Cooperative Formation Control of USVs and UAVs Based on Reinforcement Learning

Ting Wu[1], Linqi Ye[1], Xianglong Li[2], Yan Peng[1]

[1] Shanghai University, Shanghai 200444, China
wu_ting@shu.edu.cn, yelinqi@shu.edu.cn, pengyan@shu.cn

[2] Beijing University of Chemical Technology, Beijing 100029, China
lixianglong@buct.edu.cn

Abstract. We consider the scenarios that multi-agent cooperatively compete the task of collision-free formation cruise in a specific region. Considering mission complexity and real world constraints, we propose reinforcement learning-based solution for USVs and UAVs cooperative formation. Firstly, based on the curriculum learning, the complex formation control task is decomposed into two-stage training. In the first phase, the USV team is trained with PPO algorithm to track the target moving along a predetermined trajectory and avoid obstacles under the interference of waves. Subsequently, in the second stage, UAV team is trained with similar method. When UAV team is trained, the control strategies of USV team are fixed to the neural network obtained in the first stage. Combining with partial observable information, We design the reward function to make the USVs and UAVs learn policy and maintain a stable linear formation during movement. We validated the effectiveness of the proposed method with two-stage-simulation of USVs and UAVs in Unity environment. Compared to traditional control methods, the proposed method enables the agents to learn effective strategies by interacting with the environment through a relatively simple training process without accurate mathematical model. This result simplifies the complexity of formation control and provides an easier solution for multi-agent formation control.

Keywords: Multi-agent reinforcement learning · Formation control · Proximal policy optimization · Curriculum learning · Leader-follower architecture.

1 Introduction

Compared to signal agent, multi agents could complete more complex assignments, and have been widely applied to various fields, especially in industry and military. Even though there are abundant works and mature theories for multi-agent system problems, traditional control does not perform well when facing with nonlinear and higher-order model control problems, especially in dynamic environments. One of the difficulties of traditional control is that, it is intricate for traditional control method to construct the accurate models. Therefore, reinforcement learning is proposed to deal with the complicated nonlinear systems, especially for addressing the intractability in real world and solutions of Hamilton–Jacobi–Bellman (HJB) [1]. Additionally, various works and approaches based on reinforcement learning performed well in diverse assignments, such as multi-agent path planning [2], collision avoidance [3], and cooperative formation control [4].

Multi-agent cooperative formation control has become a research hotspot in recent years, because of increasing economical efficiency and widespread application. To complete cooperative formation, multiple agents move towards a specific goal or maintain geometry under constraints, such as obstacle avoidance [5, 6] with learned strategies. Leader-follower [7–9], behaviour-based method [10], virtual structure [11] and consensus-based [12] approaches are typical frameworks of formation control.

Compared to the other methods, leader-follower architecture's principle is more concise and it has been widely applied. Moreover, such architecture could alleviate computational burden, as long as the conditions of the leader are determined, it is sufficient for the follower to track the state or trajectory of the leader [13]. Additionally, there are massive efficient researches based on leader-follower framework. [14] combined a random barking mechanism and deep reinforcement learning, and designed reward function avoiding local optimum to achieve USV following a virtual leader rapidly. [15] considered the constraints in real world and designed a leader to identity the region that to be sampled and assigned the regions to followers, and then followers could optimize the policy to detect a collision-free path and maximum the information gain independently by combining Bayesian optimization and Monte Carlo simulation. [16] constructed a leader-follower model, treated the formation control problem as an optimal output regulation problem, and utilized off-policy reinforcement learning to derive a solution to the discounted performance function, which allowed the tracking error of the formation to converge to zero. [17] proposed an optimized leader-follower formation control and simplified reinforcement learning for nonlinear multi-agent system, with identifier-critic-actor framework. The proposed RL updating laws derived from negative gradient of positive function, which approximate HJB, so that simplified the solution. [18] addressed optimal containment of heterogeneous multi-agent system while followers with unknown dynamics tracked dynamic leader through a model-free reinforcement learning-based on-line ARE learning method.

Proximal Policy Optimization Algorithm has been proved to be effective and has been commonly used in multi-agent cooperation combined with few improvement [19], or other methods [20]. [21] had demonstrated that compared to off-policy, PPO algorithm could obtain competitive sample efficiency and better-performance in multi-agent cooperative benchmarks, only with few modification. [22] trained attitude control of inner control loop with Deep Deterministic Policy Gradient, Trust Region Policy Optimization, and Proximal Policy Optimization separately, and obtained the conclusion that the performance and accuracy of

PPO was the best compared to the other two algorithms. [23] proposed multi-agent proximal policy optimization, combining the global information of critic network and actor network to achieved the cooperative task and trained with course learning to improve the generalization.

Given the advantage and effectiveness of PPO and leader-follower framework, a phased reinforcement learning based on leader-follower framework method is proposed for USVs and UAVs collision-free cooperative formation while maintaining linear formation respectively. The effectiveness of the method is verified by simulation, and agents are trained with PPO. Firstly, we separate the USV team and UAV team, and set the leader and followers for both teams. Then, we leverage the curriculum learning [24, 25] to train the agents to maintain desired formations. In the first phase of training, we train USV team to track the target moving along the preset trajectory under the disturbance of waves, and the ships keep a safe distance to avoid collision and maintain a linear formation, simultaneously. In the second phase of training, the similar method is leveraged to train the UAV team to track the target while maintaining the linear formation, which is parallel to USV's. Finally, we validate the effectiveness of our approach in unity with self-designed scene.

2 Related Work

2.1 Reinforcement Learning and Markov Decision Process

Markov decision process is the theoretical framework of reinforcement learning, which could be represented by a tuple$\{S, A, P, R, \gamma\}$, where S represents the environment state observed by the agent, A delegates the action space. P is the probability of state transition under the situation that agent implements the action $a \in A$. R denotes the cumulative reward, and $\gamma \in (0, 1]$ represents the discount of reward [26, 27]. The propose of agent is to learn a policy that maximize the long-term reward. The information obtained by interacting with the environment, decision-making with unknown transition function and reward-design for optimize the expected performance are three critical factors of reinforcement learning. The action decision is relevant to the current state but not to the historical state. When there are multiple agents, the mentioned progress could be extended to Markov Game. Based on the aforementioned theory, the the decision-making of multiple agents would be interfered by others.

2.2 Proximal Policy Optimization Algorithm

Compared to policy gradient and trust region methods, proximal policy optimization performances better through alternating between data sampling and optimizing surrogate objective. Generally, PPO algorithm is based on policy gradient and constructs the loss function with KL penalty coefficient and clip surrogate objective. PPO through clip surrogate objective compares the current probability ration to the old one, which are sampled with current policy and old policy separately and decide whether it needs to be clipped through comparing to a preset threshold value, so that PPO ensure the iterative efficiency increases the entropy reward and utilize the network that shares parameters by combining policy gradient and value function [28, 29]. In this paper, ML-Agents module integrated in Unity was leveraged to invoke PPO algorithm for reinforcement learning training.

3 Method

3.1 USV Motion Model

Simplified three-degree-of-freedom motion model of USV could be described as:

$$\dot{\eta} = R(\psi)\, v \tag{1}$$

where $\eta = [x, y, \psi]^T$ denotes position vector of USV in inertial frame XOY, ψ is the heading angle, $v = [u, v, r]^T$ denotes the state of USV in the body-fixed frame, u, v and r are the forward, transverse and yaw velocity respectively.

$$R(\psi) = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

$R(\psi)$ denotes the rotation transformation matrix from body-fixed frame to inertial frame.
And the motion model of USV could be described as:

$$M\dot{v} = -C(v)\, v - D(v)\, v - g(v) + \tau \tag{3}$$

$$M = \begin{bmatrix} m_{11} & 0 & 0 \\ 0 & m_{22} & m_{23} \\ 0 & m_{32} & m_{33} \end{bmatrix} C = \begin{bmatrix} 0 & 0 & c_{13} \\ 0 & 0 & c_{23} \\ -c_{13} & -c_{23} & 0 \end{bmatrix} D = \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & d_{23} \\ 0 & d_{32} & d_{33} \end{bmatrix} \tag{4}$$

M, C and D denote the inertia, Coriolis force centripetal, damping matrix, and g is unknown dynamics. And there are $c_{13} = -m_{22}v - \frac{1}{2}(m_{23} + m_{32})\, r$, $c_{23} = m_{11}u$.

Acceleration To avoid the complexity of modelling, the forward force of the USV is shown as follows: Let I_v, I_h denote acceleration signal and horizontal input signal, which obtained from vertical and horizontal input, respectively. And the inertia factor J, turning factor K, acceleration torque factor T_a and turning torque factor T_s are 1,000, 500, 20 and 20 respectively. We applied the local coordinate forces $J \times I_v$ to the boat, so that it could move forward in its local coordinate system.

Torque on X,Y,Z axes The torque of the x, y and z directions in local coordinate system are: $I_v \times (-T_a)$, $I_h \times K$, and $I_h \times (-T_s)$.

3.2 UAV Motion Model

The six-degree-of-freedom motion model is shown as follows:

$$\dot{\zeta} = S\left(\theta,\psi,\varphi\right)\lambda \tag{5}$$

And $\zeta = \left[x, y, z\right]^T$ and $\lambda = \left[u, v, w\right]^T$ denote the position vector of UAV and state vector of UAV in body-fixed frame, where u, v, w denote the forward, lateral and vertical speed, respectively.

$$S\left(\theta,\psi,\varphi\right) = \begin{bmatrix} \cos\theta\cos\psi & \cos\psi\sin\theta\sin\varphi - \sin\psi\cos\varphi & \cos\psi\sin\theta\cos\varphi + \sin\psi\sin\varphi \\ \cos\theta\sin\psi & \sin\psi\sin\theta\sin\varphi + \cos\psi\cos\varphi & \sin\psi\sin\theta\cos\varphi - \cos\psi\sin\varphi \\ -\sin\theta & \sin\varphi\cos\theta & \cos\varphi\cos\theta \end{bmatrix} \tag{6}$$

$S\left(\theta,\psi,\varphi\right)$ denotes the rotation matrix, where φ, θ, ψ are the pitch, yaw and roll angle.

To simplify the modelling, aircraft's local coordinate force in forward direction obtained in flight is determined by: $2.5 \times T \times a$, where the T is the engine thrust of magnitude 10^5 in this article, and a denotes the boost factor, and when it is 1 for no boost and 2 for acceleration. To accelerate the training process and prevent stagnation, the boost is greater than zero in the experiment, and simplified the motion function are shown below:

Pitch:

$$x = x_c + \Delta\varphi \times \Delta T \times u \tag{7}$$

Yaw:

$$y = y_c + \Delta\theta \times \Delta T \times v \tag{8}$$

Roll:

$$z = z_c + \Delta\psi \times \Delta T \times w \tag{9}$$

where u, v, w are the pitch, yaw and roll speed, and we propose they are all 100 meters per second. x_c, y_c and z_c represent the current pitch, yaw and roll, respectively. And φ, θ, ψ are the pitch, yaw and roll angle. $\Delta\varphi$, $\Delta\theta$ and $\Delta\psi$ are the smooth increment of the pitch, yaw and roll, severally, and they would transition from current value to the corresponding action space in twice of time step ΔT, which is 0.02 in this paper. Besides, if pith and roll are greater than 180°, $\varphi = \varphi - 360$ and $\psi = \psi - 360$. Additionally, we restricted the pitch and roll in the range of $[-\pi/2, \pi/2]$.

3.3 Observation Space

The observations of USV and UAV team are the current relative distance to target, the last time relative distance to target and the current and last time relative direction to target. To ensure that the observations are within the detection range of the observer, we narrow down the aforementioned observations.

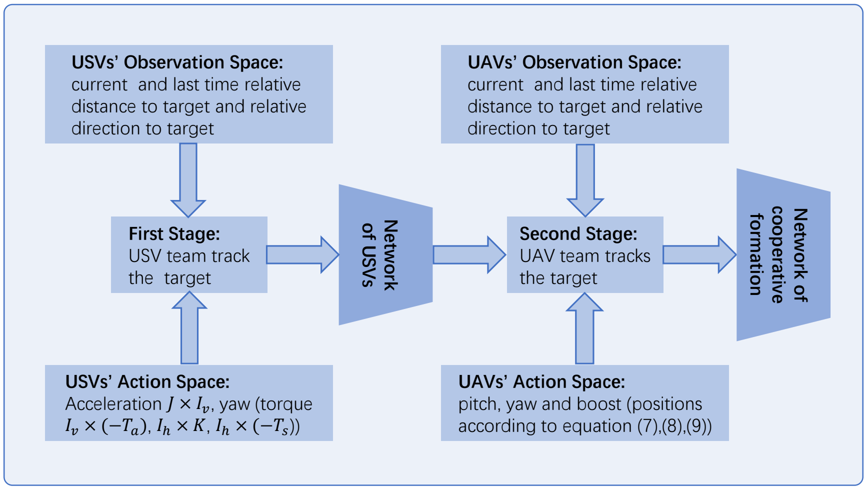

USVs' Observation Space:
current and last time relative distance to target and relative direction to target
UAVs' Observation Space:
current and last time relative distance to target and relative direction to target
First Stage:
USV team track the target
Network of USVs
Second Stage:
UAV team tracks the target
Network of cooperative formation
USVs' Action Space:
Acceleration $J \times I_v$, yaw (torque $I_v \times (-T_a)$, $I_h \times K$, $I_h \times (-T_s)$)
UAVs' Action Space:
pitch, yaw and boost (positions according to equation (7),(8),(9))

Fig. 1. Diagram of USV and UAV formation control.

3.4 Action Space

For USV team, the action space contains the acceleration and yaw. And set acceleration and yaw change signal be obtained from vertical input and horizontal input. For UAV team, we set the pitch, yaw and boost be the action space. To accelerate the training progress and improve system stability and responsiveness, the acceleration and boost are greater than zero.

3.5 Reward Function

Reward of USV team For USV team, the reward function considers the dot product of the relative position with the forward direction of the ship D_{df}, the dot product of forward directions of boat with target's forward direction D_{ff}, and the relative distance to target d. The reward function of USV team is as follow:

$$R_{USV} = \begin{cases} D_{df} \times (1 - 0.001 \times d)\,, if\ \ d > 100 \\ D_{ff} \times (0.9 + 2 - 0.02 \times d)\,, if\ \ d <= 100 \end{cases} \tag{10}$$

Reward of UAV team For the leader of UAV team, the reward function is similar to USV team's. It combines the dot product of relevant distance with forward direction of the leader A_{df}, the dot product of forward directions of boat with forward direction of target A_{ff}, and the relative distance to target d_a.

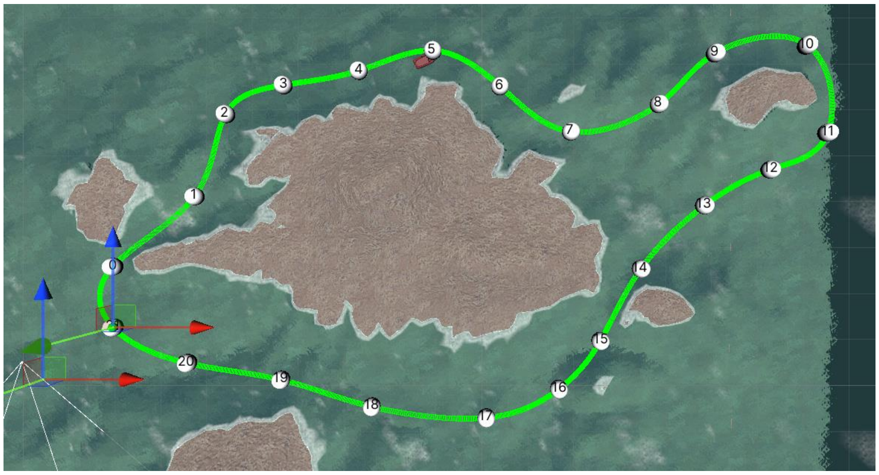
$$R_{UAV} = \begin{cases} A_{df} \times (1 - 0.0005 \times d_a)\,, if\ \ d_a > 100 \\ A_{ff} \times (0.95 + 1 - 0.01 \times d_a)\,, if\ \ d_a <= 100 \end{cases} \tag{11}$$

For the followers of UAV team, if the relative distance to target is less than 100 meters, the reward function is $R_{UAVF} = 1 - 0.025 \times d_f$, which just considers the relative distance to target d_f.

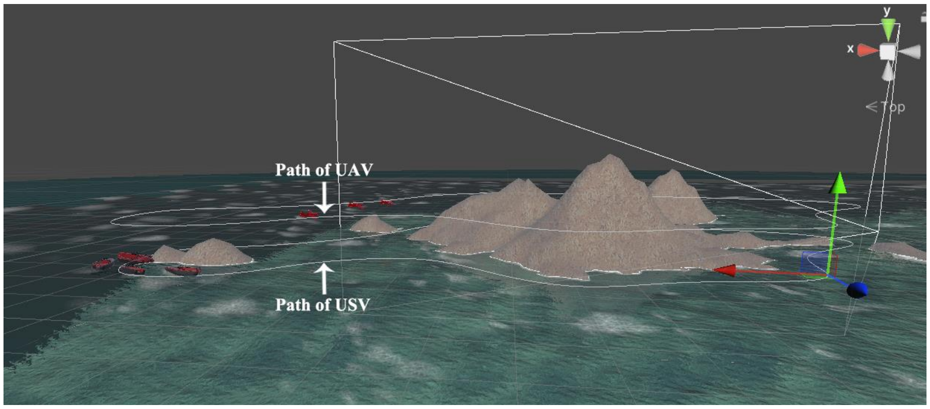
Penalty Additionally, there are penalties for all agents in both teams. If they collide with the obstacles, they would obtain penalties with values of 1.

4 Simulation and Results

The video of simulation: https://linqi-ye.github.io/video/ship24.mp4. In Unity, we design an ocean scene with multiple obstacles and two trajectories, and both trajectories have the length about 4,000 meters. The whole scene is restricted within an area about 3,000 meters long by 3,000 meters wide, where there are 22 checkpoints unevenly distributed on the trajectory. And the checkpoints are represented by numbers in the white circles. In figure 2, (a) and is the top view of predetermined trajectories of the moving targets that USV and UAV track. Since they are two parallel trajectories, the top view looks like the two trajectories overlap. And figure 2 (b) shows the side view of the aforementioned two parallel paths in simulation scene, and we leverage arrows and text to annotate.

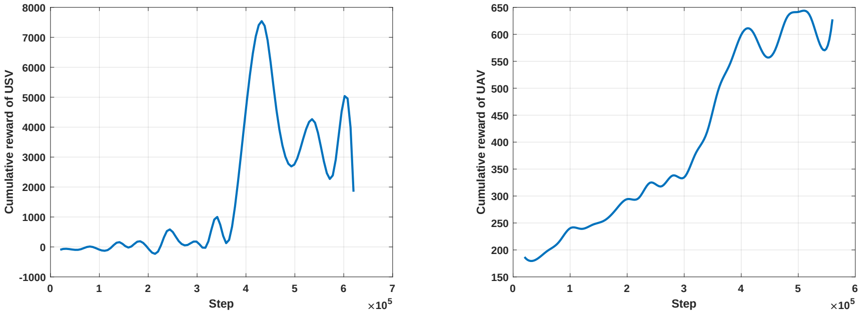


(a)



(b)

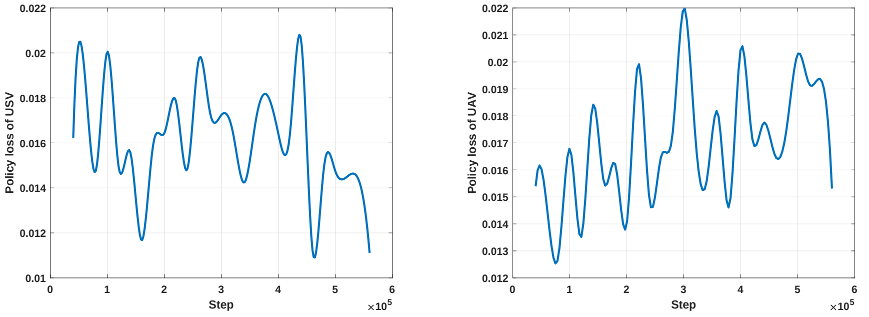
Fig. 2. Path of USV and UAV in simulation scene.



(a) (b)

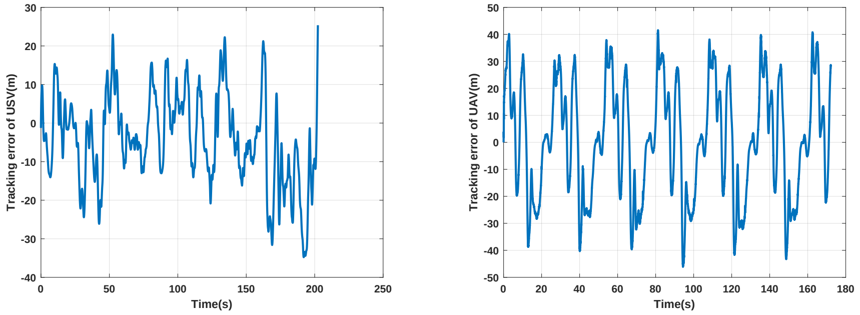
Fig. 3. Cumulative reward of USV team's formation control.

In the first stage, we train the USV team to track the target moving along the preset trajectory within 20 minutes. Similar to the first stage, in the second stage, UAV team could track the target and achieve obstacle avoidance with 20-minute training, while the network of USV is fixed to the first stage's. In figure 3, (a) and (b) are the reward diagrams of USV team and UAV team, respectively. However, the reward of USV team could not converge well, which may be caused by the following reasons: the obstacles on the surface of ocean, the collisions among multiple ships and the limited training time. Figure 4 shows the bias between current policy and expected policy is approximate zero, which infers the current policy performance well.



(a) (b)

Fig. 4. Policy loss of USV and UAV formation.



(a) (b)

Fig. 5. Tracking error of USV and UAV.

Figure 5 presents the tracking error of USV and UAV and both errors are acceptable. Tracking error of USV is around zero within a range of time, while tracking error of UAV shows more regular fluctuations. The reasons for tracking errors are as follows: the limited training time, the random parameters of models, such as action signals and large drag, and another reason is relative general reward, that is the agents seldom obtain penalty for the distance to target exceeding a threshold. Figure 6 shows the simulation of USVs and UAVs cooperative formation control form several different viewpoints. As is shown in these figures, the USV and UAV team could maintain linear formation respectively and keep distance to avoid collision.

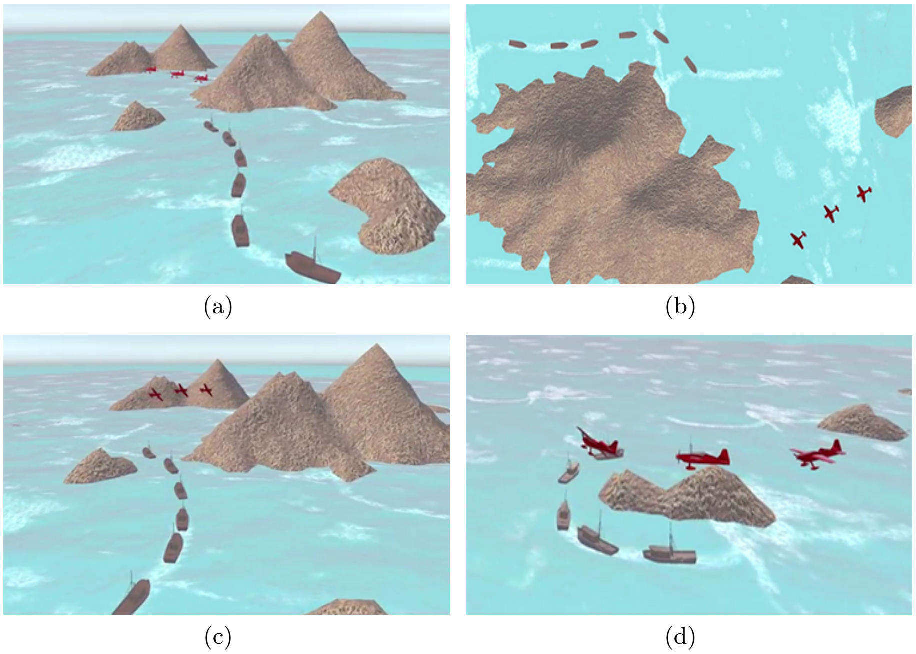


Fig. 6. Simulation diagram of USVs and UAVs formation control.

5 Conclusion

The main contributions of this paper can be summarized as follows: (1)We presented a phased reinforcement learning-based solution for controlling the formation of USV and UAV without complex modeling process compared to traditional control methods. (2) The method is based on leader-follower architecture and curriculum learning, so that the relatively complex formation task is decomposed into a two-stage training task. (3) The existing PPO algorithm performances well for cooperation formation control and none significant changes to the algorithm is required. In future work we will test the relationship between decision frequency and task difficulty and local optima, and we will consider fine-tuning rewards for getting rid of local optima faster or modifying the kinematic equations for improving the agents performance.

References

1. Zhang, Y., Chadli, M., Xiang, Z.: Prescribed-time formation control for a class of multiagent systems via fuzzy reinforcement learning. IEEE Transactions on Fuzzy Systems **31**(12), 4195–4204 (2023)
2. Liu, Z., Chen, B., Zhou, H., Koushik, G., Hebert, M., Zhao, D.: Mapper: Multi-agent path planning with evolutionary reinforcement learning in mixed dynamic

environments. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 11748–11754. IEEE (2020)

3. Huang, S., Zhang, H., Huang, Z.: Multi-uav collision avoidance using multi-agent reinforcement learning with counterfactual credit assignment. arXiv preprint arXiv:2204.08594 (2022)
4. Zhao, W., Liu, H., Lewis, F.L.: Robust formation control for cooperative underactuated quadrotors via reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems **32**(10), 4577–4587 (2020)
5. Liwei, Y., Lixia, F., Ping, L.: Summary of development of multi-agent system formation control. Electronic measurement technology **43**(24), 18–27 (2020)
6. Chang-Yin, S., Chao-Xu, M.: Important scientific problems of multi-agent deep reinforcement learning. Acta Automatica Sinica **46**(7), 1301–1312 (2020)
7. Consolini, L., Morbidi, F., Prattichizzo, D., Tosques, M.: Leader–follower formation control of nonholonomic mobile robots with input constraints. Automatica **44**(5), 1343–1349 (2008)
8. Roldão, V., Cunha, R., Cabecinhas, D., Silvestre, C., Oliveira, P.: A leader-following trajectory generator with application to quadrotor formation flight. Robotics and Autonomous Systems **62**(10), 1597–1609 (2014)
9. Zhang, Q., Lapierre, L., Xiang, X.: Distributed control of coordinated path tracking for networked nonholonomic mobile vehicles. IEEE Transactions on Industrial Informatics **9**(1), 472–484 (2012)
10. Balch, T., Arkin, R.C.: Behavior-based formation control for multirobot teams. IEEE transactions on robotics and automation **14**(6), 926–939 (1998)
11. Do, K.D.: Formation control of multiple elliptical agents with limited sensing ranges. Automatica **48**(7), 1330–1338 (2012)
12. Zhang, J., Wang, W., Zhang, Z., Luo, K., Liu, J.: Cooperative control of uav cluster formation based on distributed consensus. In: 2019 IEEE 15th International Conference on Control and Automation (ICCA). pp. 788–793 (2019). https://doi.org/10.1109/ICCA.2019.8899916
13. Wang, X., Li, X., Zheng, Z.Q.: Survey of developments on multi-agent formation control related problems. Control and decision **28**(11), 1601–1613 (2013)
14. Zhao, Y., Ma, Y., Hu, S.: Usv formation and path-following control via deep reinforcement learning with random braking. IEEE Transactions on Neural Networks and Learning Systems **32**(12), 5468–5478 (2021)
15. Di Caro, G.A., Yousaf, A.W.Z.: Multi-robot informative path planning using a leader-follower architecture. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 10045–10051. IEEE (2021)
16. Liu, H., Meng, Q., Peng, F., Lewis, F.L.: Heterogeneous formation control of multiple uavs with limited-input leader via reinforcement learning. Neurocomputing **412**, 63–71 (2020)
17. Wen, G., Chen, C.P., Li, B.: Optimized formation control using simplified reinforcement learning for a class of multiagent systems with unknown dynamics. IEEE Transactions on Industrial Electronics **67**(9), 7879–7888 (2019)
18. Zhang, H., Zhao, W., Xie, X., Yue, D.: Dynamic leader–follower output containment control of heterogeneous multiagent systems using reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2024)
19. Liu, B., Cai, Q., Yang, Z., Wang, Z.: Neural trust region/proximal policy optimization attains globally optimal policy. Advances in neural information processing systems **32** (2019)

20. Bøhn, E., Coates, E.M., Moe, S., Johansen, T.A.: Deep reinforcement learning attitude control of fixed-wing uavs using proximal policy optimization. In: 2019 international conference on unmanned aircraft systems (ICUAS). pp. 523–533. IEEE (2019)
21. Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., Wu, Y.: The surprising effectiveness of ppo in cooperative, multi-agent games. arxiv 2021. arXiv preprint arXiv:2103.01955
22. Koch, W., Mancuso, R., West, R., Bestavros, A.: Reinforcement learning for uav attitude control. ACM Transactions on Cyber-Physical Systems **3**(2), 1–21 (2019)
23. Zhan, G., Zhang, X., Li, Z., Xu, L., Zhou, D., Yang, Z.: Multiple-uav reinforcement learning algorithm based on improved ppo in ray framework. Drones **6**(7), 166 (2022)
24. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
25. Wang, X., Chen, Y., Zhu, W.: A survey on curriculum learning. IEEE transactions on pattern analysis and machine intelligence **44**(9), 4555–4576 (2021)
26. Zhang, K., Yang, Z., Başar, T.: Multi-agent reinforcement learning: A selective overview of theories and algorithms. Handbook of reinforcement learning and control pp. 321–384 (2021)
27. Gronauer, S., Diepold, K.: Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review **55**(2), 895–943 (2022)
28. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
29. Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., Madry, A.: Implementation matters in deep rl: A case study on ppo and trpo. In: International conference on learning representations (2019)