Multi-Agent Air Combat Decision-Making Based on Battlefield Attention Information

Yihuan Wang, Yaofei Ma, Jiangyun Wang, Haitao Yuan,
Meijia Wang and Hanbo Ma

November 7, 2024

Multi-agent Air Combat Decision-making Based on Battlefield Attention Information

Yihuan Wang[1], Yaofei Ma[1], Jiangyun Wang[1], Haitao Yuan[1], Meijia Wang[1] and Hanbo Ma[1]

[1]Beihang University School of Automation Science and Electrical Engineering Beijing, China

***Abstract*—With the rapid development of artificial intelligence and neural networks, deep reinforcement learning has achieved remarkable results in a series of complex sequential decision-making problems. The application of multi-agent reinforcement learning in air combat game scenarios is also booming. In the use of reinforcement learning for multi-agent air combat decision-making, the scalability and transferability of the model have become critical issues. Designing a multi-agent air combat decision-making framework with solid scalability, robustness, and rapid convergence has become a research hotspot in various countries. To address this problem, this paper proposes a multi-agent air combat decision-making framework based on attention mechanism transfer and designs a 2D air combat simulation environment for this framework. The decision-making process of this framework is divided into two stages. First, course learning is carried out in the designed essential air combat environment to enhance the aircraft's combat capability. Then, the trained strategy is transferred to a complex air combat environment for further training. Experiments have shown that this framework has better transferability and robustness.**

***Index Terms*—Air combat, Multi-Agent Reinforcement Learning, Transfer Learning, Curriculum Learning**

I. Introduction

With the development of artificial intelligence technology, multi-agent air combat decision-making technology has become increasingly popular in domestic and foreign military fields [1]. Deep reinforcement learning techniques offer a robust approach to enable the autonomous operational capabilities of UAV swarms. A collection of multi-agent training frameworks [2], such as QMIX, Q-Trans, and MAPPO, have been invented based on reinforcement deep learning. To enhance the learning efficiency and quality of multi-agent collaborative strategies [3], these learning frameworks are designed for multi-agent systems with many basic principles or structural design innovations. Based on the principle of value decomposition, the QMIX network simplifies the value distribution between individual agents and collective agents and simplifies the learning structure [4]. The Q-Trans network introduces a transformation function that allows individual agents to adjust their strategies based on the behaviour of others, effectively capturing complex interactions between agents [5]. MAPPO employs the advantage function to assess the expected benefits of each agent's actions, promoting the development of more effective independent strategies. At the same time, a coordination mechanism maintains consistency across the agent system's policies, balancing independence with coordination [6]. At the same time, Mappo introduces the most important sampling method to enable multiple agents to complete off-policy training efficiently. CLIP is introduced into the loss function to accelerate the exploration of strategies in the training process. Furthermore, The Transformer network has proven advantageous in expressing high-dimensional situations across various applications, including decision-making in multi-aircraft combat scenarios. As the costs of unmanned aerial vehicles (UAVs) decrease, the emergence of large-scale UAV combat scenarios on future battlefields becomes increasingly likely, highlighting the need for effective air combat strategy training [7]. When swarm drones, unmanned robots, and other unmanned technologies are widely used on the battlefield, it is particularly important to design a more open and intelligent UAV suitable for air combat [8]. Therefore, conducting air combat strategy training for large UAVs is important. While a multi-agent learning framework that supports different agent types is available [9], in practice, it is difficult for an agent to learn the optimal air combat decision from scratch due to the complexity of the fixed-wing aircraft model and the large exploration space.

To efficiently and reliably train a larger-scale UAV collaborative air combat strategy, this paper suggests transferring knowledge from a small-scale air combat model (2 vs. 2) to a larger-scale model (5 vs. 5). This approach leverages the understanding of the global situation developed in the small-scale model by reusing its Transformer module in the larger context. As a result, strategy convergence can be achieved with minimal training, greatly enhancing the stability and efficiency of the large-scale air combat model.

1)Design a simulation platform that can quickly realize multi-agent air combat simulation and interaction of different scales and different combat difficulties, and design basic scenarios for training using course learning.

2)Transfer learning is introduced to transfer the aircraft strategies of basic scenarios, and training and comparison are carried out in more complex scenarios. Practice has proved that this method can improve the model's convergence speed, winning rate, and robustness.

II. Problem overview

A. *Modeling Air Combat Scenarios*

The multi-aircraft air combat scenario discussed in this article refers to within visual range (WVR) engagements, where tactical maneuvers are employed to gain a superior position to fire upon the adversary. In these scenarios, the primary weapons used are short-range air-to-air missiles and machine guns, emphasizing the importance of close-range combat and real-time decision-making. The opposing sides in these air combat simulations are designated as the red side and the blue side. In this context, the red side represents our forces, with the tactical strategies being derived through advanced learning algorithms. On the other hand, the blue side represents the adversary, whose responses are guided by script-based tactics that vary in difficulty. In this article, the scenario involves red side aircraft that employ tactics learned through machine learning or other adaptive processes, while the blue side represents the opposing force, utilizing pre-determined tactical responses based on various levels of difficulty. These responses range from random action sampling to more sophisticated air combat scripts and simulated self-play techniques. The blue side's tactics are designed to challenge the red side in multiple ways, testing the adaptability and effectiveness of learned strategies. The focus of the research is on a multi-aircraft air combat environment within visual range, which uses tactical maneuvers as a key factor in achieving the firing advantage. The primary weapon systems in use are short-range missiles and machine guns, emphasizing the need for close-quarters engagement. In this framework, the red side's aircraft apply learned strategies, while the blue side represents the opponent and uses various difficulty-based tactics that include random action sampling, predefined air combat strategies, and virtual simulations against itself. The mathematical modeling of the aircraft's movement within a two-dimensional plane is provided as follows:

$$\begin{cases} \dot{x}{=}u\cos\psi{-}v\sin\psi \\ \dot{y}{=}u\sin psi{+}v\cos psi \\ \dot{u}{=}r{\cdot}v{+}\frac{1}{m}F_x \\ \dot{v}{=}-r{\cdot}u{+}\frac{1}{m}F_y \\ \dot{r}{=}\frac{M_z}{I_z} \\ \Psi{=}r \end{cases} \tag{1}$$

Based on the findings presented in the literature [12], a two-dimensional air combat aircraft model is developed, capable of deploying both air-to-air missiles and aircraft cannons as its primary weapons systems. The key characteristics of this model are depicted in Fig. 1 and are outlined as follows: - The heading angular velocity falls within a range of $[0, 5]$ degrees per second. - The aircraft's speed can vary between 200 and 500 meters per second. - For short-range missile systems, the aircraft can carry up to 8 missiles, with an effective range between 0 and 11 kilometers. The missile launch cone is centered along the aircraft's longitudinal axis, with an allowable angle range of $[-60, 60]$ degrees. Each missile has a single-shot hit probability of 0.75. - The aircraft's cannon system has a maximum ammunition capacity of 400 rounds. The effective firing range for the cannon is between 0 and 2 kilometers, with a firing cone angle range of $[-60, 60]$ degrees. Additionally, the cannon can be continuously fired for up to 200 seconds.

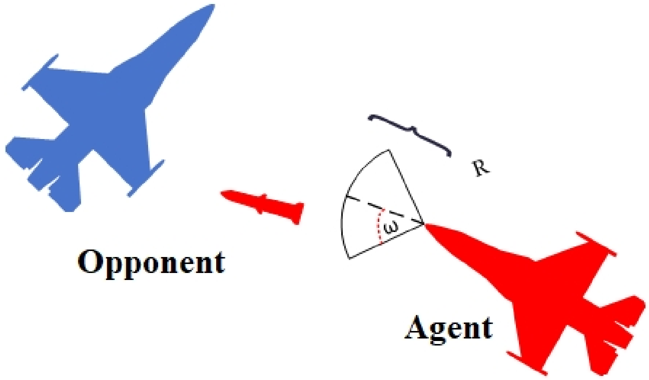


Fig. 1. Aircraft Model

B. Reinforcement Learning for Multi-Agent Air Combat

In this scenario, multi-aircraft air combat is modeled as a Partially Observable Markov Game (POMG). From the Red Army's perspective, the air combat process is defined by a six-tuple $(S, O^i(i \in n), A^i(i \in n), P, R, \gamma)$. Here, S represents the overall air combat state, which is the combined state of both the red and blue forces. O^i denotes the observable state for the i-th red aircraft, where $n \geq 1$ corresponds to the number of observable red aircraft. A^i defines the action space available to the i-th red aircraft, while the joint action $A = A^1 \times \cdots \times A^n$ signifies the collective actions of all red aircraft at a given decision-making step. The state transition probability is represented by $P : S \times A \to \triangle S$, describing the likelihood of transitioning from one state s to another state $s^{'}$ after the Red team performs a joint action a in state $s \in S$. The reward function is given by $R : S \times A \times S \to \mathbb{R}$, which defines the immediate reward the red aircraft group receives upon taking action in state s and transitioning to the new state $s^{'}$. The discount factor $\gamma \in [0, 1]$ is used to calculate and accumulate long-term rewards over time. In this model, air combat is treated as a typical sequential decision-making problem where aircraft continuously observe the environment, make tactical decisions, execute those decisions (within a simulation environment), and accumulate rewards based on their actions. The POMG framework defines the observation space for each aircraft, which is composed of three main components: 1) Self-information, expressed as:

$$\begin{aligned} O_{t,i}{=}[x, y.v, \psi, Aoff_{i,j}, AA_{i,j}, ATA_{i,j}, \\ d_{i,j}, d_{i,f}, Aoff_{i,f}, AA_{i,f}, ATA_{i,f}, s_r, c_a] \end{aligned} \tag{2}$$

In this context, the subscript i refers to the current time step, while $i\,(i \in n)$ represents the i-th aircraft of the Red side, with n being the total number of observable Red aircraft. Similarly, $j\,(j \in m)$ denotes the j-th aircraft of the Blue side, where m is the number of observable Blue aircraft. The subscript f refers to friendly aircraft. The variables x, y represent the aircraft's 2D position, v is the current speed, s_r indicates missile readiness, and c_a shows whether the aircraft is firing. The heading angle of the aircraft is denoted by ψ. Additionally, $Aoff_{i,j}$ represents the angle of attack, which is the angle between the velocity directions of the Blue aircraft j and Red aircraft. $AA_{i,j}$, or azimuth, is the angle between the vertical axis of the Blue aircraft and the distance vector between the Red and Blue aircraft. $ATA_{i,j}$ refers to the radar tracking angle, defined as the angle between the distance vector of the Blue aircraft relative to the Red aircraft and the vertical axis of the Red aircraft. Other subscripts follow the same pattern. For further details, refer to Fig. 2.

2) Friendly information, expressed as:

$$O_{t,f}{=}[v,\psi,Aoff_{f,i},AA_{f,i},ATA_{f,i},d_{f,i},s_r,c_a] \tag{3}$$

3) Opponent information, expressed as:

$$O_{t,j}{=}[v,\psi,Aoff_{j,i},AA_{j,i},ATA_{j,i},d_{j,i},s_r,c_a] \tag{4}$$

4) Combine the observation information of all red aircraft to get the global observation:

The action space is defined as $[h,e,c,b]$, where each component represents a specific control parameter for the aircraft: $h \in [-6,\ldots,6]$ is a discrete action space that adjusts the aircraft's heading angle ψ within the range of $[-180,180]$ degrees per second. $e \in [0,\ldots,9]$ is a discrete action space that controls the aircraft's velocity v, ranging from $[0,500]$ meters per second. $c \in [0,1]$ indicates whether or not to fire the aircraft's cannon. $b \in [0,1]$ determines whether to launch a missile. The reward function is designed to reflect advantageous combat scenarios, such as positioning behind the opponent's tail for an optimal attack. To encourage the agent to improve its attack efficiency, factors such as the azimuth angle (AA), distance (d), radar tracking angle (ATA), and the remaining cannon rounds and missiles are considered in the reward structure. The function incentivizes actions that increase the likelihood of successful engagement, as expressed in the following equation:

$$r_a{=}AA{+}d{+}\frac{C_{remain}}{C_{max}}{+}\frac{S_{remain}}{S_{max}} \tag{5}$$

In this equation, C_{max} and S_{max} represent the maximum number of cannon rounds and missiles the aircraft can carry, respectively. Meanwhile, C_{remain} and S_{remain} indicate the remaining cannon rounds and missiles available after engaging and defeating the opponent.

$$O_{t,full}{=}O_{t,i} \cup O_{t,f} \cup O_{t,j} \tag{6}$$

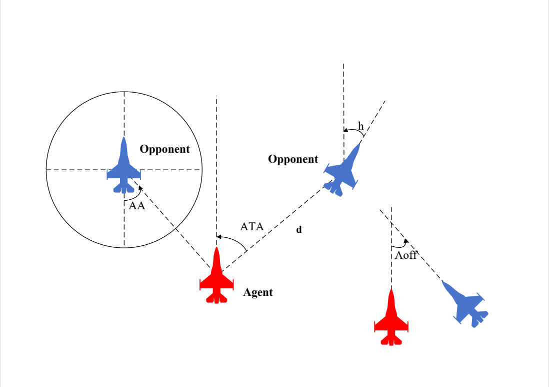


Fig. 2. Air Combat Model

III. METHOD

A. Curriculum Learning

Curriculum learning in reinforcement learning is a transfer learning approach that accelerates training by progressively increasing task difficulty. It involves creating a sequence of tasks similar to the final target, with strategies transferred across tasks to improve learning speed and performance. In multi-agent air combat decision-making, the red side (trained through reinforcement learning) faces varying blue-side strategies. Curriculum learning is applied to gradually increase the complexity of blue-side tactics. Training begins with simple strategies, and as the red side's model converges, the difficulty of blue-side strategies is raised to enhance the red aircraft's combat effectiveness and adaptability. This study defines four levels of blue-side strategies.

- **L1:** static, the blue aircraft is static.
- **L2:** random, the blue aircraft takes random actions within the allowed action space.
- **L3:** script, the blue aircraft engages the nearest Red aircraft and after engaging, moves away from the red aircraft
- **L4:** self-play,the red aircraft adopts the blue strategy of L3 training to confront the opponent.

B. Neural Network Architecture

Our network structure is based on the Actor-Critic network. The Network structure of the proposed method mainly consists of three components, as shown in Fig. 3:

1) **Actor network structure:** The actor parameterized by $\theta : \pi_i^\theta : o_i \rightarrow a$,consists of the embedding layer, Long Short-Term memory(LSTM) and Multi-head attention network.It takes partial observation o_i as input and outputs action values for making decisions. In the actor network, firstly, As a high-dimensional sparse vector, the observation space o_full is mapped to a D-dimensional real vector, which concludes the battlefield situation representation information through the embedding layer. LSTM is used to fuse the environmental embedding e_i^{t-1}at last time $t{-}1$ and the interaction embedding h_i^t at current time t,yielding the environmental embedding e_i^t at current time t, and then, the environmental embedding e_i^t through a multi-head attention mechanism network to obtain a vector att_i^t containing battlefield situation attention. Lastly, as shown in the equation:

$$f_i^t{=}[e_i^t;att_i^t] \tag{7}$$

The attention vector is concatenated with the embedding of the original environment. Then, f_i^t is fed into a policy network with FC shared layer and FC layer, which outputs action values of aircraft i.

2) **Critic network structure:** The critic parameterized by $\phi : v_i^{\phi} : s_i \rightarrow \mathbb{R}$, is similar to actor network structure. The critic network takes the aircraft global state s_i of aircraft i as inputs and outputs a scalar value for the actor training.

3) **Actor-critic network parameter sharing** The training of policies for homogeneous agents can be made more efficient through parameter sharing. In this Actor-critic network, the actor-network embedding f_i^t and critic network embedding c_i^i as inputs go through a shared FC layer to obtain situation information containing more cooperative knowledge.

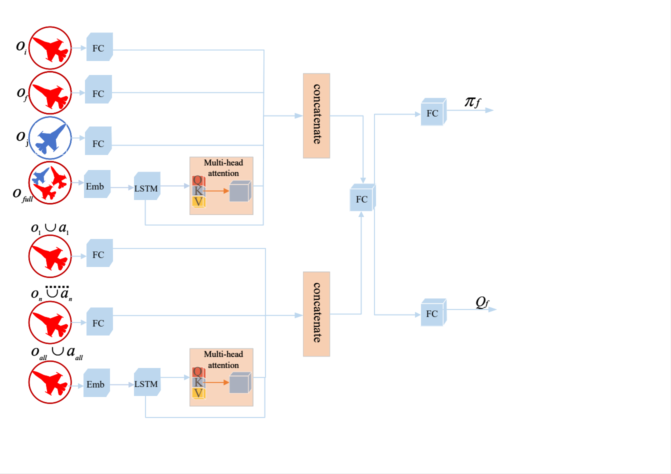


Fig. 3. Neural Network Architecture

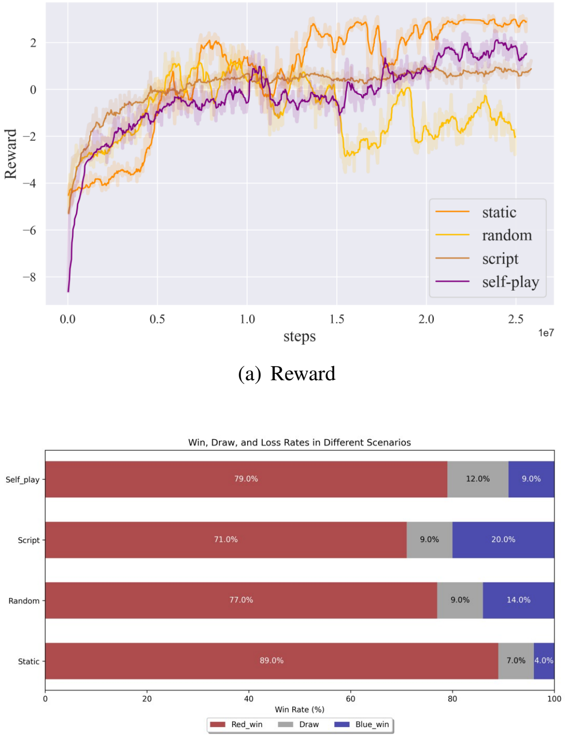
C. Transfer Learning

In the structure shown in Fig 3, the Transformer module processes global observation data to better capture global features and support decision-making for individual aircraft. The idea is that if the Transformer module has effectively learned to represent situational features in two vs. two air combat, reusing its parameters could speed up learning for larger-scale scenarios. Based on this, the paper reuses Transformer parameters from two vs. two scenarios for five vs. five air combat to enhance training efficiency without compromising quality.

IV. EXPERIMENTS

A. 2 vs. 2 curriculum training results

Training takes place within the air combat simulation environment described earlier, which allows visualization of aircraft flight paths and weapon usage. A simulation cycle concludes when either the time limit is reached or all aircraft on one side are destroyed. Aircraft are considered destroyed if hit by a shell or missile, or if they collide with the map boundary. At the start of each episode, aircraft positions and orientations are randomly assigned to opposite sides of the map. In the 2 vs. 2 training scenario, the map size is set to 5 kilometers. The red side is equipped with five air-to-air missiles and 200 cannons, while the blue side carries eight air-to-air missiles and 400 cannons. The blue side trains using four levels of strategies. The blue team's model, trained using the Script strategy, is then loaded into the simulation environment, where the self-play strategy is used for further training. This model is later compared to a model trained from scratch, as shown in Fig 4.



(a) Reward

(b) Winning rate

Fig. 4. 2 vs. 2 air combat training results

In the 2 vs. 2 curriculum training, the red plane carries 200 cannon shots and 5 missiles for each episode. The blue plane carries 400 cannon shots and 8 missiles. In this training, the opponent's combat capability is stronger than our intelligent body. In order to win this air battle, our aircraft's strategy will be more intelligent. As the blue side's strategy level increases, we increase the simulation time from L1's $T = 200$ by $\triangle t = 50$. Start training the basic model from the static strategy and gradually increase the blue side's strategy level to L4.wining rate. First, let the two planes of the blue side adopt the simplest static strategy, train the red side model, and test the winning rate of the trained model (randomly initialize the positions of the red and blue sides 100 times, and then calculate the winning rate of the red side), and then gradually upgrade the blue side strategy. The trained old red strategy is gradually loaded into the air combat environment with a high-level blue strategy for course learning.

It can be seen from Fig. 4 that the red team's intelligent agent, trained through the curriculum learning, demonstrated good combat capabilities. Fig. 5 shows the Red side's combat trajectory when the Blue side adopts self-play strategies. From Fig. 4, we can see that the red team has the highest win rate for the simplest static scene. When the blue team starts to execute the more difficult random motion strategy, the red team's win rate decreases, showing that the model strategy trained from the static scene is relatively simple. As the difficulty of the Blue Team's strategy increases, the red team's win rate gradually increases. It can be seen that when the blue team is executing the L2-L4 strategy, the course learning is effective for model training.

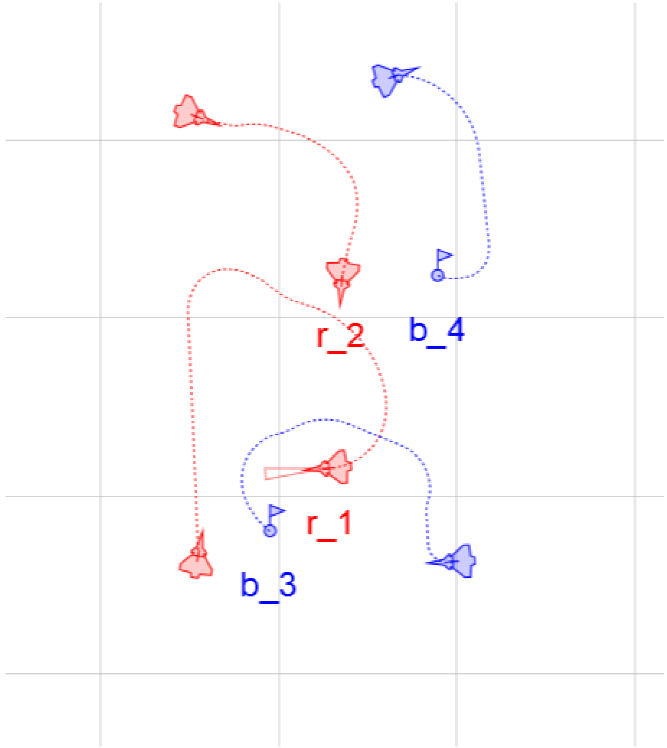
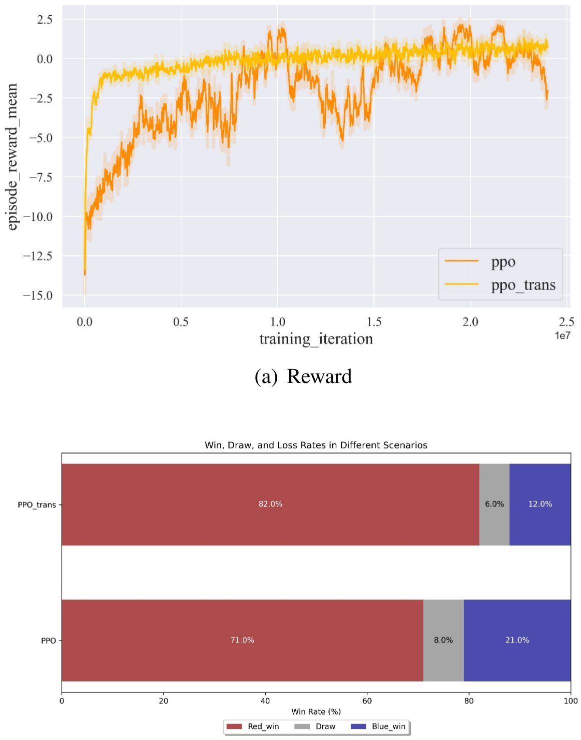


Fig. 5. 2 vs. 2 Air Combat Trajectory

B. Transfer learning based on attention parameter sharing

In this study, we designed a 5 vs. 5 air combat simulation task utilizing the aforementioned simulation platform. We reconstructed the parameters of the Fully Connected (FC) and Embedding layers in the decision network while reusing the parameters from the LSTM and Transformer networks. The model trained on the 2 vs. 2 scenario was then adapted for the 5 vs. 5 air combat task, which covers a combat range of 80x80 kilometers. The number of weapons carried by both the red and blue sides remained consistent with the previous section, allowing for a comparison with curves generated from training from scratch. The resulting reward and win rate curves are illustrated in Fig 6. From the reward curve shown in Fig 6, it is evident that the model trained with parameter sharing from the 2 vs. 2 scenario converges more quickly and exhibits a smoother reward trajectory compared to the model trained from scratch. This indicates enhanced convergence speed and robustness. Additionally, analyzing Fig 6, we found that the winning rate of the model trained from scratch was significantly lower than that of the transfer-learned model, which achieved a win rate as high as [insert specific percentage here]. This demonstrates the superior combat capability of the transferred model. Overall, the migration of the strategy model from the 2 vs. 2 air combat scenario to the more complex 5 vs. 5 environment effectively enhances both the convergence speed and robustness of agent training, resulting in improved combat capabilities and higher win rates for the red aircraft.



(a) Reward

(b) Wining rate

Fig. 6. 5 vs. 5 air combat transfer training results

Select the strategy model obtained by transfer training, randomly initialize the red and blue square position information and weapon information, load the model, and obtain the trajectory diagram:

As shown in Fig 7, the curves of the red and blue planes represent the motion trajectories of the red and blue planes from the beginning to the end of the simulation. It can be found from the trajectory diagram that, at the beginning of the simulation, due to the long distance between the two planes, the blue side will execute the motion command to approach the red plane under the control of the script. In contrast, the red side now has no detection range of the blue plane. There are only friendly forces, and the Red aircraft will randomly select actions in the specified action space to

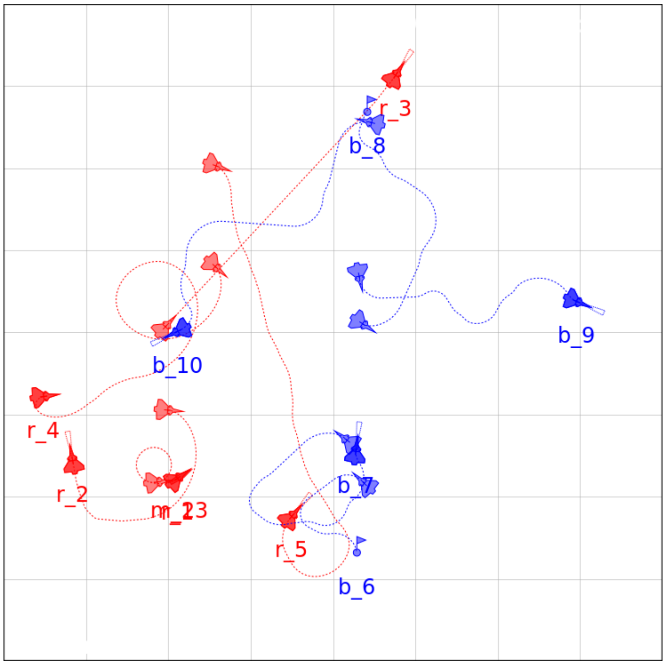


Fig. 7. 5 vs. 5 Air Combat Trajectory

explore the surrounding environment (such example, When the simulation step length increases, the blue side will move to the detection range of the red side aircraft. When the blue-side aircraft appears within the detection range of the red side, it will take action to approach the blue-side aircraft under the guidance of the distance reward function (such as direct pursuit and offensive rotation). Once within range, the red-headed and blue-square aircraft engage in a close dogfight, choosing the dominant position to attack until the aircraft is dead or the simulation ends (for example, attacking the enemy's rear). The curve analysis shows that after training in the framework of this paper, the red intelligence can learn effective tactical attack actions and destroy the blue aircraft.

V. CONCLUSIONS

This paper focuses on multi-agent air combat decision-making utilizing battlefield attention information within a two-dimensional simulation environment. We employed a curriculum learning approach to train a 2 vs. 2 air combat model, where the blue side employs various strategic levels. Subsequently, this model was transferred to the more complex 5 vs. 5 air station scenario using transfer learning, allowing for a comparison with a model trained from scratch. Results indicate that this approach significantly enhances the winning rate, convergence speed, and robustness of the multi-agent air combat decision-making process. Additionally, we analyzed the multi-head attention architecture of each aircraft during the simulations to extract attention information relevant to the current combat situation, confirming that the transfer learning method aligns more closely with real-world air combat scenarios.

In future work, we aim to combine this framework with a six-degree-of-freedom high-precision aircraft countermeasure environment to validate the effectiveness of our approach in more complex environments, and, we will further increase the number of agents in the new adversarial environment and the complexity of the blue square agent strategy, and update our transfer learning approach. Implement transfer decision learning in the complex adversarial environment of larger scale agents.

REFERENCES

[1] R. Li and H. Ma, "Research on UAV Swarm Cooperative Reconnaissance and Combat Technology," *2020 3rd International Conference on Unmanned Systems (ICUS)*, Harbin, China, 2020, pp. 996–999.

[2] Z. Ren, D. Dong, H. Li, and C. Chen, "Self-Paced Prioritized Curriculum Learning With Coverage Penalty in Deep Reinforcement Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2216–2226, Jun. 2018.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level Control Through Deep Reinforcement Learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Dec. 2015.

[4] T. Zhang, *et al.*, "Multi-UAV Cooperative Short-range Combat via Attention-based Reinforcement Learning using Individual Reward Shaping," *International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, pp. 13737–13734, 2022.

[5] S. Gronauer and K. Diepold, "Multi-agent Deep Reinforcement Learning: a Survey," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 895–943, Feb. 2021.

[6] S. Pateria, B. Subagdja, A. Tan, and C. Quek, "End-to-End Hierarchical Reinforcement Learning With Integrated Subgoal Discovery," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7778–7790, Dec. 2022.

[7] X. He, N. Jing, and C. Feng, "Air Combat Maneuver Decision Based on MCTS Method," *Journal of Air Force Engineering University (Natural Science Edition)*, vol. 18, no. 5, pp. 36–41, Mar. 2017.

[8] H. Zhang, C. Huang, Z. Zhang, X. Wang, B. Han, Z. Wei, *et al.*, "The Trajectory Generation of UCAV Evading Missiles Based on Neural Networks," *Journal of Physics Conference Series*, vol. 1486, no. 2020, pp. 22025–22040, Jan. 2020.

[9] W. Kong, D. Zhou, Y. Du, Y. Zhou, and Y. Zhao, "Reinforcement Learning for Multiaircraft Autonomous Air Combat in Multisensor UCAV Platform," *IEEE Sensors Journal*, vol. 23, no. 18, pp. 20596–20606, Sept. 2023.

[10] J. Zhang, Y. Yu, L. Zheng, Q. Yang, G. Shi, and Y. Wu, "Situational Continuity based Air Combat Autonomous Maneuvering Decision-making," *Defence Technology*, pp. 66–79, Jun. 2022.

[11] Y. Wang, T. Ren, and Z. Fan, "Autonomous Maneuver Decision of UAV Based on Deep Reinforcement Learning: Comparison of DQN and DDPG," *2022 34th Chinese Control and Decision Conference (CCDC)*, Hefei, China, 2022, pp. 4857–4860.

[12] D. Hu, R. Yang, J. Zuo, Z. Zhang, J. Wu, and Y. Wang, "Application of Deep Reinforcement Learning in Maneuver Planning of Beyond-Visual Range Air Combat," *IEEE Access*, vol. 9, pp. 32282–32297, Feb. 2021.

[13] J. Chai, *et al.*, "A Hierarchical Deep Reinforcement Learning Framework for 6-DOF UCAV Air-to-air Combat," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 5417–5429, Dec. 2023.

[14] J. Guo, Z. Wang, J. Lan, B. Dong, R. Li, Q. Yang, and J. Zhang, "Maneuver Decision of UAV in Air Combat based on Deterministic Policy Gradient," *IEEE 17th International Conference on Control Automation*, Naelps, Italy, 2022, pp. 243–248.

[15] W. Yuan, Z. Xiwen, Z. Rong, T. Shangqin, Z. Huan, and D. Wei, "Research on UCAV Maneuvering Decision Method based on Heuristic Reinforcement Learning," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1477078–1477086, Jul. 2022.

[16] Z. Fan, Y. Xu, Y. Kang, and D. Lou, "Air Combat Maneuver Decision Method based on A3C Deep Reinforcement Learning," *Machines*, vol. 10, pp. 1033–1045, May 2022.

[17] Q. Yang, J. Zhang, G. Shi, J. Hu, and Y. Wu, "Maneuver Decision of UAV in Short-range Air Combat based on Deep Reinforcement Learning," *IEEE Access*, vol. 8, pp. 816–831, May 2020.