applied sciences

MDPI

Article

Reinforcement-Learning-Based Asynchronous Formation Control Scheme for Multiple Unmanned Surface Vehicles

Jiajia Xie [1], Rui Zhou [1], Yuan Liu [1], Jun Luo [1,2], Shaorong Xie [1], Yan Peng [1,*] and Huayan Pu [1]

[1] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China; bigjjia@shu.edu.cn (J.X.); zr901122@sina.com (R.Z.); liuyuanji@shu.edu.cn (Y.L.); luojun@shu.edu.cn (J.L.); srxie@shu.edu.cn (S.X.); phygood_2001@shu.edu.cn (H.P.)

[2] State Key Laboratory of Mechanical Transmission, Engineering Department, Chongqing University, Chongqing 400030, China

* Correspondence: pengyan@shu.edu.cn

Abstract: The high performance and efficiency of multiple unmanned surface vehicles (multi-USV) promote the further civilian and military applications of coordinated USV. As the basis of multiple USVs' cooperative work, considerable attention has been spent on developing the decentralized formation control of the USV swarm. Formation control of multiple USV belongs to the geometric problems of a multi-robot system. The main challenge is the way to generate and maintain the formation of a multi-robot system. The rapid development of reinforcement learning provides us with a new solution to deal with these problems. In this paper, we introduce a decentralized structure of the multi-USV system and employ reinforcement learning to deal with the formation control of a multi-USV system in a leader–follower topology. Therefore, we propose an asynchronous decentralized formation control scheme based on reinforcement learning for multiple USVs. First, a simplified USV model is established. Simultaneously, the formation shape model is built to provide formation parameters and to describe the physical relationship between USVs. Second, the advantage deep deterministic policy gradient algorithm (ADDPG) is proposed. Third, formation generation policies and formation maintenance policies based on the ADDPG are proposed to form and maintain the given geometry structure of the team of USVs during movement. Moreover, three new reward functions are designed and utilized to promote policy learning. Finally, various experiments are conducted to validate the performance of the proposed formation control scheme. Simulation results and contrast experiments demonstrate the efficiency and stability of the formation control scheme.

Keywords: deep reinforcement; formation control; formation generation; formation maintenance learning; multi-USV system

check for updates

Citation: Xie, J.; Zhou, R.; Liu, Y.; Luo, J.; Xie, S.; Peng, Y.; Pu, H. Reinforcement-Learning-Based Asynchronous Formation Control Scheme for Multiple Unmanned Surface Vehicles. *Appl. Sci.* **2021**, *11*, 546. https://doi.org/10.3390/app11020546

Received: 2 December 2020
Accepted: 31 December 2020
Published: 8 January 2021

1. Introduction

Due to the rapid development of communication, navigation, and computer technology related to ship motion control, cooperative ship control has an extensive range of application prospects in military and production fields, including fleet cooperative combat, ocean-going replenishment, environmental monitoring, oil and gas detection, etc. Because of higher operational security, lower cost, and greater autonomy and flexibility, unmanned surface vehicles (USVs) are applied to perform extensive missions in hazardous maritime environments instead of manned vehicles [1]. Compared with a single USV, multiple USVs' cooperation has the advantages of strong adaptability and fault tolerance. The fleet can form a dynamic network during the navigation. Through division and cooperation, each USV can perceive the environmental information about the area quickly and accurately to accelerate the completion of missions and improve the efficiency of the system. Formation control is the most fundamental problem of multiple USV cooperative control. Therefore, formation control of USVs has become one of the hot issues in the research of USV motion control. A collective scheme is necessary to ensure that the USVs work together to complete a common task and coordinate in time and space. As of today, many scholars have studied the formation control problem of a multi-USV system. The formation control problems are summarized into two fundamental problems: (1) formation generation, which refers to how to form a designated formation [2,3], and (2) formation maintenance, which refers to how to keep formation unchanged in the process of movement [4].

USVs' formation originated from the study of biological cluster dynamics, which can be traced back to the Boid model proposed by Reynolds [5]. Based on this model, Olfati-Saber [6] extended the multi-agent consistency work to the usual swarm formation control field, introduced obstacle avoidance and tracking agents, and designed a distributed control framework including gradient-based term, velocity consensus term, and navigational feedback. Su et al. [7] further developed the collective formation control strategy based on the work of Olfati-Saber, using a virtual leader to replace the actual leader. Ponomarev et al. [8] proposed a consistency control method based on a predictive mechanism to accelerate the convergence speed of multi-agent consistency. Chen et al. [9] proposed the collective circular motion behavior control of heterogeneous multi-agents under arbitrarily closed curves. In the research of multi-agent formation control, most researchers treat the model as a linear system of first-order or multi-orders [6–8]. Taking into consideration the dynamic characteristics (nonlinearity, coupling, underactuation, etc.) of the robots in a multi-robot formation, it is often difficult to directly consider it as a system for analysis. To better achieve stability and efficiency, and at the same time be useful to the theoretical analysis of the multi-robot formation, researchers have proposed the following formation methods: the virtual structure method, the behavior-based method, and the leader–follower method [10]. The virtual structure method [11] is not flexible, and it is difficult to achieve obstacle avoidance with that method. It is hard to express the entire system in mathematical form and difficult to prove and guarantee the stability of the system using the behavior-based method [12]. In contrast, the leader–follower method [13] is easy to design and implement, and easy to ensure stability. The advantages of the leader–follower method are that the moving goal is only assigned to the leader to navigate the movement of the agent swarm, and each member only needs to collect information about its immediate leader instead of the whole swarm. For example, when performing seafloor terrain scanning, hydrological sampling, target search, and resource detection, a team of robots performs tasks in formation, and the path trajectory is assigned to the leader of the formation. Other robots keep a certain distance (for example, sonar detection radius) from their associates. The swarm can perform tasks with a fixed geometric structure, which improves work efficiency and safety.

The current full actuated leaderless formation control algorithms and leader–follower formation control algorithms are all based on back-stepping [14–16]. The repeated derivation of the virtual control law by a back-stepping method in practical design will bring about a sharp increase in partial derivative calculations as the order of the system increases, which obviously increases the complexity of nonlinear system design. Through interacting with the environment through a trial-and-error mechanism, reinforcement learning optimizes policy by maximizing cumulative rewards and finally achieves the optimal policy. Other existing works only use the current optimal sample update, while reinforcement learning makes full use of historical samples to get the gradient descent direction based on the cumulative discounted reward. Since the cumulative discounted reward is based on all the existing samples, the sample information is more fully utilized, and the efficiency of policy learning is significantly improved. The combination of reinforcement learning and deep learning can provide optimal decision-making strategies for complex high-dimensional multi-agent systems and can lead to efficient performance of tasks in challenging environments. The policy gradient adopted and improved in this paper is a direct parametric policy, optimizing the trajectory from the initial state to obtain an optimal policy, which is a continuous function of the state–action value and more suitable for dealing with continuous problems, such as formation control. The advantage of deterministic policy gradient is that less data need to be sampled and the algorithm efficiency is high. One way of optimizing policies is to adjust the gradient toward the direction of "good" actions. Advantage function is usually used to measure the quality of each action in each state. Therefore, we propose an improved DDPG based on advantage function to train policy for formation control of a multi-USV system.

In this paper, to solve the above problems, an asynchronous formation control scheme based on reinforcement learning and leader–follower structure is proposed for multiple USVs. First, a USV model and a novel formation shape model are established. Second, the advantage deep deterministic policy gradient algorithm (ADDPG) is proposed and used to learn a formation generation policy, which is used to generate the formation according to the control requirements. Finally, a formation maintenance policy based on the ADDPG and the designed reward function is utilized to maintain the given geometry structure of the team of USVs during movement.

To summarize, the main contributions of this paper are threefold.

- **Modeling for maritime formation control**: We introduce a USV model for underactuated USV, only considering its kinematics model. Moreover, we propose a formation shape model to describe the physical relationship between USV members, including relationship in formation, relative distance offset, and scaling coefficient.
- **Formation control scheme**: We propose an asynchronous decentralized formation control scheme for multiple cooperative USVs, in which we propose the ADDPG algorithm, and design the reward functions for the formation control problem. Then, based on the required specific geometry shape of the USV team, the decentralized formation generation policies and decentralized formation maintenance policies are trained based on the ADDPG to generate the formation and keep the geometric shape, respectively.
- **Performance validation**: Evaluation criteria are designed to evaluate the performance of the proposed scheme. Extensive simulations are conducted to verify the effectiveness of the proposed formation control scheme. The simulation results show that the proposed scheme can realize the effectiveness of formation generation and the stability of formation maintenance.

The remainder of this paper is arranged as follows. In Section 2, we review the relevant research studies. In Section 3, we describe the system model. In Section 4, we present the formation control scheme. In Section 5, we verify the performance of the proposed formation control scheme by simulation. The paper is concluded in Section 6.

2. Related Works

In this section, we review the related works, including the formation control algorithms of the multi-robot system.

Fahimi [17] studied the nonlinear model predictive about the control formation problem of USVs in the environment with obstacles. Based on the decentralized geometric control strategy in the leader–follower structure, the underactuation of USVs, and environmental obstacles, a formation controller for real-time optimization nonlinear predictive control method was designed to realize formation and obstacle avoidance. Do [18] presented a design of cooperative controllers for several agents based on the constraint of sensing ranges and collision avoidance. Then, Do [19] discussed the formation control of underactuated ships limited to collision avoidance and communication. An elliptic collision avoidance method was proposed, and the nonlinear coordinate transformation and additional control were introduced to control the underactuated ship. Simultaneously, the potential energy function was used in the controller design with collision avoidance between the ships.

Peng et al. [20] proposed a neural-network-based leader–follower underdrive UAV formation controller based on the uncertainty of leader dynamics and environmental interference. The uncertainty dynamics of the leader were approximated only by the sight distance and angle measured by the local sensor, and a control law that does not rely on an accurate model was designed. After that, an observer-based distributed formation controller was proposed. The formation controller based on neighbor information was designed using a neural network, back-stepping, and graph theory, and used to estimate the speed information [14]. Since then, to overcome problems such as model uncertainty, ocean noise, and unpredictable speed of leaders and followers, adaptive control, neural networks, high-gain observers, and minimum learning parameter algorithms have been combined into the backstepping design, a new adaptive output feedback control scheme has been proposed, which realizes the leader–follower formation based only on position and heading angle, and only two parameters need to be learned online [21,22]. Ding et al. [23] proposed a distributed adaptive cooperative formation control strategy based on a virtual leader, designed a navigation system and an adaptive neural network synchronization controller that can calculate a specified trajectory, and solved the problem of model uncertainty to achieve stable formation. Sun et al. [24] studied the formation control of USVs in a leader–follower structure, considering model uncertainty and dynamic disturbance of the environment.

Shojaei [25] proposed a leader–follower formation tracking controller for USVs affected by torque limitation and environmental noise. The saturation function was used to reduce the risk of driver saturation. The radial basis function and adaptive robust control technology were used to improve the robustness of the controller in a disturbance environment. On this basis, the formation method was developed into the three-dimensional formation control of underactuated underwater vehicles based on the neural network, and the nonlinear saturation observer was introduced to estimate the speed of the follower [26]. Sun et al. [27] considered the autonomous navigation of the leader–follower USV formation in a complex environment, and the predictive control based on the limited control set realized the USV team to reach the destination in a certain formation with internal collision avoidance under the condition of no prior knowledge of the environment and predefined trajectory.

In other respects, Breivik et al. [28] studied the leader–follower formation control problem of fully actuated ships and proposed a navigation formation control method, which uses control, navigation, and synchronization algorithms to ensure that each individual can converge and stay in the assigned formation position to achieve formation. Cui et al. [29] proposed a control method based on an approximation method to address the unknown uncertainty in the leader–follower formation control model of multi-AUV. Fan et al. [30] proposed a formation control strategy based on two-layer predictive control. One layer guarantees the leader–follower cooperative formation between USVs, and the other layer realizes the USVs' tracking of the optimal command. Park [31] aimed at the asymmetry of the quality and attenuation matrix of the underwater vehicles and the uncertainty of the hydrodynamic attenuation term, introduced additional control input to solve the underactuated control, and realized the leader–follower control when only using position information. Liu et al. [32] designed two different algorithms for formation forming and path planning based on the heading navigation fast marching algorithm, which solved the heading constraint problem of the unmanned boat and realized that the USVs can follow the planned trajectory and formation through any initial state. Sui et al. [33] presented a novel formation control with collision avoidance policy using imitation learning and reinforcement learning, but only one leader and one follower are considered.

The failure of the leader may affect the robustness of the whole swarm, so selecting the best leader from the swarm is an important issue for the study in leader–follower formation control. Hou et al. [34] proposed a leader–follower formation with multiple changeable leaders and proposed a switching distributed saturated control law, which enables the formation to work even if a leader fails. Based on the status of the multi-robot system evaluated by a fuzzy inference system, Li et al. [35] proposed an affection-based dynamic leader selection model to switch leaders autonomously. To solve the failure of the current leader, Li et al. [36] proposed a neuroendocrine system to switch and evaluate a leader autonomously. Considering the time-varying and fully-decentralized structure in the leader–follower multi-agent system, Franchi et al. [4] proposed an online leader selection strategy to periodically select the best leader for the team during the movement.

Xue et al. [37] introduced a supermodular optimization approach to fixed-size set and minimum-size set of leaders to select the optimum leader for minimizing convergence error in leader–follower formation control.

Formation control of multi-robot is to control multiple robots mainly based on preset inter-robot parameters, which determines the distance and orientation displacements among these robots [38–43] during task execution. Many formation strategies are proposed to obtain and keep the stability of formation under different formation shapes. Aranda et al. [44] achieved the local and global stability of formation, while Oh et al. [45] obtained the local stability of formation, and Lin et al. [46] ensured global stability of formation. Lin et al. [47] presented a formation method based on complex Laplacian to achieve global stability by the inter-agent relative displacement.

3. System Models

In this section, we analyze the system model of maritime cooperative formation for a team of USVs, as shown in Figure 1, including the USV model, formation shape control, and control objective.

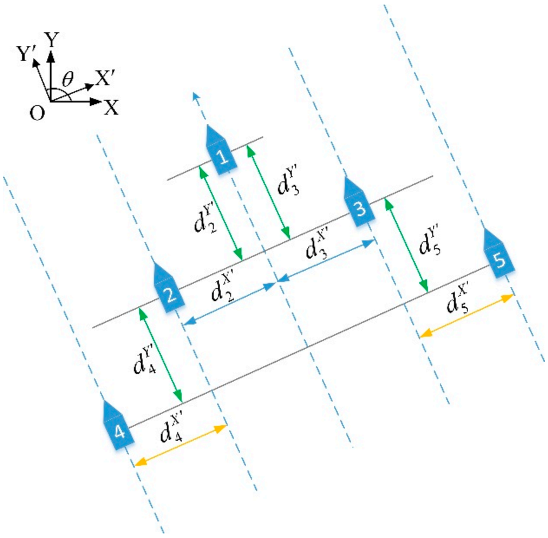


Figure 1. The standard V-shape formations for multiple unmanned surface vehicles (USVs). Every USV should keep the required distance from its immediate leader in the x-axis and y -axis according to the new coordinate frame $O - X'Y'$.

3.1. USV Model

We consider a group of N USVs for formation control in the leader–follower structure, described as U $= \{u_1, u_2, \ldots u_N\}$ with geometric shapes, such as V-shape, as shown in Figure 1. Because this paper focuses on how to design and train formation generation policies and formation maintenance policies for the USV team, we simplify the USV into a particle, mainly considering the kinematics model $\dot{p}_{u_i} = v_{u_i}$ of the particle, and temporarily ignoring the impact of the kinetics model of the particle. Each USV u_i has coordinates $p_{u_i} = (x_{u_i}, y_{u_i})$ and velocity $v_{u_i} = (v_{u_i}^x, v_{u_i}^y) = \dot{p}_{u_i}$ in the $O - XY$ coordinate frame in the two-dimensional maritime plane. At each time step, the action taken by USV u_i is the change of velocity, $a_{u_i} = (\Delta v_{u_i}^x, \Delta v_{u_i}^y)$. In a leader–follower structure, the followers should follow the leader in a geometrically balanced manner and keep a specific direction and distance from the leader. Thus, we establish a new relative coordinate frame $O - X'Y'$ that the origin is the original origin, and the direction of the Y'-axis is the heading angle of the leader. Each USV u_i has new coordinates $p'_{u_i} = (x'_{u_i}, y'_{u_i})$ in the relative coordinate frame, as calculated by the coordinate transformation matrix in Equation (1), where θ is the leader's heading angle. The formation control of a multi-robot system transforms the formation control problem into the problem of followers tracking the leader's position and direction. The classical leader–follower mode is that all the followers track a single leader with different distance offsets individually, while a chain leader–follower formation is used in this paper. The chain leader–follower is inspired by three flocking rules of Reynolds [5] that agents in a group should stay close to their neighbor agents, avoid collisions with their neighbor agents, and match speed with its neighbor agents instead of the leader of the group. In short, each agent aligns with its neighbors. In a cooperative USV team, it is more conducive to improve the team efficiency to use communication to share information than to collect information of members in the team with sensors. USVs are equipped with industrial control computers, GPS, and other sensors, so the followers obtain the coordinates of their leaders through wireless communication. For instance, the USV u_2 tracks the USV u_1 while the USV u_4 tracks the USV u_2 according to the required chain formation. USV u_2 is regarded as the immediate leader of its follower USV u_4. In addition to the leader USV, $H_i = (d_i^{X'}, d_i^{Y'})$ is the predefined offset vector, i.e., the relative positional relationship for USV u_i concerning its immediate leader in the $O - X'Y'$ coordinate frame.

Each USV takes action following its formation generation policy μ_{θ_i} to make the USV team generate the expected formation shape (see Section 4.2.1). Each USV takes action following its formation maintenance policy $\mu_{\theta_i}^{fm}$ to keep the stability of the formation structure with the predefined teammate spacing (see Section 4.2.2).

$$\begin{bmatrix} x'_{u_i} \\ y'_{u_i} \end{bmatrix} = \begin{bmatrix} \cos(\theta - \frac{\pi}{2}) & \sin(\theta - \frac{\pi}{2}) \\ -\sin(\theta - \frac{\pi}{2}) & \cos(\theta - \frac{\pi}{2}) \end{bmatrix} \begin{bmatrix} x_{u_i} \\ y_{u_i} \end{bmatrix} \tag{1}$$

3.2. Formation Shape Model

The USV team needs the geometry configuration of formation in the collective formation mission. Reasonable formation shapes of the USV team can increase the efficiency of formation task execution. To establish or maintain a specific geometric formation, it is necessary to establish a representative method of geometric formation. At present, there is no unified formation representative method to designate the formation of the USV team. In this paper, the formation shape matrix is established by combining with the formation description mode of chain guidance reference, defined as a $4 \times N$ formation shape matrix, F_s as shown in Equation (2). The matrix is adopted to represent the geometric relationship of USVs, where N is the number of USVs in the formation. In the matrix F_s, the first row denotes the number of geometric nodes in the formation. The second and third rows represent the distance offsets between the USV and its immediate leader USV in x-axis and y-axis directions. The fourth row denotes the node number of the immediate leader of the USV in each formation node. N USVs form a geometric shape. The centric USV is the leader of the team, while the other USVs form the chain tracking one by one. Follower USVs keep the distance displacement $(d_i^{X'}, d_i^{Y'})$ from their immediate leader. If one or several immediate leaders in the formation fail during the movement, it is necessary to reconstruct a formation according to the size N_t of good USVs and the first N_t columns of the formation shape matrix F_s.

$$F_s = \begin{bmatrix} 1 & 2 & 3 & \dots & N \\ 0 & \alpha d_2^{X'} & \alpha d_3^{X'} & \dots & \alpha d_N^{X'} \\ 0 & \beta d_2^{Y'} & \beta d_3^{Y'} & \dots & \beta d_N^{Y'} \\ 0 & Il(n_2) & Il(n_3) & \dots & Il(n_{N-2}) \end{bmatrix} \tag{2}$$

where $d_i^{X'}$ and $d_i^{Y'}$ are the horizontal and vertical distance of each follower–leader pair. $\alpha \in (0,1]$ and $\beta \in (0,1]$ are the expansion coefficients in horizontal and vertical directions, respectively. By adjusting the values of α and β, the horizontal and vertical expansion and contraction of the same formation can be realized. $Il(n_i)$ is the immediate leader of the USV located in the formation node n_i.

3.3. Control Objective

For formation generation, the USV team starts from the respective current locations and generates a predefined formation shape in the target location. Therefore, the goal is to minimize the sum of the length of the movement path of the team of USVs,

$$\min\left(\sum_{i=1}^{N}(\operatorname{len}(s(u_i), e(u_i)))\right) \tag{3}$$

where $s(u_i)$ is the initial position of USV u_i, and $e(u_i)$ is the final position of USV u_i in the formation. $\operatorname{len}(s(u_i), e(u_i))$ is the length of the moving path for USV u_i from position $s(u_i)$ to position $e(u_i)$ during formation generation.

For formation maintenance, in this paper, we aim at training USVs to learn policies to move in a predetermined formation and maintain the shape of the formation. All the followers keep a certain distance with their respective leaders, which reduces the stability error (i.e., distance difference between the current relative distance and the predefined distance). Thus, another control goal is defined as follows:

$$\mathrm{L}_{pg} = \min_i \left|H'_i - H_i\right|, \forall u_i \in U_{follower} \tag{4}$$

$$H'_i = p'_{u_i} - p'_{Il(u_i)} \tag{5}$$

where $Il(u_i)$ denotes the immediate leader of USV u_i. $U_{follower}$ is the set of all the followers except the only leader USV, and $p'_{Il(u_i)}$ is the position of the immediate leader $Il(u_i)$ of USV u_i. H_i is the required distance offset between USV u_i and its immediate leader, and H'_i is the current relative distance between USV u_i and its immediate leader.

4. Proposed Scheme

In this section, we introduce the problem formulation for a cooperative formation control followed by a formation control algorithm based on reinforcement learning in which policies for formation generation and formation maintenance are presented.

4.1. Problem Formulation

For the formation control of USVs in the leader–follower structure, the main problem is the question of how to form the predetermined formation shape with collision-free movement and maintain the global formation shape.

An asynchronous formation control scheme based on reinforcement learning for multiple USVs is proposed to address the problems mentioned above, making the USV team generate a formation with a minimum total length of movement path and maintain the formation. The proposed scheme contains two parts: formation generation policy and decentralized formation maintenance policy. In formation generation, each USV observes the positions and velocities of all USVs by sensors and communication, and the positions of points in formation shape are given to all USVs, and then learn a formation generation policy $\mu^*_{\theta_i}$ based on cost function J_i to make the team form the formation quickly with a series of optimal actions a^*, as shown in Figure 2. However, only the leader collects the information about the team's goal; each follower obtains its immediate leader's position and velocity by communication and follows the formation maintenance policy $\mu^{fm}_{\theta_i}$ to track the leader, as shown in Figure 3.

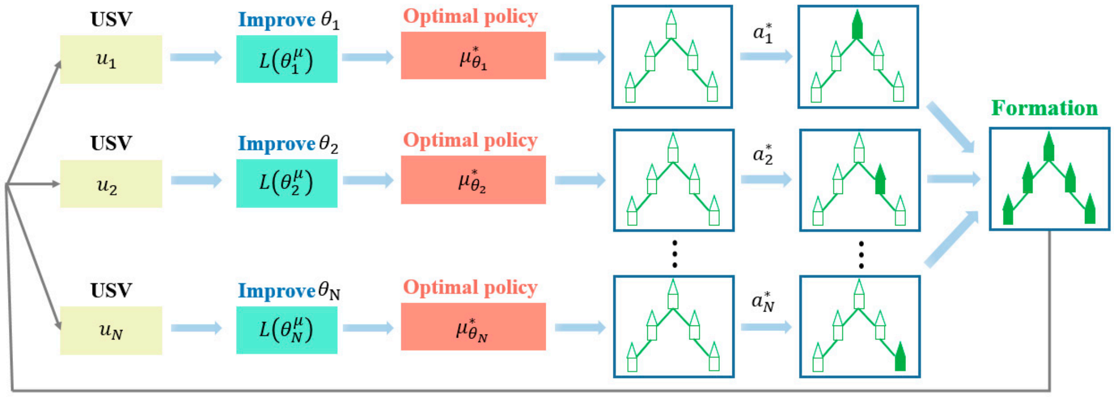


Figure 2. Formation generation.

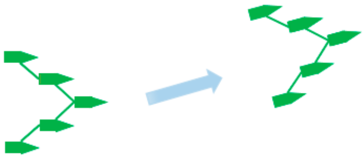


Figure 3. Formation maintenance.

4.2. Formation Control Algorithm Based on Reinforcement Learning

4.2.1. Formation Generation Policy

In a sophisticated maritime environment, the formation control problem can be regarded as a Markov decision process (MDP) and described as $< S, A, P, r >$. S describes a set of the possible states s of each USV. A is a set of actions a that a USV can take. The transition probability distribution for each pair of state s and action a is expressed as $P : S \times A \times S \to [0, 1]$. The expected reward for each state–action pair is computed as $r : S \times A \to \mathbb{R}$. Moreover, a deterministic policy $\mu : S \to \mathbb{R}^{|A|}$ is defined to output a deterministic action a in state s that will obtain a reward $r(s, a)$ and make the environment change to a new state s' with an environmental transition probability $P(s'|s, a)$. Policy optimization is realized by maximizing the cumulative return $R_i = \sum_{t=0}^{T} \gamma^t r_i^t$ of each USV, where $\gamma \in [0, 1]$ is the discount factor.

We propose an improved deep deterministic policy gradient based on the advantage function and deep deterministic policy gradient (DDPG) proposed in [48], named the advantage deep deterministic policy gradient (ADDPG). The multi-USV system for formation control considered in this paper is decentralized, and each USV runs its policy independently. We use the ADDPG to train the decentralized formation generation policy set $\boldsymbol{\mu} = \{\mu_{\theta_1}, \mu_{\theta_2}, \ldots, \mu_{\theta_N}\}$ for the team of USVs, as described in Algorithm 1. We use the advantage function of state–action value instead of the state–action value to calculate the policy gradient, which can make the policy update toward the direction of the larger action value and accelerate the efficiency of policy learning. For each USV, its observation includes information about its velocity and position, the relative distance of other USVs, the predefined parameters of the formations. Its action is the change of velocity. At each time step t of formation generation, each USV u_i obtains its observation s_t^i, uses its policy μ_{θ_i} to generate an action a_t^i, and receive a reward r_{t+1}^i from the environment. After the USV executes the action a_t^i, the environment state s_t^i transfer to the next state s_{t+1}^i contains. The transition experience $e_t^i = \left(s_t^i, a_t^i, r_{t+1}^i, s_{t+1}^i\right)$ of all USVs is collected and stored in the shared experience replay buffer D_s and used to train the formation generation policy.

The network structure of the ADDPG is illustrated in Figure 4. Inspired by the target network in DQN, we introduce target networks and actor-critic paradigm in the proposed scheme to address continuous actions and action-value estimation and improve the stability of learning. Thus, there are four neural networks in the ADDPG. The critic is used to train the state value network to approximate the value of the state–action, including the current critic network and target critic network, which are three-layer multilayer perceptron (MLP) with parameters θ^V and $\theta^{V'}$, respectively. The actor is designed to train the formation policy to output the action that should be taken in the current state, containing current actor network and target actor network, with parameters θ^μ and $\theta^{\mu'}$, respectively. The target networks' parameters $\theta^{V'}$ and $\theta^{\mu'}$ use the parameters from some previous iteration of θ^V and θ^μ. We use an advantage function to evaluate the relative advantage of each action in a state and accelerate the learning of policies. The actor part based on advantage function uses the DPG method; the critic part uses the TD error method to update the parameters. During training, each USV in the team has independent networks with different parameters and independent optimization for its formation control policy.

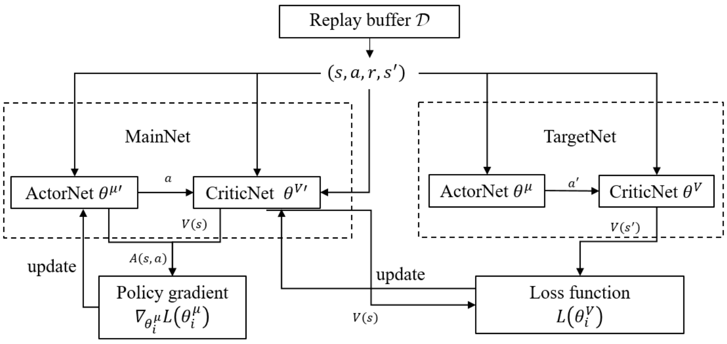


Figure 4. The network structure of each USV's policy.

The advantage function describes the advantage of selecting an action a in the state s, compared with other actions under the state s, as denoted by

$$A(s_i, a_i) = Q(s_i, a_i) - V(s_i) \tag{6}$$

For the critic part, the loss function is denoted as follows:

$$L\left(\theta_i^V\right) = \left(R_i - V_i\left(s_i|\theta_i^V\right)\right)^2 \tag{7}$$

where

$$R_i = r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots + \gamma^{n-1} r_{n-1} + \gamma^n V_i'\left(s_n|\theta_i^{V'}\right) \tag{8}$$

The critic is updated by minimizing the loss,

$$\nabla_{\theta_i^V} L\left(\theta_i^V\right) = -\mathrm{E}_{s,a,r,s'}\left[\left(R_i - V_i\left(s_i|\theta_i^V\right)\right)\nabla_{\theta_i^V} V_i\left(s_i|\theta_i^V\right)\right] \tag{9}$$

The actor is updated by the gradient of Q-value and advantage function:

$$\nabla_{\theta_i^\mu} L\left(\theta_i^\mu\right) = \mathrm{E}_s\left[\nabla_{a_i}(R_i - V(s_i))|_{a_i=\mu_i(s_i)}\nabla_{\theta_i^\mu}\mu_i\left(s_i|\theta_i^\mu\right)\right] \tag{10}$$

The interactions between the USV team and the ocean environment can be divided into separate episodes. An episode starts in a random state of the team and ends at a terminal state or after a specified number of time steps. The number of episodes is set to 20,000, and the maximum episode length is 100—that is, each episode has up to 100 timesteps. At each time step in every episode, every USV interacts with the environment, selects its action a_t according to the current state s_t, and dynamically calculates the reward r_{t+1} generated by the environment in real time according to the reward function r_1 and r_2. The reward function r_1 is defined to be the minimum distance between the USVs and the geometric points of formation. If a collision occurs between members during the formation control process, a collision penalty is given—that is, the negative reward value r_2.

$$r_1 = -\sum_{j=1}^{N} (\min_{u_i} Dis(p'_{u_i} - p'_{n_j})) \tag{11}$$

$$r_2 = -c_1 \tag{12}$$

where p'_{n_j} is the position of the formation node n_j in the coordinate frame $O - X'Y'$, which can be calculated by the predefined location of the leader and the parameters in the formation shape matrix F_s. $Dis\left(p'_{u_i}, p'_{n_j}\right)$ is the Euclidean distance between the USV u_i and the formation node n_j. c_1 is a positive constant.

Algorithm 1: Formation Generation Policy Based on the Advantage Deep Deterministic Policy Gradient (ADDPG)

```
Input Reward Function $r_1$,$r_2$ for formation generation scenario
Input The predefined formation shape $F_s$
Output formation generation policies $\boldsymbol{\mu} = \{\mu_{\theta_1}, \mu_{\theta_2} \ldots \mu_{\theta_N}\}$ for the USV team
Initial experience replay buffer $D_s$
1:  for $episode = 1 : M$ do
2:      Initialize a random process $N_{rp}$ for action exploration
3:      Receive environment state
4:      for $t = 1 : T$ do
5:          for $i = 1 : N$ do
6:              Select action $a_t = \mu_i\left(s_t \middle| \theta_i^\mu\right) + N_{rp}$
7:              Execution actions $a_t$ and observe reward $r_t$ and new state $s_{t+1}$
8:              Add $e_t^i = \left(s_t^i, a_t^i, r_t^i, s_{t+1}^i\right)$ into replay buffer $D_s$
9:              Sample a random minibatch of transitions $e_j$ from $D_s$
10:             Calculate reward in real-time $R_i = \begin{cases} 0 & \text{for terminal } s_t \\ V(s_t) & \text{for non} - \text{terminal } s_t \end{cases}$
11:             for $m = \{t-1, t-2, \ldots, t_{end}\}$ do
12:                 $R_i = r_m + \gamma R_i$
13:                 Update the critic by (9):
14:                     $\theta_i^V \leftarrow Adam\left(\theta_i^V, \nabla_{\theta_i^V} L(\theta_i^V)\right)$
15:                 Update the formation generation policy for USV $u_\mathrm{i}$ by (10):
16:                     $\theta_i^\mu \leftarrow Adam\left(\theta_i^\mu, \nabla_{\theta_i^\mu} L\left(\theta_i^\mu\right)\right)$
17:                 Update the "soft" target networks for the actor and critic:
18:                     $\theta_i^{\mathrm{V}\,\prime} \leftarrow \tau\theta_i^{\mathrm{V}} + (1-\tau)\theta_i^V$
19:                     $\theta_i^{\mu'} \leftarrow \tau\theta_i^\mu + (1-\tau)\theta_i^{\mu'}$
20:             end for
21:         end for
22:     end for
23: end for
```

The details of the proposed formation generation policies are shown in Algorithm 1, where M is the number of episodes, T is the maximum episode length, and N is the number of USVs. The inputs are the designed reward r_1,r_2 for the formation generation scenario and the formation shape F_s. The output are the formation generation policies for the USV team. In each episode, every USV adopts a random process N_{rp} to achieve sufficient exploration and collects experiences $e_t^i = (s_t^i, a_t^i, r_t^i, s_{t+1}^i)$ that are stored in the replay buffer D_s (lines 8–9) and sampled randomly to update the policies. The value of each state–action pair is estimated by the cumulative discount reward $\mathrm{R_i}$ (lines 10 and 12), which is used to calculate the advantage of action. A random minibatch of samples in the replay buffer is sampled by each USV to improve the formation generation policies (lines 13–16). The "soft" updates are used to make the target function change more slowly and improve the stability of learning (lines 18–19).

4.2.2. Decentralized Formation Maintenance Policy

The leader–follower method requires the followers to maintain a specific position and direction offset from the leader, so the structure is simple and robust engineering. We consider a chain mode in the leader–follower structure, in which each follower tracks its immediate leader instead of the mode in which all followers track the same leader. The advantage of this method is that the communication pressure of the leader is reduced, and the stability of the formation structure is realized by minimizing the tracking error for each USV.

We adopt the proposed ADDPG in this paper to train the decentralized formation maintenance policy set $\boldsymbol{\mu}^{fm} = \left\{\mu_{\theta_1}^{fm}, \mu_{\theta_2}^{fm}, \ldots, \mu_{\theta_N}^{fm}\right\}$ for the decentralized multi-USV system, similar to the training of formation generation policy. For each USV, the state s contains information about its velocity and position and the relative distance from its immediate leader. The action a contains the change of velocity. At each time step j of formation maintenance, each USV u_i obtains its observation s_j^i, uses its policy $\mu_{\theta_i}^{fm}$ to generate an action a_j^i and receive a reward r_{j+1}^i from the environment. After the USV executes the action a_j^i, the environment state s_j^i transfers to the next state s_{j+1}^i. The reward function r_3 is defined to be the distance difference, which is the error between the current distance H'_i and the expected distance H_i between the USV and its immediate leader. The transition experience $e_j^i = \left(s_j^i, a_j^i, r_{j+1}^i, s_{j+1}^i\right)$ of the USV u_i is stored in the experience replay buffer D_i and sampled randomly to update the policy. The reward r_3 is used to measure the performance of the formation maintenance policy, aiming to reduce the difference between the current formation and the expected formation. All UAVs choose and execute actions asynchronously and are not limited to synchronous operations.

$$r_3 = -\left|H'_i - H_i\right| \tag{13}$$

5. Experiment and Analysis

In this section, we design comparative experiments to evaluate the proposed scheme and analyze the simulation results.

5.1. Experimental Setting

We design a formation generation scenario based on the simulation platform designed by [49]. A V-shape formation is used in the experiments. In the formation generation scenario, N USVs are moving in the two-dimensional maritime surface, which is considered as a square with a side length of 2. Only the kinematics model of USVs is considered. In a formation generation scenario, the formation generation is to control the team of USVs to form a predefined formation F_s by following the formation generation policies $\boldsymbol{\mu} = \{\mu_{\theta_1}, \mu_{\theta_2}, \ldots, \mu_{\theta_N}\}$. The input of the formation generation policy of each USV u_i is a row vector $s_{u_i} = (p_{u_i}, v_{u_i}, re_{\text{point}}, re_{oth})$, including 1×2 position vector p_{u_i}, 1×2 velocity vector v_{u_i}, $1 \times 2N$ relative position vector re_{point} between USV u_i and N formation points, and $1 \times 2(N-1)$ relative position vector re_{oth} between USV u_i and $N-1$ other USVs. The output $a_{u_i} = \Delta v_{u_i} = (\Delta v_{u_i}^x, \Delta v_{u_i}^y)$ is the velocity change of USV u_i. The main hyperparameters of the generation policies are shown in Table 1. The cumulative discounted return and the average length of the movement path of the team are used to evaluate the performance of the policies. However, for the formation maintenance task, all the followers follow the leader's movement while maintaining the whole formation geometry. In a formation maintenance scenario, the goal is to minimize the error between the current formation shape and the expected formation shape by following the decentralized formation maintenance policies $\boldsymbol{\mu}^{fm} = \left\{\mu_{\theta_1}^{fm}, \mu_{\theta_2}^{fm}, \ldots, \mu_{\theta_N}^{fm}\right\}$. The input of the formation generation policy of each USV u_i is a row vector $s_{u_i} = (p_{u_i}, v_{u_i}, re_{ileader})$, including the 1×2 position vector p_{u_i}, 1×2 velocity vector v_{u_i}, and 1×2 relative position vector $re_{ileader}$ between USV u_i and its immediate leader. The output is the change of velocity $a_{u_i} = \Delta v_{u_i} = (\Delta v^x_{u_i}, \Delta v^y_{u_i})$. The cumulative discounted return and the average of the error are utilized to measure the performance of the maintenance policies.

Table 1. The summary of the main hyperparameters and their values.

Hyperparameter	Value
Minibatch size	1024
Replay buffer size	10^6
Discount factor	0.95
Learning rate	0.01
Maximum episode length	100
Number of episodes	20,000
Number of units in the multilayer perceptron (MLP)	64

5.2. Results Analysis

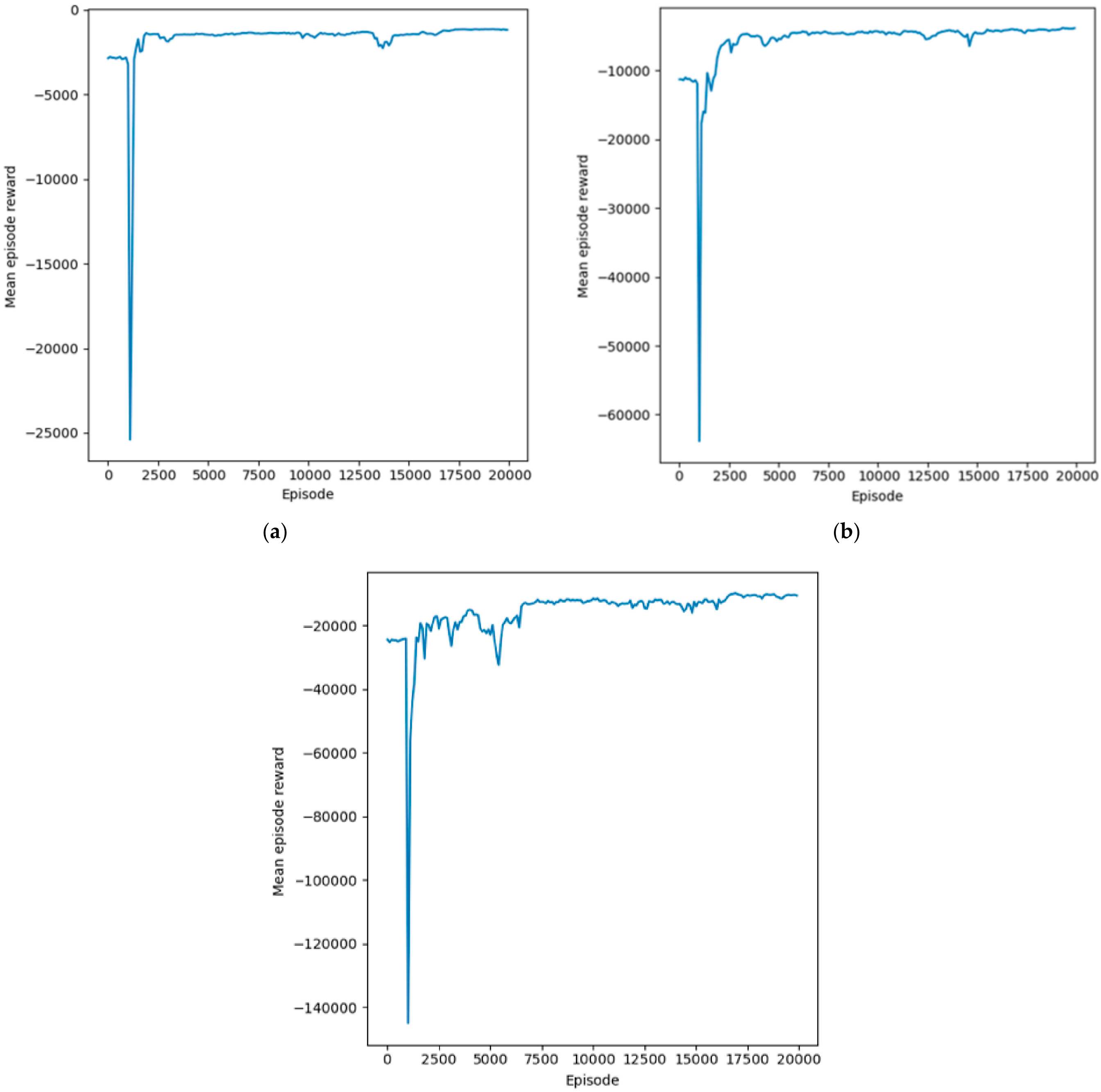
5.2.1. Formation Generation

We compare the performance of the proposed scheme with the following other schemes through simulation results and analysis of the results:

- The deep deterministic policy gradient (DDPG) scheme: In this scheme, USV learns formation generation policy based on the deep deterministic policy gradient.
- The deep Q-learning (DQN) scheme: In this scheme, USV learns formation generation policy based on deep Q-learning.

We train the decentralized formation maintenance policies based on the control objective in Equation (3). We evaluate the proposed scheme by averaging the episode reward for every 100 episodes. Figure 5 shows the mean episode reward of the USV team with different configurations of formations and different team sizes over 20,000 episodes when using the proposed formation generation policies. Figure 5a–c show the performance of the team with 3 USVs, 5USVs, and 7 USVs, respectively. The proposed scheme can learn effective formation generation policies for different USV team sizes. As shown in Figure 5, with the policy training going on, the formation generation policies based on deep reinforcement learning is continuously optimized, and the cumulative discounted return increases until the policy converges. Figure 5 shows that the proposed formation control scheme can perform formation generation with different team sizes effectively.

Next, we study the performance comparison of the proposed scheme and other existing schemes. Figure 6 shows the total length $len_{fg} = \sum_{i=1}^{N}(s(u_i) - e(u_i))$ and average length $Mlen_{fg} = \frac{1}{N}\sum_{i=1}^{N}(s(u_i) - e(u_i))$ of the moving path of the team with different team sizes. It illustrates that in the case of changing team size, the proposed scheme can still form an expected formation shape with the shortest path length. Moreover, the proposed scheme can obtain better performance than other schemes. Figure 6a shows that the total length of the USV team's moving path increases as the size of the team increases, and the proposed scheme has the shortest total length during the formation generation. Figure 6b shows that the proposed scheme has the shortest average length of the moving path during the formation generation. The reasons for these are as follows: first, the proposed scheme is used directly to parameterize the whole policy and find the optimal policy, which can get rid of the limitation of discrete action space. Second, the designed reward function is designed to drive all USVs to reach nodes in the predefined formation as soon as possible to maximize the cumulative discounted return of the team. Third, the advantage accelerates the learning of policies. As for the DDPG and DQN, the target locations in the generated formation for each USV maybe not the optimal, and the average moving path is not the shortest. Consequently, the proposed scheme is better compared with other schemes, as shown in Figure 6.

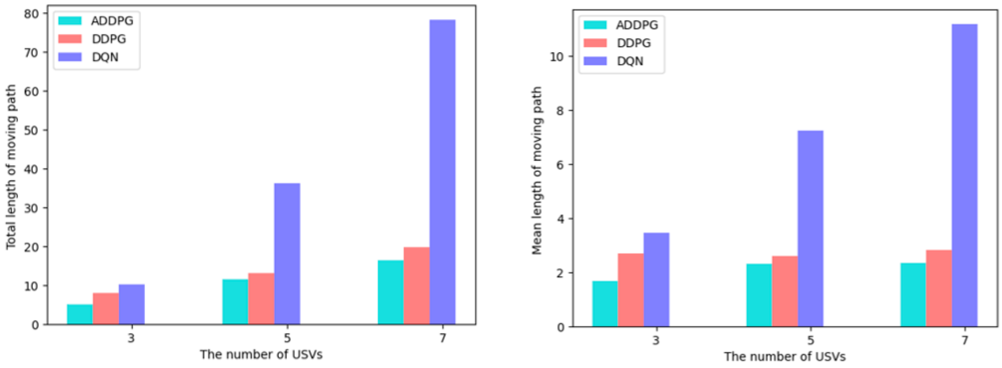


(**a**)

(**b**)

(**c**)

Figure 5. The reward of multiple cooperative USVs with formation generation policies for 40,000 episodes with different numbers of USVs: (**a**) mean episode reward for the team of 3 USVs during formation generation policies training over 20,000 episodes; (**b**) mean episode reward for the team of 5 USVs during formation generation policies training over 20,000 episodes; (**c**) mean episode reward for the team of 7 USVs during formation generation policies training over 20,000 episodes.



(**a**) (**b**)

Figure 6. The total length and the average length of the moving path with the different number of multiple cooperative USVs and different schemes: (**a**) comparison of total moving length of the USV team with different team sizes; (**b**) comparison of the total average length of the USV team with different team sizes.

5.2.2. Formation Maintenance

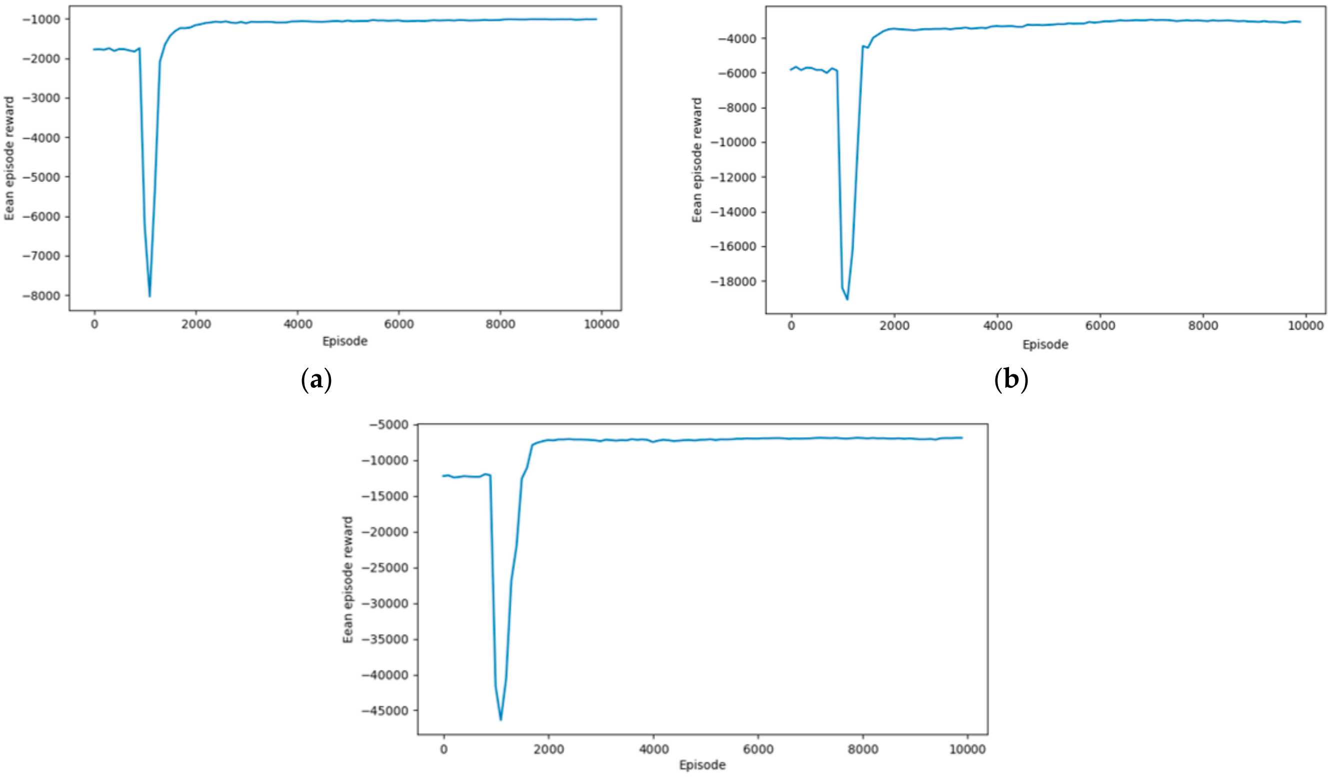
Aiming at the formation maintenance problem of the dynamic moving UAV team, we train the path planning algorithm for the leader and design the formation maintenance algorithm for the followers so that the whole formation team can move toward the target with a relatively stable geometric structure. Thus, we evaluate the performance of the proposed scheme according to the following evaluation criteria:

- Cumulative discounted reward during training
- The final distance between the leader and the team goal in each episode
- The stability difference of the whole team

We train the decentralized formation maintenance policies based on the reward function in Equations (12) and (13) and the control objective in Equation (4). Figure 7 shows the reward of the USV team with decentralized formation maintenance policies with different numbers of USVs and over 10,000 episodes. Figure 7a–c show the mean episode reward for the team of 3 USVs, 5USVs, and 7 USVs, respectively. We can see that the policies quickly converge to stable optimized policies.

In the formation maintenance scenario, the leader of the formation guides the movement of the USV team. Hence, the path planning algorithm of the leader determines the success of the formation task directly. As shown in Figure 8a–c, we test the performance of the formation maintenance policy of the leader in the team of 3 USVs, 5 USVs, and 7 USVs. The mean episode distance between the leader and the team goal all become near zero from about 2000 episodes. That means the leader in these teams can successfully reach the mission target quickly.

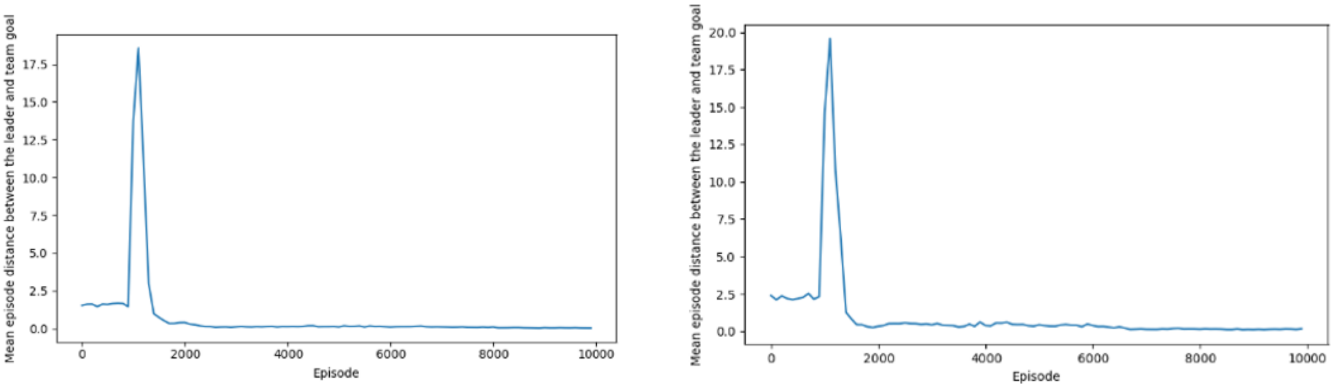
We adopt a stability error function $SF(U) = \frac{1}{T\times n}\sum_{j=0}^{T}\sum_{i=1}^{N}|H'_{i} - H_{i}|$ to measure the formation stability when following the leader. Comparison simulations are conducted amongst the proposed decentralized formation maintenance policies, DDPG, and DQN. The stability error of formation with different team sizes is selected for analysis, as shown in Figure 9. We can see that the average stability error in the proposed scheme is lower than that in other schemes. The results show that the followers can track their respective immediate leaders and maintain the predefined distance and direction from their immediate leaders. The proposed scheme can obtain the best performance for maintaining the formation compared with other schemes.



(**a**) (**b**)

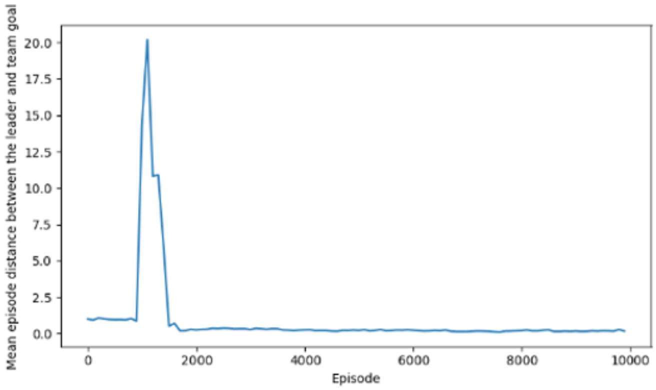
(**c**)

Figure 7. The reward of multiple cooperative USVs with formation generation policies for 40,000 episodes with different numbers of USVs: (**a**) mean episode reward for 3 USVs during formation generation policies training over 10,000 episodes; (**b**) mean episode reward for 5 USVs during formation generation policies training over 20,000 episodes; (**c**) mean episode reward for 7 USVs during formation generation policies training over 10,000 episodes.



(a) (b)

Figure 8. *Cont.*



(c)

Figure 8. The mean distance between the leader and the team goal with the proposed scheme and different team sizes during formation generation policies training: (**a**) mean distance for the leader in the team of 3 USVs; (**b**) mean distance for the team of 5 USVs; (**c**) mean distance for the team of 7 USVs.

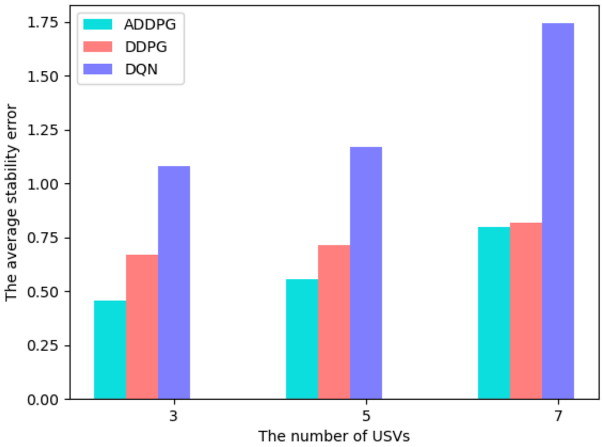


Figure 9. The average stability difference with the different number of multiple cooperative USVs and different schemes.

6. Conclusions

In this paper, we have proposed an asynchronous formation control scheme based on reinforcement learning and leader–follower structure for multiple USVs in a complex maritime environment. First, a specific USV model and a novel formation shape model have been established, where the formation shape model provides the parameters for formation generation and maintenance. Second, the formation control policies have been proposed for the cooperative USVs to generate the predefined formation shape with minimum moving path while the decentralized formation maintenance policies have been presented to maintain the stability of the geometric formation structure by minimizing the stability error between the real-time relative distances and the expected relative distances for all the USVs. Finally, simulation results have demonstrated that the proposed scheme can generate the required formation shape and maintain the geometry structure of the formation effectively compare with other schemes. In future work, we will take the communication interruption and the disturbance of wind, wave, and current into consideration.

Author Contributions: Conceptualization, J.L. and S.X.; methodology, Y.P. and J.X.; validation, J.X.; formal analysis, H.P.; investigation, Y.L. and R.Z.; writing and editing, J.X.; funding acquisition, S.X. All authors have read and agreed to the published version of the manuscript.

Funding: Research supported by the National Natural Science Foundation of China (grant no 61625304, 91748116 and 61922053), Project of Shanghai Municipal Science and Technology Commission (grant no 17DZ1205000), and Shanghai Innovation Action Plan of Science and Technology, China (grant no 19DZ1207305).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: We declare no conflict of interest.

References

1. Savitz, S.; Blickstein, I.; Buryk, P.; Button, R.W.; DeLuca, P.; Dryden, J.; Mastbaum, J.; Osburg, J.; Padilla, P.; Potter, A. *U.S. Navy Employment Options for Unmanned Surface Vehicles (USVs)*; RAND Corporation: Santa Monica, CA, USA, 2013.
2. Low, C.B. A flexible virtual structure formation keeping control design for nonholonomic mobile robots with low-level control systems, with experiments. In Proceedings of the 2014 IEEE International Symposium on Intelligent Control (ISIC), Juan Les Pins, France, 8–10 October 2014.
3. Zhang, B.; Zong, Q.; Dou, L.; Tian, B.; Wang, D.; Zhao, X. Trajectory Optimization and Finite-Time Control for Unmanned Helicopters Formation. *IEEE Access* **2019**, *7*, 93023–93034. [CrossRef]
4. Franchi, A.; Giordano, P.R. Online Leader Selection for Improved Collective Tracking and Formation Maintenance. *IEEE Trans. Control Netw. Syst.* **2018**, *5*, 3–13. [CrossRef]
5. Reynolds, C.W. Flocks, herds and schools: A distributed behavioral model. *Comput. Graph.* **1987**, *21*, 25–34. [CrossRef]
6. Olfati-Saber, R. Flocking for multi-agent dynamics systems algorithms and theory. *IEEE Trans. Autom. Control* **2006**, *51*, 401–420. [CrossRef]
7. Su, H.; Wang, X.F.; Lin, Z. Flocking of multi-agents with a virtual leader. *IEEE Trans. Autom. Control* **2009**, *54*, 293–307. [CrossRef]
8. Ponomarev, A.; Chen, Z.; Zhang, H.T. Discrete-Time Predictor Feedback for Consensus of Multiagent Systems with Delays. *IEEE Trans. Autom. Control* **2018**, *63*, 498–504. [CrossRef]
9. Chen, Z.; Zhang, H.T. A remark on collective circular motion of heterogeneous multi-agents. *Automatica* **2013**, *49*, 1236–1241. [CrossRef]
10. Yuanchang, L.; Richard, B. A survey of formation control and motion planning of multiple unmanned vehicles. *Robotica* **2018**, *36*, 1019–1047.
11. Tan, K.H.; Lewis, M.A. Virtual structures for high-precision cooperative mobile robotic control. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS ’96), Osaka, Japan, 8 November 1996.
12. Balch, T.; Arkin, R.C. Behavior-based formation control for multirobot teams. *IEEE Trans. Robot. Autom.* **1998**, *14*, 926–939. [CrossRef]
13. Chen, Y.Q.; Wang, Z. Formation control: A review and a new consideration. In Proceedings of the2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AL, Canada, 2–6 August 2005; pp. 3181–3186.
14. Peng, Z.; Wang, D.; Liu, H.H.; Sun, G.; Wang, H. Distributed robust state and output feedback controller designs for rendezvous of networked autonomous surface vehicles using neural networks. *Neurocomputing* **2013**, *115*, 130–141. [CrossRef]
15. Wang, J.; Liu, J.Y.; Yi, H. Formation Control of Unmanned Surface Vehicles with Sensing Constraints Using Exponential Remapping Method. *Math. Probl. Eng.* **2017**, *2017*, 7169086. [CrossRef]
16. Ghommam, J.; Saad, M. Adaptive Leader-Follower Formation Control of Underactuated Surface Vessels under Asymmetric Range and Bearing Constraints. *IEEE Trans. Veh. Technol.* **2017**, *67*, 852–865. [CrossRef]
17. Fahimi, F. Non-linear model predictive formation control for groups of autonomous surface vessels. *Int. J. Control* **2007**, *80*, 1248–1259. [CrossRef]
18. Do, K.D. Formation control of multiple elliptical agents with limited sensing ranges. *Automatica* **2011**, *48*, 1330–1338. [CrossRef]
19. Do, K.D. Formation control of underactuated ships with elliptical shape approximation and limited communication ranges. *Automatica* **2011**, in press. [CrossRef]
20. Peng, Z.; Wang, D.; Hu, X. Robust adaptive formation control of underactuated autonomous surface vehicles with uncertain dynamics. *IET Control Theory Applications* **2011**, *5*, 1378–1387. [CrossRef]
21. Lu, Y.; Zhang, G.; Qiao, L.; Zhang, W. Adaptive output-feedback formation control for underactuated surface vessels. *Int. J. Control* **2020**, *93*, 400–409. [CrossRef]
22. Lu, Y.; Zhang, G.; Sun, Z.; Zhang, W. Robust adaptive formation control of underactuated autonomous surface vessels based on MLP and DOB. *Nonlinear Dyn.* **2018**, *94*, 503–519. [CrossRef]
23. Ding, F.G.; Wang, B.; Ma, Y.Q. Adaptive coordinated formation control for multiple surface vessels based on virtual leader. In Proceedings of the 2016 35th Chinese Control Conference (CCC), Chengdu, China, 27–29 July 2016; pp. 7561–7566.
24. Sun, Z.; Zhang, G.; Lu, Y.; Zhang, W. Leader-follower formation control of underactuated surface vehicles based on sliding mode control and parameter estimation. *ISA Trans.* **2018**, *72*, 15–24. [CrossRef]

25. Shojaei, K. Leader–follower formation control of underactuated autonomous marine surface vehicles with limited torque. *Ocean Eng.* **2015**, *105*, 196–205. [CrossRef]
26. Shojaei, K. Observer-based neural adaptive formation control of autonomous surface vessels with limited torque. *Robot. Auton. Syst.* **2016**, *78*, 83–96. [CrossRef]
27. Sun, X.; Wang, G.; Fan, Y.; Mu, D.; Qiu, B. A Formation Collision Avoidance System for Unmanned Surface Vehicles with Leader-Follower Structure. *IEEE Access* **2019**, *7*, 24691–24702. [CrossRef]
28. Breivik, M.; Hovstein, V.E.; Fossen, T.I. Ship Formation Control: A Guided Leader-Follower Approach. *IFAC Proc. Vol.* **2008**, *41*, 16008–16014. [CrossRef]
29. Cui, R.; Sam Ge, S.; Voon Ee How, B.; Sang Choo, Y. Leader–follower formation control of underactuated autonomous underwater vehicles. *Ocean Eng.* **2010**, *37*, 1491–1502. [CrossRef]
30. Fan, Z.; Li, H. Two-layer model predictive formation control of unmanned surface vehicle. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 6002–6007.
31. Park, B.S. Adaptive formation control of underactuated autonomous underwater vehicles. *Ocean Eng.* **2015**, *96*, 1–7. [CrossRef]
32. Liu, Y.; Bucknall, R. The angle guidance path planning algorithms for unmanned surface vehicle formations by using the fast marching method. *Appl. Ocean Res.* **2016**, *59*, 327–344. [CrossRef]
33. Sui, Z.; Pu, Z.; Yi, J.; Wu, S. Formation Control With Collision Avoidance through Deep Reinforcement Learning Using Model-Guided Demonstration. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**. [CrossRef]
34. Hou, Z.; Fantoni, I. Leader-follower formation saturated control for multiple quadrotors with switching topology. In Proceedings of the 2015 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS), Cancún, Mexico, 23–25 November 2015; pp. 8–14.
35. Li, F.; Ding, Y.; Zhou, M.; Hao, K.; Chen, L. An affection-based dynamic leader selection model for formation control in multirobot systems. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *47*, 1217–1228. [CrossRef]
36. Li, F.; Ding, Y.; Hao, K. A neuroendocrine inspired dynamic leader selection model in formation control for multi-robot system. In Proceedings of the 2017 29th Chinese Control and Decision Conference (CCDC), Chongqing, China, 28–30 May 2017; pp. 5454–5459.
37. Xue, L.; Cao, X. Leader selection via supermodular game for formation control in multiagent systems. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3656–3664. [CrossRef]
38. Zelazo, D.; Franchi, A.; Bülthoff, H.H.; Robuffo Giordano, P. Decentralized rigidity maintenance control with range measurements for multi-robot systems. *Int. J. Robot. Res.* **2015**, *34*, 105–128. [CrossRef]
39. Sun, Z.; Helmke, U.; Anderson, B.D.O. Rigid formation shape control in general dimensions: An invariance principle and open problems. In Proceedings of the 2015 54th IEEE Conference on Decision and Control (CDC), Osaka, Japan, 15–18 December 2015.
40. Mou, S.; Belabbas, M.-A.; Morse, A.S.; Sun, Z.; Anderson, B.D.O. Undirected Rigid Formations Are Problematic. *IEEE Trans. Autom. Control* **2016**, *61*, 2821–2836. [CrossRef]
41. Chen, X.; Belabbas, M.A.; Başar, T. Global stabilization of triangulated formations. *SIAM J. Control Optim.* **2017**, *55*, 172–199. [CrossRef]
42. Sun, Z.; Mou, S.; Anderson, B.D.O.; Cao, M. Exponential stability for formation control systems with generalized controllers: A unified approach. *Syst. Control Lett.* **2016**, *93*, 50–57. [CrossRef]
43. Sun, Z.; Park, M.-C.; Anderson, B.D.O.; Ahn, H.-S. Distributed stabilization control of rigid formations with prescribed orientation. *Automatica* **2017**, *78*, 250–257. [CrossRef]
44. Aranda, M.; Lopez-Nicolas, G.; Sagues, C.; Zavlanos, M.M. Distributed Formation Stabilization Using Relative Position Measurements in Local Coordinates. *IEEE Trans. Autom. Control* **2016**, *61*, 3925–3935. [CrossRef]
45. Oh, K.-K.; Ahn, H.-S. Distance-based undirected formations of single-integrator and double-integrator modeled agents in n-dimensional space. *Int. J. Robust Nonlinear Control* **2014**, *24*, 1809–1820. [CrossRef]
46. Lin, Z.; Wang, L.; Chen, Z.; Fu, M.; Han, Z. Necessary and Sufficient Graphical Conditions for Affine Formation Control. *IEEE Trans. Autom. Control* **2016**, *61*, 2877–2891. [CrossRef]
47. Lin, Z.; Wang, L.; Han, Z.; Fu, M. Distributed Formation Control of Multi-Agent Systems Using Complex Laplacian. *IEEE Trans. Autom. Control* **2014**, *59*, 1765–1777. [CrossRef]
48. Timothy, P.L.; Jonathan, J.H.; Alexander, P.; Nicolas, H.; Tom, E.; Yuval, T.; David, S.; Daan, W. Continuous control with deep reinforcement learning. In Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
49. Mordatch, I.; Abbeel, P. Emergence of grounded compositional language in multi-agent populations. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.