IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 February 2023, accepted 23 March 2023, date of publication 29 March 2023, date of current version 17 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3263237

RESEARCH ARTICLE

Modeling Collective Behavior for Fish School With Deep Q-Networks

PENGYU CHEN, FANG WANG, SHUO LIU, YIFAN YU, SHENGZHI YUE, YANAN SONG, AND YUANSHAN LIN

College of Information Engineering, Dalian Ocean University, Dalian 116023, China
Liaoning Key Laboratory for Marine Information Technology, Dalian 116023, China
Key Laboratory of Environment Controlled Aquaculture, Ministry of Education, Dalian Ocean University, Dalian 116023, China

Corresponding author: Yuanshan Lin (linyuanshan2008@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61603067, in part by the Liaoning Province Natural Science Foundation under Grant 2020-KF-12-09, in part by the Foundation of Liaoning Educational Committee under Grant QL202016 and Grant LJKZ0730, and in part by the Liaoning Key Research and Development Program under Grant 2020JH2/10100043.

ABSTRACT Modeling collective behavior is a way to better understand the mechanisms that govern collective animal behaviors. Traditional rule-based modeling methods rely heavily on human prior knowledge and may not provide a proper explanation of the phenomenon of collective behaviors. This paper proposes a Deep Q-Networks (DQN)-based modeling method for fish school. Firstly, an individual's state (continuous value) is expressed by the angle between its direction and the average direction of its perceived neighbors. An individual's action is represented with discretized turning angle. Then, the reward function is constructed with the change in the number of neighbors. And finally, the neural network structure is constructed to represent the Q-value function and is trained by the DQN algorithm. The proposed approach is tested in two scenarios: single-learner and multi-learner. Results show that in both scenarios the proposed method can gradually converge and finally obtain a model that can produce collective behavior. On this basis, this paper also deeply analyzes the learned model from the perspectives of average order parameter and collective behavior patterns. It verifies that the behavior pattern generated by the learned model is a highly ordered collective behavior. In addition, we make a comparison between our proposed approach and the Q-Learning algorithm. The results show that our approach not only has a stronger ability to express policy and is better at handling continuous states but also has a more stable learning performance in training.

INDEX TERMS Collective behavior, collective behavior model, Deep Q-Networks (DQN), fish school.

I. INTRODUCTION

In nature, we often observe gregarious groups of organisms exhibiting various coordinated and orderly collective behaviors while flying, cruising, or moving. Fish collective behavior, as one of the most typical cases, has aroused great interest in statistical physics and theoretical biology [1], [2], [3], [4]. Despite a large number of studies on collective behavior for fish school, by what mechanism can those relatively simple fish perform such sophisticated behaviors remains an open question [5]. Collective behavior modeling is an important means to understand the interaction mechanism between individuals and the relationship between individual behavior and collective movement patterns. In the past few decades, many collective behavior models have been proposed. The main idea of these models is that individuals are reduced to featureless particles and subject to a fixed set of rules: alignment, cohesion, and separation. The three rules can be implemented in different ways, leading to different models. Although these models have provided an effective way to study the interaction among individuals, they almost always assume that fixed sets of rules are known in advance. Actually, obtaining these rules is often challenging and requires deep insight and domain knowledge. Furthermore, animals like fish do not seem to be governed by simple laws of nature like particles; they should be viewed as agents,

The associate editor coordinating the review of this manuscript and approving it for publication was Abderrahmane Lakas.

learning from their environment and adapting their behavior accordingly.

Therefore, with the development of reinforcement learning (RL), we proposed a Deep Q-Networks (DQN)-based collective behavior modeling method for fish school in this paper. Firstly, the continuous state representation and action representation of individuals are given. And then a reward function is constructed with the change in the number of neighbors. Finally, a neural network structure is constructed to represent the Q-value function and is trained with the DQN algorithm. The experimental results show that our proposed method can obtain a fish behavior policy that produces collective behavior in both single-learner and multi-learner situations. On this basis, this paper further analyzes the learned model in terms of order parameter and collective behavior patterns. In addition, the comparison results with the Q-Learning algorithm with higher state discrete resolution show that the proposed method not only has stronger policy expression ability, is better at dealing with continuous states, but also has stable learning performance during training.

II. RELATED WORK

Collective behavior models can be mainly divided into two categories: rule-based models and learning-based models. As one of earliest rule-based models, the Boids model was proposed in 1987 [6]. The author concluded that biological collective behavior has three basic characteristics: separation, alignment, and cohesion. Among them, the separation makes the individual avoid collision with the neighboring individuals in the group; the alignment makes the motion of the individual keep consistent with that of the neighboring individuals; the cohesion makes the individual approach the average position of the neighboring individuals. The simulation results show that using these three simple rules can produce collective behavior similar to fish school. Based on the above three basic rules, the Boids model was further developed. The individual-centered perception range is divided from the inside to the outside into three non-overlapping parts: zone of repulsion, zone of orientation, and zone of attraction [7]. Among them, the zone of repulsion has the highest priority, the focal individual is only subject to repulsion from other individuals in the zone once other individuals appear in this zone, producing a collision avoidance effect; if there are no other individuals in the zone of repulsion, the focal individual will be affected by other individuals in the zone of orientation and the zone of attraction at the same time, i.e., the orientation and attraction movements will take effect at the same time, making the focal individual keep the same direction with the individuals in the zone of orientation and move towards the individuals in the zone of attraction. The simulation results show that group formation with such rules has the ability to self-organizing. Following minimalism and trying to reveal the weakest condition of collective behavior, the Vicsek model was proposed by constructing a ''minimum model'' [8]. The individuals in this model only follow the alignment rule mentioned above, i.e., the direction of individual motion only depends on the average direction of all neighbors within the perception range. Its simulation results show that this minimalist model only with alignment can also generate collective behavior. Subsequently, many researchers have also proposed lots of extensions of Vicsek model [9], [10], [11], [12], [13], [14], [15], [16]. Furthermore, recent empirical studies have collected large datasets of animals' movement to infer the rules underlying their emergent collective behavior and provided some evidence for the above traditional models [17], [18], [19], [20], [21]. Although the above-mentioned traditional rule-based models meet the research needs of understanding the internal mechanism of collective behavior to a certain extent. However, rule-based models are inherently limited in that the pre-fixed rules rely heavily on human prior knowledge, and are very difficult to be proposed and constructed.

On the other hand, reinforcement learning methods have driven impressive advances in artificial intelligence in recent years, surpassing human performance in many domains [22], [23], [24], [25], [26], [27]. More recently, some researchers have begun to use reinforcement learning to model collective motion in a learning way [28], [29], [30], [31]. An RL-based collective behavior model was proposed for the self-organized grouping of individuals, in which the reward function was constructed based on the distances between individuals [32]. Multi-agent reinforcement learning was attempted to derive individual behavior [33]. Specifically, with the data obtained from Reynolds' Boids model, the RL model was trained by minimizing the entropy difference between the rule-based model and the RL model by using natural evolution strategies (NES). Likewise, a collective behavior model based on RL was obtained by maximizing the group-level objective function (total reward during a simulated episode) representing the desired collective configuration [34]. They showed that collective behavior models learned through different reward functions can make fish schools form different motion patterns. The above model can generate a certain collective pattern and avoid the problem of imposing mechanical laws on individuals like traditional models. However, they were done by shaping the reward signal according to the three rules of alignment, cohesion, and separation. Essentially, it is still not out of the shackles of the rules.

In order to get rid of the shackles of the rules, the agent's reward function was designed in different ways in some studies. Under the assumption that a simple goal can make a school of fish form a collective behavior, the predator-prey model and the cooperative observation model based on RL were proposed in [35] and [36]. In the predator-prey model, prey agents are encouraged to find strategies that allow them to survive longer by rewarding them at each time step. In the cooperative observation model, the reward obtained by an agent at each time step depends linearly on its distance from the observation target. The goal of the agent is to follow the target object for as long as possible. In the event that the agent loses visual contact with the target object, the agent will orient itself to its nearby agents and eventually form a

collective behavior. In contrast, an agent received a negative reward signal in the form of a cost for losing neighbors within its perception [37]. Results showed that collective behavior emerges spontaneously in an RL process from the minimization of the rate of neighbor loss. In this study, the authors used the Q-learning to obtain the collective behavior model. Specifically, the movement policy of a fish individual was represented with a table and the state of a fish individual was represented with a discretized angle between the individual's moving direction and the average moving direction of its neighbors. Such state representation may not accurately describe the real state of fish individuals for decision-making, and a simple table may not be able to fully express the fish behavior decision with highly nonlinear time-varying characteristics in the real world.

This paper proposes a fish collective behavior modeling method that solves the problem that individual states are expressed as discrete values, uses a neural network with stronger expressive ability to express the agent's policy, and designs a reward function based on the change of the number of neighbors to get rid of the restrictions of the fixed rules.

III. METHODS

It is well known that complex collective behavior can emerge from simple local interactions of fish who lack the ability to observe or directly control the collective. The local interactions of fish can be viewed as the policies by that fish respond to their observations. So, modeling collective behavior can be reduced to find the fish's policies that can generate complex collective behavior patterns commonly observed in nature. Therefore, in this paper, the fish individual is modeled as a learning agent, and the DQN algorithm is used to obtain its movement policy. To better illustrate the method in this paper, we first introduce the movement model of the fish school.

A. KINEMATIC MODEL OF A FISH INDIVIDUAL

In the real world, fish can freely swim by beating their fins in certain three-dimensional spaces, such as ponds, rivers, lakes, or marine environments, and usually form collective behavior patterns. For the sake of simplicity, here we consider the artificial fish move in a two-dimensional box with periodic boundary conditions. It means that a fish individual leaves the box from a certain boundary and enters the box from the opposite boundary (see Fig. 1). And we assume that a fish school consists of many homogeneous fish individuals with the same perceptual and motor ability.

Here, it is assumed that a fish individual has first-order kinematics during its swimming. That is, its position is determined by its velocity. So, the position update of an individual at each time step can be expressed as:

$$\boldsymbol{r}_i^{t+1} = \boldsymbol{r}_i^t + \Delta t v_0 \boldsymbol{v}_i^t \tag{1}$$

where, $\boldsymbol{r}_i^t$ is the position of the individual i at time step t; v_0 is the linear velocity of the individual; Δt is the time step; $\boldsymbol{v}_i^t$ is the moving direction of the individual i at time step t, and its update method is defined by:

$$\boldsymbol{v}_i^{t+1} = \boldsymbol{v}_i^t + \left[cosa_i^t, sina_i^t\right] \tag{2}$$

where, a_i^t is the angular velocity at time step t. It should be noted that $\left\|\boldsymbol{v}_i^t\right\| = 1$.

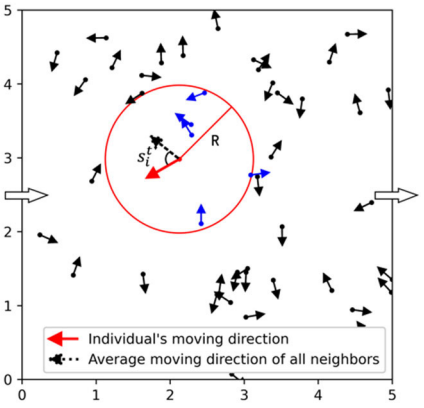

FIGURE 1. The perceptual area of a fish individual and its environment.

B. STATE REPRESENTATION OF FISH INDIVIDUAL

It has been hypothesized that vision is a primary mode of perception for fish. Its monocular field of view is greater than 160°, and the compound field of vision intersecting the binocular field of view is 20° ∼ 30°. A field of vision close to 360° can be achieved by beating its fin, but the perception distance of the individual fish is limited. So, the fish individual can perceive the underwater environment within a limited range of view with an angle of 360° centered on it. Thus, we can think of the perceptual area of the fish individual as a circle of radius R (see Fig. 1). In Fig. 1, the arrows represent the moving direction of individuals, the red circle is the perception range of the focal fish (red individual), and blue arrows are the neighbor individuals perceived by the focal fish.

Therefore, the contextual information of a fish individual includes its pose and its neighbors' poses. For simplicity, the contextual information is represented by the angle between the individual's moving direction and the average moving direction of all neighbors within its perception range, defining the current state of the individual. This kind of state representation is based on the mean-field theory. It means that the movement of each individual in the fish school is affected by the movement of all its neighbors in the visual field. The purpose of using averages is to reduce the complexity of interaction. Following this way, the state of the i-th individual at time step t can be expressed as:

$$s_i^t = \boldsymbol{P}_i^t \cdot \left(\boldsymbol{v}_i^t\right)_\perp arccos\left(\boldsymbol{P}_i^t \cdot \boldsymbol{v}_i^t / \left\|\boldsymbol{P}_i^t\right\|\right) \tag{3}$$

where, $\boldsymbol{P}_i^t$ is the average moving direction of all neighbors within the perception range of the i-th individual at time step

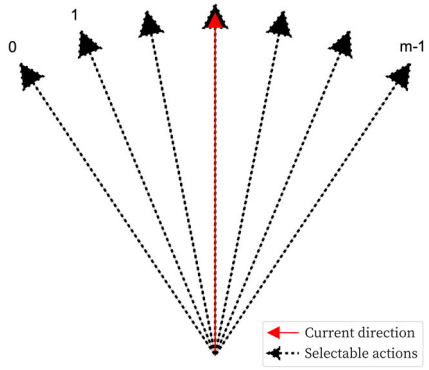

FIGURE 2. Action representation of a fish individual.

t, and it was determined by:

$$\boldsymbol{P}_i^t = \left(\sum_{\left|r_j^t - r_i^t\right| < R} \boldsymbol{v}_j^t\right) \Big/ n_i \tag{4}$$

where, $\boldsymbol{r}_j^t$ is the position of neighbor j in the environment at time step t; R is the radius of the individual perception range; n_i is the number of neighbors within the perception range of i-th individual; and $\boldsymbol{v}_j^t$ is the moving direction of neighbor j at time step t. $arccos\left(\boldsymbol{P}_i^t \cdot \boldsymbol{v}_i^t / \left\|\boldsymbol{P}_i^t\right\|\right)$ is the angle between the individual moving direction $\boldsymbol{v}_i^t$ and the average moving direction $\boldsymbol{P}_i^t$ of neighbors within the perception range. $\left(\boldsymbol{v}_i^t\right)_\perp$ is obtained by rotating $\pi/2$ counterclockwise in the individual moving direction $\boldsymbol{v}_i^t$. It should be pointed out that the value range of s_i^t in (3) is $s_i^t \in [-\pi, \pi]$. And the state s_i^t is positive when the average moving direction of neighbors is in the left of the individual's moving direction; otherwise, it is negative. Specifically, the sign of the state s_i^t is determined by $\boldsymbol{P}_i^t \cdot \left(\boldsymbol{v}_i^t\right)_\perp$.

C. ACTION REPRESENTATION OF FISH INDIVIDUAL

Linear velocity and angular velocity are the two control variables of the movement. As in [32], we assume that the linear velocity is constant. Thus, the angular velocity becomes the only control variable for a fish individual. Since a fish usually has a limited angle of rotation, notated as θ_{max}, its angular velocity is in $[-\theta_{max}, \theta_{max}]$. In the RL language, the range of $[-\theta_{max}, \theta_{max}]$ is viewed as the action space of an agent. For computational simplicity, we discretize the action space by dividing $[-\theta_{max}, \theta_{max}]$ into m equally spaced elements (Fig. 2). As shown in Fig. 2, the red arrow is the current moving direction of the fish individual; the black dotted arrows are discretized actions, which can be chosen by the fish individual.

At each time step, each fish individual decides on whether to keep the current heading direction or perform a turn. The decision-making process is based on the state (sensorial input) of the fish individual, which corresponds to the angular difference between the individual's moving direction and the average moving direction of its neighbors.

D. DESIGN OF THE REWARD FUNCTION

The reward function is an incentive mechanism that tells the agent what is correct and what is wrong using reward and punishment. Maximizing the total rewards according to certain reward functions is the goal of agents in RL. Sometimes in order to maximize the total rewards agents need to sacrifice immediate rewards. Therefore, the reward function plays important role in RL. If the reward function is "better behaved", the agents will learn better and faster.

Since there are many neighbors within a fish's surroundings when fish form collective behavior, we design the reward function based on the change in the number of neighbors, which can be described as:

$$reward_i^{t+1} = \begin{cases} 1 & other \\ 0 & if\ n_i^{t+1} < n_i^t \end{cases} \tag{5}$$

The idea behind this design is that no matter what kind of collective behavior, one individual is prone to have more and more neighbors when fish schooling. A fish individual gets a reward of 0 if the number of its neighbors decreases; otherwise, it gets a reward of 1 if the number of its neighbors remains the same or increases. Such reward function encourages individuals to choose actions that can increase the number of neighbors.

E. NEURAL NETWORK OF Q-FUNCTION

In this paper, the movement policy of a fish individual is approximated by a multi-layered neural network that outputs Q-values of all executable actions for a given state s_i^t. Here, the neural network is a function from a one-dimensional state space to an action space containing m actions. And the neural network is trained with DQN. To reduce correlations with the target value, we introduce two neural networks: an online network (Q-network) and a target network (Q'-network). The online network is used to estimate action value, whose parameters are updated at every learning step. The target network is used to calculate the target value, whose parameters are updated every K learning steps and kept fixed on all other steps. The online network and the target network have the same architecture.

Theoretically, the more layers and neurons in the neural network, the stronger the expressive ability of the neural network. However, a complex neural network structure will increase the number of neural network parameters, which will increase the computational time for updating the neural network parameters. Therefore, the exact architecture of the online network and the target network is designed as shown in Fig. 3. The input to the neural network is the state of a fish individual, represented by one neuron. The hidden layer is fully-connected and consists of 10 rectifier units. And the output layer is a fully-connected linear layer with a single output for each valid action.

F. TRAINING

The pseudo-code of the DQN-based learning algorithm for fish individuals is shown as Algorithm 1. In the beginning, the algorithm initializes the Q-network and Q'-network with random weights. Then the learning algorithm performs a series of episodes of interaction with the environment to collect

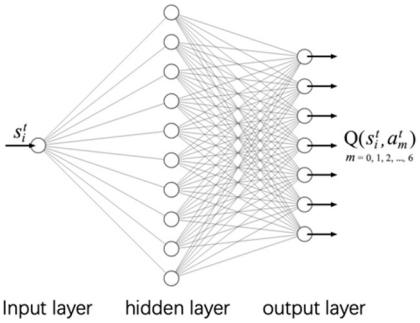

FIGURE 3. The architecture of Q-network.

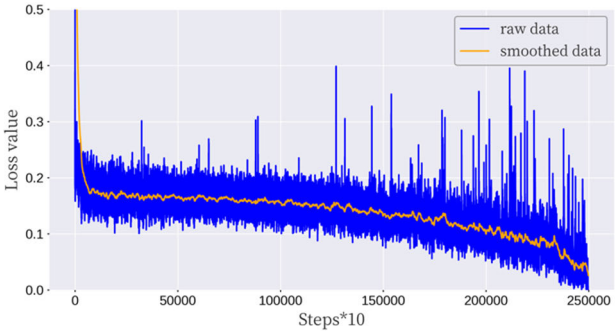

FIGURE 4. Loss curve for single-learner experiments.

training data and updates the parameters of the networks. The termination condition of each episode is the number of time steps. The positions and moving directions of all fish individuals are randomly initialized at the start of each episode. At each time step t, each fish individual obtains its state s_i^t and feeds it to Q-network. The individual samples its action a_i^t according to the Q-values output by the Q-network. After the action a_i^t is performed, the algorithm can get the interaction data of this step, represented with a 4-tuple $\left(s_i^t, a_i^t, reward_i^{t+1}, s_i^{t+1}\right)$, and put it into the replay buffer. Once the replay buffer is full, the update operation of the Q-network parameters is started. The existence of the replay buffer enables the use of experience during training, which improves the efficiency of data utilization and increases the diversity of training data.

In order to make a good balance between exploitation and exploration during the learning process, we use the ε-*greedy* strategy to select actions:

$$a_i^t = \begin{cases} \underset{a'}{argmax}\, Q\left(s_i^t, a'\right) & with\ probability\ 1-\varepsilon \\ random\ action & with\ probability\ \varepsilon \end{cases} \tag{6}$$

where, $\varepsilon \in [0, 1]$, the value of ε gradually decreases from 1 to 0 during the training process. And as the training goes on, the probability of selecting random actions gradually decreases.

The time complexity of the algorithm is $O\left(E \times T\right)$, where E is the number of episodes and T is the number of time steps in each episode.

IV. EXPERIMENTAL RESULTS

We evaluate our approach with two sets of experiments. The first set of experiments focuses on the single-learner case, in which only one individual acts as a learner and the others act as teachers equipped with a fixed policy. The single-learner experiments aim to verify whether the single-learner can learn the teachers' policy to form collective behaviors. The second set of experiments discusses the multi-learner case, in which all individuals act as learners, and each is influenced by its neighbors. The multi-learner case aims to verify whether all individuals can simultaneously learn policies that form collective behavior under dynamic uncertainty.

A. SINGLE-LEARNER EXPERIMENTS

In the single-learner experiments, the single-learner aims to learn a policy from teachers so that they can form collective behaviors. All teachers have the same policy. And their policy is fixed as the Vicsek model [8]. The policy representation is defined by:

$$a_i^t\left(s_i^t\right) = s_i^t \tag{7}$$

In this set of experiments, we break the training session into 500 training episodes of equal prescribed duration of 5000 time steps. And the size of replay buffer is set to 2000, the batch size for extracting data from the buffer is set to 32, and the learning rate is set to 0.01. The parameters of the Q-network are updated every time step, and the parameters of the target network Q′ are updated every 100 time steps. At the start of each episode, both the position and moving direction of each individual are randomly initialized. And then all teachers use their policy to move in the successive time steps. The learners start to learn when teachers form collective behavior after 100-time steps.

1) TRAINING PROCESS

During the training process, we calculated the average loss value every 10 steps. Fig. 4 shows the average loss curve. The average loss value went down with the increase of training steps, which means that the model can converge. It can also be seen that the variance is larger in the later stage of training, which is mainly because the model has learned the policy in the middle and later stages. And in the later stage of training, because the individual follows the teacher group movement for a long time, the model is better at dealing with states around 0°, which weakens the ability to deal with other states.

The order parameter, calculated by averaging all $\boldsymbol{v_i}$ at time step t, is one of the indicators to evaluate whether a fish school has formed a collective behavior:

$$\Psi\left(t\right) = \frac{1}{N}\left\|\sum_{i=1}^{N} \boldsymbol{v_i^t}\right\| \tag{8}$$

it ranges from 0 to 1. The closer the order parameter is to 1, the more identical their directions are. The closer the order parameter is to 0, the more different their directions are.

Algorithm 1 Q-Network Parameter Update

Initialize the parameters of the Q-network
Assign the parameters of the Q-network to the Q'-network
for $episode = 1, E$ do
  Initialize the positions and movement directions of all individuals
  Get the initial state s_i^t
  for $t = 1, T$ do
    With probability ε select random action a_i^t,
    otherwise select $a_i^t = \arg\max Q\left(s_i^t, a'\right)$
    Execute action a_i^t and get $reward_i^{t+1}, s_i^{t+1}$
    Store $\left(s_i^t, a_i^t, reward_i^{t+1}, s_i^{t+1}\right)$ into the buffer pool
    if buffer pool is full then
      Sample random minibatch of data $\left(s_j, a_j, reward_{j+1}, s_{j+1}\right)$ from buffer pool
      Get $Q\left(s_j, a_j\right)$ by inputting s_j into the Q-network
      Get $Q'\left(s_{j+1}, a'\right)$ by inputting s_{j+1} into the Q'-network
      Get target value $reward_{j+1} + \gamma Q'\left(s_{j+1}, a'\right)$
      Perform a gradient descent step on $\left(reward_{j+1} + \gamma Q'\left(s_{j+1}, a'\right) - Q\left(s_j, a_j\right)\right)^2$ with respect to the Q-network parameters
    end if
    if t mod $K = 0$ then
      Update Q'-network parameters by copying Q-network parameters
    end if
    $s_i^t \leftarrow s_i^{t+1}$
  end for
end for

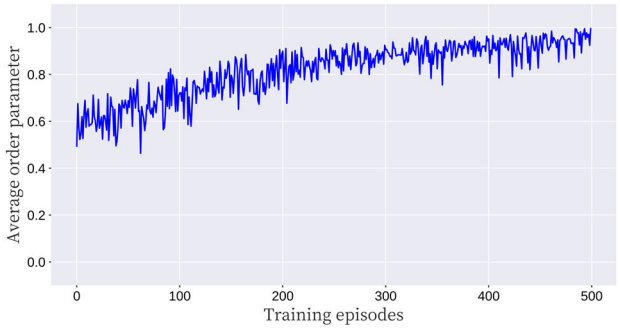

FIGURE 5. Average order parameter curve for single-learner experiments.

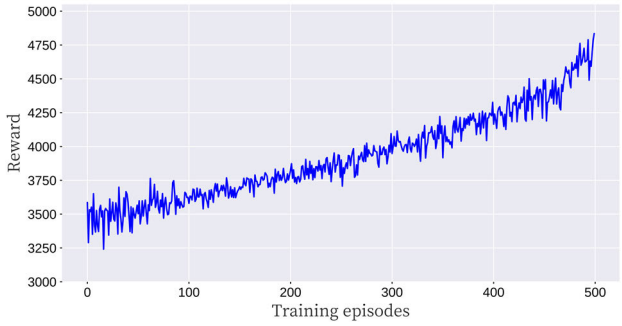

FIGURE 6. Reward curve for single-learner experiments.

In this experiment, we are concerned with the order parameter in the learner's perception zone. If there are no neighbors, the order parameter is set to 0. The average order parameter of each episode (5000 time steps) is selected as the statistic, and its curve is shown in Fig. 5. It can be seen from Fig. 5 that it went up with the increase of episodes. In the first 100 episodes of training, the average order parameter is between 0.5 and 0.8. After 300 episodes, it reaches above 0.9, which indicates that the learner and its neighbors have similar directions. They move orderly and form a pattern of collective behaviors.

Aggregation is another indicator for characterizing collective behaviors. Fig. 6 shows the learner's rewards curve in the training process. As the training goes on, the rewards increase gradually. According to reward function (5) the increase of reward is due to the increase of neighbors gathering around the learner. It follows that the learner has learned a policy to form a pattern of collective behavior with as many neighbors as possible.

2) EFFECTIVENESS OF THE LEARNED POLICY

To verify the effectiveness of the learned policy, three experiments with different learner-teacher ratios were conducted. The ratios were set to 1:49, 25:25, and 50:0, respectively. In each experiment, all learners were equipped with the learned policy. We snapshot all individuals' configurations at various time step t within an episode. Fig. 7-9 show the test results, respectively, in which the yellow arrows indicate the learners, and the black arrows indicate the teachers.

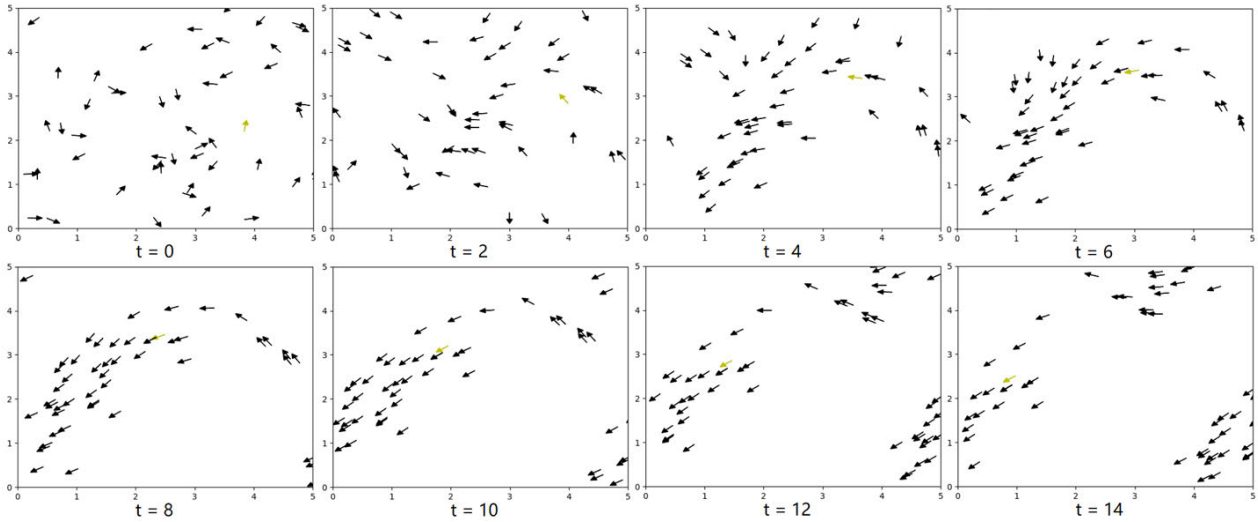

FIGURE 7. All individuals' configurations for learner-teacher ratio 1:49.

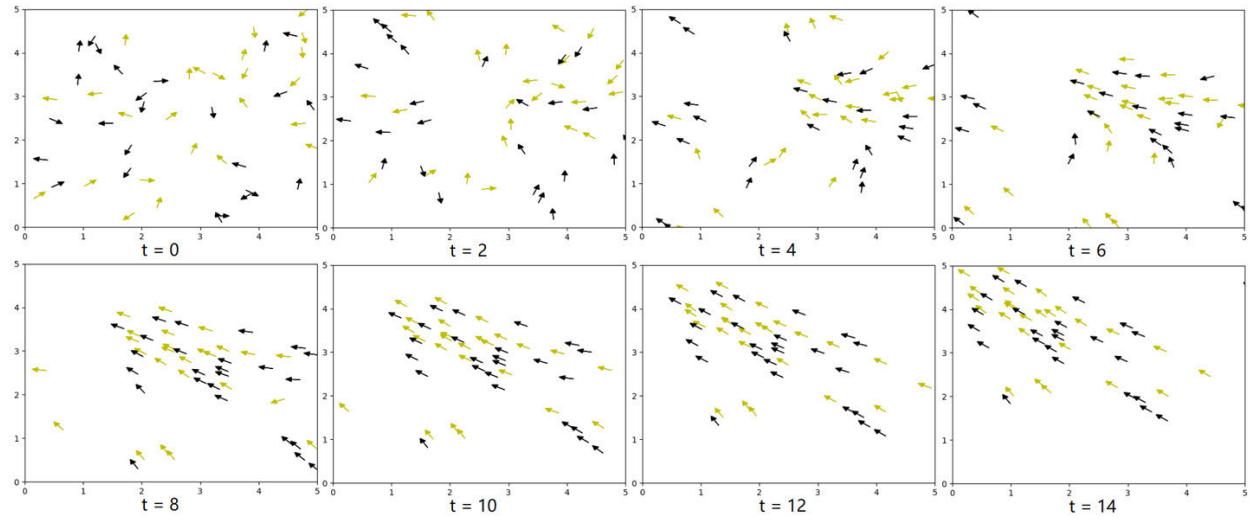

FIGURE 8. All individuals' configurations for learner-teacher ratio 25:25.

From Fig. 7-9, we can see that fish individuals autonomously form highly parallel group from the initially disordered configuration, i.e., the fish individuals can always form a pattern of collective behaviors no matter what the learner-teacher ratio is. It proves that the policy learned by the single-learner is effective.

The results of convergence, order parameter, and effectiveness of the learned policy above suggest that the single-learner can learn a policy to form a collective behavior from teachers.

B. MULTI-LEARNER EXPERIMENTS

In the multi-learner experiments, all fish individuals are learners. There are no teachers in the environment. The learners share the same Q-network and learn a common policy from each other. All learners learn at the same time at the start of each episode, and the rest of the experimental settings are the same as the single-learner experiments.

1) TRAINING PROCESS

Fig. 10 shows the curve of the average loss during the training process in the multi-learner experiment. We observed that the average loss goes down with the increase of training steps, which indicates that the common policy learned by all individuals is converge.

In this experiment, we are concerned with the order parameter of the whole group in the environment. The curve of the average order parameter is shown in Fig. 11. In the first 100 episodes of training, the average order parameter is below 0.2, indicating that the learners are moving in disorder. It reveals that they have not learned a policy that enables them to form a pattern of collective behavior. As the training went

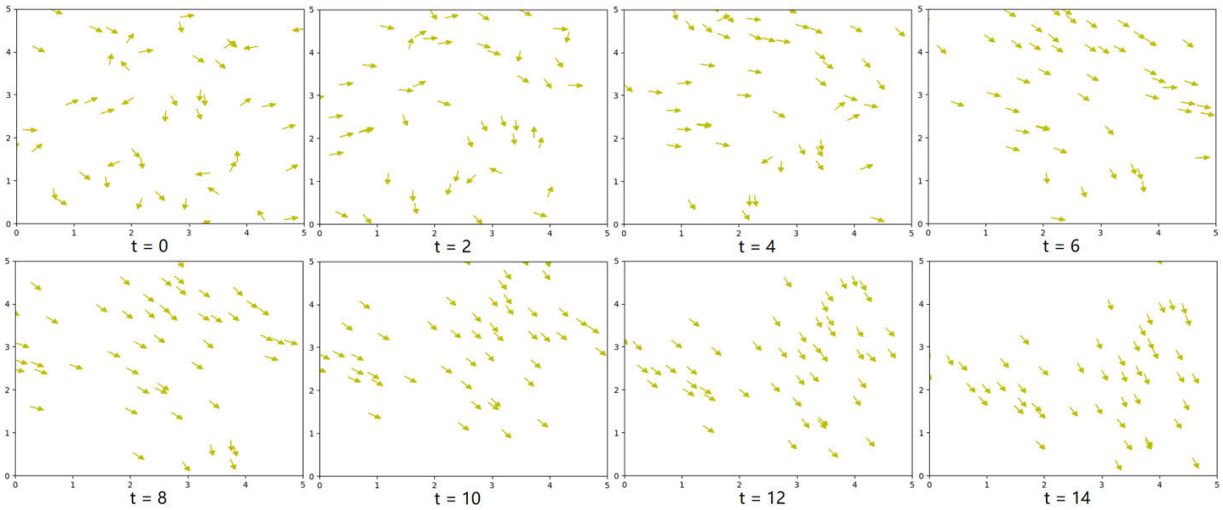

FIGURE 9. All individuals' configurations for learner-teacher ratio 50:0.

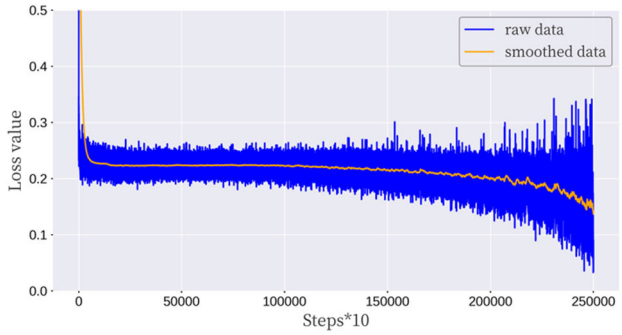

FIGURE 10. Loss curve for multi-learner experiments.

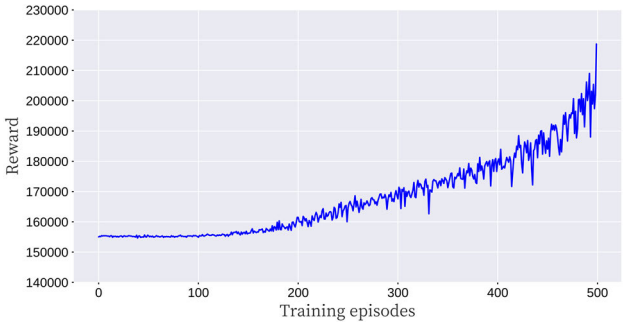

FIGURE 12. Reward curve for multi-learner experiments.

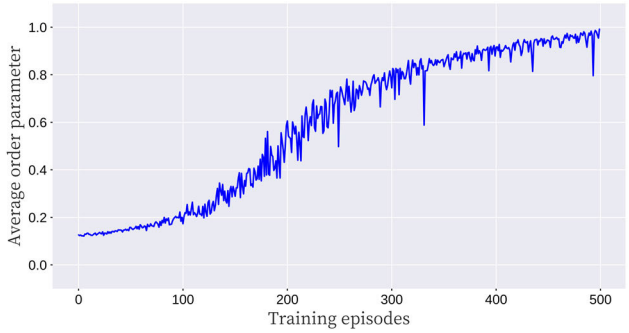

FIGURE 11. Average order parameter curve for multi-learner experiments.

on, the average order parameter increases gradually. After 400 episodes, it reaches above 0.9, indicating that all learners can move in order with their neighbors within their perception range.

Fig. 12 shows the curve of rewards during training process in the multi-learner case. As the training goes on, the rewards increase gradually, which suggests that all learners have more and more neighbors around them to form collective behavior. Compared with Fig. 6, the total reward did not increase until the middle stage of the training, which suggests that multi-learners are more difficult to learn a policy than single-learner. This is because when all individuals are learners and they learn at the same time, the dynamic change of each learner's policy leads to the dynamic change of each learner's environment. Even if some learners have obtained the optimal policy, they will be changed, which has a great influence on the convergence of policy. Fortunately, they learned the policy of forming a collective behavior in the end.

2) EFFECTIVENESS OF THE LEARNED POLICY

To verify the effectiveness of the policy learned by the learners, we conduct an experiment. In this experiment, 50 learners were equipped with the learned policy. We snapshot all individuals' configurations at various time step t within an episode (see Fig. 13). We start recording from time step t=0, and record once every 2 time steps. It can be seen from Fig. 13 that 50 learners with the learned policy can gradually form an orderly state from a disordered state, which indicates that the learned policy is effective for forming collective behavior. From t=16 to t=30, after 50 learners form collective behavior, they present a rotating trend.

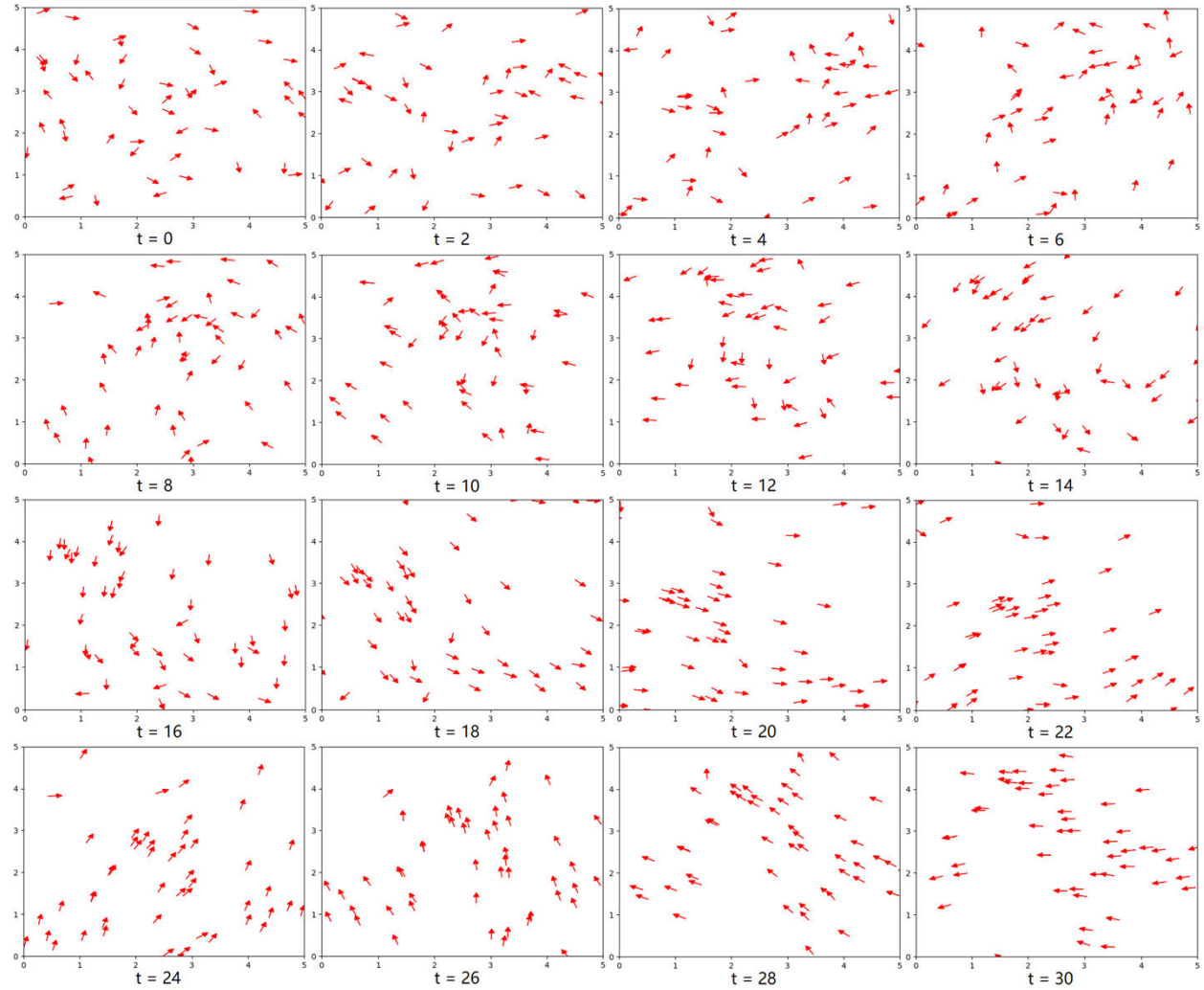

FIGURE 13. Test results of multi-learner experiments.

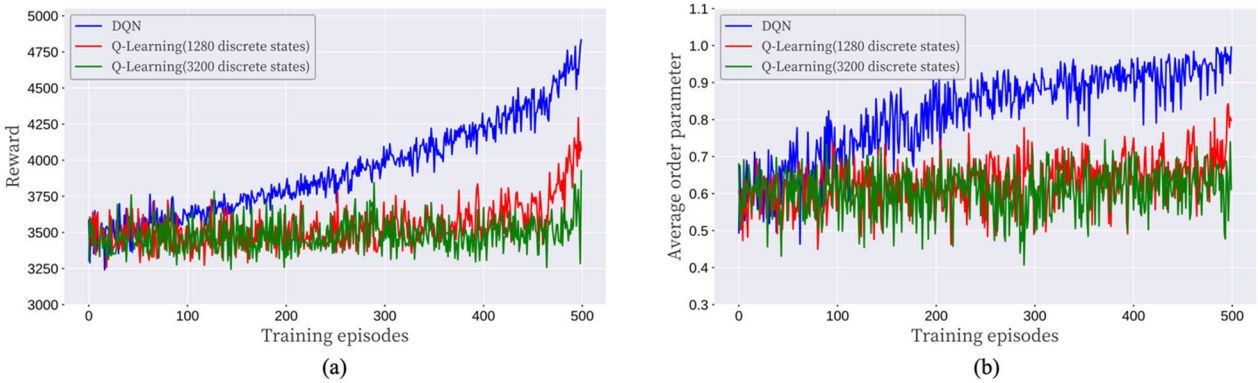

FIGURE 14. Single-learner experiment on DQN vs. Q-Learning, (a) reward curve, (b) average order parameter curve.

C. TRAINING PERFORMANCE EVALUATION

In this section, we compared our method with that in [32]. In order to approximate the continuous state, the individual's perceived zone was discretized into 1280 states and 3200 states respectively. The single-learner experiments and the multi-learner experiments were all conducted in the same settings as above. Fig. 14 shows the reward curve and the average order parameter in the case of a single learner. Fig. 15 shows the reward curve and the average order parameter in the case of multi-learner.

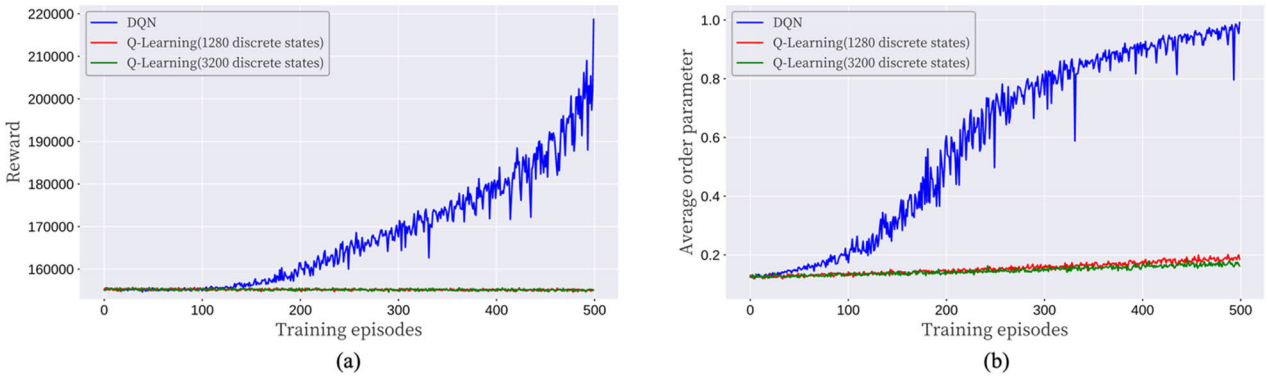

FIGURE 15. Multi-learner experiment on DQN vs. Q-Learning, (a) reward curve, (b) average order parameter curve.

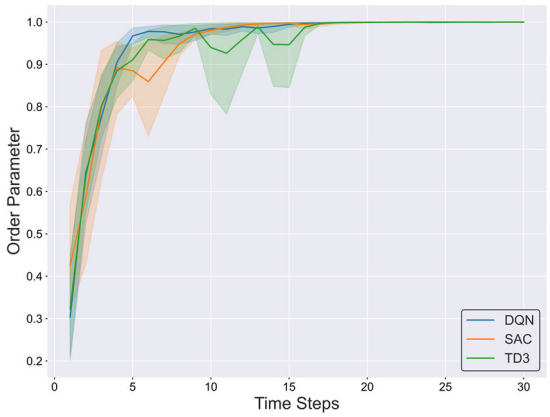

FIGURE 16. Comparison of order parameter for models obtained by three algorithms in testing experiments.

In the case of single-learner, we observed from Fig. 14(a) that rewards of DQN increase faster than those of Q-learning. At the end of the training, the reward can reach more than 4750. However, it is not the case with Q-Learning. The reward of Q-Learning with 1280 discrete states does not start to increase until the last 50 episodes. In the end, the reward reaches more than 4250. The reward of Q-Learning with 3200 discrete states has no growth trend. And it fluctuates all the time, never exceeding 4000 during the whole training. From Fig. 14(b), we observe that when using the DQN algorithm, the average order parameter can reach more than 0.9 after 300 episodes of training. However, when using the Q-Learning algorithm with discrete 1280 states and 3200 states, the average order parameter cannot reach 0.9.

In the case of multi-learners, we observe the same results as in the case of single-learner. The rewards of DQN increase faster than those of Q-learning, see Fig. 15(a). At the end of the training, the reward can reach 220,000. However, the reward of Q-Learning with both 1280 and 3200 discrete states does not throughout the training. From Fig. 15(b), we observe that when using the DQN algorithm, the average order parameter can reach 1.0 at the end of training. However, when using the Q-Learning algorithm with discrete 1280 states and 3200 states, the average order parameter cannot reach 0.2.

The above experiments tell us that when the resolution is too high, the DQN-based method proposed in this paper can learn a policy with higher rewards to form orderly collective behaviors, but the method in [32] fails. It follows that our method is better than that in [32], no matter in the case of single-learner or multi-learner.

V. DISCUSSION

Although our research mainly focuses on whether the collective behavior of fish can be modeled through DQN, we actually also have attempted to use TD3 and SAC to get the policy of fish. From the experimental results, we found that these methods can all obtain individual behavior policies that generate collective behavior, with similar learning processes and similar behavior policies. The final learned models can all make individuals form effective collective behavior. However, when we tested the final learned models of the three algorithms, we found that the model obtained by the DQN algorithm can form collective behavior more quickly and has more stable performance than those obtained by the other two algorithms. Fig. 16 shows a comparison of the order parameter changes of the models obtained by the three different algorithms in the test experiment. Each model was tested for 10 episodes with 30 time steps per episode. It can be seen that all three algorithms can eventually make individuals form highly ordered collective behavior, but the model obtained by the DQN algorithm can quickly increase the order parameter to above 0.95 and maintain it steadily above 0.95 in subsequent time steps. We think that this result may be due to the fact that the output of the DQN algorithm's policy is discrete actions, and we have set the action of maintaining straight motion in its action space, which makes individuals form collective behavior faster and better maintain it.

Since the output of the policy learned by DQN algorithm is a discrete value, we also believe that this policy is easier to analyze and explore the motion patterns of individuals forming collective behavior. In addition, DQN is a classical and mature model-free deep reinforcement learning algorithm, making it easier to implement. Based on the above content, we finally chose the DQN algorithm for our experiments.

The experimental results show that our proposed DQN-based method for modeling collective behavior of fish can obtain individual movement policies that form collective behavior, proving that deep reinforcement learning has great potential as a new method for analyzing and modeling collective behavior.

VI. CONCLUSION

This paper proposed DQN-based method to model the collective behavior for fish. The movement policy of a fish individual can be represented with a neural network, whose input was a continuous state of a fish individual, the relative angle between its direction and the average direction of its perceived neighbors. And the change of the number of neighbors was used as a reward signal to guide individuals' learning. Two classes of experiments (single-learner and multi-learner) were conducted. The results showed that the proposed method can obtain models that can generate collective behavior in both the single-learner case and the multi-learner case. It revealed that the policy can be obtained via RL even though the observation was represented with a continuous state and the reward function was designed only with the number of neighbors. In addition, we compared the proposed method with the Q-learning in both single-learner case and multi-learner case. The results of comparison showed that the proposed method has a more stable training performance. In a word, this study demonstrated that deep reinforcement learning is a potential powerful tool for analysis and modeling of collective behavior.

REFERENCES

[1] F. Ginelli, F. Peruani, M.-H. Pillot, H. Chaté, G. Theraulaz, and R. Bon, "Intermittent collective dynamics emerge from conflicting imperatives in sheep herds," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 41, pp. 12729–12734, Sep. 2015.

[2] J. G. Puckett, D. H. Kelley, and N. T. Ouellette, "Searching for effective forces in laboratory insect swarms," *Sci. Rep.*, vol. 4, no. 1, pp. 1–5, Apr. 2014.

[3] M. Ballerini, "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1232–1237, 2008.

[4] T. Vicsek and A. Zafeiris, "Collective motion," *Phys. Rep.*, vol. 517, nos. 3–4, pp. 40–71, Aug. 2012.

[5] P. Yang, M. Liu, X. Peng, and X. Lei, "Progress of theoretical modelling and empirical studies on collective motion," *Chin. Sci. Bull.*, vol. 59, no. 25, pp. 2464–2483, Sep. 2014.

[6] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," in *Proc. 14th Annual Conf. Comput. Graph. Interact. Techn.*, 1987, pp. 25–34.

[7] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, "Collective memory and spatial sorting in animal groups," *J. Theor. Biol.*, vol. 218, no. 1, pp. 1–11, Sep. 2002.

[8] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, "Novel type of phase transition in a system of self-driven particles," *Phys. Rev. Lett.*, vol. 75, no. 6, p. 1226, 1995.

[9] M. Nagy, I. Daruka, and T. Vicsek, "New aspects of the continuous phase transition in the scalar noise model (SNM) of collective motion," *Phys. A, Stat. Mech. Appl.*, vol. 373, pp. 445–454, Jan. 2007.

[10] F. Ginelli and H. Chaté, "Relevance of metric-free interactions in flocking phenomena," *Phys. Rev. Lett.*, vol. 105, no. 16, Oct. 2010, Art. no. 168103.

[11] M. Aldana, H. Larralde, and B. Vázquez, "On the emergence of collective order in swarming systems: A recent debate," *Int. J. Modern Phys. B*, vol. 23, no. 18, pp. 3661–3685, Jul. 2009.

[12] M. Aldana, V. Dossetti, C. Huepe, V. M. Kenkre, and H. Larralde, "Phase transitions in systems of self-propelled agents and related network models," *Phys. Rev. Lett.*, vol. 98, no. 9, Mar. 2007, Art. no. 095702.

[13] H. Chaté, F. Ginelli, and R. Montagne, "Simple model for active nematics: Quasi-long-range order and giant fluctuations," *Phys. Rev. Lett.*, vol. 96, no. 18, May 2006, Art. no. 180602.

[14] G. Grégoire, H. Chaté, and Y. Tu, "Moving and staying together without a leader," *Phys. D, Nonlinear Phenomena*, vol. 181, nos. 3–4, pp. 157–170, Jul. 2003.

[15] J. A. Pimentel, M. Aldana, C. Huepe, and H. Larralde, "Intrinsic and extrinsic noise effects on phase transitions of network models with applications to swarming systems," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 77, no. 6, Jun. 2008, Art. no. 061138.

[16] H. Duan and P. Li, "A flocking model base on selective attention mechanism," *SCIENTIA SINICA Technolog.*, vol. 49, no. 9, pp. 1040–1050, Sep. 2019.

[17] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, "Inferring individual rules from collective behavior," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 28, pp. 12576–12580, Jun. 2010.

[18] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and R. Tavarone, "From empirical data to inter-individual interactions: Unveiling the rules of collective animal behavior," *Math. Models Methods Appl. Sci.*, vol. 20, no. 1, pp. 1491–1510, Sep. 2010.

[19] J. E. Herbert-Read, A. Perna, R. P. Mann, T. M. Schaerf, D. J. T. Sumpter, and A. J. W. Ward, "Inferring the rules of interaction of shoaling fish," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 46, pp. 18726–18731, Nov. 2011.

[20] Y. Katz, K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin, "Inferring the structure and dynamics of interactions in schooling fish," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 46, pp. 18720–18725, Nov. 2011.

[21] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, "Statistical mechanics for natural flocks of birds," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 13, pp. 4786–4791, Mar. 2012.

[22] S. P. R. Banoth, P. K. Donta, and T. Amgoth, "Dynamic mobile charger scheduling with partial charging strategy for WSNs using deep-Q-networks," *Neural Comput. Appl.*, vol. 33, no. 22, pp. 15267–15279, 2021.

[23] D. Silver, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[24] O. Vinyals, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[25] J. Lee and M. Mitici, "Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics," *Rel. Eng. Syst. Saf.*, vol. 230, Feb. 2023, Art. no. 108908.

[26] A. Namdari, M. A. Samani, and T. S. Durrani, "Lithium-ion battery prognostics through reinforcement learning based on entropy measures," *Algorithms*, vol. 15, no. 11, p. 393, Oct. 2022.

[27] G. Basile, D. G. Lui, A. Petrillo, and S. Santini, "Deep deterministic policy gradient-based virtual coupling control for high-speed train convoys," in *Proc. IEEE Int. Conf. Netw., Sens. Control (ICNSC)*, Dec. 2022, pp. 1–6.

[28] X. Wang, J. Cheng, and L. Wang, "A reinforcement learning-based predator-prey model," *Ecolog. Complex.*, vol. 42, Mar. 2020, Art. no. 100815.

[29] P. Sunehag, G. Lever, S. Liu, J. Merel, N. Heess, J. Z. Leibo, E. Hughes, T. Eccles, and T. Graepel, "Reinforcement learning agents acquire flocking and symbiotic behaviour in simulated ecosystems," in *Proc. Conf. Artif. Life*, 2019, pp. 103–110.

[30] A. López-Incera, K. Ried, T. Müller, and H. J. Briegel, "Development of swarm behavior in artificial learning agents that adapt to different foraging environments," *PLoS ONE*, vol. 15, no. 12, Dec. 2020, Art. no. e0243628.

[31] K. Ried, T. Müller, and H. J. Briegel, "Modelling collective motion based on the principle of agency: General framework and the case of marching locusts," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0212044.

[32] K. Morihiro, H. Nishimura, and T. Isokawa, "Learning grouping and anti-predator behaviors for multi-agent systems," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.* Berlin, Germany: Springer, 2008, pp. 426–433.

[33] K. Shimada and P. Bentley, "Learning how to flock: Deriving individual behaviour from collective behaviour with multi-agent reinforcement learning and natural evolution strategies," in *Proc. Genetic Evol. Comput. Conf. Companion*, Jul. 2018, pp. 169–170.

[34] T. Costa, A. Laan, F. J. H. Heras, and G. G. de Polavieja, "Automated discovery of local rules for desired collective-level behavior through reinforcement learning," *Frontiers Phys.*, vol. 8, pp. 1–13, Jun. 2020.

[35] C. Hahn, T. Phan, T. Gabor, L. Belzner, and C. Linnhoff-Popien, "Emergent escape-based flocking behavior using multi-agent reinforcement learning," 2019, *arXiv:1905.04077*.
[36] C. Hahn, F. Ritz, P. Wikidal, T. Phan, T. Gabor, and C. Linnhoff-Popien, "Foraging swarms using multi-agent reinforcement learning," in *Proc. Conf. Artif. Life*, 2020, pp. 333–340.
[37] M. Durve, F. Peruani, and A. Celani, "Learning to flock through reinforcement," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 102, no. 1, pp. 1–8, Jul. 2020.

PENGYU CHEN received the B.S. degree in computer science and technology from Dalian Ocean University, Dalian, China, where he is currently pursuing the M.S. degree in computer science and technology.

His current research interest includes artificial intelligence, such as reinforcement learning and collective behavior modeling.

FANG WANG received the B.S. degree in computer science and technology from Northeast Normal University and the Ph.D. degree in computer science and technology from Jilin University.

She is currently an Assistant Professor with the School of Information Science and Engineering, Dalian Ocean University. Her main research interests include knowledge representation and automated reasoning, machine learning, and path planning.

SHUO LIU received the B.S. degree in information management and information systems from Shandong Jiaotong University, Jinan, China. He is currently pursuing the M.S. degree in electronic and information engineering with Dalian Ocean University, Dalian, China.

His current research interest includes artificial intelligence, such as reinforcement learning and multi-target tracking.

YIFAN YU received the B.S. degree in automation from Dalian Ocean University, Dalian, China, where he is currently pursuing the M.S. degree in computer science and technology.

His current research interest includes artificial intelligence, such as multi-agent reinforcement learning and formation control.

SHENGZHI YUE received the B.S. degree in automation from Dalian Ocean University, Dalian, China, where he is currently pursuing the M.S. degree in computer science and technology.

His current research interest includes artificial intelligence, such as reinforcement learning and multi-target tracking.

YANAN SONG received the B.S. degree in software engineering from the City Institute, Dalian University of Technology, Dalian, China, where she is currently pursuing the M.S. degree in computer science and technology.

Her current research interest includes artificial intelligence, such as reinforcement learning.

YUANSHAN LIN received the M.S. and Ph.D. degrees from the School of Computer Science, Dalian University of Technology (DLUT), Dalian, China, in 2008, and 2013, respectively.

He is currently an Associate Professor with the School of Information Science and Engineering, Dalian Ocean University (DLOU). His current research interest includes robotics and control, in particular, robot learning, motion planning, autonomous navigation, and machine learning.