, the initial state of the UAV and the target in the training episode are selected randomly within a wide range, and in order to ensure the uniformity of evaluation, the initial situation is fixed in the evaluation episode. For example, when performing the first item, the initial situation of training episode and evaluation episode are shown in Table 4 and Figure 7.

Figure 8 shows the change of the maximum episode value during the training of item 1. It can be seen that the UAV updates the maneuver policy through interaction training, and the maximum episode value continues to increase. Figure 9 shows the maneuvering trajectory of an evaluation episode once the training of item 1 is completed. It can be seen from the figure that the UAV starts to chase the target from the

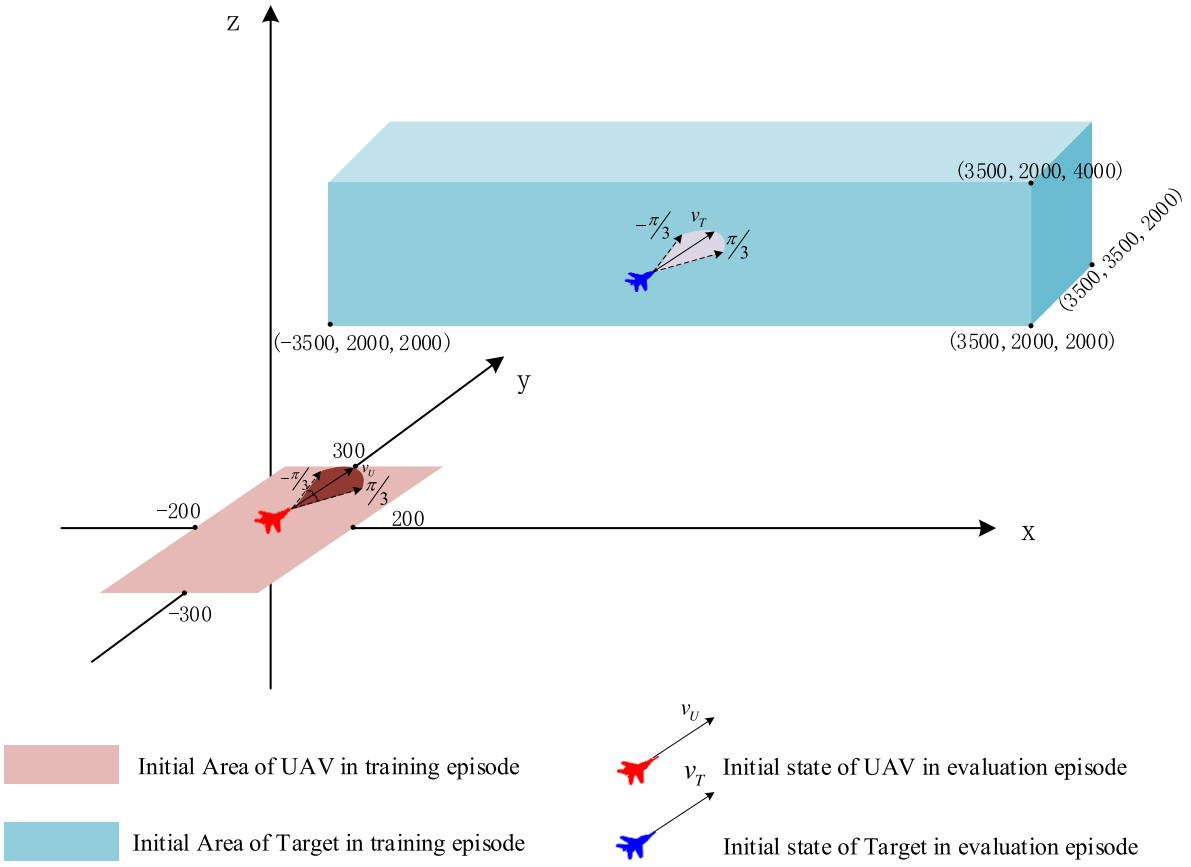


FIGURE 7. Initial state settings for training item 1.

TABLE 4. Initial state settings for training item 1.

Initial state		x (m)	y (m)	z (m)	v (m/s)	γ	ψ
Training episode	UAV	[-200,200]	[-300,300]	3000	280	0	$[-\frac{\pi}{3}, \frac{\pi}{3}]$
	Target	[-3500,3500]	[2000,3500]	[2000,4000]	[100,300]	0	$[-\frac{\pi}{3}, \frac{\pi}{3}]$
Evaluation episode	UAV	0	0	3000	180	0	0
	Target	3000	3000	3000	180	0	0

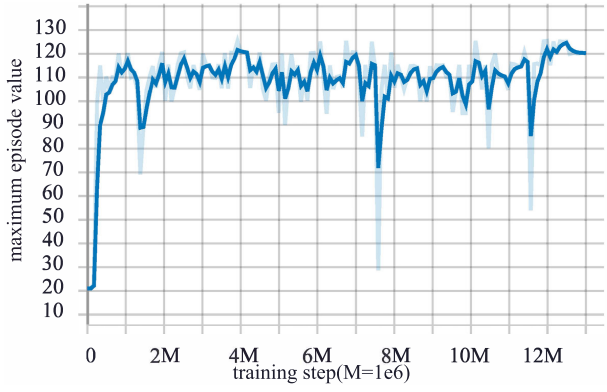


FIGURE 8. The maximum episode value during the training of item 1 (Dark line is the result of smoothing the light line, with a smoothing rate of 0.5).

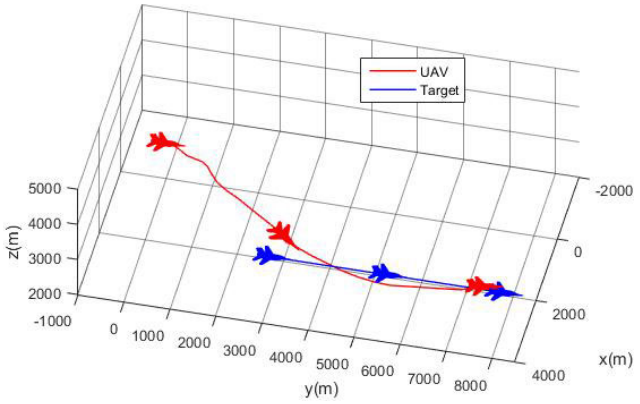


FIGURE 9. Maneuvering trajectory after training item 1.

left rear side of the target, continuously adjusts the heading and speed, and maintains the tail chasing situation, so that the target is always in the intercepted area of the missile.

Figure 10 shows the change of the maximum episode value during a training process of item 3. It can be seen from the figure that the UAV is at a disadvantage in the initial

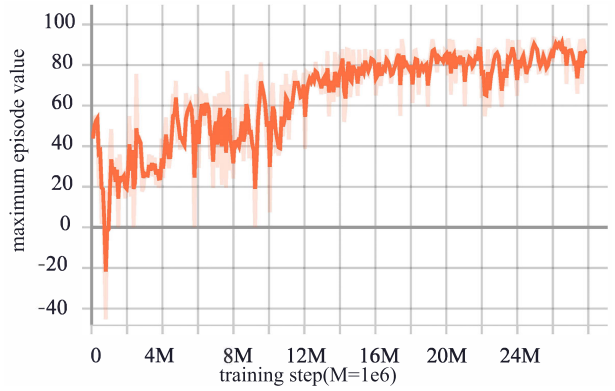


FIGURE 10. The maximum episode value during the training of item 3 (Dark line is the result of smoothing the light line, with a smoothing rate of 0.5).

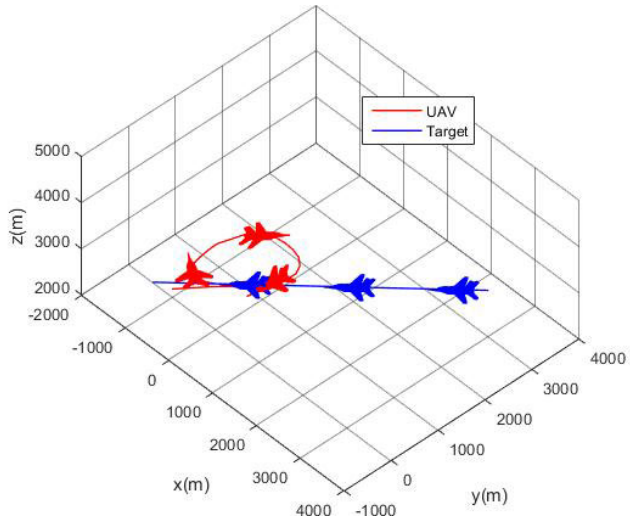


FIGURE 11. Maneuvering trajectory after training item 3.

situation, so the maximum return value is lower at the beginning of training, but with the development of the training, the UAV gradually grasps the maneuver policy of getting rid of the disadvantage and transferring to the advantage, thus making the maximum episode value rise. Figure 11 shows the maneuvering trajectory of an evaluation episode once the training of item 3 is completed. It can be seen that the UAV is at a disadvantage in front of the target at the initial moment, then starts to turn right, and faces the target, and finally turns right into the target tail, and adjusts the speed to catch up with the target, keeping the interception of the target.

2) CONFRONTATION TRAINING

After completing all the learning items in Table 2, confrontation training with smart target is implemented. UAV adopts basic trained DQN model as its maneuver policy, and the target adopts the statistic principle-based method as its policy, and the performance of each methods is verified by confrontation simulation.

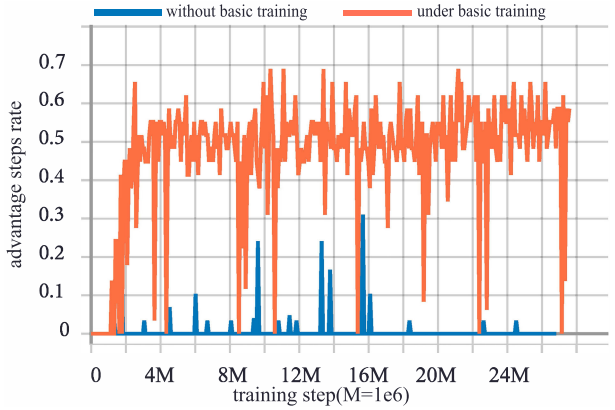


FIGURE 12. The advantage steps rate during the confrontation training with balance initial situation.

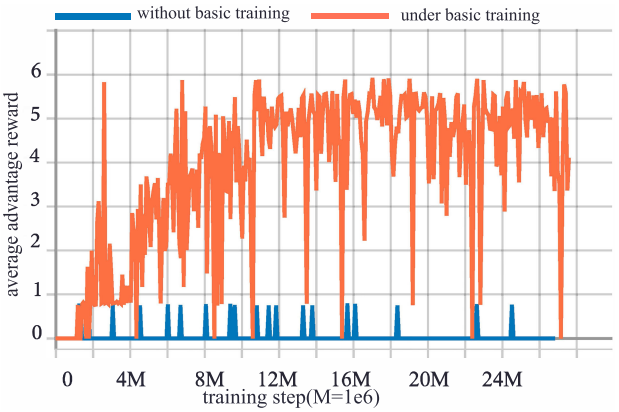


FIGURE 13. The average advantage reward during the confrontation training with balance initial situation.

In confrontation training, the UAV uses the ϵ-greedy method to gradually explore the maneuver policy based on the results of the basic training. Since the state of confrontation is more complicated, in order to increase the sample space, the size of the replay buffer is increased from $10^5$to 10^6.

In order to ensure the diversity of combat state and the generalization of maneuver policy, the initial state of the UAV and the target in the training episode are randomly generated within a certain range, and the confrontation training is carried out under the condition that the initial situation of the UAV is balance and disadvantage. Table 5 shows the initial state of the training in the balance initial situation.

In the training process with balance initial situation, the changes of the advantage steps rate, the average advantage reward and the maximum episode value are shown in Figure 12, Figure 13 and Figure 14, respectively. In the three figures, the blue line shows the change process of the indicator in confrontation training without basic training, and the yellow line indicates the confrontation training process under basic training. Since the UAV has a fundamental flight policy after the basic training, there will be no low-level errors such as flying out of the boundary. Therefore, it can be seen from the figure that the maximum episode value rarely

TABLE 5. Setting of the balance initial situation in confrontation training.

Initial state		x (m)	y (m)	z (m)	v (m/s)	γ	ψ
Training episode	UAV	[-200,200]	[-300,300]	3000	280	0	$[-\frac{\pi}{3},\frac{\pi}{3}]$
	Target	[-3500,3500]	[2000,3500]	[2000,4000]	[100,300]	0	$[-2\frac{\pi}{3},4\frac{\pi}{3}]$
Evaluation episode	UAV	0	0	3000	180	0	$\frac{\pi}{4}$
	Target	3000	3000	3000	180	0	$5\frac{\pi}{4}$

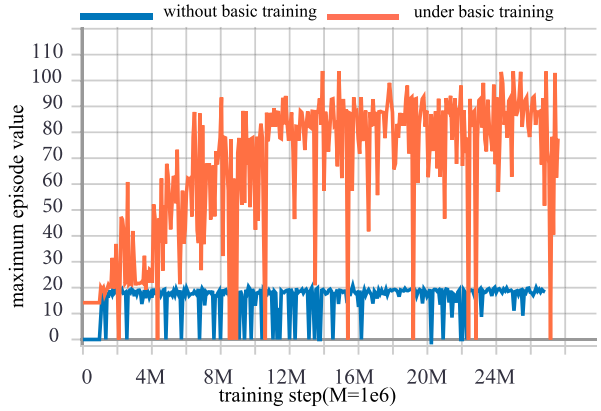


FIGURE 14. The maximum episode value during the confrontation training with balance initial situation.

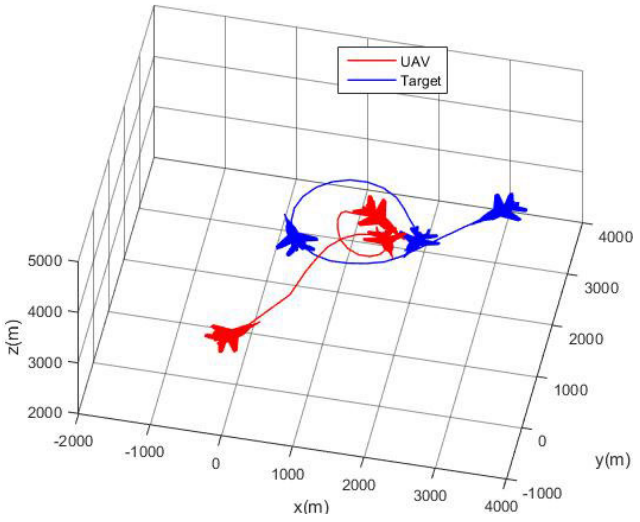


FIGURE 15. Confrontation maneuvering trajectory under balance initial situation.

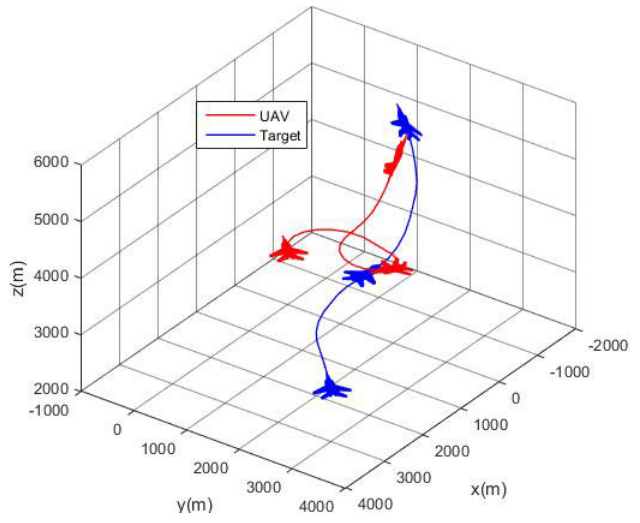


FIGURE 16. Confrontation maneuvering trajectory under disadvantage initial situation.

appears 0 value due to the episode interruption in the initial stage of the confrontation training. The UAV without basic training has no experience in the air combat environment, and it is easy to exceed the limitation boundary in the process of maneuver exploration, so that the maximum episode value has many zero values during the confrontation training. In addition, since the target aircraft is smarter than the UAV at the beginning of the training, it has more opportunities to fire weapon to the UAV, resulting in many negative values for the maximum episode value. As the training continues, the basically trained UAV gradually masters the target's maneuver policy and explores the maneuver policy that can defeat the target. Therefore, the three indicator values gradually increase with the development of training. This proves that UAV's maneuver policy allows itself to move from the balance situation to the advantage situation as quickly as possible and continue to maintain its advantage. In contrast, during the same confrontation training period, the three indicators of the UAV without basic training do not rise and converge steadily, and the maximum episode value still shows a large negative value at the end of the training period, indicating that the UAV does not get maneuver policy from the training to get advantage to the target.

Figure 15 shows the maneuvering trajectory in an evaluation episode after the confrontation training with balance initial situation. The two sides start to fly head-on from the initial position. After reaching a certain distance, the UAV flies to the right side of the target. The target turns to the right to pursue the UAV. Then the UAV reduces the speed and the turning radius, so that the target rushes to the front of the UAV, and thus the UAV enters an advantage position.

Figure 16 shows the maneuvering trajectory in an evaluation episode after the confrontation training with disadvantage initial situation. At the initial moment, the target forms a situation of chasing the UAV from its tail, so it is constantly maneuvering in the direction of the UAV, intending to reduce the distance and let the UAV enter the missile interception area. UAV turns to the right immediately, intending to get rid of the unfavorable situation of being chased, and constantly adjust the speed and heading. After meeting with the target, the UAV quickly turns right and climbs up. When the target is doing the barrel rolling maneuver and intending to turn back, the UAV cuts into the back side of the

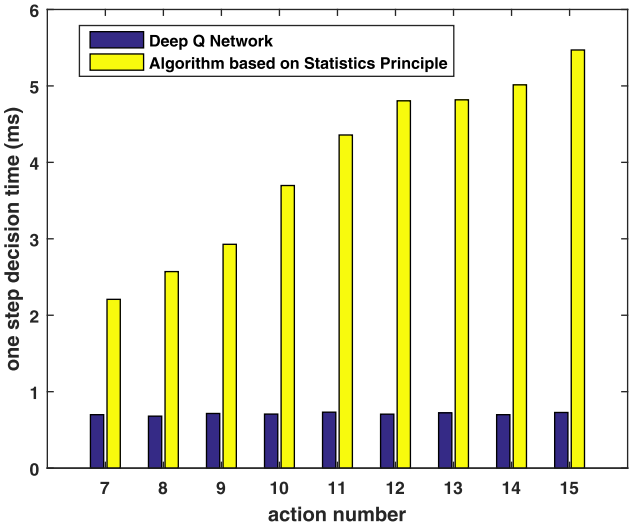


FIGURE 17. One step decision time performance.

target, realizes the tail chase to the target, and obtains the fire chance.

Based on the above-mentioned confrontation training simulation, the UAV short-range air combat maneuver decision model based on deep reinforcement learning established in this paper is proved to be able to obtain the maneuver policy through autonomous learning and defeat the target with statistic principle-based maneuver policy.

a: DECISION TIME PERFORMANCE

Air combat has extremely strict real-time performance requirement for maneuver decision. It is therefore necessary to test the real-time performance of the maneuver decision model. According to the size of the maneuver library, 9 groups of tests were performed. The number of maneuver actions of each group was increased from 7 to 15. And the average one-step decision time of DQN model and algorithm based on statistics principle in each test were calculated through 1000 decision steps, and the experimental results are shown in Figure 17. It can be seen from the figure that as the number of maneuver action increases, the one-step decision time of the DQN model remains at around 0.6ms, while the decision time of algorithm based on statistics principle increases from about 2ms to nearly 6ms.

The experimental results show that the real-time performance of the DQN decision model is better than that of the algorithm based on statistics principle. The algorithm based on statistics principle only performs one traversal calculation, and the calculation time is relatively short. Other optimization algorithms such as genetic algorithms require a large number of loop iteration calculations, and real-time performance is more difficult to meet the requirements of online decision-making, so the author of [9] believes that the purpose of this optimization is not to achieve online control, but to find some meaningful new maneuvers to carry out tactical research. So, it can be concluded that the real-time performance of the model established in this paper is better than that of iterative optimization algorithms.

3) MAN-MACHINE AIR COMBAT CONFRONTATION

Although in the confrontation training, UAV can defeat the target with certain maneuver policy through learning. However, this kind of policy of target is relatively fixed, the randomness is not strong, and it is easy to be mastered and cracked, so it cannot reflect the complexity of the opponent's maneuver policy in real air combat. In order to further verify the self-learning ability of the reinforcement learning and the correctness of the acquired maneuver policy, the target aircraft should be controlled by people, so a man-machine air combat confrontation system is developed.

The UAV self-learning module is constructed by using the above-mentioned UAV air combat maneuver decision model. The main function is to update and improve its maneuver policy according to the air combat data of man-machine confrontation. As shown in Figure 18 and Figure 6 (b), the manned aircraft operation simulation module provides the operator with simulation pictures of flight attitude and air combat situation. At the same time, the module provides the operator with the HOTAS joystick, realizing the real-time control of the aircraft.

The main function of the air combat environment module is to receive the flight status information of the UAV and the manned aircraft, and then display the three-dimensional situation of the current air combat, and evaluate the current air combat situation, determine whether one of the two sides is shot down, and then output the air combat situation information and the evaluation value to the both sides.

In the model, the flight performance of the manned aircraft and the UAV is exactly the same. During the training process, the UAV firstly conducts online real-time confrontation with the manned aircraft based on the policy learning from confrontation training. After a certain amount of confrontation training vs manned aircraft, the saved flight path state data of manned aircraft is randomly intercepted to simulate the target trajectory for UAV off-line reinforcement learning, and then the maneuver policy is improved, and then based on this policy, the manned aircraft is confronted. Continue to iterate the confrontation-learning process in turn.

Figure 19 is a trajectory diagram of a confrontation before confrontation policy update is performed. It can be seen that the manned aircraft defeat the UAV. Figure 20 is a trajectory diagram of a confrontation after the confrontation policy update. It can be seen that the UAV defeat the manned aircraft after off-line training.

Through the simulation experiment of man-machine air combat confrontation, it is proved that the UAV short-range air combat maneuver decision model based on deep reinforcement learning can self-learn and update the maneuver policy to gain advantages in air combat confrontation.

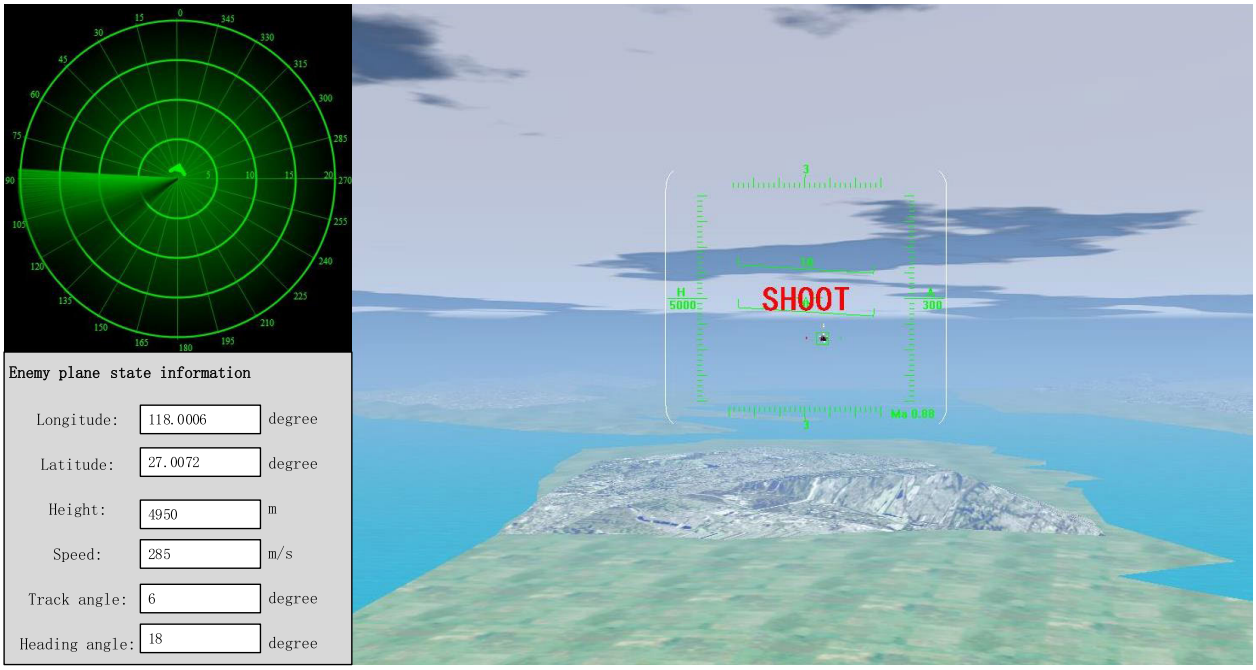


FIGURE 18. Interaction interface of manned aircraft operation simulation module.

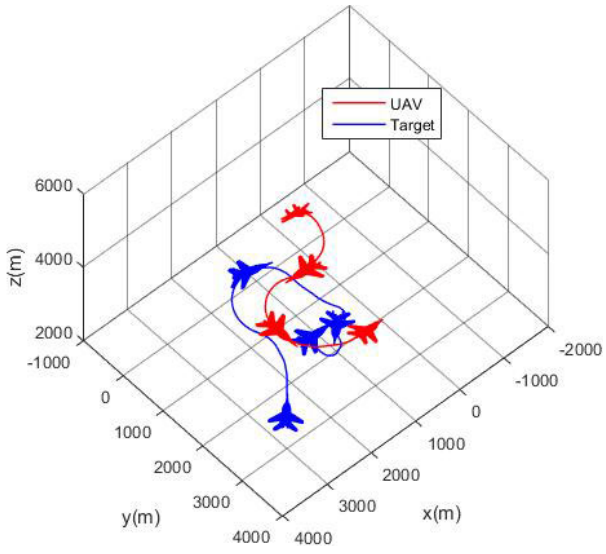


FIGURE 19. Maneuver trajectory of a confrontation before the confrontation policy update.

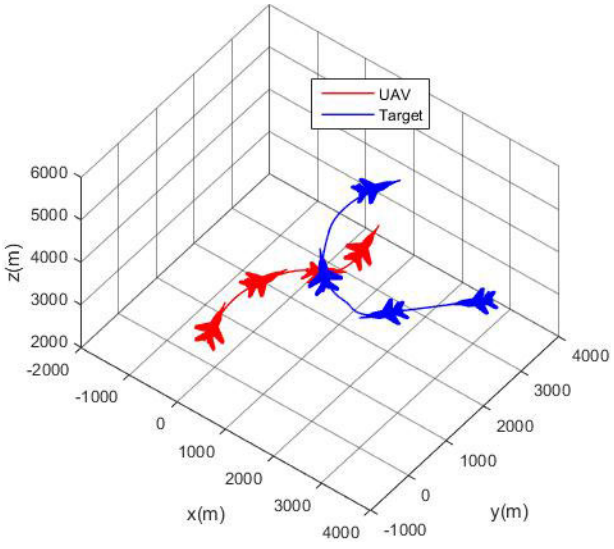


FIGURE 20. Maneuver trajectory of a confrontation after the confrontation policy update.

4) DQN vs DQN

Finally, an interesting exploratory experiment is carried out. In this experiment, UAV and target are set to use the same DQN maneuver decision model to conduct training simulation. Both sides use the basic training model parameters, and 1,000,000 episodes training with the balance initial situation are performed. The difference between the maximum episode value of UAV and target is added as the evaluation index. As shown in Figure 21, in the early stage of training, due to the randomness of model exploration, the index oscillates back and forth around 0. As the training progresses, the amplitude of the oscillation gradually slows down and finally converges to 0, indicating that the policies of the two sides are becoming consistent. As shown in Figure 22, after the training is completed to form a balancing policy, the two sides pursue each other in the confrontation, forming an equilibrium situation in which the advantage cannot be obtained.

From the experimental results, it can be seen that if the target's maneuver policy is deterministic, the DQN model can learn to get a maneuver policy to gain the advantage in air combat, and the simulation gets the unbalanced result,

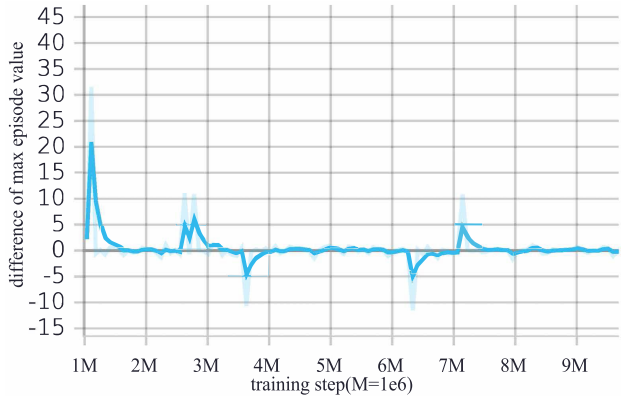


FIGURE 21. **The difference between the maximum episode value of UAV and target.**

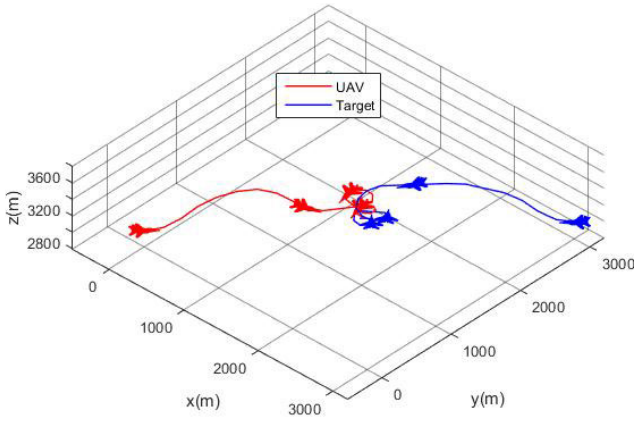


FIGURE 22. **Maneuver trajectory of a confrontation after DQN vs DQN training.**

because the purpose of the reinforcement learning algorithm is to obtain the maximum return value. If both sides use the DQN model, it means that both sides could adopt the same policy. In theory, the two sides will form an equilibrium state.

V. CONCLUSION

Based on the reinforcement learning theory, a maneuver decision model for UAV short-range air combat was established. Improved state dimension of the air combat maneuver decision-making environment made the state description of air combat maneuver decision more realistic, and the action space was expanded more comprehensive.

Aiming at the problem of low learning efficiency and local optimization due to the large state space of air combat, this paper proposed a model training method based on the principle of going to confrontation training from basic training. The simulation results proved that the training method could effectively improve the efficiency of the UAV learning confrontation maneuver policy. It also proved that the UAV short-range air combat maneuver decision model based on deep reinforcement learning could realize the self-learning and update policy until the target was defeated.

Due to the limitation of time and equipment resources, this paper does not conduct more detailed experimental analysis on some issues, such as the impact of the division of the action space on the effectiveness of the decision. If the action space is more detailed, the larger size of the maneuver library will undoubtedly improve the accuracy of the UAV control, but too many output units will weaken the recognition ability of the deep neural network. Therefore, in the case of determining the network structure, the optimal number of actions should be found through a large number of experiments, which can be considered as a problem for future research. In addition, when designing the basic training program, the impact of the specific contents of the basic training program on the learning efficiency of the confrontation training is not analyzed, and only the fact that the basic training can improve the learning efficiency of the confrontation training is proved. In the later stage, the experimental analysis can be continued on the optimization of the basic training design.

REFERENCES

[1] E. Skjervold and Ø. T. Hoelsreter, "Autonomous, cooperative UAV operations using COTS consumer drones and custom ground control station," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2018, pp. 1–6.

[2] Y. Qiming, Z. Jiandong, and S. Guoqing, "Modeling of UAV path planning based on IMM under POMDP framework," *J. Syst. Eng. Electron.*, vol. 30, no. 3, pp. 545–554, 2019.

[3] J. S. Mcgrew, J. P. How, and B. Williams, "Air-combat strategy using approximate dynamic programming," *J. Guid. Control Dyn.*, vol. 33, no. 5, pp. 1641–1654, 2010.

[4] G. Xu, S. Wei, and H. Zhang, "Application of situation function in air combat differential games," in *Proc. 36th Chin. Control Conf. (CCC)*, 2017, pp. 5865–5870.

[5] H. Park, B. Y. Lee, and M. J. Tahk, "Differential game based air combat maneuver generation using scoring function matrix," *Int. J. Aeronaut. Space Sci.*, vol. 17, no. 2, pp. 204–213, 2015.

[6] R.-Z. Xie, J.-Y. Li, and D.-L. Luo, "Research on maneuvering decisions for multi-UAVs Air combat," in *Proc. 11th IEEE Int. Conf. Control Autom.*, Jan. 2014, pp. 767–772.

[7] Z. Lin, T. Minq'an, Z. Wei, and Z. Shcnquun, "Sequential maneuvering decisions based on multi-stage influence diagram in air combat," *J. Syst. Eng. Electron.*, vol. 18, no. 3, pp. 551–555, Sep. 2007.

[8] S. Zhang, Y. Zhou, and Z. Li, "Grey wolf optimizer for unmanned combat aerial vehicle path planning," *Adv. Eng. Softw.*, vol. 99, pp. 121–136, Sep. 2016.

[9] R. E. Smith, B. A. Dike, R. K. Mehra, B. Ravichandran, and A. El-Fallah, "Classifier systems in combat: Two-sided learning of maneuvers for advanced fighter aircraft," *Comput. Methods Appl. Mech. Eng.*, vol. 186, nos. 2–4, pp. 421–437, 2016.

[10] H. Changqiang, D. Kangsheng, H. Hanqiao, T. Shangqin, and Z. Zhuoran, "Autonomous air combat maneuver decision using Bayesian inference and moving horizon optimization," *J. Syst. Eng. Electron.*, vol. 29, no. 1, pp. 86–97, 2018.

[11] Q. Pan, D. Zhou, J. Huang, X. Lv, Z. Yang, K. Zhang, and X. Li, "Maneuver decision for cooperative close-range air combat based on state predicted influence diagram," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Jul. 2017, pp. 726–731.

[12] D. Wang, W. Zu, H. Chang, and J. Zhang, "Research on automatic decision making of UAV based on plan goal graph," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2016, pp. 1245–1249.

[13] W. Yuan, H. Changqiang, and T. Chuanlin, "Research on unmanned combat aerial vehicle robust maneuvering decision under incomplete target information," *Adv. Mech. Eng.*, vol. 8, no. 10, pp. 1–12, 2016.

[14] H.-F. Guo, M.-Y. Hou, Q.-J. Zhang, and C.-L. Tang, "UCAV robust maneuver decision based on statistics principle," *Acta ARMAM*, vol. 38, no. 1, pp. 160–167, 2018.

[15] W. X. Geng, F. Kong, and D. Q. Ma, "Study on tactical decision of UAV medium-range air combat," in *Proc. 26th Chin. Control Decision Conf. (CCDC)*, 2014, pp. 135–139.

[16] F. Li and X. Huaifu, "An UAV air-combat decision expert system based on receding horizon control," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 41, no. 11, pp. 1994–1999, 2015.
[17] W. S. Roger and E. B. Alan, "Neural network models of air combat maneuvering," Ph.D. dissertation, New Mexico State Univ., Las Cruces, NM, USA, 1992.
[18] D. Linjing and Y. Qiming, "Research on air combat maneuver decision of UAVs based on reinforcement learning," *Avionics Technol.*, vol. 49, no. 2, pp. 29–35, 2018.
[19] P. Liu and Y. Ma, "A deep reinforcement learning based intelligent decision method for UCAV air combat," in *Proc. Asian Simul. Conf.* Singapore: Springer, 2017, pp. 274–286.
[20] J. Zuo, R. Yang, Y. Zhang, Z. Li, and M. Wu, "Intelligent decision-making in air combat maneuvering based on heuristic reinforcement learning," *Acta Aeronautica Et Astronautica Sinica*, vol. 38, no. 10, pp. 321168-1–321168-14, 2017.
[21] Z. Xianbing, L. Guoqing, Y. Chaojie, and W. Jiang, "Research on air confrontation maneuver decision-making method based on reinforcement learning," *Electronics*, vol. 7, no. 11, p. 279, 2018.
[22] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, p. 1054, Sep. 1998.
[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.
[24] A. Fred, C. Giro, F. Michael, H. Hinz, and M. Lewis, "Automated maneuvering decisions for air-to-air combat," in *Proc. AIAA Guid. Navigat. Control Conf.* Monterey, CA, USA: AIAA, 1987.
[25] T.-Y. Sun, S.-J. Tsai, Y.-N. Lee, S.-M. Yang, and S.-H. Ting, "The study on intelligent advanced fighter air combat decision support system," in *Proc. IEEE Int. Conf. Inf. Reuse Integr.*, Waikoloa Village, HI, USA, Sep. 2006, pp. 39–44.

QIMING YANG was born in Xining, Qinghai, China, in 1988. He received the master's degree from Northwestern Polytechnical University, Xi'an, China, in 2013, where he is currently pursuing the Ph.D. degree in electronic science and technology. His main research interests are artificial intelligence and its application on control and decision of UAV.

JIANDONG ZHANG was born in Yantai, Shandong, China, in 1974. He received the M.S. and Ph.D. degrees in system engineering from Northwestern Polytechnical University, China.

He is currently an Associate Professor with the Department of System and Control Engineering, Northwestern Polytechnical University. He has published more than 20 refereed journal and conference papers. His research fields and interests include modeling simulation and effectiveness evaluation of complex systems, development and design of integrated avionics system, system measurement, and test technologies.

GUOQING SHI was born in Xi'an, Shaanxi, China, in 1974. He received the M.S. and Ph.D. degrees in system engineering from Northwestern Polytechnical University, China.

He is currently an Associate Professor with the Department of System and Control Engineering, Northwestern Polytechnical University. He has published more than ten refereed journal and conference papers. His research fields and interests include integrated avionics system measurement and test technologies, development and design of embedded real-time systems, and modeling simulation and effectiveness evaluation of complex systems.

JINWEN HU received the bachelor's and master's degrees from Northwestern Polytechnical University, in 2005 and 2008, respectively, and the Ph.D. degree from Nanyang Technological University, in 2013. He was a Research Scientist with the Singapore Institute of Manufacturing Technology, from 2012 to 2015. He is currently an Associate Professor with the School of Automation, Northwestern Polytechnical University. His research interests include multiagent systems, distributed control, unmanned vehicles, information fusion, and process control.

YONG WU was born in Xi'an, Shanxi, China, in 1964. He received the M.S. degree in system engineering from Northwestern Polytechnical University, China.

He is currently a Professor with the Department of System and Control Engineering, Northwestern Polytechnical University. He has published more than 20 refereed journal and conference papers. His research fields and interests include modeling simulation and effectiveness evaluation of complex systems, development and design of integrated avionics systems, system measurement, and test technologies. Prof. Wu received four awards of the national defense science and technology progress in 2004, 2005, and 2011, respectively.