2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

Multi-UAV Cooperative Short-Range Combat via Attention-Based Reinforcement Learning using Individual Reward Shaping*

Tianle Zhang[1,2], Tenghai Qiu[1,2](✉), Zhen Liu[1,2], Zhiqiang Pu[1,2], Jianqiang Yi[1,2], Jinying Zhu[2], Ruiguang Hu[3]

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9982096

Abstract— **In this paper, we propose a novel distributed method based on attention-based deep reinforcement learning using individual reward shaping, for multiple unmanned aerial vehicles (UAVs) cooperative short-range combat mission. Specifically, a two-level attention distributed policy, composed of observation-level and communication-level attention networks, is designed to enable each UAV to selectively focus on important environmental features and messages, for enhancing the effectiveness of the cooperative policy. Moreover, due to the high complexity and stochasticity of the UAV combat mission, the learning of UAVs is tricky and low efficient. To embed knowledge to accelerate the policy learning, a potential-based individual reward function is constructed by implicitly translating the individual reward into the specific form of dynamic action potentials. In addition, an actor-critic training algorithm based on the centralized training and decentralized execution framework is adopted to train the policy network of UAV maneuver decision. We build a three-dimensional UAV simulation and training platform based on Unity for multi-UAV short-range combat missions. Simulation results demonstrate the effectiveness of the proposed method and the superiority of the attention policy and individual reward shaping.**

I. INTRODUCTION

With the characteristics of low cost, strong mobility, high concealment and no pilot control, unmanned aerial vehicles (UAVs) are more and more widely used to replace manned aircraft to perform military missions such as detection, monitoring, and air combat [1], [2]. In recent years, UAV air combats have been explored because of the high complexity and uncertainty [3]. Moreover, multi-UAV cooperative air combats have become a research hotspot due to the limitations of single UAV's battle capabilities [4]. In particular, each UAV need to cooperate with the allies to automatically make maneuver decisions according to the situation faced, for realizing multi-UAV autonomous cooperative combat. Because of the highly dynamic and uncertain maneuvers of enemies and the complexity of UAV's maneuver model, the multi-UAV cooperative combats remains a great challenge. Furthermore, multi-UAV cooperative short-range combat (MUSC) is one of the most challenging application direction, because the two sides in the short-range combat perform the most violent maneuvers, making the situation change very rapidly. In addition, due to the lower cost and stronger maneuver compared with fixed wing UAV, a large number of quadrotor UAV combats will be a potential possibility of short-range air combats in the future [5], such as urban warfare [6].

The existing methods of autonomous maneuver decisions for UAV air combat can be divided into two categories: rule-based methods and learning-based methods. The former mainly makes decisions according to the given maneuver rules in air combat, including game theory algorithm [7], inference method [8], expert system method [9], etc. However, many of these rule-based methods require prior models and have poor real-time performance, which are difficult to adapt to the complex and highly dynamic air combat scenarios requiring autonomous and intelligent decision-making.

Due to the limitation of the rule-based methods, the learning-based methods show great potential by introducing deep reinforcement learning (DRL) [10] to generate autonomous maneuver policies for UAV air combat. Yang et al. [11] propose an autonomous maneuver decision model based on deep Q network for the UAV short-range air combat. Wang et al. [12] propose an autonomous maneuver strategy of UAV swarms in beyond visual range air combat based on DRL. However, they only realize one-to-one or multi-to-one UAV combats instead of multi-to-multi UAV cooperative air combat. Fortunately, Zhang et al. [13] build a multi-UAV cooperative air combat maneuver decision model based on DRL, and use bidirectional recurrent neural networks to achieve communication among UAVs. But, they make each UAV equally treat the observations and communication messages from neighbors, which ignores that the importance or influence of different neighbors to the UAV is different. For instance, each UAV should pay attention to information from the neighbors that have the greatest impact on it. To address this issue, one efficient way is to utilize attention mechanisms. Through attention mechanisms, each UAV can focus on the features of important neighbors using the assigned weights.

Besides, due to the high complexity and stochasticity of the Multi-UAV air combat in three-dimensional space, the policy learning of UAVs based on DRL is often tricky

*This work was supported by the National Key Research and Development Program of China (No. 2018AAA0102405 and No. 2018AAA0102402), the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA27030204), Fundamental Research Program (No.XXXX2019XXXX029).

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

[2]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

[3]National Key Laboratory of Science and Technology on Aerospace Intelligent Control, Beijing Aerospace Automatic Control Institute, China (e-mail: tianle-zhang@outlook.com, tenghai.qiu@ia.ac.cn, liuzhen@ia.ac.cn, zhiqiang.pu@ia.ac.cn, jianqiang.yi@ia.ac.cn, jinying.zhu@ia.ac.cn, rghu258@163.com).

and low efficient, especially in large-scale UAV combat scenarios. To deal with this problem, one natural idea is to embed available behavior knowledge about air combats to speed up the policy learning of UAVs [14]. However, the direct embedding of knowledge will change the original goal of team optimization, which lead to the generation of suboptimal policies. In particular, the direct embedding of individual behavior knowledge will cause individual UAV selfish behaviors, which will destroy the cooperation of the team.

Motivated by the aforementioned discussions, we propose a new distributed method based on attention-based DRL using individual reward shaping, named MUSC-ADRL-IRS, to generate autonomous maneuver policies for the MUSC mission. In particular, a two-level attention distributed policy, composed of observation-level and communication-level attention networks, is designed to enable each UAV to selectively focus on important environmental features and messages. Moreover, to embed knowledge for accelerating the policy learning under the guarantee of optimization-objective invariance, a potential-based individual reward function is built by implicitly translating the individual reward into the specific form of dynamic action potentials. Besides, an actor-critic training algorithm based on the centralized training and decentralized execution (CTDE) framework is used to train the policy of UAV maneuver decision. In this paper, the main contributions are listed as follows:

- Differing from simply integrating multiple rule-based algorithms, a new distributed method based on attention-based DRL using individual reward shaping is proposed to produce autonomous maneuver policies for MUSC.
- A observation-level and communication-level attention policy is designed to make each UAV selectively focus on important features and messages from its neighbors, instead of treating these information equally.
- Differing from embedding directly individual knowledge to speed up the policy training, a potential-based individual reward function is constructed with the reward shaping method.
- A three-dimensional UAV simulation and training platform based on Unity for MUSC is built, and simulation results demonstrate the effectiveness and superiority of the proposed method.

II. PROBLEM FORMULATION

As shown in Fig. 1, a MUSC mission, where n blue ally UAVs need to battle m red enemy UAVs in a short-range three-dimensional space and win (i.e., all the red enemy UAVs are destroyed), is investigated in this paper. For simplicity, the kinematic model [11] of each UAV i ($i \in 1, ..., n$) is based on quadrotors [15], which is defined in the ground coordinate system by,

$$\begin{cases} \dot{x}_i = u_i cos\phi_i sin\psi_i, \\ \dot{y}_i = u_i cos\phi_i cos\psi_i, \\ \dot{z}_i = u_i sin\phi_i, \end{cases} \tag{1}$$

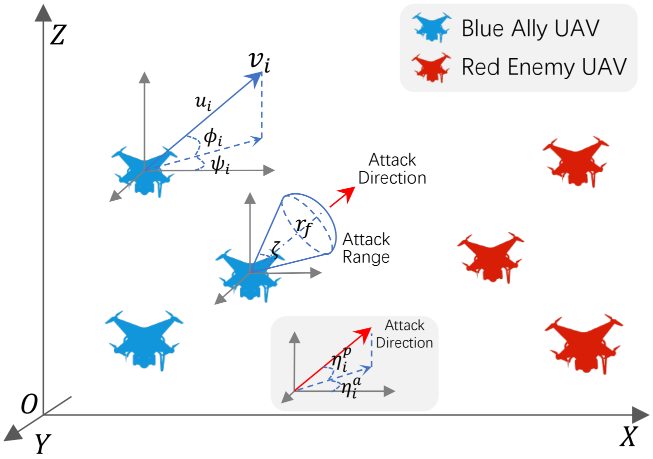


Fig. 1. Illustration of multi-UAV cooperative short-range combat mission.

where x_i, y_i, z_i represent the position coordinates of UAV i in the coordinate system, u_i represents the speed and $\dot{x}_i, \dot{y}_i, \dot{z}_i$ represent the values of the velocity v_i on the three coordinate axes, i.e. $v_i = [v_i^x, v_i^y, v_i^z] = [\dot{x}_i, \dot{y}_i, \dot{z}_i]$. The pitch angle ϕ_i represents the angle between the velocity vector and the horizontal plane $O-X-Y$. The heading angle ψ_i represents the angle between the projection of the velocity vector on the $O-X-Y$ plane and the OX axis. The position vector is recorded as $p_i = [x_i, y_i, z_i]$. Moreover, the attack range of UAV i is defined as a geometry, which is formed by a sector with angle ζ and radius r_f rotating 360 degrees about its central axis. The vertex position of the geometry is the position of the UAV. Meanwhile, the attack direction can be moved through two angle control variables, attack pitch angle η_i^p and attack azimuth angle η_i^a. Besides, to simulate real combat situations, each UAV i has a health point b_i, which represents the health of the UAV and indicates how well the UAV is and how much damage the UAV can take before crashing. If the health point of UAV i becomes 0 or less, the UAV will crash. Enemies in the attack range of UAV i will lose their health points continuously and vice versa. According to the above description, $[u_i, \phi_i, \psi_i]$, $[\eta_i^p, \eta_i^a]$ are sets of motion and attack control variables, respectively. Thus, the action of UAV i can be denoted as $a_i = [u_i, \phi_i, \psi_i, \eta_i^p, \eta_i^a]$, which can control the maneuvering of the UAV.

In the MUSC mission, each UAV can only observe the positions, velocities, health points of enemy UAVs and other ally UAVs within its visual area with radius D^O, and communicate with ally UAVs within its communication range with radius D^C. In this distribution environment, the ally UAVs need to cooperate with each other to annihilate the enemy UAVs through controlling their respective decision variables, $u_i, \phi_i, \psi_i, \eta_i^p, \eta_i^a$.

This MUSC task can be formulated as a partially observable Markov decision process [16] in a reinforcement learning framework. At each timestep, UAV i can obtain a partial observation set $o_i = \{c_k | k \in \mathcal{N}_i^O\}$, where $c_k = [p_k, v_k, b_k]$, N_i is some neighborhood (including ally and enemy UAVs) within the visual area of UAV i. Then, the UAV based on its own observation interacts with neighbor ally UAVs and selects an appropriate action a_i for defeating the enemies cooperatively. Subsequently, each UAV can obtain a team reward R from the environment after all ally UAVs take the action. The team rewards reflect whether the joint actions taken are conducive to the victory of allies. This paper aims to design an optimal policy $\pi_i : o_i \to a_i$ for UAV i that maximizes its expected accumulated discounted return $\mathbb{E}[\sum_{t=0}^{T} \gamma^t R(t)]$, where $\gamma \in [0, 1]$ is a discount factor and T is the time horizon. The optimal policy enables UAV i to complete the MUSC task and win efficiently.

III. METHOD

A. Overall Structure

The overall structure of the proposed method mainly consists of three components as shown in Fig. 2: 1) a two-level attention distributed policy network structure, composed of observation-level and communication-level attention networks; 2) a defined potential-based individual reward function, which embeds individual behavior knowledges to speed up the learning of the policy under keeping the optimization objective unchanged; 3) an actor-critic training algorithm, which trains the policy network for completing the MUSC mission.

B. Two-Level Attention Distributed Policy Network Structure

The distributed policy network structure is a actor-critic structure. The actor parameterized by θ, $\pi_i^\theta : o_i \times m_i \mapsto a_i$, consists of observation-level and communication-level attention, gate recurrent unit (GRU) and policy head networks. It takes partial observation o_i and received communication message m_i of UAV i as input and outputs action values for making decisions. The critic parameterized ϕ, $v_i^\phi : s_i \mapsto \mathbb{R}$, composed of a value head network with FC layers and GRU, takes the UAV-specific global state $s_i = c_i \cup \{c_k - c_i | k \in 1, ..., n, 1, ..., m \textit{ and } k \neq i\}$ of UAV i as inputs and outputs a scalar value for the actor training. Especially, for the actor, in the MUSC miscellaneous information environment, the observation-level and communication-level attention networks are designed to make each UAV selectively focus on important environmental features and teammate messages, for improving the effectiveness of the cooperative policy.

Observation-Level Attention Network: In the complex MUSC mission, each UAV can obtain miscellaneous partial observation information. However, not all information needs to be valued by the UAV. UAVs need to pay attention to the information that can promote the completion of the mission. Therefore, an observation-level attention network based on Transformer [17] is designed to extract important features in the partial observation of each UAV.

Firstly, each neighbor state in the observation o_i of UAV i is encoded as an encoding embedding, $\hat{c}_k = \boldsymbol{W}_N c_k$, $k \in \mathcal{N}_i^O$ and $k \neq i$; $\hat{c}_i = \boldsymbol{W}_S c_i$, where $\boldsymbol{W}_N, \boldsymbol{W}_S$ are learnable parameter matrices. Then, UAV $k \in \mathcal{N}_i^O$ computes a key $E_k^O = \boldsymbol{W}_E^O \hat{c}_k$, query $Q_k^O = \boldsymbol{W}_Q^O \hat{c}_k$ and $V_k^O = \boldsymbol{W}_V^O \hat{c}_k$ vectors where $\boldsymbol{W}_E^O, \boldsymbol{W}_Q^O, \boldsymbol{W}_V^O$ are other learnable parameter matrices. Next, after receiving query-value pair (Q_i^O, E_k^O), UAV i assigns weights to observed neighbors,

$$\alpha_{ik} = softmax(\frac{(Q_i^O)^T E_k^O}{d_E}), \qquad (2)$$

where d_E is the dimensionality of the key vector. Then, according to the assigned weights, the UAV aggregates the states of the observed neighbors and computes an aggregated embedding, $h_i = \boldsymbol{W}_{out}^O \sum_{k \in \mathcal{N}_i^O} \alpha_{ik} V_k^O$, where $\boldsymbol{W}_{out}^O$ is another learnable parameter. Finally, the UAV updates the aggregated embedding h_i by doing a non-linear transformation of h_i^k concatenated with $\hat{c}_i$ by using a one fully-connected (FC) layer network. Herein, the aggregated embedding h_i implicitly encodes the important features from the observed neighbors by selective aggregation with the assigned weights.

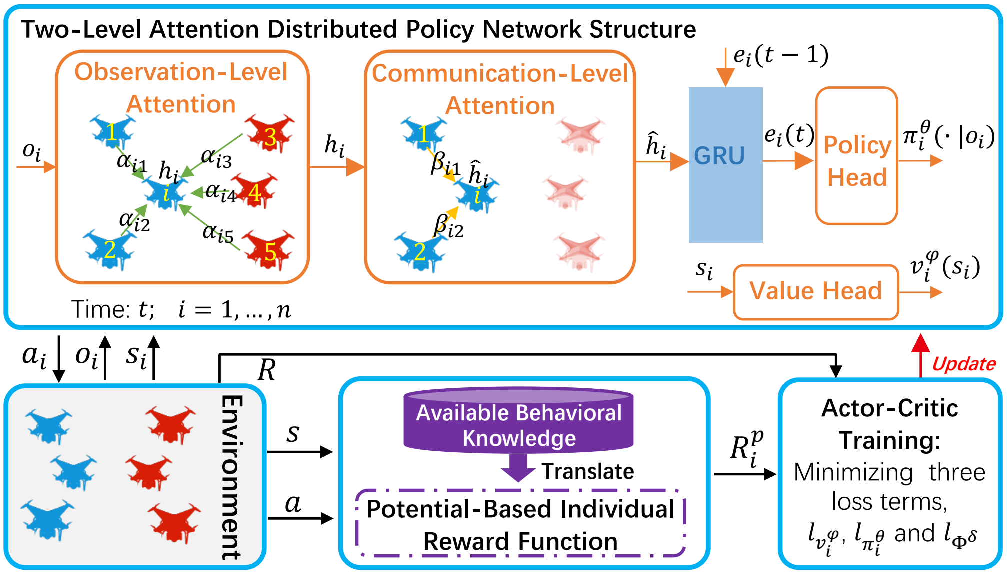
Communication-Level Attention Network: Communication is an common and important method to promote the cooperation among UAVs through message transmission. However, useless messages will interfere with the decision-making of UAVs. UAVs need to focus on useful messages that can facilitate the completion of the MUSC mission, and ignore useless messages from neighbor allies. Hence, a communication-level attention network based on Transformer is designed to extract useful messages from neighbor allies for efficient cooperation.

We firstly define a communication topology graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, where each node denotes an ally UAV, and there exists an edge between two nodes if the nodes are in their respective communication range. The aggregated embedding h_i is used as the interactive message to be transmitted in this graph. Hence, the received message set m_i of UAV i is defined by $m_i = \{h_l | l \in \mathcal{N}_i^C\}$, where N_i^C is some neighborhood within the communication range of UAV i. Similarly, each UAV $j \in \mathcal{V}$ calculate a key $E_j^C = \boldsymbol{W}_E^C h_j$, query $Q_j^C = \boldsymbol{W}_Q^C h_j$ and $V_j^C = \boldsymbol{W}_V^C h_j$ vectors where $\boldsymbol{W}_E^C, \boldsymbol{W}_Q^C, \boldsymbol{W}_V^C$ are learnable parameter matrices. Then, UAV i assigns weights to each of the incoming messages after receiving query-value pair (Q_i^C, E_k^C),

$$\beta_{ij} = softmax(\frac{(Q_i^C)^T E_j^C}{d_E}). \qquad (3)$$

It then aggregates all the messages by calculating a weighted sum of the values of its neighbors and follows a linear transformation, yielding a interaction embedding, $\hat{h}_i = \boldsymbol{W}_{out}^C \sum_{j \in \mathcal{N}_i^C} \beta_{ij} V_j^C$, where $\boldsymbol{W}_{out}^C$ is another learnable parameter. The interaction embedding implicitly encodes the useful messages from the communication neighbors, and is a state representation of the surrounding environment of UAV i at current time.

After observation-level and communication-level attention network, GRU is used to fuse the environmental embedding $e_i(t-1)$ at last time $t-1$ and the interaction embedding $\hat{h}_i$ at current time t, yielding the environmental embedding $e_i(t)$ at current time t. Then, the environmental embedding $e_i(t)$ is fed into a policy head network with two FC layers, which outputs action values of UAV i.



Two-Level Attention Distributed Policy Network Structure
Observation-Level Attention
o_i
α_{i1}
α_{i2}
α_{i3}
α_{i4}
α_{i5}
h_i
1
2
3
4
5
i
h_i
Communication-Level Attention
β_{i1}
β_{i2}
$\hat{h}_i$
$\hat{h}_i$
$e_i(t-1)$
GRU
$e_i(t)$
Policy Head
$\pi_i^\theta(\cdot|o_i)$
s_i
Value Head
$v_i^\varphi(s_i)$
Time: t; $i = 1, \dots, n$
a_i
o_i
s_i
R
Update
Environment
s
a
Available Behavioral Knowledge
Translate
Potential-Based Individual Reward Function
R_i^p
Actor-Critic Training:
Minimizing three loss terms,
$l_{v_i^\varphi}$, $l_{\pi_i^\theta}$ and l_{Φ^δ}

Fig. 2. Overall structure of MUSC-ADRL-IRS.

C. Potential-Based Individual Reward Function

In this paper, the optimization objective is to maximize the expectation of discount cumulative team return $\mathbb{E}[\sum_{t=0}^{T}\gamma^t R(t)]$ for each ally UAV. In the MUSC mission, the team reward R is firstly defined by, $R = \lambda_1(\sum_{j=1}^{m}(b_j(t) - b_j(t+1)))/\sum_{j=1}^{m} b_j^{max}$, which denotes the normalization of the sum of the enemies' health point loss at time t, where b_j^{max} denotes the maximum health point of the enemy UAV j and λ_1 is a hyper parameter. Meanwhile, we define the individual available behavior knowledge denoted as $R_i^{bk} = \lambda_2(\sum_{k\in\mathcal{N}_i^E}(b_k(t) - b_k(t+1)))/\sum_{j=1}^{m} b_j^{max}$, where $\mathcal{N}_i^E(t)$ denotes a set of enemies attacked by UAV i at time t and λ_2 is another hyper parameter.

If we directly embed the individual behavior knowledge for the learning of each UAV, it will lead to the selfishness of individual behavior and the change of the original optimization objective of the team. Therefore, to speed up the learning of the policy network, we indirectly embed the individual behavior knowledge by translating it into a potential-based individual reward function under the guarantee of optimization-objective invariance [18].

Firstly, we define a potential function $\Phi^\delta : s_i \times a_i \mapsto \mathbb{R}$ parameterized by δ using FC layer networks. The potential function takes the UAV-specific global state s_i and the action a_i of UAV i as input and outputs a state-action value about the individual behavior knowledge, which can be regarded as an auxiliary action-state value function for R_i^{bk}. Moreover, the target value of the potential value is defined as [19]

$$\hat{\Phi}^\delta(s_i^t, a_i^t) = -R_i^{bk}(t) + \gamma\Phi^\delta(s_i^{t+1}, a_i^{t+1}). \tag{4}$$

The potential function is updated by minimizing the loss $l_{\Phi^\delta} = \mathbb{E}_{(s_i,a_i),t}[(\hat{\Phi}^\delta(s_i(t), a_i(t)) - \Phi^\delta(s_i(t), a_i(t)))^2]$. Therefore, according to the potential function embedding individual behavior knowledge, a potential-based individual reward function R_i^p is defined by

$$R_i^p(t) = \gamma\Phi^\delta(s_i^{t+1}, a_i^{t+1}) - \Phi^\delta(s_i^t, a_i^t). \tag{5}$$

The received reward of UAV i can be denoted as $R_i(t) = R_i^p(t) + R(t)$ for the training of the policy, which is a sum of the team reward and the potential-based individual reward.

Individual Reward Shaping in Expectation: In order to investigate the relationship between the individual behavior knowledge R_i^{bk} and the potential-based individual reward R_i^p, the expectation (w.r.t. the transition matrix and the action a_i' at next time) of R_i^p is calculated by,

$$\begin{aligned}\mathbb{E}_{(s_i',a_i')}[R_i^p] &= \mathbb{E}_{(s_i',a_i')}[\gamma\Phi^\delta(s_i', a_i') - \Phi^\delta(s_i, a_i)] \\ &= \mathbb{E}_{(s_i',a_i')}[\gamma\Phi^\delta(s_i', a_i') + R_i^{bk} - \gamma\mathbb{E}_{(s_i',a_i')}[\Phi^\delta(s_i', a_i')]] \\ &= R_i^{bk}.\end{aligned} \tag{6}$$

Hence, R_i^p in expectation and R_i^{bk} are equivalent.

Optimization Objective Introducing Individual Reward Shaping: The optimization objective introducing individual reward shaping is to maximize the expected accumulated discounted return about the received reward R_i,

$$\begin{aligned}&\mathbb{E}[\sum_{t=0}^{T}\gamma^t(R(t) + R_i^p(t))] \\ &= \mathbb{E}[\sum_{t=0}^{T}\gamma^t(R(t) + \gamma\Phi^\delta(s_i^{t+1}, a_i^{t+1}) - \Phi^\delta(s_i^t, a_i^t))] \\ &= \mathbb{E}[\sum_{t=0}^{T}\gamma^t R(t)] + \mathbb{E}[\sum_{t=1}^{T}\gamma^t\Phi^\delta(s_i^t, a_i^t)] - \mathbb{E}[\sum_{t=0}^{T}\gamma^t\Phi^\delta(s_i^t, a_i^t)] \\ &= \mathbb{E}[\sum_{t=0}^{T}\gamma^t R(t)] - \Phi^\delta(s_i^0, a_i^0).\end{aligned} \tag{7}$$

Thus, when the potential function is initialized to 0, the optimization objective introducing individual reward shaping remains unchanged.

D. Actor-Critic Training Algorithm

In this paper, we utilize the CTDE framework [20] to learn a centralized critic to update the distributed policy network of UAV i during training. Moreover, the proximal policy optimization (PPO) [21] algorithm based on actor-critic style is used to update the parameter of the policy network composed of the actor and critic by minimizing two loss terms,

$$\begin{aligned} l_{v_i^{\phi}} &= \mathbb{E}[(y_i - v_i^{\phi}(s_i)^2], \\ l_{\pi_i^{\theta}} &= \mathbb{E}[min(\frac{\pi_i^{\theta}(\cdot|o_i)}{\pi_i^{\theta_{old}}(\cdot|o_i)}A_i(o,a), \\ & clip(\frac{\pi_i^{\theta}(\cdot|o_i)}{\pi_i^{\theta_{old}}(\cdot|o_i)}, 1-\epsilon, 1+\epsilon)A_i(o,a)], \end{aligned} \tag{8}$$

where $\pi_i^{\theta_{old}}(\cdot|o_i)$ is the actor before the update or the sampling actor, $y_i = R_i + \gamma v_i^{\phi}(s_i)$ is the temporal-difference (TD) target, $\epsilon = 0.2$ is a hyper parameter and $A_i(o,a)$ is an advantage function, which is estimated through the generalized advantage estimator (GAE) method [22]. Besides, the potential function network Φ_δ is trained by minimizing the loss l_{Φ^δ} simultaneously.

IV. SIMULATIONS

A. Simulation Settings

We build a three-dimensional UAV simulation and training platform based on Unity shown in Fig. 3. This platform simulates real UAV flight environment conditions based Unity, including terrain condition, wind interference, and UAV control model. Moreover, the training part in the platform adopts the Pytorch structure in Python for building and training the policy networks of UAVs. In this platform, we design a multi-UAV cooperative combat mission environment. Specifically, all blue and red ally UAVs are initially randomly positioned on $\{600 \leq x \leq 650, 300 \leq y \leq 700, 500 \leq z \leq 550\}$ and $\{350 \leq x \leq 400, 300 \leq y \leq 700, 500 \leq z \leq 550\}$ space areas, respectively. This can make the red and blue sides on one side and form a combat situation. The blue ally UAVs are controlled by the proposed method, while the red enemy UAVs are controlled through the traditional knowledge method composed of clustering [23], target allocation [24] and optimal reciprocal collision avoidance (ORCA) [25] algorithms. The maximum speed u_i^{max} of each UAV is set to 200 $unit/s$. The mission ends when one of the red and blue sides is destroyed or the running time exceeds a fixed period $T_{max} = 150$ timesteps. The time interval of each step is 0.1.

The proposed method MUSC-ADRL-IRS and the following baseline methods are implemented for performance evaluation,

- MUSC-NODRL: This is a centralized traditional knowledge method, which is not DRL-based method. The method integrates the clustering, target allocation and ORCA algorithm for making decisions [23], [24], [25], which is the same as the strategy of the enemies;
- MUSC-MAPPO: The method adopts the PPO algorithm learn a centralized critic to update the decentralized policy network without attention mechanism [26];
- MUSC-ADRL: The method learns an observation-level and communication-level attention policy network using the CTDE framework, without individual reward shaping.

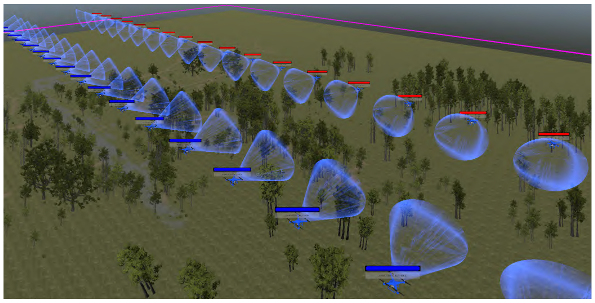


Fig. 3. Multi-UAV cooperative combat platform based on Unity.

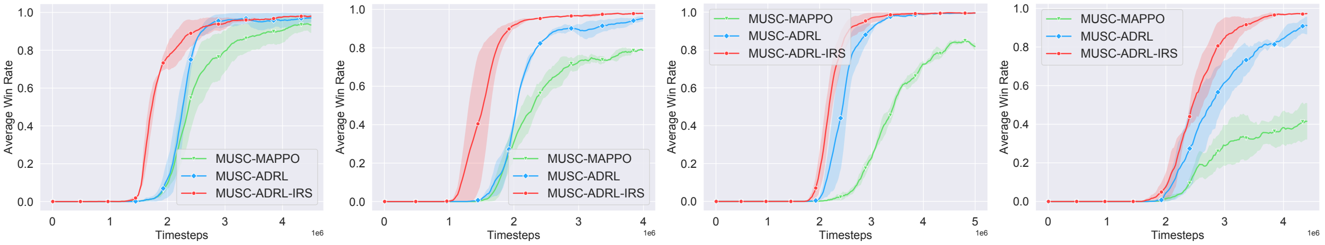
In addition, three metrics are set to evaluate the performance of different methods. The performance metrics of each method are obtained by testing 500 episodes with 5 different random seeds, 1) Win Rate (WR%): Percentage of the blue ally UAVs won in all test episodes (the condition for the blue side to win is that all the enemies are destroyed, and at least one blue UAV is alive); 2) Mean Episode Reward (MER): Mean of episode cumulative team rewards in all test episodes; 3) Mean Episode Length (MEL): Mean of length of winning episode the blue UAVs in all test episodes.

B. Implementation Specifications

In the observation-level and communication-level attention, the learnable parameters are set with $\boldsymbol{W}_N, \boldsymbol{W}_S \in \mathcal{R}^{7X64}$, $\boldsymbol{W}_E^O, \boldsymbol{W}_Q^O, \boldsymbol{W}_V^O, \boldsymbol{W}_{out}^O \in \mathcal{R}^{64X64}$ and $\boldsymbol{W}_E^C, \boldsymbol{W}_Q^C, \boldsymbol{W}_V^C, \boldsymbol{W}_{out}^C \in \mathcal{R}^{64X64}$. The hidden layer size of GRU is 64. The value head network outputs a scalar value, while the policy head network outputs move-level and attack-level action values.

Moreover, the action space of each UAV is divided into move-level and attack-level action spaces. The move-level actions are discretized into 121 move actions: stop, and 3 speeds exponentially spaced between $(0, u_i^{max}]$ and 8 heading angles evenly spaced between $[0, 2\pi)$ and 5 pitch angles evenly spaced between $[-\pi/2, \pi/2]$. The attach-level actions are discretized into 40 attack actions: 8 attack azimuth angles evenly spaced between $[0, 2\pi)$ and 5 attack pitch angles evenly spaced between $[-\pi/2, \pi/2]$.

Beside, the visual radius D^O and communication radius D^C of each UAV both are set as 90. The maximum health point b_i^{max} of each UAV is set to 100. Enemies within the allies' attack range lose 50 health points every time step, as do the allies within enemies' attack range. Meanwhile, the angle $\zeta = 1.0$ and radius $r_f = 30$ of the attach range are set. $\lambda_1 = 20$ and $\lambda_2 = 10$.



(a) 10 ally UAVs vs. 10 enemy UAVs(b) 10 ally UAVs vs. 15 enemy UAVs(c) 15 ally UAVs vs. 15 enemy UAVs(d) 15 ally UAVs vs. 20 enemy UAVs

Fig. 4. Training results of various methods in the MUSC mission: (a-d) are the training curves of average win rates vs. training steps.

TABLE I

TEST RESULTS OF OUR METHOD AND BASELINE METHODS IN THE MUSC MISSION

Methods	10 vs. 10	10 vs. 15	15 vs. 15	15 vs. 20
	WR(%) / MER / MEL	WR(%) / MER / MEL	WR(%) / MER / MEL	WR(%) / MER / MEL
MUSC-NODRL	$27.5_{\pm 1.50}$/$15.7_{\pm 0.03}$/$\mathbf{23.6}_{\pm 0.16}$	$0.0_{\pm 0.00}$/$9.8_{\pm 0.35}$/–	$27.0_{\pm 1.00}$/$15.9_{\pm 0.23}$/$26.1_{\pm 0.17}$	$0.5_{\pm 0.50}$/$11.3_{\pm 0.08}$/$\mathbf{29.0}_{\pm 0.00}$
MUSC-MAPPO	$98.5_{\pm 0.50}$/$19.8_{\pm 0.01}$/$49.7_{\pm 0.29}$	$81.0_{\pm 2.00}$/$18.3_{\pm 0.14}$/$88.8_{\pm 2.08}$	$87.0_{\pm 1.00}$/$18.9_{\pm 0.01}$/$57.4_{\pm 1.19}$	$47.5_{\pm 5.50}$/$15.7_{\pm 0.40}$/$80.5_{\pm 1.55}$
MUSC-ADRL	$99.5_{\pm 0.50}$/$19.9_{\pm 0.04}$/$30.8_{\pm 0.47}$	$99.5_{\pm 0.50}$/$19.9_{\pm 0.03}$/$38.2_{\pm 0.47}$	$99.0_{\pm 1.00}$/$19.8_{\pm 0.07}$/$29.1_{\pm 0.34}$	$90.0_{\pm 1.00}$/$19.4_{\pm 0.04}$/$53.2_{\pm 4.97}$
MECA-ADRL-IRS (ours)	$\mathbf{99.9}_{\pm 0.10}$/$\mathbf{19.9}_{\pm 0.10}$/$33.3_{\pm 0.38}$	$\mathbf{99.5}_{\pm 0.50}$/$\mathbf{19.9}_{\pm 0.03}$/$\mathbf{24.9}_{\pm 0.74}$	$\mathbf{99.0}_{\pm 0.00}$/$\mathbf{19.9}_{\pm 0.03}$/$\mathbf{22.6}_{\pm 0.09}$	$\mathbf{97.0}_{\pm 1.00}$/$\mathbf{19.8}_{\pm 0.04}$/$31.4_{\pm 0.42}$

TABLE II

GENERALIZATION RESULTS OF OUR METHOD AND BASELINE METHODS IN THE MUSC MISSION

Methods	5 vs. 8	10 vs. 20	15 vs. 25	20 vs. 30
	WR(%) / MER / MEL	WR(%) / MER / MEL	WR(%) / MER / MEL	WR(%) / MER / MEL
MUSC-NODRL	$0.0_{\pm 0.00}$/$8.1_{\pm 0.12}$/–	$0.0_{\pm 0.00}$/$6.9_{\pm 0.00}$/–	$0.0_{\pm 0.00}$/$8.9_{\pm 0.07}$/–	$0.0_{\pm 0.00}$/$9.3_{\pm 0.40}$/–
MUSC-MAPPO	–	–	–	–
MUSC-ADRL	$97.0_{\pm 1.00}$/$19.8_{\pm 0.06}$/$35.8_{\pm 0.17}$	$90.5_{\pm 0.50}$/$19.3_{\pm 0.02}$/$45.0_{\pm 0.53}$	$92.5_{\pm 0.50}$/$19.5_{\pm 0.07}$/$42.7_{\pm 0.03}$	$94.0_{\pm 1.00}$/$19.6_{\pm 0.02}$/$40.5_{\pm 0.51}$
MECA-ADRL-IRS (ours)	$\mathbf{98.0}_{\pm 1.00}$/$\mathbf{19.8}_{\pm 0.09}$/$\mathbf{22.2}_{\pm 0.48}$	$\mathbf{96.5}_{\pm 0.05}$/$\mathbf{19.7}_{\pm 0.08}$/$\mathbf{26.9}_{\pm 0.14}$	$\mathbf{98.0}_{\pm 1.00}$/$\mathbf{19.9}_{\pm 0.04}$/$\mathbf{23.9}_{\pm 0.58}$	$\mathbf{99.0}_{\pm 1.00}$/$\mathbf{19.9}_{\pm 0.04}$/$\mathbf{22.8}_{\pm 0.58}$

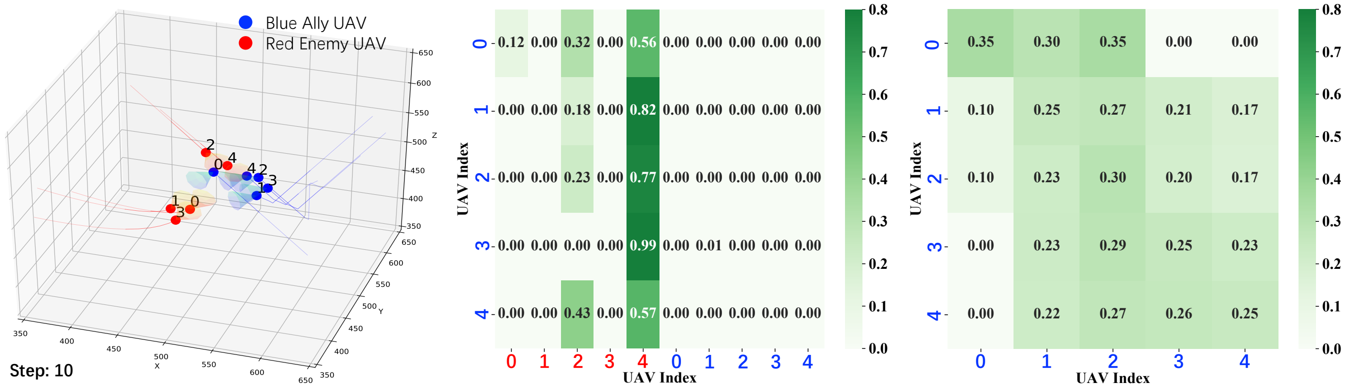
C. Simulation Results

1) Effectiveness: To fully evaluate the effectiveness of the proposed method, we conduct four combat scenarios: two symmetric combat scenarios (i.e., 10 ally UAVs vs. 10 enemy UAVs; 15 ally UAVs vs. 15 enemy UAVs), and two difficult asymmetric combat scenarios (i.e., 10 ally UAVs vs. 15 enemy UAVs; 15 ally UAVs vs. 20 enemy UAVs) where requires a high level of cooperation between teammates for the ally UAVs due to the disadvantage of quantity.

The training results are shown in Fig. 4. The results show that the proposed method has better performance than the baseline methods in terms of average win rate and convergence rate. Especially, due to the addition of the individual reward shaping, our method realizes efficient training and converges to a higher winning rate faster than the other methods. Moreover, compared with MUSC-MAPPO, MUSC-ADRL with observation-level and communication-level attention has better performance, especially in difficult asymmetric combat scenarios with higher requirements for cooperation. This also confirms that the attention mechanism is conducive to the cooperation among UAVs for battling the enemies together, because of its selective and focused characteristics.

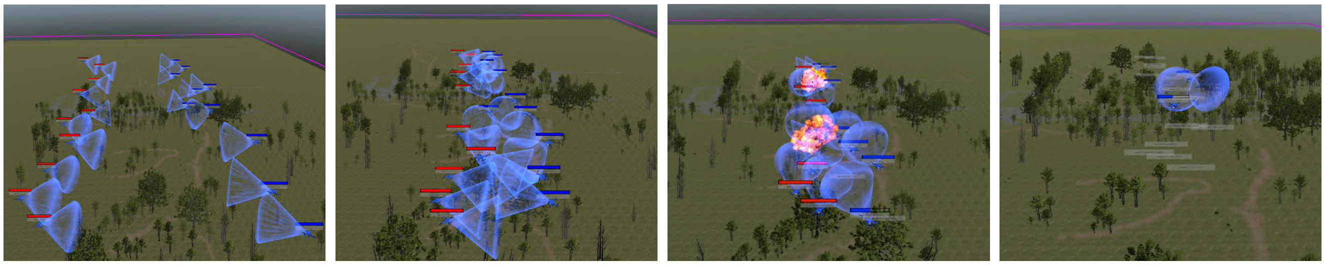
In addition, we test the policy learned by different methods. The test results are shown in Table I. Similarly, the proposed method MUSC-ADRL-IRS has better performance than the other methods in terms of WR, MER and MEL. On the contrary, MUSC-NODRL fails. Besides, MUSC-ADRL-IRS and MUSC-ADRL have little difference in the performance of the combat scenarios: 10*vs*.10, 10*vs*.15 and 15*vs*.15. This is because they can all converge to good results due to the breakability of the enemies' strategy. However, in the more difficult combat scenario, 15*vs*.20, MUSC-ADRL-IRS greatly outperforms MUSC-ADRL, which is because the individual reward shaping can can help and accelerate the learning of effective policies of UAVs in the 15*vs*.20 scenario. Moreover, MUSC-ADRL-IRS aims to add the individual reward shaping on MUSC-ADRL to improve training efficiency, which is verified in the training results. By contrary, the performance of MUSC-MAPPO without the attention and individual reward shaping is always not good. In general, the training and test results fully demonstrate the effectiveness and superiority of the proposed method.

2) Generalization: In order to verify the generalization of the proposed method, the learned policy is evaluated in new scenarios without any fine-tuning. The learned policy is obtained in the combat scenario where 10 ally UAVs battle 15 enemy UAVs, i.e., 10 vs. 15. Four new hard and asymmetric combat scenarios are selected, including 5 vs. 8; 10 vs. 20; 15 vs. 25; 20 vs. 30. The generalization results of our method and the baseline methods is shown in Table. II. As we expected, the proposed method outperforms the baseline methods in terms of win rate, episode reward and length, especially in more difficult scenarios. On the



(a) Keyframe of the 5vs.5 scenario: Step equals 10 (b) Observation-level attention weights, i.e., α_{ik} (c) Communication-level attention weights, i.e., β_{ij}

Fig. 5. Keyframe analysis with observation-level and communication-level attentions in the 5vs.5 MUSC scenario.



(a) Initial stage (b) Encounter stage (c) Fierce battle stage (d) End with blue wining

Fig. 6. Snapshots of multi-UAV cooperative short-range air combat in the 10vs.10 scenario.

contrary, MUSC-MAPPO can not adapt to the new scenarios different from the training and fails since its policy network input size is fixed, while the designed two-level attention policy can process the dynamic input size. Beside, although MUSC-ADRL-IRS and MUSC-ADRL have a small gap in the test results, MUSC-ADRL-IRS is highly superior to MUSC-ADRL in the generalization results. This is due to the individual reward shaping, which also shows that the embedding of individual behavior knowledge helps to produce more generalized policies.

3) Ablation Analysis: We further investigate and analyze the effectiveness of key components of the proposed method. MUSC-ADRL is an ablation version of our method, without the individual reward shaping. The main difference between MUSC-ADRL and MUSC-MAPPO is mainly on the policy network. According to the above simulation results shown in Fig. 4 and Table I II, MUSC-ADRL outperforms MUSC-MAPPO, which demonstrates the superiority of the designed attention policy network. Moreover, MUSC-ADRL-IRS has better performance than MUSC-ADRL, which verifies the effectiveness of the individual reward shaping for efficient training.

D. Qualitative Analysis

To analyze the effectiveness of the proposed two-level attention policy network, as shown in Fig. 5, we present keyframe analysis with observation-level and communication-level attentions in the 5vs.5 MUSC scenario. Fig. 5(a) shows a combat situation where 5 blue ally UAVs $(0, 1, 2, 3, 4)$ battle 5 red enemy UAVs $(0, 1, 2, 3, 4)$ at 10 timestep, and Fig. 5(b-c) present the observation-level and communication-level attention weights of the blue side in this combat situation. In the observation-level attention, the blue side mainly focuses on the two red enemy UAVs close to them, red UAVs 2 and 4. This is very helpful for the blue side to quickly destroy the enemy UAVs 2 and 4. In the communication-level attention, blue UAVs $1, 2, 3, 4$ use almost equal attentions to communicate with each other, while blue UAV 0 only communicate with blue UAVs 1 and 2 close to it due to the limitation of the communication range. The results of keyframe analysis show that the blue UAVs can learn reasonable and effective attention weights to focus on important observation information or communication messages through using the two-level attention policy, and demonstrate the effectiveness of the proposed policy.

Aside from the attention analysis above, we further observe the whole process of the MUSC mission. The four snapshots of the MUSC mission in the 10vs.10 scenario are shown in Fig. 6. The four snapshots show the four stages of the combat: initial, encounter, fierce battle and end stages, respectively. The results demonstrate that the blue UAVs can win quickly by using the policies obtained through the proposed method.

V. CONCLUSION

In this paper, MECA-ADRL-IRS is proposed to learn cooperative policies of UAVs for completing the MUSC mission through attention-based DRL using individual reward shaping. Specifically, an observation-level and communication-level attention distributed policy is designed to enable each UAV to selectively focus on important environmental features and messages. Moreover, to embed knowledge to accelerate the policy learning without changing the optimization objective, a potential-based individual reward function is constructed by implicitly translating the individual reward into the specific form of dynamic action potentials. Besides, an actor-critic training algorithm based on the CTDE framework is adopted to train the policy network of UAV maneuver decision. We build a three-dimensional UAV simulation and training platform based on Unity for realizing the MUSC mission. Simulation results confirm the effectiveness and generalization of the proposed method.

References

[1] L. Yue, Q. Xiaohui, L. Xiaodong, and X. Qunli, "Deep reinforcement learning and its application in autonomous fitting optimization for attack areas of ucavs," *Journal of Systems Engineering and Electronics*, vol. 31, no. 4, pp. 734–742, 2020.

[2] C. Fan, H. Liu, B. Li, C. Zhao, and S. Mao, "Adversarial game against hybrid attacks in uav communications with partial information," *IEEE Transactions on Vehicular Technology*, 2021.

[3] W. Kong, D. Zhou, Z. Yang, Y. Zhao, and K. Zhang, "Uav autonomous aerial combat maneuver strategy generation with observation error based on state-adversarial deep deterministic policy gradient and inverse reinforcement learning," *Electronics*, vol. 9, no. 7, p. 1121, 2020.

[4] Y. Ma, G. Wang, X. Hu, H. Luo, and X. Lei, "Cooperative occupancy decision making of multi-uav in beyond-visual-range air combat: A game theory approach," *IEEE Access*, vol. 8, pp. 11 624–11 634, 2019.

[5] X. Zhou, J. Zhu, H. Zhou, C. Xu, and F. Gao, "Ego-swarm: A fully autonomous and decentralized quadrotor swarm system in cluttered environments," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4101–4107.

[6] A. King, *Urban Warfare in the Twenty-first Century*. John Wiley & Sons, 2021.

[7] H. Park, B.-Y. Lee, M.-J. Tahk, and D.-W. Yoo, "Differential game based air combat maneuver generation using scoring function matrix," *International Journal of Aeronautical and Space Sciences*, vol. 17, no. 2, pp. 204–213, 2016.

[8] H. Changqiang, D. Kangsheng, H. Hanqiao, T. Shangqin, and Z. Zhuoran, "Autonomous air combat maneuver decision using bayesian inference and moving horizon optimization," *Journal of Systems Engineering and Electronics*, vol. 29, no. 1, pp. 86–97, 2018.

[9] L. Fu, F. Xie, G. Meng, and D. Wang, "An uav air-combat decision expert system based on receding horizon control," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 41, no. 11, p. 1994, 2015.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] Q. Yang, J. Zhang, G. Shi, J. Hu, and Y. Wu, "Maneuver decision of uav in short-range air combat based on deep reinforcement learning," *IEEE Access*, vol. 8, pp. 363–378, 2019.

[12] L. Wang, J. Hu, Z. Xu, and C. Zhao, "Autonomous maneuver strategy of swarm air combat based on ddpg," *Autonomous Intelligent Systems*, vol. 1, no. 1, pp. 1–12, 2021.

[13] Z. Jiandong, Y. Qiming, S. Guoqing, L. Yi, and W. Yong, "Uav cooperative air combat maneuver decision based on multi-agent reinforcement learning," *Journal of Systems Engineering and Electronics*, vol. 32, no. 6, pp. 1421–1438, 2021.

[14] G. Zhang, Y. Li, X. Xu, and H. Dai, "Efficient training techniques for multi-agent reinforcement learning in combat tasks," *IEEE Access*, vol. 7, pp. 109 301–109 310, 2019.

[15] G. Torrente, E. Kaufmann, P. Föhn, and D. Scaramuzza, "Data-driven mpc for quadrotors," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3769–3776, 2021.

[16] M. T. Spaan, "Partially observable markov decision processes," in *Reinforcement Learning*. Springer, 2012, pp. 387–414.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99, 1999, pp. 278–287.

[19] A. Harutyunyan, S. Devlin, P. Vrancx, and A. Nowé, "Expressing arbitrary reward functions as potential-based advice," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[20] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in neural information processing systems*, 2017, pp. 6379–6390.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[22] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[23] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[24] G. A. Mills-Tettey, A. Stentz, and M. B. Dias, "The dynamic hungarian algorithm for the assignment problem with changing costs," *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*, 2007.

[25] J. v. d. Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics research*. Springer, 2011, pp. 3–19.

[26] C. Yu, A. Velu, E. Vinitsky, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative, multi-agent games," *arXiv preprint arXiv:2103.01955*, 2021.