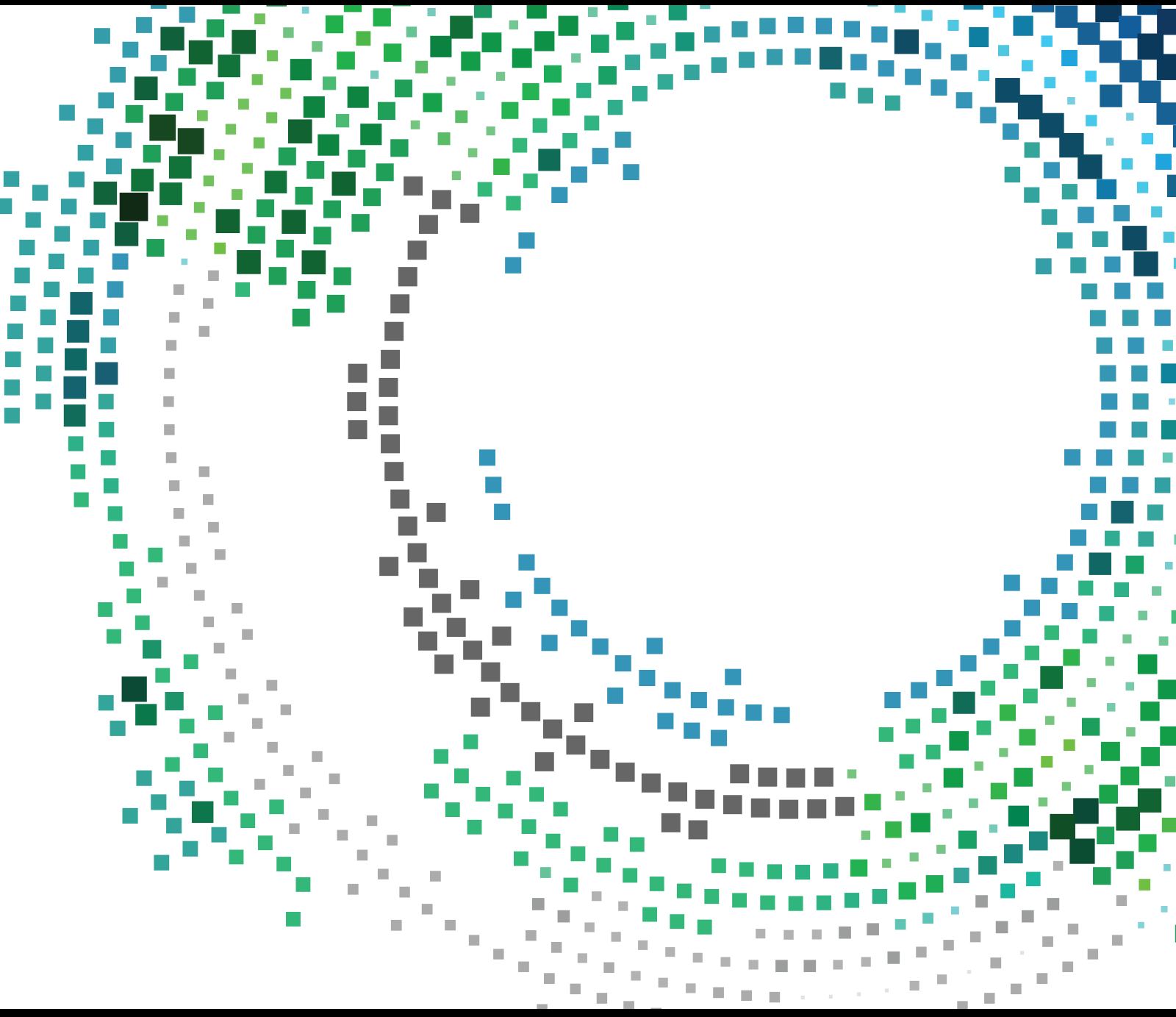
Mobile Information Systems

Intelligent Reflecting Surface Empowered 6G Networks

Lead Guest Editor: Sun Mao
Guest Editors: Ming Xiao, He Li, and Lei Liu



Intelligent Reflecting Surface Empowered 6G Networks

Mobile Information Systems

Intelligent Reflecting Surface Empowered 6G Networks

Lead Guest Editor: Sun Mao
Guest Editors: Ming Xiao, He Li, and Lei Liu

Chief Editor

Academic Editors

Contents

Hindawi
Mobile Information Systems
Volume 2023, Article ID 4916127, 10 pages
https://doi.org/10.1155/2023/4916127

Research Article

Intelligent Path Planning for AGV-UAV Transportation in 6G Smart Warehouse

Weiya Guo and Shoulin Li

Qingdao Agricultural University, Qingdao, China

Correspondence should be addressed to Shoulin Li; shoulin_li@126.com

Received 28 November 2022; Revised 24 April 2023; Accepted 9 May 2023; Published 31 May 2023

Academic Editor: He Li

Recently, deep reinforcement learning (DRL) has attracted increasing interest in the field of intelligent navigation and path planning in smart warehousing. The latest imitation augmented DRL (IADRL) model has achieved good performance for the cooperative transportation tasks of automatic guided vehicles (AGVs) and unmanned aerial vehicles (UAVs). However, this model cannot always transport target cargoes with the optimized policy due to premature convergence. Therefore, we propose an intelligent path planning model for AGV-UAV transportation in this paper. The proposed model utilizes the proximal policy optimization with covariance matrix adaptation (PPO-CMA) in the imitation learning and DRL networks, which enables the AGV-UAV coalition to plan the optimal transportation route at a lower cost. Experiments conducted in simulation warehousing scenarios demonstrated the proposed model and improved the accumulated training reward by more than 10%, outperforming the existing state-of-the-art models in terms of effectiveness and efficiency.

1. Introduction

With the rapid deployment of 5G networks worldwide, 6G and its applications in the industry have attracted more and more attention from researchers [1–3]. The number of materials stored in smart warehouses has increased significantly recently. Maximizing warehouse space utilization is one way to make 6G smart warehouses more common in the future. In modern intelligent warehousing, the transportation of goods is mainly completed by automatic guided vehicles (AGVs) [4]. Due to the limited reachable height of the AGVs, it is impossible to transport goods at higher positions, which constrains the height of the goods storage racks in the warehouse, resulting in a waste of warehouse space. When the quantity of goods exceeds the affordability of the warehouse, additional warehouse space can only be opened to store the goods, and the new warehouse space means an increase in cost and a decrease in profit.

With the development of hardware devices, more and more unmanned aerial vehicles (UAVs) have been developed for various operations, e.g., monitoring, ground target tracking, optical remote sensing, and precision agriculture [5–8]. The most significant advantage of UAVs is that they can conduct tasks at high positions. Applying UAVs to cargo transportation tasks can overcome the limitation of AGVs. However, the energy consumption of UAVs is much higher than that of AGVs. Thus, the working time and operational distance of UAVs are compromised, making it impossible to carry out long-distance transportation tasks [9]. Therefore, we cannot directly replace AGVs with UAVs in cargo transportation tasks.

Based on the previous facts, we intend to combine AGV and UAV to form a cooperative AGV-UAV transportation for cargo transportation tasks and solve problems they cannot complete alone. During transportation, UAVs can target goods at higher positions, while AGVs can target those at lower positions. For goods at a long distance and a high position, the AGV can carry the UAV to the location of the goods, and then, the UAV can fly to process the goods. In this way, the UAV makes up for the height limitation of the AGV, improving the space utilization of the warehouse

effectively, and reducing the working time and the power consumption of the UAV.

For AGV-UAV transportation, path planning is an essential part of its navigation process. Selecting the shortest transportation route during transportation can reduce transportation costs in terms of time and energy. Recent years have witnessed the emergence of various path planning algorithms, including traditional path planning algorithms (e.g., Dijkstra algorithm [10], A* algorithm [11], artificial potential field algorithm [12]), and intelligent path planning algorithms (e.g., genetic algorithm [13], particle swarm algorithm [14], and ant colony algorithm [15]). These algorithms achieved specific achievements in path planning but are easily disturbed by environmental factors and cannot process data in large-scale state space. With the popularity of artificial intelligence, deep reinforcement learning (DRL) is playing an increasingly important role in intelligent navigation and path planning due to its excellent perception and decision-making capabilities [16]. Particularly, Zhang et al. proposed an imitation augmented deep reinforcement learning (IADRL) model for transportation tasks in complex environments [17]. Compared with the traditional algorithms, IADRL enables the AGV-UAV coalition to accomplish cargo tasks at a lower cost.

However, IADRL may converge in advance and fall into a local optimum in the training process [18]. To target the previous problem, we propose an intelligent path planning model for AGV-UAV transportation in this paper. By introducing the proximal policy optimization with covariance matrix adaptation (PPO-CMA) [19] into the policies of the imitation learning (IL) and DRL networks, our model can not only learn the latent behavioral features of the AGV-UAV coalition from the demonstration data but also provides behavioral decisions for the coalition with better optimization policy. Experimental results show that our model is superior to its rivals by solving the premature convergence problem, enabling the AGV-UAV coalition to complete the transportation task at a lower cost.

The remainder of this paper is organized as follows. Section 2 discusses the related work, and the proposed approach is detailed in Section 3. Section 4 presents the experimental results and Section 5 concludes this paper.

2. Related Works

Path planning has recently been a hot issue in robotics research, and the core requirement is to find an optimal path from the starting point to the endpoint with the lowest cost (e.g., distance, time, and energy). Existing algorithms can be mainly divided into three categories: (1) traditional algorithms, (2) intelligent algorithms, and (3) DRL-based algorithms.

2.1. Traditional Algorithms. Traditional path planning algorithms include the Dijkstra algorithm [10], A* algorithm [11], and artificial potential field algorithm [12]. The Dijkstra algorithm is a classic algorithm in the field of path planning, which uses a greedy policy to expand one node at a time to traverse the nodes in the environment to achieve the shortest path from the start to the end. Based on the Dijkstra's algorithm, the A* algorithm adds heuristic rules to converge faster when nodes expand. Although the A* algorithm has been widely used in many fields, the application scenarios of the A* algorithm are limited to discrete spaces. The artificial potential field algorithm sets the gravitational force between the agent and the target and the repulsion force between the agent and the obstacle so that the agent can reach the target position along the direction of the resultant force. However, the force ratio for different scenes can only be manually coordinated, making the optimal configuration difficult to obtain, which limits its applications in complex environments.

2.2. Intelligent Algorithms. Intelligent path planning algorithms are a series of algorithms produced by observing natural phenomena and animal habits, including the genetic algorithm [13], particle swarm optimization (PSO) algorithm [14], and ant colony algorithm [15]. The genetic algorithm imitates the selection and genetic mechanism of nature to seek the optimal solution. However, it depends on the initial population selection, and its convergence speed is slow when solving large-scale problems. The ant colony algorithm and the PSO algorithm imitate the swarm intelligence behavior of ant colonies and bird swarms and have good parallelism and fast convergence speed. Nevertheless, the parameter setting affects the performance of these two algorithms, making them easily fall into the local optimal solution.

2.3. DRL-Based Algorithms. Reinforcement learning can optimize the agent's action policy by maximizing long-term returns without background knowledge. It can find the optimal path through continuous trial and error in a completely unknown environment [20]. Therefore, researchers applied DRL to target path planning problems. Mirowski et al. proposed a DRL method to train agents to navigate within large and visually rich environments by introducing memory and auxiliary learning targets [21]. Sallab et al. presented the DQN algorithm for the discrete actions and deep deterministic actor-critic algorithm for continuous actions to lane keeping assist [22]. Chen et al. designed a time-efficient navigation policy based on socially aware collision avoidance with DRL, which can enable fully autonomous navigation of a robotic vehicle in an environment with many pedestrians [23]. Kendall et al. applied the DRL to a full-sized autonomous vehicle, which can learn a policy for lane following in a handful of training episodes via a single monocular image as input [24]. By combining imitation learning (IL) and DRL, Zhang et al. proposed an IADRL model for the AGV-UAV coalition [17] to cooperatively and cost-effectively accomplish tasks. However, the IADRL model suffers from the local optimum problem due to the convergence in advance. Therefore, there is still space to enhance the path planning performance in AGV-UAV transportation tasks.

3. The Proposed Approach

3.1. Motivation and Challenges. As discussed in Introduction Section, the IADRL model combines deep reinforcement learning and imitation learning to learn the cooperative and complementary behavior mode of AGV-UAV transportation alliance from expert data and interactive data. As the action policy adopted by IADRL, however, the defects of proximal policy optimization (PPO) itself may cause IADRL to fall into local optimum in the learning process and thus be unable to find the optimal path.

To better analyze the shortcomings of PPO, we create an environment only containing two-dimensional actions, which facilitates us to visualize the distribution of the actions chosen by the policy during the iterative process. In this environment, the reward is negatively correlated with the sum of squares of the actions chosen by the policy so that the policy reaches the optimum when both actions chosen by the policy are zero. In Figure 1, we visualized the distribution of actions selected by different policies at different iterations, where green represents positive-advantage actions and red represents negative-advantage actions.

In the first row of Figure 1, when the policy performs multiple minibatch gradient descent with the same data in PPO style without considering the clipping loss, the actions chosen by the policy at 9 iterations deviate from the optimal point. Such a situation happens because the negative-advantage actions push the policy away from the negative-advantage actions. In contrast, the positive-advantage actions pull the policy towards the positive-advantage actions. Each step of the updating process moves the policy away from the negative-advantage actions, eventually causing the strategy to deviate from the optimal point.

As shown in the second row of Figure 1, compared with the first row, PPO does not deviate during the iterative process but approaches the optimal point as the iteration proceeds. However, the final policy still does not exactly reach the optimal point. This is because PPO limits the update range of the policy through the clipping loss to prevent the policy's deviation. But the clipping loss also causes the policy to converge early and fall into the local optimum [22].

Based on our research on reinforcement learning algorithms, we noted that PPO-CMA [25] can solve the previously mentioned problems of PPO well. PPO-CMA prevents the early convergence of policy by using the standard policy gradient loss instead of clipping loss and updating the policy's variance and mean with separate networks, respectively. Moreover, PPO-CMA avoids the policy deviation problem caused by negative-advantage actions by converting negative-advantage actions to positive ones through a mirroring method. As seen in the third row of Figure 1, PPO-CMA starts to converge only when it is close to the optimal point and finally reaches the optimal point of the strategy exactly.

All these observations inspired us to propose a new model based on PPO-CMA to solve the premature convergence problem presented in IADRL and to provide path planning for AGV-UAV alliances in transportation tasks.

3.2. The Proposed Model. To deal with the problem of premature convergence in IADRL, we propose a new model for path planning of the AGV-UAV alliance using PPO-CMA as the action policy. Specifically, the clipping loss is first replaced by the standard policy gradient loss to prevent premature convergence. Afterward, the mean and variance of the policy are updated separately using separate networks to further extend the variance in the optimal search direction. Moreover, the negative-advantage action is turned into a positive-advantage action by a mirroring method.

The AGV-UAV transportation coalition can be described by the tuple $\langle \epsilon, \mathbf{o}, \mathbf{a}, r, \gamma, M \rangle$, where ϵ represents the environment, r is the reward function, $\gamma \in (0, 1]$ is the discount factor for future rewards, and M is the complementary cooperation model of the AGV-UAV. The $\mathbf{o} = (o_1, o_2)$ represents the observed values of the coalition on the environment, consisting of o_1 for the observation value of the AGV and o_2 for the observation value of the UAV. The $\mathbf{a} = (a_1, a_2) \sim M$ means the action of the transportation coalition, which consists of the action a_1 taken by the AGV and the action a_2 taken by the UAV. The goal is to learn a joint value-action function $Q_c^\pi(\mathbf{o}, \mathbf{a}; \theta)$ that enables the AGV-UAV coalition to achieve maximum overall reward (or minimum overall cost), while accomplishing various tasks.

According to the generative adversarial imitation learning (GAIL) model [25], the IL model in this paper includes a generator G and a discriminator D. The generator G, also the policy π in the DRL model, is responsible for producing actions closer to the distribution of expert data based on a given observation o to pass the detection of the discriminator D. The discriminator D distinguishes the expert data from the data obtained by the generator G. During the training process, the value function should be maximized, described as follows [17]:

$$\begin{aligned} V(\omega) = {} & \mathbb{E}_\pi[\log(D(\mathbf{o}, \mathbf{a}; \omega))] \\ & + \mathbb{E}_{\tau_E}[\log(1 - D(\mathbf{o}, \mathbf{a}; \omega))] - \lambda H(\pi). \end{aligned} \tag{1}$$

Here, ω is the weight of the D, $H(\pi)$ is the entropy of the policy π [26], $\lambda \geq 0$ is the discount factor for H, and τ_E is the expert policy provided by the demonstrated data.

The value function Q_c^π in the DRL model is used to process the received rewards and evaluate the current action selected by policy π. The training of the DRL model aims to maximize the value function Q_c^π of the AGV-UAV, defined by

$$Q_c^\pi(\mathbf{o}, \mathbf{a}; \theta) = \mathbb{E}\left[r_{au}(\mathbf{o}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{a}' \sim \pi}\left[Q_c^\pi(\mathbf{o}', \mathbf{a}')\right]\right], \tag{2}$$

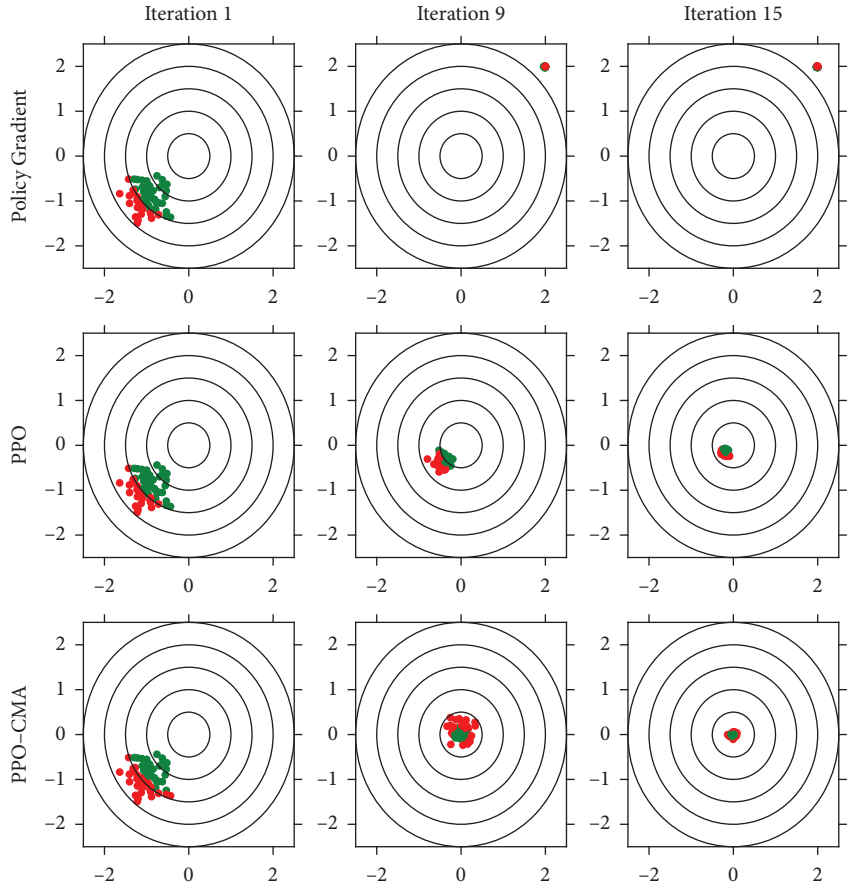


Figure 1: Multiple minibatch iterations of different policies in a two-dimensional action environment.

where θ is the parameter of the function Q_c^{π}, $\gamma \in (0, 1]$ is the discount factor for future rewards, and r_{au} is the augmented reward function.

To prevent premature policy convergence, the following standard policy gradient loss is used as the loss function of the policy π instead of the clipping loss.

$$J(\varphi) = \frac{1}{K}\sum_{i=1}^{K} A^{\pi}(o_i, a_i) \sum_{j}\left[\frac{\left(a_{i,j} - \mu_{j;\varphi}(o_i)\right)^2}{\left(c_{j;\varphi}(o_i)\right)} + 0.5\log c_{j;\varphi}(o_i)\right], \tag{3}$$

where φ is the parameter of the value function J_{φ}, i is the mini-batch sample index, j indexes the operand variables, and K is the number of sample batches. $A^{\pi}(o_i, a_i)$ represents the advantage function for measuring the payoff of taking action a_i in state o_i.

In addition, the mean and variance of the policy are generated using separate networks so that the variance can be updated before the mean is updated. This allows the policy to find the optimal point more quickly by elongating the exploration distribution along the optimal search direction rather than converging the variance prematurely [27].

Considering that negative-advantage actions may cause policy deviation, a mirroring technique is employed to convert negative-advantage actions into positive ones. Given the linearity of advantage around the current policy mean $\mu(s_i)$, it is possible to mirror negative-advantage actions into positive-advantage actions about the mean. Specifically, we set $a_i' = 2\mu(s_i) - a_i$, $A^{\pi}(a_i') = -A^{\pi}(a_i)\psi(a_i, s_i)$, where $\psi(a_i, s_i)$ is a Gaussian kernel that assigns less weight to actions far from the mean.

4. Experimental Results and Analysis

In this section, we first conducted the experiment of PPO and PPO-CMA in the gym environment provided by OpenAI. After that, we built an experimental environment for the AGV-UAV problem and detailed the environment configuration. Based on this, we demonstrated the effectiveness and superiority of the proposed model by comparing the experimental results with other models.

4.1. Gym Experiment. From Figure 1, we can see that PPO-CMA solves the problem of PPO's early convergence, and it is no longer disturbed by negative-advantage actions. To better demonstrate the advantages of PPO-CMA, we further compare the two algorithms in the gym environment.

As can be seen from Figure 2, the experiments in MountainCar-v0 and BipedalWalker-v3 show that PPO-CMA can achieve higher rewards in the experiment, which shows that PPO-CMA is superior to PPO.

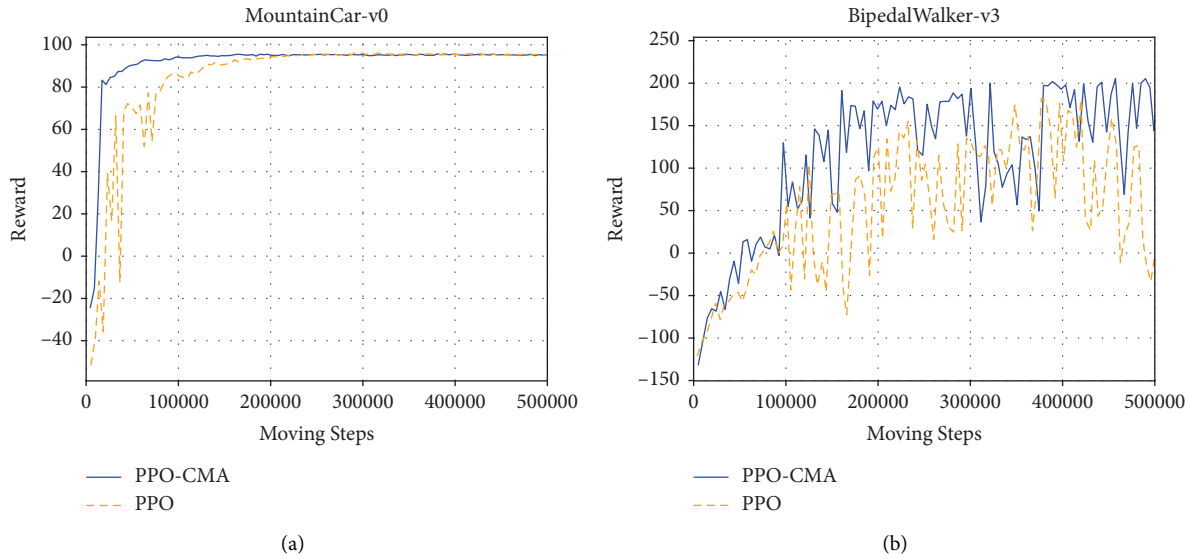


Figure 2: Rewards of PPO-CMA and PPO algorithms in MountainCar-v0 and BipedalWalker-v3 environments.

Moreover, PPO-CMA is obviously faster than PPO in convergence speed.

Figure 3 gives the variance of the two policies in the training process. It can be seen that the sampling variance of PPO-CMA reduces to the minimum value more slowly than that of PPO, which effectively expands the exploration variance, prevents the policy from falling into local optimum, and achieves a better final training effect of PPO-CMA.

4.2. AGV-UAV Transportation Experiment

4.2.1. Experimental Configuration. We designed a virtual simulation scenario for the proposed model based on the Unity3D ML-Agents platform [28], and we deployed an AGV-UAV coalition with the size of $50\,\text{m} \times 50\,\text{m} \times 10\,\text{m}$, and the mission of the coalition was to complete the transportation of goods in the shortest path. As shown in Figure 4, the cyan-blue squares represent the AGV, the yellow square represents the UAV, and the green, red, and purple spheres represent the target cargoes at different heights and positions.

In the experiments, each agent's ray-cast sensor provided by Unity3D collects the environment states. The ray-cast sensor casts rays into the surrounding environment and the position of all detected objects and their distances can be obtained. The ray of the AGV only detects the environment in the horizontal direction, while the ray of the UAV swings up and down 45 degrees to detect the environment. The detection range of all rays is set to 20 meters. The observation o of an AGV-UAV coalition is a vector containing environmental information combined with all its detected ray returns.

The action of the AGV is expressed as $a_1 = [a_x, a_y]$, and the action of the UAV is expressed as $a_2 = [a_z]$, where a_x, a_y, and a_z represent the agent's acceleration in the x, y, and z directions. The action of the AGV-UAV coalition is composed of the action of the AGV and the UAV, $a = (a_1, a_2)$.

In the proposed model, the discriminator is set up with two hidden layers of 128 neural units each. Meanwhile, the value function is set up with three hidden layers with 512 units per layer, and the policy π is set up with three hidden layers with 512 units per layer. In addition, the initial positions of the AGVs, the UAVs, and the target cargoes are random.

The environmental reward is designed based on the situation that the AGV-UAV coalition may encounter. For the coalition to learn the least expensive path, we set a small penalty of 0.01 for each step of the coalition. Since the battery life of the AGV is 5 to 10 times that of the UAV, we set the penalty for each step of the UAV to be 6 times that of the AGV. Therefore, under normal circumstances, the UAV should be carried by the AGV to the destination, and then, the UAV starts to work. We set the reward for each goal to be 120 to encourage the coalition to complete the task. Considering that there may be obstacles in the actual situation, we set up obstacles in the scene and made a large penalty of −30 for the coalition to collide with the obstacles. The final reward of 120 is obtained when the coalition has achieved all objectives.

In the experiments, we can manually control the agent to complete some simple tasks and record the data to train the model as expert data. We collected the running data of the agent for 10,000 steps, where the data includes all basic scenarios of AGV and UAV cooperating to complete the task. It should be noted that the expert data enables the model to learn the cooperative and complementary relationship between AGV and UAV, not to learn the optimization policy of the path. Therefore, our demo data only needs to reflect the behavioral characteristics of the AGV-UAV coalition. That is to say, the AGV first carries the UAV to the target position, and

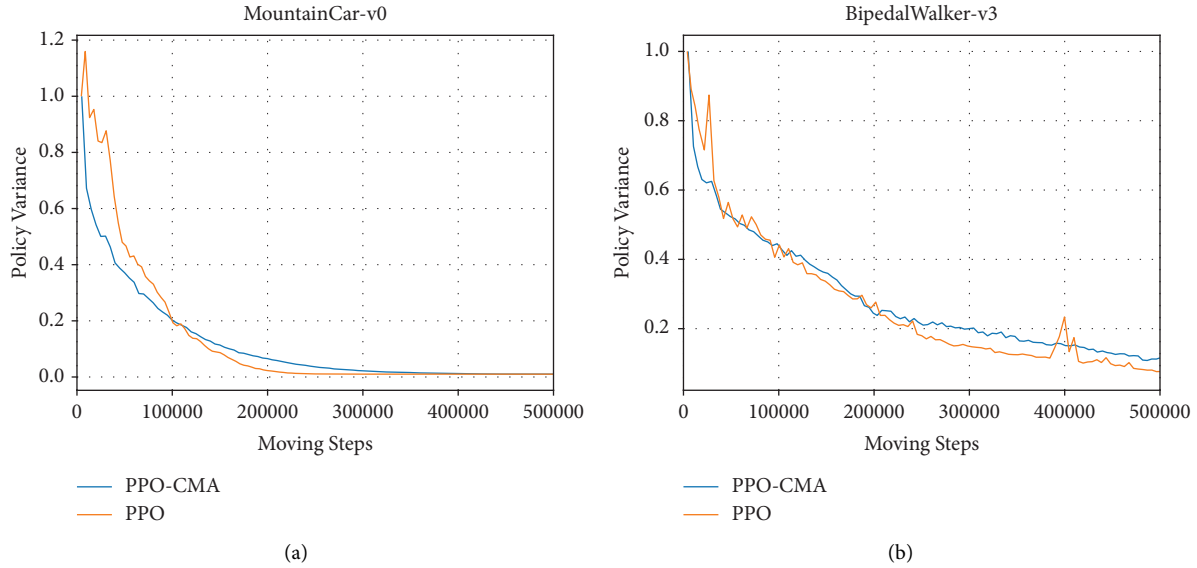


Figure 3: Exploration variances of PPO-CMA and PPO algorithms in MountainCar-v0 and BipedalWalker-v3 environments.

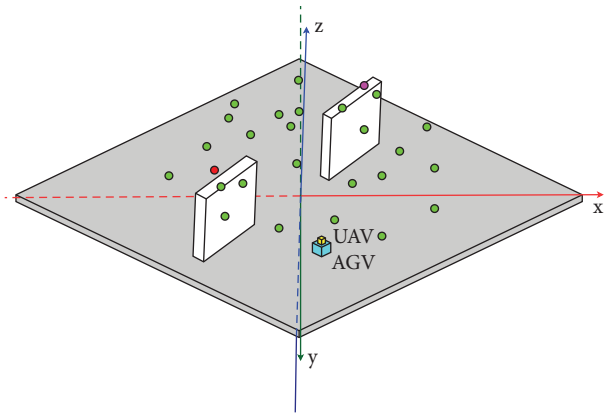


Figure 4: Simulation scenario for AGV-UAV transportation tasks. The AGV-UAV coalition consists of the AGVs in cyan-blue, the UAV in yellow, and target cargoes in light green, red, and purple.

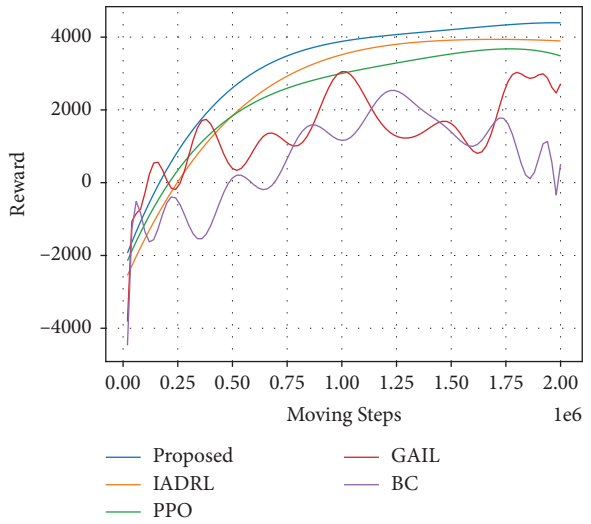


Figure 5: Accumulated training rewards values for the PPO, GAIL, IADRL, BC, and proposed models.

then, the UAV takes off and starts to work. Moreover, there is no need to artificially optimize the route from the coalition to the target.

4.2.2. Experimental Results. In the AGV-UAV transportation task, the maximum training step for each episode is set to 20,000. If the coalition gets all the goods, the episode terminates immediately, otherwise, training continues until the agent runs out of the maximum step. In the experiments, we compare the proposed model with four models including PPO, behavior cloning (BC) [29], GAIL, and IADRL for performance evaluation. To ensure a fair comparison, we use the same parameters, i.e., the number of targets, the learning rate, and the maximum step, for all models.

Figure 5 first compares the rewards obtained by all five models. Obviously, the proposed model has the highest rewards, indicating the best optimization ability of path planning. According to the results, the highest reward of the proposed model is 4400, but the highest reward of IADRL is less than 4000, resulting in more than 10% improvement. The IADRL outperforms the PPO, GAIL, and BC models due to the combination of IL and DRL. The PPO model can learn policies based on the environment, so it can quickly learn to avoid obstacles at the beginning of the training process. However, without the guidance of demonstration data, it cannot learn the behavior characteristics of the AGV-UAV coalition, resulting in its training speed and final reward being lower than IADRL and the proposed model. In addition, it can be seen that the PPO, IADRL, and the proposed models tend to converge in the end. But the GAIL and BC models fail, which is basically consistent with the theoretical conjecture that GAIL and BC only replicate the

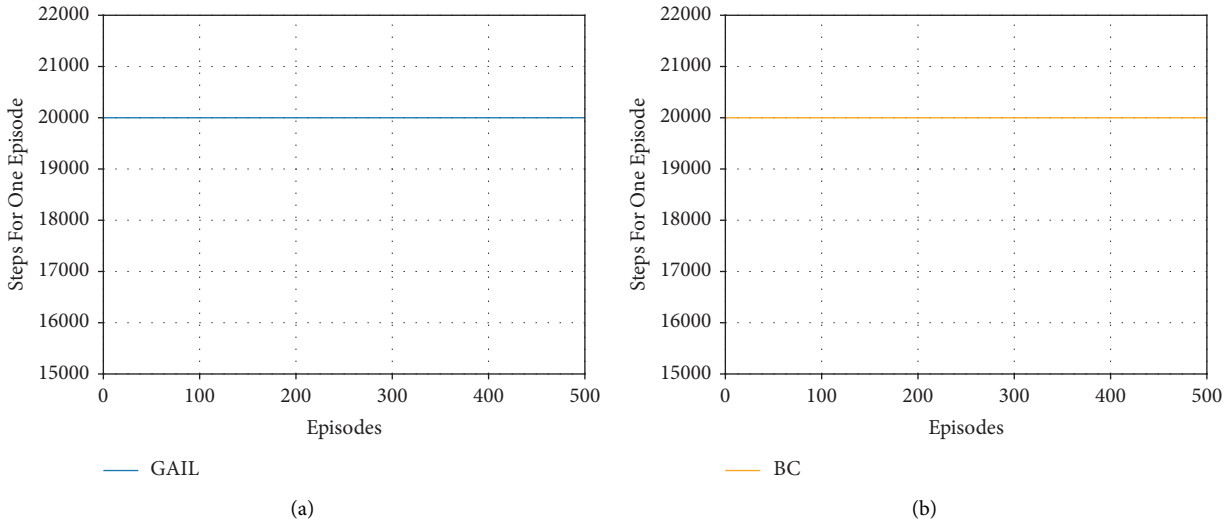


Figure 6: The number of running steps in each episode of the GAIL and BC models. (a) GAIL. (b) BC.

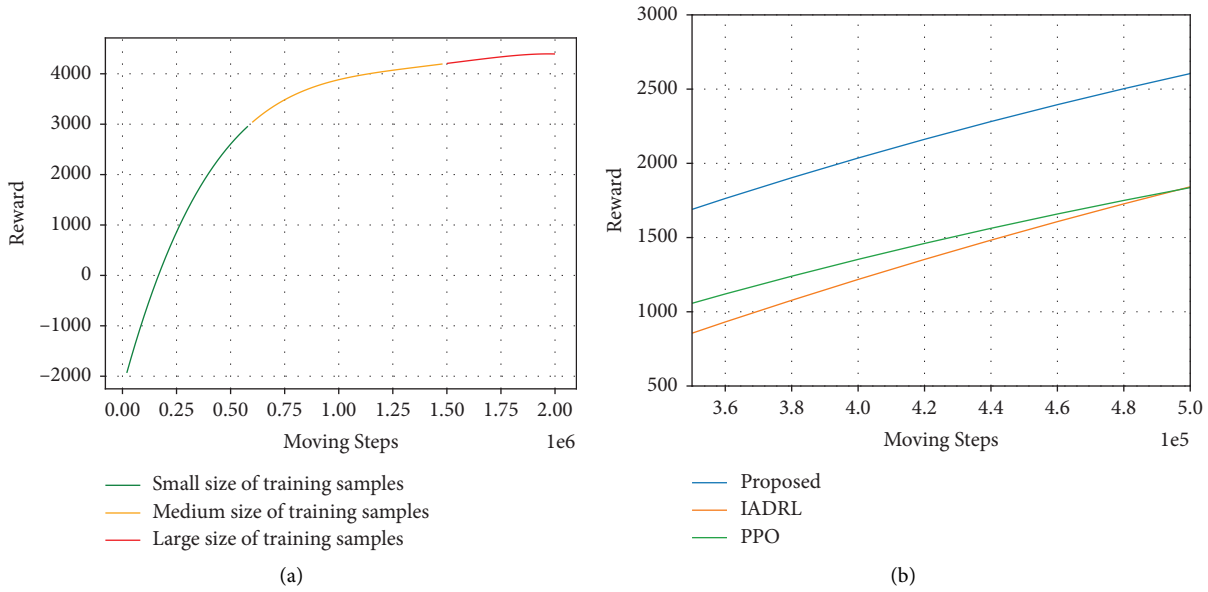


Figure 7: (a) Accumulated training rewards values for different training data sizes: green for the period with a small size of data, orange for a medium size, and red for a large size. (b) Performance comparison of the proposed model with IADRL and PPO when the training data size is small.

actions and policies provided by the demonstration data, rather than obtaining optimal policies by obtaining higher rewards. As shown in Figure 6, the number of moving steps in each episode of the BC and GAIL models is always equal to the maximum step of 20000, which means that both of them fail to complete the transportation task of all targets. This is because these two models depend highly on expert data and cannot be adaptively suitable for complicated environments.

During the training process, the agents collect data by continuously interacting with the environment. The training samples can get more and more with the increment of moving steps. To evaluate how well the proposed model performs in different sizes of training samples, Figure 7(a) shows the accumulated training reward values for different periods according to the training sample size. In particular, we colored the reward curve in different periods: green for the period with a small size of training samples, orange for a medium size, and red for a large size. It can be seen that the reward is low but increases faster when it has a small size of training samples. In this case, our model still outperforms the IADRL and PPO models with a higher reward, as shown in Figure 7(b).

Figure 8 shows the number of collisions between the AGV-UAV coalition and obstacles in each episode. It can be seen that the PPO, IADRL, and proposed models can quickly reduce the number of collisions to a minimum after training with many collisions in the early stage of training. But the BC and GAIL models keep a high number of collisions due to lacking environmental rewards.

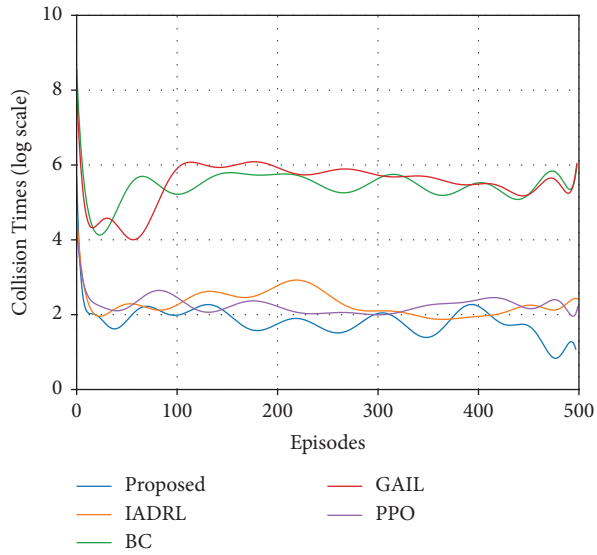


Figure 8: The number of collisions between the AGV-UAV coalition and obstacles in each episode. For clarity, the y-axis is plotted as a log scale and smoothed.

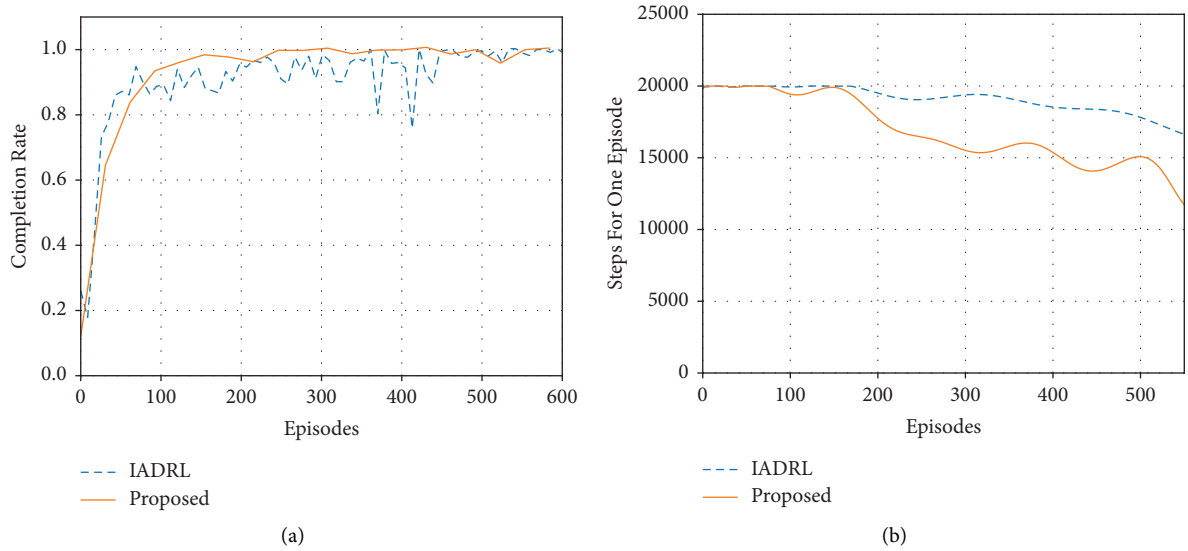


Figure 9: Performance comparison of the IADRL and proposed models. (a) The comparison of task completion rate in each episode. (b) The comparison of the moving step number in each episode.

To better show the superiority of the proposed model, we use the following metrics for evaluation: (1) the mission completion rate, that is, the percentage of goals reached by the coalition in each episode to the total number of goals and (2) the number of steps required to complete an episode of tasks. The IADRL and proposed models are further compared in Figure 9, where (a) shows the task completion rate and (b) shows the number of moving steps. It can be seen that the completion rate of the proposed model is superior to that of the IADRL model. However, after about 250 episodes, the completion rate of the proposed model reaches one and keeps stable for the following episodes. According to Figure 9(b), in the early learning stage, it is difficult for the models to complete the transportation of all goods without a suitable policy. Therefore, the number of steps consumed by the models in each episode reaches a maximum of 20000. With gradual training, the policy is gradually optimized, and the number of steps for completing an episode decreases. It can be seen that the proposed model outperforms the IADRL in terms of task completion rate and the number of moving steps.

4.2.3. Discussion. In this paper, we have taken into account the energy constraints of the UAV and outlined the operational guidelines for the AGV-UAV alliance. The AGV will initially transport the UAV to the desired location, where it will then take off to the required altitude to complete the task. This approach restricts the UAV's operational radius to only the

area directly above the AGV. Unfortunately, when the target is beyond the AGV's reach or can only be accessed via a lengthy detour, the UAV's limited range of motion will increase the overall cost of completing the mission for the alliance.

For example, as shown in Figure 4, the purple target is located directly above the obstacle. Although the UAV can reach the height where this target is located yet it cannot complete the transportation mission because the AGV cannot reach directly below the target. In addition, the red target in Figure 4 is located on the other side of the obstacle, which requires the AGV to go around the obstacle to reach directly below the target before the UAV can take off to handle the target. In this case, if the UAV can move horizontally, then as soon as the AGV reaches the vicinity of the obstacle, the UAV can take off to handle the target and the alliance can accomplish the task with less cost.

5. Conclusion

In this paper, an intelligent path planning model was proposed for the AGV-UAV transportation task in 6G smart warehouse environments. The proposed model utilizes PPO-CMA in the IL and DRL networks to prevent premature convergence of policy. This enables the AGV-UAV coalition to learn behavior patterns and complete transportation tasks at a lower cost. The experiments conducted in a simulated warehouse environment demonstrate that the proposed model outperforms the baselines. In the future, the focus will be on enabling the AGV-UAV alliance to accomplish transport missions in complementary and cooperative working modes. In addition, exploring ways to allow the UAV to move horizontally to further reduce costs will be a topic of interest.

Data Availability

The data used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank TopEdit (https://www.topeditsci.com) for its linguistic assistance during the preparation of this manuscript. This work was supported by the Social Science Planning Program of Qingdao under Grant QDSKL2201278 and the Qingdao City "Government-Industry-University-Research Fund Service" Innovation and Entrepreneurship Community Project under Grand 22-7-5-gtt-2-gx.

References

[1] J. Feng, L. Liu, Q. Pei, and K. Li, "Min-max cost optimization for efficient hierarchical federated learning in wireless edge networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 11, pp. 1–2700, 2022.

[2] L. Liu, M. Zhao, M. Yu, M. Jan, D. Lan, and A. Taherkordi, "Mobility-aware multi-hop task offloading for autonomous driving in vehicular edge computing and networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1–14, 2022.

[3] S. Mao, L. Liu, N. Zhang et al., "Reconfigurable intelligent surface-assisted secure mobile edge computing networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6647–6660, 2022.

[4] E. Oyekanlu, A. Smith, W. Thomas et al., "A review of recent advances in automated guided vehicle technologies: integration challenges and research areas for 5g-based smart manufacturing applications," *IEEE Access*, vol. 8, pp. 202312–202353, 2020.

[5] J. Martinez, M. Gheisari, and L. Alarcón, "Uav integration in current construction safety planning and monitoring processes: case study of a high-rise building construction project in Chile," *Journal of Management in Engineering*, vol. 36, no. 3, Article ID 05020005, 2020.

[6] B. Li and Y. Wu, "Path planning for uav ground target tracking via deep reinforcement learning," *IEEE Access*, vol. 8, pp. 29064–29074, 2020.

[7] E. Alvarez-Vanhard, T. Corpetti, and T. Houet, "Uav & satellite synergies for optical remote sensing applications: a literature review," *Science of remote sensing*, vol. 3, Article ID 100019, 2021.

[8] G. Messina and G. Modica, "Applications of uav thermal imagery in precision agriculture: state of the art and future research outlook," *Remote Sensing*, vol. 12, no. 9, p. 1491, 2020.

[9] Q. Yu, Z. Shen, Y. Pang, and R. Liu, "Proficiency constrained multi-agent reinforcement learning for environment-adaptive multi uav-ugv teaming," in *Proceedings of the 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, pp. 2114–2118, Lyon, France, August 2021.

[10] M. Enayattabar, A. Ebrahimnejad, and H. Motameni, "Dijkstra algorithm for shortest path problem under interval-valued pythagorean fuzzy environment," *Complex & Intelligent Systems*, vol. 5, no. 2, pp. 93–100, 2019.

[11] S. Erke, D. Bin, Y. Nie, Q. Zhu, L. Xiao, and D. Zhao, "An improved a-star based path planning algorithm for autonomous land vehicles," *International Journal of Advanced Robotic Systems*, vol. 17, no. 5, Article ID 1729881420962263, 2020.

[12] U. Orozco-Rosas, K. Picos, and O. Montiel, "Acceleration of path planning computation based on evolutionary artificial potential field for non-static environments," in *Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications*, pp. 271–297, Springer, Berlin, Germany, 2020.

[13] M. Nazarahari, E. Khanmirza, and S. Doostie, "Multi-objective multi-robot path planning in continuous environment using an enhanced genetic algorithm," *Expert Systems with Applications*, vol. 115, pp. 106–120, 2019.

[14] X. Liu, D. Zhang, J. Zhang, T. Zhang, and H. Zhu, "A path planning method based on the particle swarm optimization trained fuzzy neural network algorithm," *Cluster Computing*, vol. 24, no. 3, pp. 1901–1915, 2021.

[15] S. Mirjalili, J. Song Dong, and A. Lewis, "Ant colony optimizer: theory, literature review, and application in auv path planning," *Nature-inspired optimizers*, Springer, Berlin, Germany, pp. 7–21, 2020.

[16] H. Bayerlein, M. Theile, M. Caccamo, and D. Gesbert, "Multi-uav path planning for wireless data harvesting with deep

reinforcement learning," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1171–1187, 2021.

[17] J. Zhang, Z. Yu, S. Mao, S. C. Periaswamy, J. Patton, and X. Xia, "Iadrl: imitation augmented deep reinforcement learning enabled ugv-uav coalition for tasking in complex environments," *IEEE Access*, vol. 8, pp. 102335–102347, 2020.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, https://arxiv.org/abs/1707.06347.

[19] P. Hämäläinen, A. Babadi, X. Ma, and J. Lehtinen, "PPO-CMA: proximal policy optimization with covariance matrix adaptation," in *Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, Espoo, Finland, September 2020.

[20] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Foundations and Trends® in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.

[21] P. Mirowski, R. Pascanu, F. Viola et al., "Learning to navigate in complex environments," 2016, https://arxiv.org/abs/1611.03673.

[22] A. Sallab, M. Abdou, E. Perot, and S. Yogamani, "End-to-end deep reinforcement learning for lane keeping assist," 2016, https://arxiv.org/abs/1612.04340.

[23] Y. Chen, M. Everett, M. Liu, and J. How, "Socially aware motion planning with deep reinforcement learning," in *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1343–1350, IEEE, Vancouver, Canada, September 2017.

[24] A. Kendall, J. Hawke, D. Janz et al., "Learning to drive in a day," in *Proceedings of the 2019 International Conference on Robotics and Automation (ICRA)*, pp. 8248–8254, IEEE, Montreal, Canada, May 2019.

[25] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[26] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the International Conference on Machine Learning*, pp. 1889–1897, PMLR, Lille, France, July 2015.

[27] N. Hansen, "The cma evolution strategy: a tutorial," 2016, https://arxiv.org/abs/1604.00772.

[28] A. Juliani, E. Teng, A. Cohen et al., "Unity: a general platform for intelligent agents," 2018, https://arxiv.org/abs/1809.02627.

[29] A. Edwards, H. Sahni, Y. Schroecker, and C. Isbell, "Imitating latent policies from observation," in *Proceedings of the International Conference on Machine Learning*, pp. 1755–1763, PMLR, Long Beach, CA, USA, June 2019.