electronics

MDPI

Article

Reinforcement Learning-Based UAVs Resource Allocation for Integrated Sensing and Communication (ISAC) System

Min Wang [1], Peng Chen [1,*], Zhenxin Cao [1] and Yun Chen [2]

[1] State Key Laboratory of Millimeter Waves, Southeast University, Nanjing 210096, China; simin@seu.edu.cn (M.W.); caozx@seu.edu.cn (Z.C.)
[2] State Key Laboratory of ASIC and System, Fudan University, Shanghai 200433, China; chenyun@fudan.edu.cn
* Correspondence: chenpengseu@seu.edu.cn

Abstract: Due to the limited ability of a single unmanned aerial vehicle (UAV), group unmanned aerial vehicles (UAVs) have attracted more attention in communication and radar fields. The use of an integrated sensing and communication (ISAC) system can make communication and radar modules share a radar module's resources, coupled with efficient resource allocation methods. It can effectively solve the problem of inadequate UAV resources and the low utilization rate of resources. In this paper, the resource allocation problem is addressed for group UAVs to achieve a trade-off between the detection and communication performance, where the ISAC system is equipped in group UAVs. The resource allocation problem is described by an optimization problem, but with group UAVs, the problem is complex and cannot be solved efficiently. Compared with the traditional resource allocation scheme, which needs a lot of calculation or sample set problems, a novel reinforcement-learning-based method is proposed. We formulate a new reward function by combining mutual information (MI) and the communication rate (CR). The MI describes the radar detection performance, and the CR is for wireless communication. Simulation results show that compared with the traditional Kuhn Munkres (KM) or the deep neural network (DNN) methods, this method has better performance with the increase in problem complexity. Additionally, the execution time of this scheme is close to that of the DNN scheme, and it is better than the KM algorithm.

Keywords: group UAVs; resources allocation; reinforcement learning; integrated sensing and communication (ISAC) system

Citation: Wang, M.; Chen, P.; Cao, Z.; Chen, Y. Reinforcement Learning-Based UAVs Resource Allocation for Integrated Sensing and Communication (ISAC) System. *Electronics* **2022**, *11*, 441. https://doi.org/10.3390/electronics11030441

Academic Editor: Nurul I. Sarkar

Received: 31 December 2021
Accepted: 30 January 2022
Published: 1 February 2022

1. Introduction

In recent years, due to the ability limitation of a single unmanned aerial vehicle (UAV), group UAVs have been proposed for complex applications. With the increasing popularity of modernization and intelligence, the intelligence of UAVs has attracted more attention. Group UAVs have prominent advantages such as high autonomy, multiple functions, timeliness, strong anti-damage ability and low cost. Their applications include logistics distribution, agricultural plant protection, reconnaissance and raids, electronic countermeasures, communication and navigation [1–6]. Group UAVs mainly use public resources to perform these tasks at present, which causes information leakage and communication interference between UAVs. With the exponential growth of the number of UAVs, expanding resources and improving the utilization rate of resources have become an urgent demand. Additionally, Group UAVs are supposed to have the abilities of detection, localization and communication according to their tasks. However, the weight and cost of equipment for UAVs limit the available resources of communication and detection. For the above consideration, an integrated sensing and communication (ISAC) system with the characteristics of low-cost, lightweight and a high level of integration has been proposed by using the same hardware [7].

The ISAC protocol, system architecture design [8], signal-sharing [8–13], time-sharing [14], array-sharing [15], spectrum-sharing [16] and power-sharing [17] have been studied to

realize the ISAC system. In signal-sharing, due to no significant difference between radar and communication in the time and spatial domains, both the radar detection and the wireless communication use the same waveform [8–13]. However, the interference between radar and communication is relatively large. Time-sharing means that different tasks are allocated at different time slots, and either the communication task or the detection task is performed in one time slot [14]. Array-sharing is used in array system, where the different functions are realized in different sub-arrays [15]. Recently, a distributed dual-function radar-communication (DFRC), multiple input and multiple output (MIMO) system was proposed using both signal and power sharing [17]. Integrating cognitive radio (CR-o) and cognitive radar (CR-r) paradigms, as spectrum-sharing paradigms, achieve the intelligent utilization of spectrum resources [18]. Additionally, power-sharing is based on game theory [19]. However, existing resource allocation methods are for a single type of resource or a single system, so it cannot be applied to group UAVs directly.

Information sharing and target detection among group UAVs greatly expand the applications. Improving the stability and reliability of the communication and the accuracy of target detection among UAVs is important. Due to resource limitations and the task importance, both communication and detection tasks need a trade-off in the allocation of resources. With the development of artificial intelligence in the fields of communication and radar, reinforcement learning algorithms can be introduced. In the fields of radar and communication, modulation identification, signal detection, radar target identification [20–24], cluster cooperative control [25–30] and resource allocation [31–33] are included. Additionally, policy-based reinforcement learning and action-critical deep deterministic policy gradient algorithms are used to allocate the energy resources of cellular network reasonably [34], and the resource allocation method for radar detection [35] is given.

In recent years, there have been many studies on resource optimization by reinforcement learning in vehicle-mounted networks. Ref. [36] proposed a vehicle to vehicle (V2V) distributed resource allocation mechanism based on deep reinforcement learning, which can make power and channel allocation decisions without global network information. Ref. [37] studied the vehicle spectrum sharing problem based on multi-agent reinforcement learning to solve the spectrum and power distribution problem in the scene where the channel conditions in vehicle-mounted network change rapidly and CSI cannot be accurately obtained. Ref. [38] proposed a communication resource allocation method based on deep reinforcement learning to ensure the reliability and delay constraint of ultra-high reliability and low delay communication service on the internet of vehicles. However, the above research focuses on communication resource allocation, not radar. The reinforcement learning they adopt is mainly Q-learning and DQN. Given the low complexity of Q-learning and whether the convergence rate of DQN can be further optimized, this paper proposes an improved DQN model: Dueling DQN.

In this paper, with the limited ability of a single UAV, group UAVs has attracted more attention in communication and radar fields. Under the condition of limited group UAVs resources, the resource allocation problem is addressed for the group UAVs to achieve a trade-off between the detection and communication performance, where ISAC is equipped in group UAVs. Due to the group UAVs, the resource allocation problem is complex and cannot be solved efficiently. A novel reinforcement-learning-based method is proposed, including the Q-Learning, state action reward state action (SARSA), deep Q network (DQN) and dueling deep Q network (Dueling DQN) reinforcement learning algorithms. We formulate a new reward function by combining both the mutual information (MI) and the communication rate (CR), where the MI is used to describe the performance of the radar detection and the CR is for that of the wireless communication. The contributions of this paper are given as follows:

- A novel low-cost, weight-light, and integration-high degree group UAV system is proposed. The ISAC system is mounted on UAVs, and the radar resources are used for communication and target detection together. It not only broadens the use of

UAV resources but also improves the anti-jamming ability of UAVs from the common frequency ban.

- A novel reinforcement-learning-based method is proposed to solve the complex problem, where we formulate a new reward function by combining both the MI and the CR. The MI describes the radar detection performance, and the CR is for wireless communication.

The rest of this paper is organized as follows: In Section 2, the system model is formulated. In Section 3, the design of the algorithms is introduced. In Section 4, the analysis and discussion of simulation results of different algorithms in different group UAVs under different environments are given. The conclusions are presented in Section 5.

Notations: lower-case boldface letters denote vectors. $\xi\{\cdot\}$ is the nomination operation.

2. Groups UAVs Resource Allocation Model for ISAC System

Consider the ISAC system for group UAVs as shown in Figure 1, where group UAVs in number K are sharing information among each other and detecting targets. Three kinds of resources such as the beam in number X, the power in number Y and the channel in number Z are allocated among group UAVs to achieve a trade-off between the wireless communication and the target detection. According to these resources, due to its low cost, light weight and high integration characteristics, the ISAC system is employed in each UAV. In the case of limited resources, how to allocate resources usage reasonably can optimize the balancing of the two tasks.

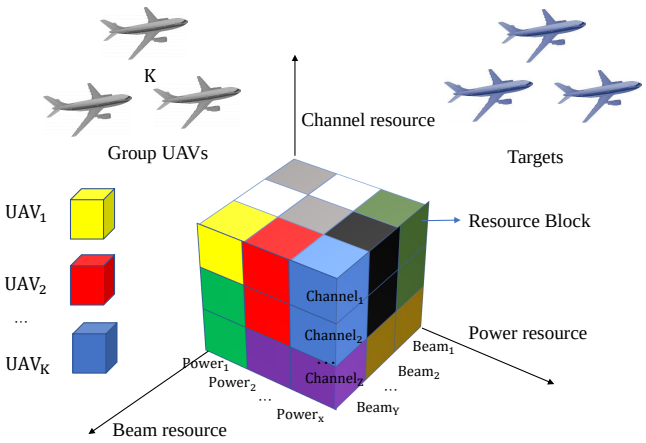

Figure 1. Resource allocation diagram for the ISAC system with group UAVs.

The problem model can be described to find the maximum of the reward of the ISAC system for group UAVs $R(\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{p})$ by allocating resources usage reasonably:

$$\max_{\boldsymbol{b},\boldsymbol{c},\boldsymbol{p}} \quad R(\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{p}), \tag{1}$$

where $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{p}$ represent the set of beam values, channel values and power values, respectively, for the corresponding n-th task and k-th UAV. The definitions of $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{p}$ are given as:

$$\boldsymbol{b} := \{b_{n,k} | n = 1, 2; k = 1, 2, \cdots, K\}, \tag{2}$$

$$\boldsymbol{c} := \{c_{n,k} | n = 1, 2; k = 1, 2, \cdots, K\}, \tag{3}$$

$$\boldsymbol{p} := \{p_{n,k} | n = 1, 2; k = 1, 2, \cdots, K\}, \tag{4}$$

where we use $n = 1$ and $n = 2$ to represent the information communication task and the radar detection task for the individuals of group UAVs, respectively. $b_{n,k}$, $c_{n,k}$ and $p_{n,k}$ represent the beam, the channel and the power of the k-th UAV performing the n-th task, respectively.

The domain values of $\boldsymbol{b}$,$\boldsymbol{c}$ and $\boldsymbol{p}$ are defined as follows:

$$b_{1,k} \neq b_{2,k} \tag{5}$$

which shows that the k-th UAV perform different tasks at different beam.

$$\begin{aligned} &\sum_{n}\sum_{k} c_{n,k} \leq \delta_1 \\ &\sum_{n} c_{n,k} \leq \delta_2 \\ &c_{n,k} \neq c_{n',k'} \end{aligned} \tag{6}$$

δ_1 and δ_2 represents the all bandwidth of UAVs and the k-th UAV, respectively. $c_{n,k} \neq c_{n',k'}$ shows that the k-th UAV perform different tasks at different channels.

$$\begin{aligned} &\sum_{n}\sum_{k} p_{n,k} \leq \delta_3 \\ &\sum_{n} p_{n,k} \leq \delta_4 \end{aligned} \tag{7}$$

δ_3 and $delta_4$ represent the power of all UAVs and the k-th UAV, respectively.

The definition of $R(\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{p})$ is given as:

$$\begin{aligned} &R(\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{p}) \\ &:= \sum_{n=1,k} R_{1,k}(b_{1,k}, c_{1,k}, p_{1,k}) + \sum_{n=2,k} R_{2,k}(b_{2,k}, c_{2,k}, p_{2,k}), \end{aligned} \tag{8}$$

where $R_{1,k}(b_{1,k}, c_{1,k}, p_{1,k})$ represents the reward of the ISAC system for the k-th UAV which performs the 1-th task. It can be represented by the reward of the communication system [18]. The communication reward is the communication rate. $R_{2,k}(b_{2,k}, c_{2,k}, p_{2,k})$ represents the reward of the radar communication integrated system for the k-th UAV which performs the 2nd task. It can be represented by the reward of the radar system. The radar reward is mutual information. Specific equations about $R_{1,k}(b_{1,k}, c_{1,k}, p_{1,k})$ and $R_{2,k}(b_{2,k}, c_{2,k}, p_{2,k})$ will be seen as follows.

The communication rate is taken as the reward of communication performance. The communication rate can represent the performance of the communication link and is defined as:

$$\begin{aligned} &R_{1,k}(b_{1,k}, c_{1,k}, p_{1,k}) \\ &= (1-\lambda)\xi\left\{ c_{1,k} \log_2\left(1 + \frac{\zeta_k^2 p_{1,k}}{\gamma_k + \kappa \bar{T} c_{1,k}}\right)\right\}, \end{aligned} \tag{9}$$

where λ is the parameter to achieve a trade-off between the communication and detection performance. $\xi\{\cdot\}$ is the nomination operation. ζ_k represents the channel loss of the k-th UAV of group UAVs. γ_k represents the communication interference caused to the individual of group UAVs when group UAVs share information. κ is the Boltzmann constant, and $\bar{T}$ represents the system noise temperature.

On the basis of information theory, target detection can be regarded as a non-cooperative communication problem, which means that the detected target is reluctant to transmit information to the radar. Mutual information is proposed to measure the ability of radar to acquire target information and is defined as:

$$\begin{aligned} &R_{2,k}(b_{2,k}, c_{2,k}, p_{2,k}) \\ &= \lambda\xi\left\{ T_k \int_{c_{2,k}} \ln\left(1 + \frac{2\left|\sqrt{p_{2,k}} s_k(f)\right|^2 \sigma^2(f)}{(w_k(f) + v_k) T_k}\right) \mathrm{d}f \right\}. \end{aligned} \tag{10}$$

where T_k is the pulse duration, $s_k(f)$ denotes the normalized baseband signal in the frequency domain, $\sigma(f)$ denotes the frequency response of the target corresponding to transmitted signal, $w_k(f)$ is the noise power in the frequency domain and v_k is the interference caused by other UAVs.

3. Reinforcement-Learning-Based UAVs Resource Allocation Method

To optimize the resources among UAVs for the ISAC system, a reinforcement learning-based method is proposed. As shown in Figure 2, at each time t, the group UAVs, as the agents, observe the state $\mathcal{E}^{[t]}$ from the state space S and take an action $\mathcal{A}^{[t]}$, selecting the beam, channel and transmission power from the action space $a_m^{[t]}$ according to the current environment $\mathcal{E}^{[t]}$. Based on the action taken by the agent, the environment transforms to a new state, $\mathcal{E}^{[t+1]}$, and the agent receives a reward, $r^{[t]}$, from the environment. The overall algorithm design is mainly composed of an environment state, anagent action, a reward given by the environment and the corresponding reinforcement learning algorithm model design.

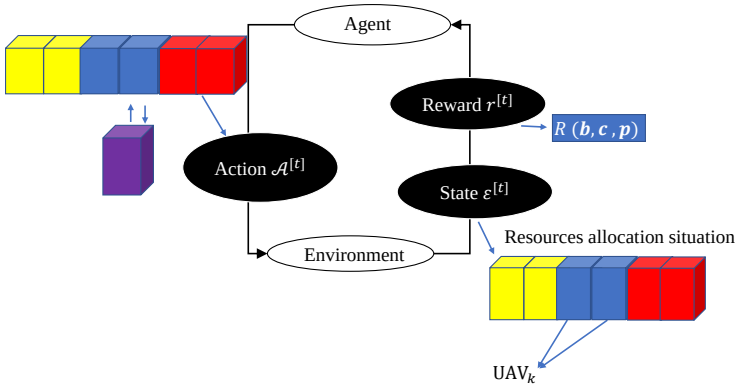

Figure 2. The flow chart of algorithms.

Due to the discrete characteristics, a resource allocation problem can be described as the interaction between resource allocation and the environment, and the corresponding resource allocation scheme can be described by the reward function. It is consistent with reinforcement learning. Therefore, in order to solve problem (1), we propose a resource allocation method based on reinforcement learning. As shown in Figure 2, according to each group UAV's perceived business request and available resource, the group UAVs take the corresponding resource allocation. The reward value can be obtained according to the strategy, so as to describe the performance of resource allocation. Different from traditional resource allocation, reinforcement learning methods can effectively solve the problem of high complexity and unrealization of UAVs' resource allocation. Each section is described in detail below:

1. Environment State
   Since the state is the mapping and representation of the environment and also the basis for agents to take agents, the design of the environment state is very meaningful. During the t-th state, it is defined as the current resource allocation strategy and the status of the UAVs, such as their locations $\boldsymbol{u}^{[t]}$ and velocity $\boldsymbol{v}^{[t]}$, the locations of targets $\boldsymbol{w}^{[t]}$, available energy $\boldsymbol{b}_r^{[t]}, \boldsymbol{c}_r^{[t]}, \boldsymbol{p}_r^{[t]}$ and last time reward $l^{[t]}$:

$$\mathcal{E}^{[t]} \triangleq \left\{ \boldsymbol{b}^{[t]}, \boldsymbol{c}^{[t]}, \boldsymbol{p}^{[t]}, \boldsymbol{u}^{[t]}, \boldsymbol{w}^{[t]}, \boldsymbol{v}^{[t]}, \boldsymbol{b}_r^{[t]}, \boldsymbol{c}_r^{[t]}, \boldsymbol{p}_r^{[t]}, l^{[t]}, \right\} \in \mathbb{R}^{9K+W+1}, \tag{11}$$

   where $\mathbb{R}$ denotes the real number set. W denotes the number of targets.

2. Agent Actions
Actions are the outputs of an agent and the inputs to the environment. Group UAVs allocates resources reasonably according to group UAV task requests and the available resource status of the system. Therefore, action $\mathcal{A}^{[t]}$ can be defined as below:

$$\begin{aligned} \mathcal{A}^{[t]} &= \left\{ a | a \in \mathcal{M}^{[t]} \right\} \\ \mathcal{M}^{[t]} &\triangleq \left\{ a_m^{[t]} : m = 1, 2, 3, \ldots, M^{[t]} \right\}. \end{aligned} \tag{12}$$

where all the possible resource allocation strategies during the t-th state from an action set is $\mathcal{M}^{[t]}$, the number of actions is $M^{[t]}$ and the m-th action is defined as: $a_m^{[t]} = \{\boldsymbol{b}_m^{[t]}, \boldsymbol{c}_m^{[t]}, \boldsymbol{p}_m^{[t]}\}$. $\boldsymbol{b}_m^{[t]}$, $\boldsymbol{c}_m^{[t]}$ and $\boldsymbol{p}_m^{[t]}$ denote the channel and power allocation strategy in the m-th action, respectively. Notably, considering the channel interference, once a channel is used, it cannot be selected the next time.

3. Reward
Reward refers to the feedback after the agent taking action according to certain environmental states. The reasonability is closely related to the income that can be obtained by the agent and the performance of the dynamic resource allocation algorithm. In the ISAC system for group UAVs with dynamic resource allocation, it is necessary to give certain rewards to learn the optimal resource allocation strategy. According to the current allocation strategy $\mathcal{E}^{[t]}$, the reward $r^{[t]}$ is defined as below:

$$r^{[t]} = \begin{cases} R(b^{[t]}, c^{[t]}, p^{[t]}), & \mathcal{E}^{[t]} = \mathcal{E}_{\text{terminal}} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

$\mathcal{E}_{\text{terminal}}$ denotes that resources have been allocated for this episode. Set the reward to 0 when the resource allocation has not ended and to $R(b^{[t]}, c^{[t]}, p^{[t]})$ when the resource allocation has ended. Finally, we try to maximize the reward $r^{[t]}$.

According to the above definition of environment state, action and reward, four reinforcement learning algorithms including Q-Learning, SARSA, DQN and Dueling DQN are used to realize the RL-based resource allocation strategy. Q-Learning, SARSA, DQN and Dueling DQN are value-based and model-free reinforcement learning algorithms. DQN and Dueling DQN have changes relative to Q-Learning and SARSA, mainly in three aspects [39,40]:

(1) DQN and Dueling-DQN use convolutional neural networks (CNN) to approximate value functions. The reinforcement learning has entered the stage of deep reinforcement learning.
(2) DQN and Dueling-DQN use replay buffer training to strengthen the learning process.
(3) DQN and Dueling-DQN set up the target network independently to deal with the deviation in timing difference separately.

DQN and Dueling-DQN can break out the fact that state space and action space are discrete and cannot be too large from Q-Learning and SARSA.

The differences of Q-Learning and SARSA are as follows. The purpose of Q-Learning is to learn the value of a specific action in a specific state. Create a Q-table with state rows and action columns, and update the Q-Table with rewards for each action. Q-Learning is off-policy. A different strategy means that the action strategy and the evaluation strategy are not the same strategy. In Q-Learning, the action strategy is ε-greedy, and the strategy to update the Q-table is greedy.

SARSA stands for state-action-reward-state-action. The Q-table is also used to store the action value function. Moreover, the decision-making part is the same as Q-Learning, which also adopts the ε-greedy strategy. The differences are as follows:

1. SARSA is an update of on-policy, and its action strategy and evaluation strategy are ε-greedy.

2. SARSA takes action first and updates later:

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma Q(s^{'},a^{'}) - Q(s,a)] \tag{14}$$

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a^{'}} Q(s^{'},a^{'}) - Q(s,a)] \tag{15}$$

As can be seen above, $Q(s,a)$ and $Q(s^{'},a^{'})$ denote the Q-value at the current moment and the next moment, respectively. α indicates the learning rate. Formula (14) is the updated formula of the SARSA Q-value, which performs the action using the ε-greedy strategy, then updates the value function based on the action being performed. The Formula (15) is the updated formula of the Q-Learning Q-value, which assumes the next step to select the maximum reward action and update the value function. Then, the action is selected using the ε-greedy strategy.

To compare DQN and Dueling-DQN, the algorithm flow is shown the Figure 3. Firstly, based on the greedy strategy, the main network generates action $a^{[t]}$ according to the corresponding environment $\mathcal{E}^{[t]}$. Then, with the action, the reward $R(\boldsymbol{b}^{[t]}, \boldsymbol{c}^{[t]}, \boldsymbol{p}^{[t]})$ can be obtained, and the environment is updated as $\mathcal{E}^{[t+1]}$.

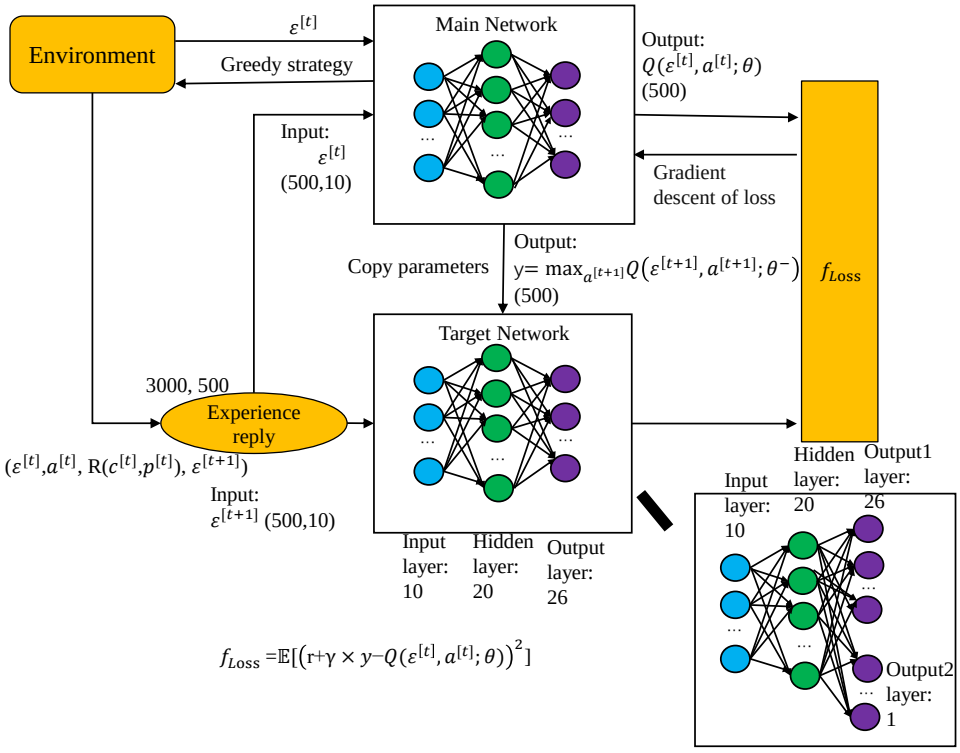

Figure 3. The process structure of DQN and dueling-DQN.

The trajectory $(\mathcal{E}^{[t]}, a^{[t]}, R(\boldsymbol{b}^{[t]}, \boldsymbol{c}^{[t]}, \boldsymbol{p}^{[t]}), \mathcal{E}^{[t+1]})$ is stored in the experience replay with the maximum size being 3000. The network begins to learn when the quantity of storage reaches 500. Thus, 500 experiences are used as the input of the main network, with the input layer being 10, the hidden layer being 20 and the output layer being 26. The target network has the same network structure as the main network.

To train the network, the loss function is defined as:

$$f_{\text{Loss}} = \mathbb{E}\Big\{ R(\boldsymbol{c}^{[t]}, \boldsymbol{c}^{[t]}, \boldsymbol{p}^{[t]}) + \gamma \times y - Q(\mathcal{E}^{[t]}, a^{[t]}; \theta) \Big\}, \tag{16}$$

$$y = \max_{a^{[t+1]}} Q(\mathcal{E}^{[t+1]}, a^{[t+1]}; \theta^{-}). \tag{17}$$

The main network generates the Q-value as $Q(\mathcal{E}^{[t]}, a^{[t]}; \theta)$ during the state $\mathcal{E}^{[t]}$ with performing action $a^{[t]}$. Additionally, the target network generates the max Q-value as y with performing a different action $a^{[t+1]}$.

4. Simulation Results

The simulation results of the ISAC system for group UAVs under different resource methods are carried out in a personal computer with 16 GByte DDR4 Intel Core i7-8750H, 6 GByte Nvidia GeForce GTX 1060 with Max-Q Design and stated in this chapter. Specific simulation parameters are shown in Table 1. With a direct component in the channel, the received signal is the superposition of complex Gaussian signals and direct components, so the channel fading model adopted in this paper is the rice channel fading model. As given in Table 1, we use 5 UAVs to form a group and 24 channels, 3 beams and 6 power grades to develop resource block resources allocation.

Table 1. Simulation parameters.

Parameter	Value
Number of UAVs	5
Total number of available channels	24
Total number of beams	3
Total number of power grades	6
Communication transmitting power	100 W
Radar transmitting power	100 KW
System noise temperature	290 K
Boltzmann constant	1.38×10^{-23}
The wavelength	0.1 m

First, we show the system performance with different reinforcement learning algorithms, where the system performance is measured by the reward defined as $R(\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{p})$, which is a combination of MI and CR and has been used in many papers. Specific simulation parameters are shown in Table 2 for Q-Learning, SARSA, DQN and Dueling DQN used in resource allocation. The DQN and Dueling DQN algorithms belong to deep reinforcement learning algorithms. They contain a neural network framework structure, including a neural network input layer, hidden layer and output layer. Linear stands for Linear functions at the input and output levels. Relu is the activation function that breaks linearity. The size of the input layer is the dimension of state S, and the dimension of the hidden layer is 20. The output layer of the former is the corresponding dimension of behavior A, while the output layer of the latter mainly corresponds to dimensions one and A.

Table 2. Specific parameters of reinforcement learning algorithms.

Parameter	Q-Learning	SARSA	DQN	Dueling DQN
Learning rate	0.01	0.1	0.001	0.1
$\varepsilon_{\text{init}}$	0.7	0.7	0.7	0.7
ε_{gap}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
ε_{end}	0.999	0.9	0.9	0.999
γ	0.9	0.9	0.9	0.9
Input layer	-	-	Linear, $\|S\|$	Linear, $\|S\|$
Hidden layer	-	-	ReLU, 20	ReLU, 20
Output layer 1	-	-	Linear, $\|A\|$	state function layer:Linear, $\|A\|$
Output layer 2	-	-	-	advantage function layer:Linear, 1

In this paper, the convergence performance of each reinforcement learning algorithm is verified by simulation. It is shown in Figure 4 that group UAVs' resource allocation is

simulated in a joint PyCharm and Matlab platform, where λ is 0.1, which means the gravity of target detection task, and the gravity of the information sharing task is 0.9. At this point, the normalized value of CR is 1000 Mbit/s, and the normalized value of MI is 5000 bit/s. The CR is the specific gravity times communication the normalized value times the total reward. MI is the specific gravity times the target detection normalized value times the total reward.

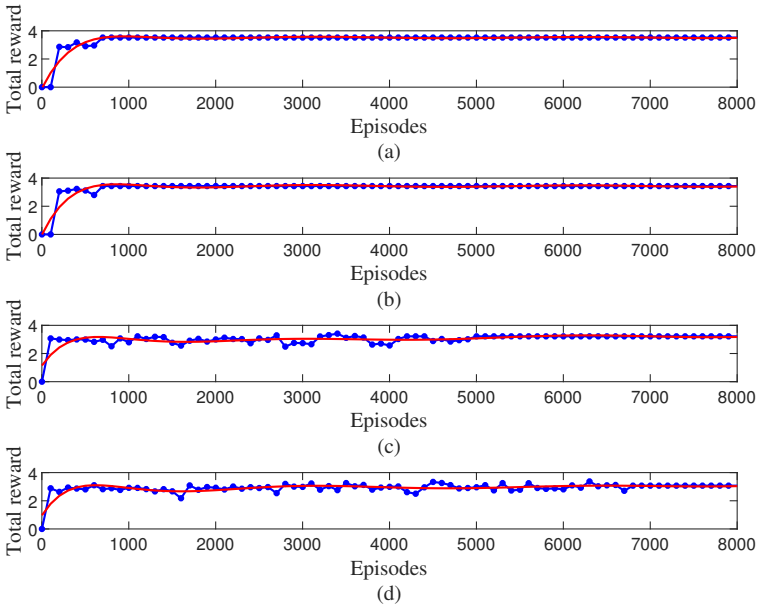

Figure 4. Comparison of convergence under different kinds of algorithms: (**a**) Dueling DQN; (**b**) DQN; (**c**) Q-Learning; (**d**) SARSA.

As can be seen from Figure 4, in the process of group UAVs learning, the four reinforcement learning algorithms have a relatively small number of learning iterations and low total reward values in the early stage, which is because group UAVs have little awareness of the environment in the early stage of learning and are in the exploratory stage. With the increase in environmental awareness, group UAVs gradually learn the optimal strategy, which makes the total reward tend to be stable. As shown in the data fitting curves, Q-Learning and SARSA converged at round 5000 and 7000, respectively, while DQN and Dueling DQN converged at round 500. Compared with the Q-Learning and SARSA algorithms, DQN and Dueling-DQN are modified in three aspects based on Q-Learning: using the DL approximation function, using experience reply to train the learning process of reinforcement learning and independently establishing target networks to deal with TD deviation in the time difference algorithm. This dramatically solves the problem of too much moving space and breaks the correlation between experiences. The number of episodes has also been significantly reduced. Additionally, due to too much moving, the system performance of DQN and Dueling DQN is more excellent than Q-Learning and SARSA algorithms.

Figure 5 shows comparison of loss under DQN and Dueling DQN with the same learning rate. With the increase in episodes, the loss of Dueling DQN tended to be 0, while that of DQN tended to be 1.5. Obviously, Dueling DQN had lower loss than DQN. Additionally, the curve of Dueling DQN fluctuates wildly while that of DQN tended to be flat. Dueling DQN had better performance on convergence than DQN. As shown in Figure 6, under the condition of the same learning rate, Dueling DQN requires fewer episodes than DQN, although the complexity of the problem they solved is increasing. Therefore, the overall performance of Dueling DQN is better than DQN.

Second, we show the system performance under different methods. The proposed method is compared with two benchmark methods:

- KM method [41]: The traditional Kuhn Munkres (KM) iterative optimization method is used for resource allocation by iterating over each resource to optimize allocation.
- DNN method [42]: The Deep Neural Network(DNN) method based on supervised learning is used for resource allocation by fitting initial data sets.

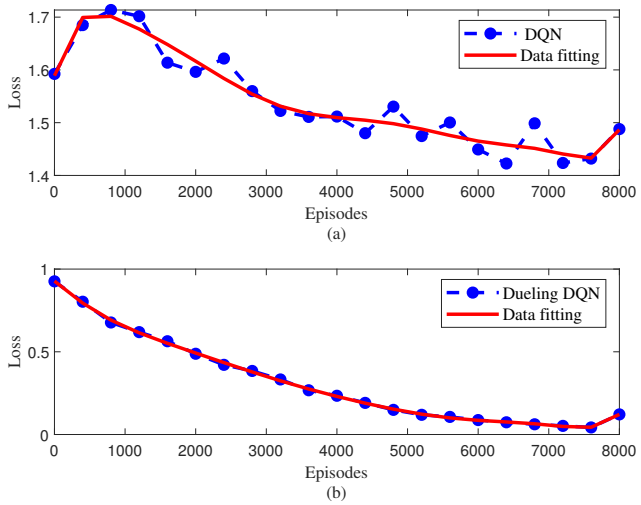

Figure 5. Comparison of loss under different kinds of algorithms: (**a**) DQN; (**b**)Dueling DQN.

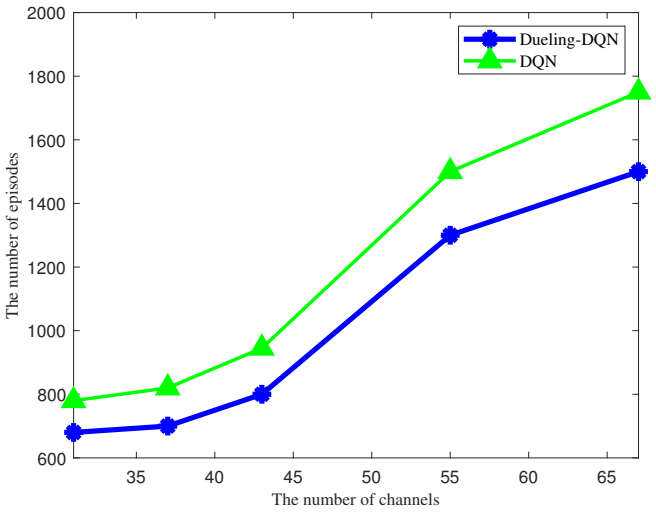

Figure 6. Comparison of episodes of algorithms.

As shown in Figures 7–11, there are four resource allocation methods: DQN, Dueling DQN, KM and the DNN allocation algorithm. The four figures show the system performance of the four methods in different environments. As shown in Figure 7, a reasonable distribution of resources is carried out under different channel numbers. Channel values range from 21 to 33. As the number of channels increases, the total reward performance curves of three algorithms also gradually improve, which is caused by the gradual increase in the system channel resources. It can be further seen from the figure that under the same number of channels, when the number of channels is small, the performance of Dueling DQN and DQN approaches the KM algorithm and is better than the DNN algorithm. However, with the increase in the channel number, the performance of Dueling DQN and DQN is obviously better than that of the KM algorithm and the DNN algorithm. This is mainly because the KM algorithm is based on multiple resource iteration. Compared with the DNN algorithm, it mainly depends on its fitting data set.

As shown in Figure 8, the proper allocation of resources is carried out under other power grades numbers. Power grades range from 6 to 18. As the number of power levels increases, the total reward performance curves of three algorithms gradually increase because it is highly likely to approach the optimal power as the number of power levels increases. It can be further seen from the figure that under the same number of power levels, the KM algorithm will perform better than the Dueling DQN and DQN algorithms when the number of levels is small. However, the performance of the KM algorithm grows slowly with the increase in the number of power levels, while the performance of the Dueling DQN and DQN algorithms is better. The DNN algorithm has inferior performance to the Dueling DQN and DQN algorithms from beginning to end.

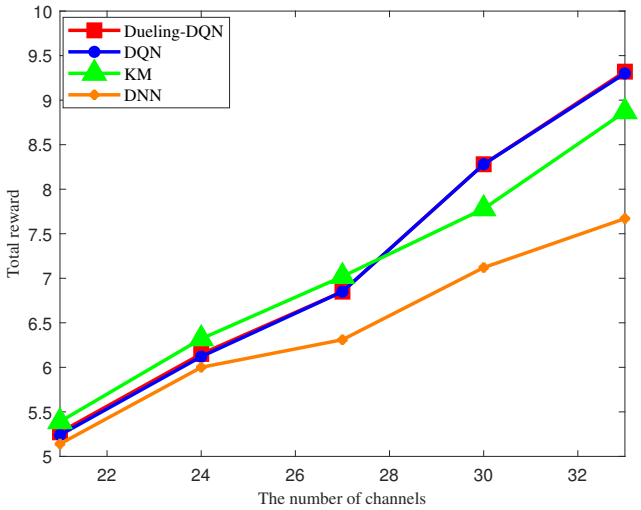

Figure 7. Comparison of the return values of different methods under different channels number.

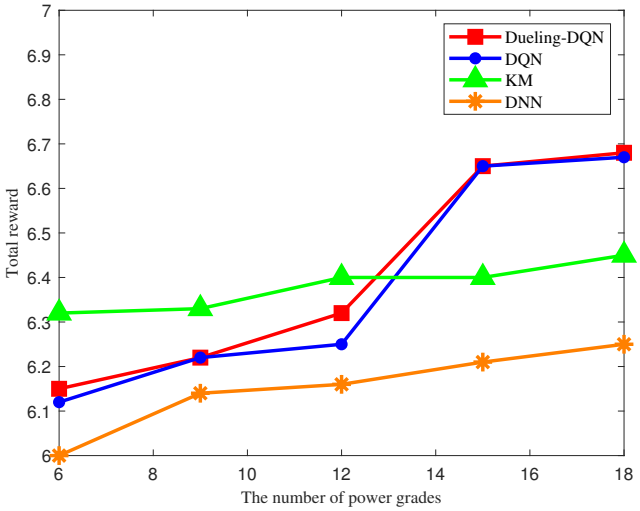

Figure 8. Comparison of the return values of different methods under different power grades.

As shown in Figure 9, proper allocation of resources is carried out under different beam number. Beam values range from 3 to 7. With the increase of beam number, the total reward performance curves of four algorithms gradually improve. This is mainly due to the increase of beam resources. It can be further seen from the figure that the performance of Dueling DQN algorithm and DQN algorithm are significantly better than that of KM algorithm and DNN algorithm with the increase of beam number.

As shown in Figure 10, proper allocation of resources is carried out under other UAVs numbers. UAVs numbers range from 5 to 25. As the number of users increases, so does the

number of channels. The total reward performance curves of four algorithms gradually increase. However, it can be clearly seen from the figure that when the number of users and channels is small, the performance of KM algorithm is slightly higher than that of Dueling DQN algorithm and DQN algorithm and the performance of DNN algorithm is the lowest. However, the performance of Dueling DQN algorithm and DQN algorithm show a rapid growth with the increase of the number of users while the KM algorithm and DNN algorithm grow slowly. In addition, Dueling DQN algorithm and DQN algorithm have the best performance.

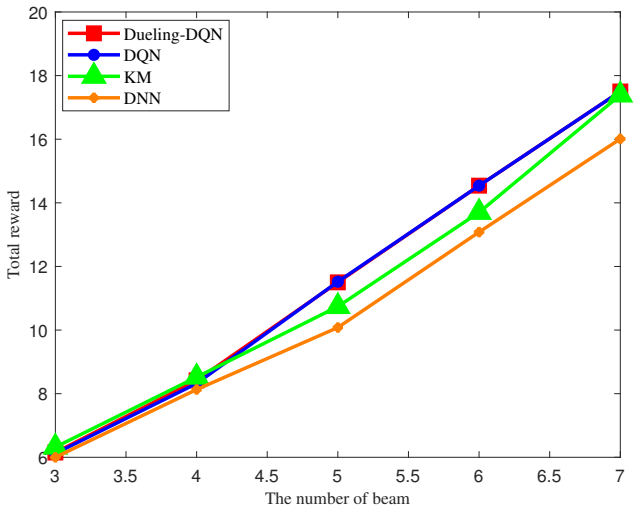

Figure 9. Comparison of the return values of different methods under different beam numbers.

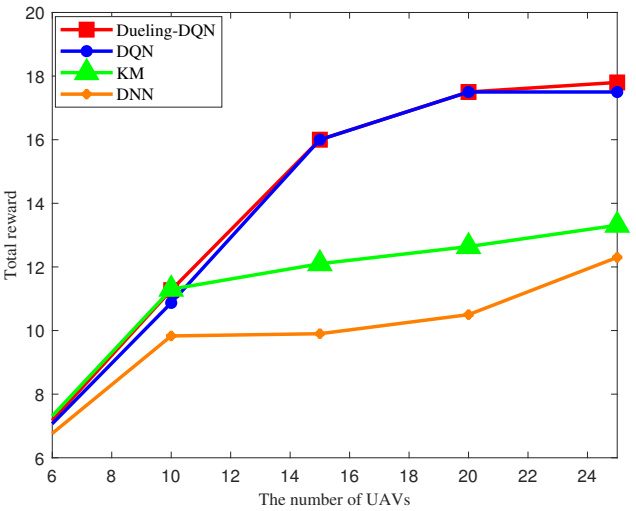

Figure 10. Comparison of the return values of different methods under different UAV numbers.

As shown in Figure 11, a reasonable distribution of resources is carried out under different λ. λ ranges from 0.1 to 0.9. As can be seen from the figure, it is changing lambda to compare the performance of the four algorithms under the condition that the resource situation remains unchanged. The DNN algorithm has the worst performance. the Dueling DQN algorithm and the DQN algorithm have the best performance. The KM algorithm is average. It can be seen from the figures that the DQN and Dueling DQN resource allocation algorithms are superior to the KM and DNN allocation algorithms with the increase in environment complexity.

To show the computational complexity clearly, the computational time is given in Figure 12. The computational time of the proposed method is shorter than KM and close to DNN with the increase in environment complexity. When the environment becomes more

and more complex, the time consumption of the KM algorithm increases obviously while those of the Dueling DQN, DQN and DNN algorithms increase slowly.

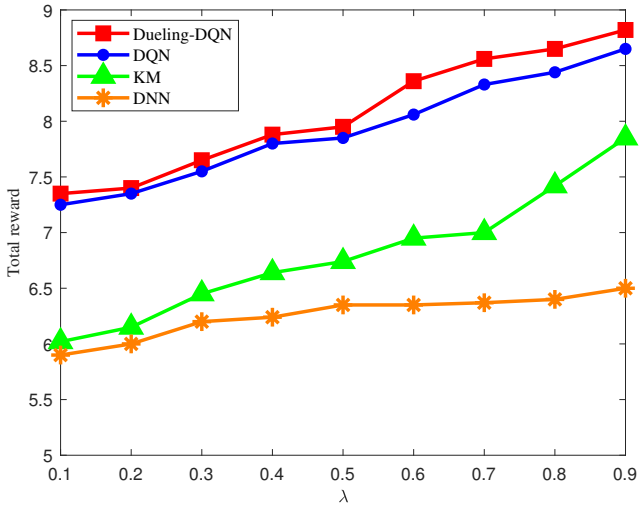

Figure 11. Comparison of the return values of different methods under different λ.

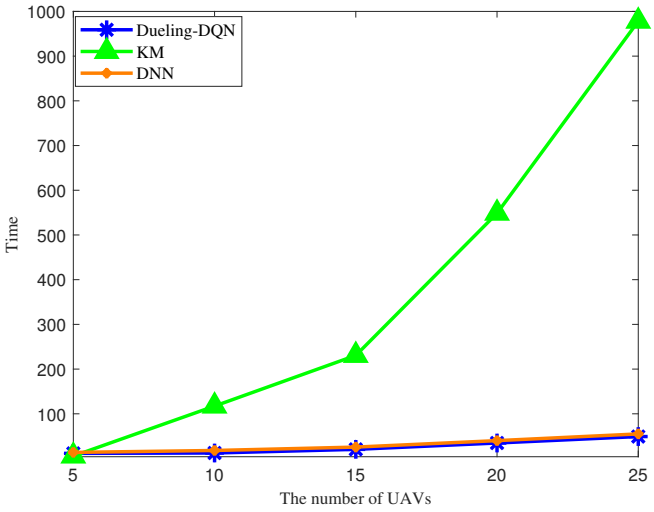

Figure 12. The computational time of the different methods.

5. Summary and Prospect

For the flexible mobilization of the ISAC system, the traditional fixed resource allocation can no longer satisfy the effective allocation of resources according to the real-time situation, resulting in the low utilization of resources. In order to solve this problem, this paper has firstly summarized and analyzed the resource allocation technology of the ISAC system and then introduced the related resource allocation technology. Secondly, with the considering of complexity for the resource allocation problem, reinforcement-learning-based methods including Q-Learning, SARSA, DQN and Dueling DQN have been proposed under a novel reward combining both the MI and the CR, and the reasons for its combination have been analyzed. Finally, the allocation of resources under different environments has been introduced. Simulation results show that compared with the KM method, the resource allocation method based on reinforcement learning has better performance and lower time complexity with the increase in environment complexity. Compared with the DNN method, this method does not require prior data set preparation. Additionally, when the time complexity of both methods is almost the same, the system performance of this method is better. The resource allocation problem for the ISAC system is studied based on the reinforcement learning method and the simulation results are used

to show the advantages of the proposed method. In future work, we will realize a hardware system to test the proposed method.

Author Contributions: Conceptualization, M.W. and P.C.; methodology, M.W. and P.C.; software, M.W.; validation, M.W.; formal analysis, M.W.; investigation, M.W. and P.C.; resources, M.W. and P.C.; writing—original draft preparation, M.W.; writing—review and editing, P.C. and Y.C.; visualization, M.W. and P.C.; supervision, P.C. and Z.C.; project administration, P.C. and Z.C.; funding acquisition, P.C. and Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (Grant No. 61801112), the Equipment Pre-Research Field Foundation, the Industry-University-Research Cooperation Foundation of The Eighth Research Institute of China Aerospace Science and Technology Corporation (Grant No. SAST2021-039), the National Key R&D Program of China (Grant No. 2019YFE0120700) and the Natural Science Foundation of Jiangsu Province (Grant No. BK20180357).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xiong, Z.; Zhang, Y.; Lim, W.; Kang, J.; Niyato, D.; Leung, C.; Miao, C. UAV-assisted wireless energy and data transfer with deep reinforcement learning. *IEEE Trans. Cogn. Commun. Netw.* **2021**, *7*, 85–99. [CrossRef]
2. Peer, M.; Bohara, V.; Srivastava, A. Multi-UAV placement strategy for disaster-resilient communication network. In Proceedings of the 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), Victoria, BC, Canada, 18 November–16 December 2020; pp. 1–7.
3. Li, K.; Wang, C.; Lei, M.; Zhao, M.M.; Zhao, M.J. A local reaction anti-jamming scheme for UAV swarms. In Proceedings of the 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), Victoria, BC, Canada, 18 November–16 December 2020; pp. 1–6.
4. Altan, A. Performance of metaheuristic optimization algorithms based on swarm intelligence in attitude and altitude control of unmanned aerial vehicle for path following. In Proceedings of the 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 22–24 October 2020; pp. 1–6.
5. Chen, R.; Yang, B.; Zhang, W. Distributed and collaborative localization for swarming UAVs. *IEEE Internet Things J.* **2021**, *8*, 5062–5074. [CrossRef]
6. Shen, J.; Wang, S.; Zhai, Y.; Zhan, X. Cooperative relative navigation for multi-UAV systems by exploiting GNSS and peer-to-peer ranging measurements. *IET Radar Sonar Navig.* **2021**, *15*, 21–36. [CrossRef]
7. Zhang, R.; Ishikawa, A.; Wang, W.; Striner, B.; Tonguz, O.K. Using reinforcement learning with partial vehicle detection for intelligent traffic signal control. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 404–415. [CrossRef]
8. Zhang, S.; Zhou, Y.; Zhang, L.; Zhang, Q.; Du, L. Target detection for multistatic radar in the presence of deception jamming. *IEEE Sens. J.* **2021**, *21*, 8130–8141. [CrossRef]
9. Thornton, C.; Kozy, M.; Buehrer, R.; Martone, A.F.; Sherbondy, K.D. Deep reinforcement learning control for radar detection and tracking in congested spectral environments. *IEEE Trans. Cognit. Commun. Netw.* **2020**, *6*, 1335–1349. [CrossRef]
10. Krishnamurthy, V.; Angley, D.; Evans, R.; Moran, B. Identifying cognitive radars—Inverse reinforcement learning using revealed preferences. *IEEE Trans. Signal Process.* **2020**, *68*, 4529–4542. [CrossRef]
11. Yuan, T.; Neto, W.; Rothenberg, C.; Obraczka, K.; Barakat, C.; Turletti, T. Dynamic controller assignment in software defined internet of vehicles through multi-agent deep reinforcement learning. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 585–596. [CrossRef]
12. Aznar, J.; Garcia, A.; Egea, E.; Garcia-Haro, J. MDPRP: A q-learning approach for the joint control of beaconing rate and transmission power in VANETs. *IEEE Access* **2021**, *9*, 10166–10178. [CrossRef]
13. Jang, J.; Yang, H. Deep reinforcement learning-based resource allocation and power control in small cells with limited information exchange. *IEEE Trans. Veh. Technol.* **2020**, *69*, 13768–13783. [CrossRef]
14. James, S.; Raheb, R.; Hudak, A. Impact of packet loss to the motion of autonomous UAV swarms. In Proceedings of the 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 11–15 October 2020; pp. 1–9.
15. Majidi, M.; Erfanian, A.; Khaloozadeh, H. Prediction-discrepancy based on innovative particle filter for estimating UAV true position in the presence of the GPS spoofing attacks. *IET Radar Sonar Navig.* **2020**, *14*, 887–897. [CrossRef]
16. He, C.; Yu, B.; Yi, Q. A cooperative positioning method for micro UAVs in challenge environment. In Proceedings of the 2020 3rd International Conference on Unmanned Systems (ICUS), Harbin, China, 27–28 November 2020; pp. 1157–1160.
17. Chen, H.; Xian, W.; Liu, J.; Wang, J.; Ye, W. Collaborative multiple UAVs navigation with GPS/INS/UWB jammers using sigma point belief propagation. *IEEE Access* **2020**, *8*, 193695–193707. [CrossRef]
18. Meng, F.; Chen, P.; Wu, L.; Cheng, J. Power allocation in multi-user cellular networks: Deep reinforcement learning approaches. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 6255–6267. [CrossRef]
19. Hussien, Z.; Sadi, Y. Flexible radio resource allocation for machine type communications in 5G cellular networks. In Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4.

20. Maurício, W.; Araújo, D.; Maciel, T.; Lima, F.R.M. A framework for radio resource allocation and SDMA grouping in massive MIMO systems. *IEEE Access* **2021**, *9*, 61680–61696. [CrossRef]
21. Chen, X.; Chen, B.; Guan, J.; Huang, Y.; He, Y. Space-range-doppler focus-based low-observable moving target detection using frequency diverse array MIMO radar. *IEEE Access* **2018**, *6*, 43892–43904. [CrossRef]
22. Mou, X.; Chen, X.; Guan, J.; Chen, B.; Dong, Y. Marine target detection based on improved faster R-CNN for navigation radar PPI images. In Proceedings of the 2019 International Conference on Control, Automation and Information Sciences (ICCAIS), Chengdu, China, 23–26 October 2019; pp. 1–5.
23. Bae, Y.; Shin, J.; Lee, S.; Kim, H. Field experiment of photonic radar for low-RCS target detection and high-resolution image acquisition. *IEEE Access* **2021**, *9*, 63559–63566. [CrossRef]
24. Zhu, S.; Li, X.; Yang, R.; Zhu, X. A low probability of intercept OFDM radar communication waveform design method. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 15–17 January 2021; pp. 653–657.
25. Zhang, W.; Zhang, H. The design of integrated waveform based on MSK-LFM signal. In Proceedings of the 2020 15th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 6–9 December 2020; pp. 565–569.
26. Sanson, J.; Tomé, P.; Castanheira, D.; Gameiro, A.; Monteiro, P.P. High-resolution delay-doppler estimation using received communication signals for OFDM radar-communication system. *IEEE Trans. Veh. Technol.* **2020**, *69*, 13112–13123. [CrossRef]
27. Ma, Q.; Lu, J.; Maoxiang, Y. Integrated waveform design for 64QAM-LFM radar communication. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021; pp. 1615–1625.
28. Zhang, Z.; Zhu, G.; Sabahi, M. Poster abstract: Array resource allocation based on KKT optimization for radar and communication integration. In Proceedings of the 2019 18th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Montreal, QC, Canada, 15–18 April 2019; pp. 307–308.
29. Rihan, M.; Huang, L. Optimum co-design of spectrum sharing between MIMO radar and MIMO communication systems: An interference alignment approach. *IEEE Trans. Veh. Technol.* **2018**, *67*, 11667–11680. [CrossRef]
30. Ahmed, A.; Zhang, Y.; Himed, B. Distributed dual-function radar-communication MIMO system with optimized resource allocation. In Proceedings of the 2019 IEEE Radar Conference (RadarConf), Boston, MA, USA, 22–26 April 2019; pp. 1–5.
31. Zhang, X.; Wang, X. Investigation on range sidelobe modulation effect of co-designed radar-communications shared waveforms. In Proceedings of the 2019 International Applied Computational Electromagnetics Society Symposium—China (ACES), Nanjing, China, 8–11 August 2019; pp. 1–2.
32. Shi, C.; Wang, F.; Salous, S.; Zhou, J. Low probability of intercept-based optimal OFDM waveform design strategy for an integrated radar and communications system. *IEEE Access* **2018**, *6*, 57689–57699. [CrossRef]
33. Nijsure, Y.; Chen, Y.; Yuen, C.; Chew, Y.H. Location-aware spectrum and power allocation in joint cognitive communication-radar networks. In Proceedings of the 2011 6th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), Yokohama, Japan, 1–3 June 2011; pp. 171–175.
34. Mishra, K.; Martone, A.; Zaghloul, A. Power allocation games for overlaid radar and communications. In Proceedings of the 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC), New Delhi, India, 9–15 March 2019; pp. 1–4.
35. Wang, F.; Li, H. Joint power allocation for radar and communication co-existence. *IEEE Signal Process Lett.* **2019**, *26*, 1608–1612. [CrossRef]
36. Ye, H.; Li, G.Y.; Juang, B.H.F. Deep reinforcement learning based resource allocation for V2V communications. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3163–3173. [CrossRef]
37. Liang, L.; Ye, H.; Li, G.Y. Spectrum sharing in vehicular networks based on multi-agent reinforcement learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2282–2292. [CrossRef]
38. Yang, H.; Xie, X.; Kadoch, M. Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks. *IEEE Trans. Veh. Technol.* **2019**, *68*, 4157–4169. [CrossRef]
39. Liu, Y.; Zhang, C. Application of Dueling DQN and DECGA for Parameter Estimation in Variogram Models. *IEEE Access* **2020**, *8*, 38112–38122. [CrossRef]
40. Ban, T. An Autonomous Transmission Scheme Using Dueling DQN for D2D Communication Networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 16348–16352. [CrossRef]
41. Li, X.; Zhu, B. Full-duplex D2D user clustering resource allocation scheme based on Kuhn-Munkres algorithm. *Comput. Eng. Des.* **2019**, *40*, 959–963.
42. Bhandari, S.; Kim, H.; Ranjan, N.; Zhao, H.P.; Khan, P. Optimal cache resource allocation based on deep deural detworks for fog radio access networks. *Internet Technol.* **2020**, *21*, 967–975.