IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 18, 2020, accepted April 27, 2020, date of publication May 4, 2020, date of current version May 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992127

A Novel On-Demand Charging Strategy Based on Swarm Reinforcement Learning in WRSNs

ZHEN WEI[1,2,3], MENG LI[1], ZHENCHUN WEI[1,2,3], LEI CHENG[1,2,3], ZENGWEI LYU[1,2,3], AND FEI LIU[1]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China
[2]Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education, Hefei 230601, China
[3]Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei 230601, China

Corresponding authors: Zhenchun Wei (weizc@hfut.edu.cn) and Zengwei Lyu (lvzengwei@mail.hfut.edu.cn)

This work was supported in part by the National Key Research Development Program of China under Grant 2018YFC0604404, in part by the Key Research and Development Project in Anhui Province, China, under Grant 201904a07020030, and in part by the Fundamental Research Funds for the Central Universities, China, under Grant PA2019GDPK0079.

ABSTRACT The charging issue in Wireless Rechargeable Sensor Networks (WRSNs) is a popular research problem. With the help of wireless energy transfer technology, electrical energy can be transfer from Wireless Charging Equipment (WCE) to the sensor nodes, providing a new paradigm to prolong the network lifetime. Existing research usually takes the periodical and deterministic charging approach, but ignore the limited energy of the WCE and the influences of non-deterministic factors such as topological changes and node failures, making them unsuitable for real networks. In this study, we aim to minimize the number of dead sensor nodes while maximizing energy utilization of WCE under the limited energy of the WCE. Furthermore, the Swarm Reinforcement Learning (SRL) method is firstly introduced to achieve the autonomous planning ability of WCE. Moreover, to solve the problem of insufficient search in existing SRL algorithm, we improve the SRL by firefly algorithm. And a novel charging algorithm, named Swarm Reinforcement Learning based on Firefly Algorithm (SRL-FA), is proposed for the on-demand charging architecture. To evaluate the performance of the proposed algorithm, SRL-FA is compared with the existing swarm reinforcement learning algorithms and classic on-demand charging algorithms in two network scenarios. The Extensive simulation shows that the proposed algorithm can achieve promising performance in energy utilization of WCE, charging success rate and other performance metrics.

INDEX TERMS Wireless rechargeable sensor networks, on-demand charging algorithm, swarm reinforcement learning, firefly algorithm.

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are widely used in military, intelligent transportation, human health monitoring and so on [1]–[3]. These application scenarios require WSN to work continuously. However, the network lifetime is restricted by the limited battery capacity of sensor nodes. So the energy problem of sensor node has become a bottleneck in the research of WSNs. To solve this problem, scholars have conducted a lot of research. The existing reports can be divided into three categories, namely energy saving [4], energy harvesting [5] and Wireless Energy Transfer (WET) [6], [7]. The energy saving scheme extends the life of sensor nodes by reducing the energy consumption per unit of time or workload. Whereas the energy of sensor nodes is still limited, this method cannot solve the problem fundamentally. The energy harvesting scheme restores energy through environments (eg., solar energy and wind energy). However, the great influence by the environments and unpredictability in the amount of harvested energy make energy harvesting scheme unreliable. The main idea of WET is to charge the sensor nodes using the magnetic resonant coupling. And WET can provide a stable energy supply by controllable charging power. With the help of promising WET technique, researchers have proposed a new concept of Wireless Rechargeable Sensor Networks (WRSNs) [8], [9]. In WRSNs, the sensor nodes can be charged by the Wireless Charging Equipment (WCE).

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

Hence WCE charging schedule becomes a prominent issue in WRSNs. And different perspectives on charging schedule have been investigated, including path planning, system performance optimizing and so on.

In existing literatures, charging strategies are two-folds: periodic strategies and on-demand strategies. In the former strategies, the WCE usually follows a fixed charging path to charge all the sensor nodes in the networks [10]–[12]. However, due to the interaction with the surrounding environment, the energy consumption rate of the sensor nodes in the networks was demonstrated significantly different [13]. So the sensor nodes have different energy requirements. It is not necessary to charge all the sensor nodes in the networks. Moreover, the energy consumption profiles of the sensor nodes are high uncertainty. Therefore this charging manner is not suitable for the dynamic nature of WRSNs. In contrast to this, the sensor node in the on-demand strategies sends a charging request when its energy below a known threshold value. Upon the reception of a request, WCE inserts it to the charging list, and then charges the sensor nodes according to the charging strategies. So the on-demand strategies are more suitable for WRSNs.

As for on-demand strategies, there are many unknown information, such as the number of charged sensor nodes. So determining the order of charging the sensor nodes is difficult. Most studies take the greedy method. They set a charging priority for the sensor nodes and select the sensor node with the highest priority in each step. Such a local, impromptu decision bears a low overhead (no global request is necessary). Unfortunately, it usually means no global optimality. In this case, if the WCE can learn and adjust the charging path by interacting with the environment, the WCE can charge more efficiently and obtain a better charging path with consideration of global information. Based on this idea, the on-demand charging strategy with autonomous planning for WCE is studied.

Solving independent path planning for robots is an important branch of Reinforcement Learning (RL) application. Moreover, RL has been verified to be effective in solving charging path planning problem in WRSNs. Therefore, RL is considered to solve the problem in this study. Most reports use ordinary RL. However, in ordinary RL, only an agent learns to achieve goal. The agent essentially learns through trial and error, therefore ordinary RL takes much computation time to acquire the solution and causes inadequate search for optimal solution. To solve these problems, RL is improved by swarm methods, called Swarm Reinforcement Learning (SRL). There are multiple agents in SRL algorithm. Moreover, the agents learn through their respective experiences and the information exchanged among them. SRL algorithm has been recognized that it is able to rapidly find the global optimal solution. Therefore, SRL algorithm is introduced into this study.

The performance of SRL algorithm highly depends on the method of exchanging the information of Q-value. The existing methods calculate Q-value directly. Therefore, they are mostly used to solve continuous problems, which have limitations on solving discrete problems. FA [14], [15] is an optimization method inspired from behavior of firefly movement. FA can solve discrete problem and is similar to method used in PSO-Q (a kind of SRL). Therefore, to maintain the advantages and to overcome the disadvantages of existing SRL algorithms, we improve the SRL algorithm by Firefly Algorithm (FA), named SRL-FA.

Most on-demand charging strategies design the charging path according to the greedy method and cannot obtain the global optimal solution. Moreover, they ignore the limited energy of WCE and the charging strategies is not practical. To solve these problems, the WCE in this study can independently design global optimal charging path through interacting with the network. Furthermore, we aim to ensure the stability of the system while maximizing energy utilization of WCE under the limited energy of the WCE. The main contribution of this paper are as follows.

1) The improved Swarm Reinforcement Learning is introduced into the on-demand charging problem. By interacting with the network, the WCE can independently select the charged sensor nodes and the charging path.
2) SRL is improved according to Firefly Algorithm (FA), named SRL-FA. The performance of SRL highly depends on the information exchanging methods, which are mostly used to solve continuous problems and have limitations on discrete problems. Therefore, the information exchanging method is redesigned.
3) Comprehensive simulations are conducted to compare the performance of our SRL-FA with other charging strategies (named First Come First Serve (FCFS), Nearest Job Next with Preemption (NJNP)) and other swarm reinforcement learning algorithms (named BEST-Q, AVG-Q and PSO-Q). Then salient features of SRL-FA are demonstrated in comparison.

The remainder of study is organized as follows. Section II gives a brief overview of charging strategies on WRSNs and reinforcement learning. In Section III and IV, we detail system model, problem statement, as well as learning model. Algorithm descriptions are given in Section V. Evaluations and comparisons are shown in Section VI and we conclude this study in Section VII.

II. RELATED WORK

In this section, works about this study are introduced, including on-demand charging strategies and reinforcement learning. In the reinforcement learning part, the application of reinforcement learning in WRSNs is also introduced.

A. ON-DEMAND CHARGING STRATEGIES

The on-demand charging strategies can be divided into two categories. One focuses on the performance of the networks, and the other improves the performances of both networks and the WCE. In the former research, the First Come First Serve (FCFS) algorithm is proposed [16]. FCFS schedules the incoming charging requests based on their temporal property and can lead to the back-and-forth charger movement in the space. To overcome the drawback of FCFS, He *et al.* propose a charging Strategy based on the Nearest-Job-Next with Preemption (NJNP) discipline [17], which can increase throughput by always selecting the spatially closest requesting sensor nodes as the next charging node, but it ignores sensor nodes in urgent need of charging. To balance the fairness of charging, Kaswan *et al.* [18] consider both temporal and spatial priorities of the sensor nodes. They present a Linear Programming (LP) formulation for the WCE scheduling problem and a charging strategy based on gravitational search algorithm is presented to solve this problem. Zhu *et al.* [19] present a charging strategy that chooses the sensor nodes which make the least number of other request nodes suffer from energy depletion as the charging candidates. Lin *et al.* [20] present a Primary and Passer-by Scheduling (P2S) algorithm for large-scale WRSNs. After choosing the sensor nodes to be charged, they use a local searching algorithm to find surrounding sensor nodes and add them to the charging path. However, the above literatures do not consider the limited energy of WCE, therefore ignoring the charging cost of the WCE.

To optimize the WCE charging performance at the same time, Fu *et al.* [21] consider different network parameters, such as the travel distance of the WCE and the energy received by the sensor nodes. And then they construct a set of nested TSP tours based on the energy consumption of the sensor nodes, and only sensor nodes with low remaining energy are involved in each charging round. With the similar network parameter considerations, Zhao *et al.* [22] propose to jointly optimize the charging scheduling and charging time allocation. Tomar *et al.* [23] propose a fuzzy logic based scheduling scheme to maximize the survival ratio and energy usage efficiency. Unlike the above charging mode, the WCE in [24] can charge the sensor nodes by one-to-more manner. Not fully charging the sensor nodes, Xu *et al.* [25] propose a charging strategy that only supplements the sensor nodes with partial energy, and then these two articles designed the charging path with a priority strategy that can maximize the sum of sensor lifetime and minimize the traveling distance of the WCE. Although the above reports can improve the performances of the WCE and the networks, none of them consider the autonomy of the WCE.

B. REINFORCEMENT LEARNING (RL)

The main idea of RL is to achieve experience through interaction between the agent and the environment [26], [27]. As shown in Fig.1, there are three representations. Firstly, the state represents the decision-making factors under consideration being observed by an agent. Secondly, the action represents an optimal action being selected by the agent, which may change or affect the state and reward. Thirdly, reward represents the gains or losses in network performance for taking an action on a particular state.

It is assumed that every state update of agent is a time step. At any time step t, the agent observes state $x(t)$ and learns the long-term reward of each state-action pair, decides and carries out an appropriate action $a(t)$ on the environment in a trial-and-error manner. And then the agent reaches the next state $x(t+1)$ and receives the reward $r(t)$. Next, the agent updates the Q-value of this state-action pair according to Eq.(1) [28]. Repeats this operation until the agent reaches the final state.

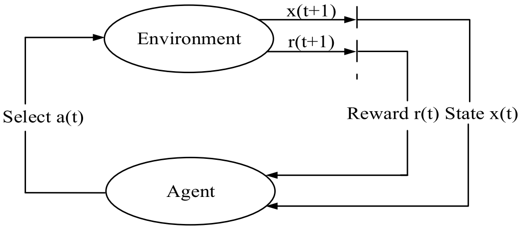


FIGURE 1. **Diagram of reinforcement learning framework.**

$$Q(x(t), a(t)) = \psi r(t) + \psi \left(\gamma \max_{a(t+1)} Q(x(t+1), a(t+1)) - Q(x(t), a(t)) \right) \quad (1)$$

where ψ is the update factor, and γ $(0 < \gamma < 1)$ is the discount factor.

In recent years, RL is widely used in path planning, especially in robot path planning. The robot, treated as the agent, has its own ‘‘brain’’ to plan a path in an environment [29]. To improve the ability of WCE’s autonomous path planning in WRSNs, Wei *et al.* [30] proposes a novel charging strategy called CSRL. CSRL uses Simulated Annealing (SA) to select the action and original RL to obtain the charging path for all the sensor nodes in the networks. However, Original RL uses an agent to learn which may cause a slow convergence speed. To solve the problem of original RL, Iima and Kuroe [31] proposes Swarm Reinforcement Learning (SRL) in which multiple agents are set and they learn through not only their respective experiences but also exchanging Q-value among them. And SRL has been recognized that they are able to rapidly find an optimal solution than original RL. Therefore, we consider using SRL in this study.

To adapt to the changeable network environment as well as improve the autonomous planning capability of the WCE in on-demand charging strategies, The SRL is introduced into this study. And to overcome the drawback that the existing SRL often falls into local optimum, firefly algorithm is used to improve the SRL.

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. NETWORK MODEL

As shown in Fig.2. A WRSN consists of N sensor nodes, a Charging Service Station (CS) and a WCE, which is deployed over a 2-D monitored area. The set of the sensor nodes is denoted as $V_s = \{n_1, n_2 \cdots n_i \cdots n_N\}$. All the positions of the sensor nodes are fixed and the sensor nodes are powered by the same type of battery that the capacity is $E_{\max}$.

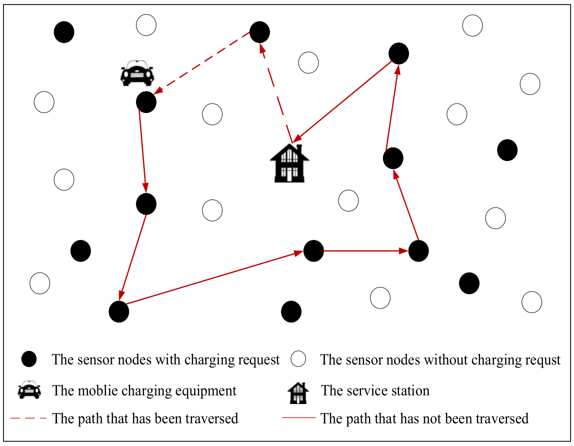

● The sensor nodes with charging request
○ The sensor nodes without charging requst
The moblie charging equipment
The service station
The path that has been traversed
The path that has not been traversed

FIGURE 2. Diagram of network topology.

TABLE 1. Symbol and definition.

Symbol	Definition
V_s	Set of sensor nodes
N	The number of the sensor nodes
$E_{\max}$	Maximum battery capacity of the sensor node
$E_{\min}$	The lower energy threshold of the sensor node
$E_c^{\max}$	The energy of the WCE for charging the sensor nodes
$E_d^{\max}$	The energy of the WCE for driving
R	Energy threshold for Sending a Charging request
RM_i	The charging request of the sensor node n_i
p_i	Energy consumption rate of the sensor node n_i
M	The number of the charging requests
V_c	Set of sensor nodes which send the charging requests
π_0	The charging service station
EU_c	The energy used by the WCE to charge the sensor nodes
EU_d	The driving energy used by the WCE
ER_c	The remaining charging energy of the WCE
ER_d	The remaining driving energy of the WCE
W	A charging path of the WCE
π_j	The jth sensor node in W
ec_j	The energy consumed by the WCE to charge π_j
ed_j	The energy consumed by the WCE to drive to π_j
μ	The driving energy consumption rate of the WCE
v	The driving speed of the WCE
ρ	The charging loss rate
U	The charging power of the WCE
K	The unit distance value
ψ	The update factor
γ	The discount factor
na	The number of agents

The WCE can provide energy for sensor node one-to-one. The energy for charging the sensor nodes and the energy for driving are $E_c^{\max}$, $E_d^{\max}$ respectively. When the remaining energy of the WCE is insufficient, it will return to the CS for energy replenishment. The WCE stays at the CS as a **vocation**, and the time at the CS is the **vocation period**.

The symbols used in this study are shown in TABLE 1.

B. CHARGING MODEL

The sensor node will die if its energy is lower than $E_{\min}$. To prolong the survival time, n_i will send a charging request $RM_i = (n_i, p_i, t_{r,i})$ to the WCE when its energy is below a threshold R. RM_i contains the time point $t_{r,i}$ issuing the request, the sensor node ID n_i and its energy consumption rate p_i.

The WCE accepts the charging request and stores the request in the order of $t_{r,i}$. When the vocation period ends, the WCE accepts M charging requests and puts the corresponding sensor nodes to V_c, then designs a charging path for the sensor node in the V_c.

After designing the charging path, the WCE sets off from the CS to charge the sensor nodes, and then goes back to the CS. This period is defined as a **charging round**.

C. PROBLEM FORMULATION

The problem in this study is to determine which sensor nodes will be charged and the charging strategy for the WCE, so as to improve both the number of sensor nodes that will be charged and the charging efficiency of the WCE under the limited energy.

Firstly, to measure the charging efficiency of WCE, the concept of WCE energy utilization η is introduced. We assume the charging path $W = (\pi_0, \pi_1 \cdots \pi_L, \pi_0)$. π_0 represents the CS, L is the number of the sensor nodes in the charging path. Therefore, η can be calculated as Eq.(2).

$$\eta(W) = EU_c/EU_d \tag{2}$$

EU_c is the energy used by the WCE to charge the sensor nodes in a charging round. EU_d is the driving energy used by WCE. EU_c, EU_d satisfy Eq.(3) and Eq.(4).

$$EU_c = \sum_{j=1}^{L} ec_j, \quad EU_c \leq E_c^{\max} \tag{3}$$

$$EU_d = \sum_{j=1}^{L} ed_j, \quad EU_d \leq E_d^{\max} \tag{4}$$

The WCE should ensure that the sensor node π_q has been charged before the energy of π_q is below Emin. We assumed ρ is the charging loss rate and U is the charging power of WCE. Then, we have:

$$0 \leq (1-\rho)\, ec_j \leq E_{\max} - E_{\min} \tag{5}$$

Each sensor node can only be charged at most once in a charging round. Assume the duration of a charging round is T. The energy of the sensor node in a charging round should satisfy the Eq.(6).

$$(1-\rho)\, ec_j - p_j T \geq 0 \tag{6}$$

To ensure the WCE can reach the sensor node π_q and return to the CS after charging the sensor node π_q, the driving energy of the WCE should meet Eq.(7).

$$E_d^{\max} - \mu \frac{\sum_{j=0}^{q} l_{j,j+1}}{v} - \mu \frac{l_{q+1,0}}{v} \geq 0 \tag{7}$$

It is assumed that v is the driving speed and μ is the driving energy consumption rate of WCE. $l_{j,j+1}$ is the distance between π_j and π_{j+1}. And π_0 represents the CS.

Then, the On-demand Charging Planning Problem (OCPP) can be formulated as follows,

$$\begin{aligned} cObj: arg \max_{W} \ \eta(W) \\ \text{s.t. } (2) - (7) \end{aligned}$$

IV. LEARNING MODEL OF THE WCE

As described above, the On-demand Charging Planning Problem (OCPP) in this study is to plan a charging path for the WCE in WRSNs, the objective is optimizing networks and the WCE performance. And OCPP is similar to the mobile robot path planning problem. As the mobile robot path planning problem is to search an optimal path from the start point to the end point with the goal of no collision. And solving the mobile robot path planning problem is an important branch of the Reinforcement Learning (RL) application. With the help of RL, the mobile robot has "brain", it can autonomously learn the path. In the OCPP, the charging nodes and charging path for each charging round is uncertain. Hence, the WCE can charge more efficiently and the charging strategy is flexible if WCE can autonomously learn and adjust the charging path. Therefore, the RL is introduced to solve the OCPP in this study.

And the relationship between RL and OCPP are as follows: WRSNs is considered as the environment in RL; The WCE in WRSNs is considered as the agent in RL; the state of WRSNs and the WCE is considered as the state in RL; the action of the WCE to the next charging sensor node is considered as the action in RL. Therefore, the learning model in WRSNs can be represented by a triple $\langle X, A, R\rangle$. X is the state space, which represents the state of the WRSNs and WCE. A is the action space, which represents the action set of the WCE. R is the reward generated by actions of the WCE.

A. STATE MODEL

Due to the different states, the location of the WCE, the remaining travel energy of the WCE, and the energy of the sensor nodes of the network will change. Therefore, the definition of the state space considers both the WCE and the state of the sensor nodes in the network. The state space is defined as a two-tuple $X = \langle X_{WCE}, X_{network}\rangle$.

$$\begin{aligned} X_{WCE} &= \langle x, ER_c, ER_d\rangle \\ X_{network} &= \left\langle \overrightarrow{e}, \overrightarrow{d}, \overrightarrow{trd} \right\rangle \end{aligned}$$

$X_{WCE} = \langle x, ER_c, ER_d\rangle$ indicates the state of the WCE. Renumber the sensor nodes in V_c to $(1, 2 \cdots M)$, we assume $N_c = \{0, 1, \cdots M, N+1\}$. x represents the current location states of the WCE, $x = 0$ indicates that WCE has not yet left the CS, $x = N+1$ indicates the WCE has completed a charging task and back to the service station for the energy replenishing. $x = m\,(1 \le m \le M)$ indicates that the WCE is charging the sensor node numbered m in V_c. ER_c is the remaining charging energy and ER_d is the remaining driving energy of the WCE. Initially, $ER_c^1 = E_c^{\max}$, $ER_d^1 = E_d^{\max}$.

$X_{network} = \left\langle \overrightarrow{e}, \overrightarrow{d}, \overrightarrow{trd} \right\rangle$ means the state of the sensor nodes in the networks. $\overrightarrow{e} = (e_1, e_2 \cdots e_i \cdots e_N)$ means the current energy state of the networks. $\overrightarrow{d} = (d_1 \cdots d_M)$ means the distance between the WCE and the sensor nodes. $\overrightarrow{trd} = (trd_1 \cdots trd_m \cdots trd_M)$ indicates the flag of the sensor node in V_c. If the sensor node numbered m in V_c is traversed by the WCE, the flag trd_m is set to 1. Otherwise, the value of trd_m is 0.

B. ACTION MODEL

For the WCE, to select an action is to determine the next sensor node to be charged. We assumed that all sensor nodes are reachable in the network. Because the sensor nodes can only be charged at most once in a charging round, the WCE can only select the sensor node in V_c whose flag is 0. In this paper, the action space is defined as $A = \{a | a \in N_c\}$. $a = m$ means the next sensor node to be charged is the sensor node numbered m in V_c.

C. STATE TRANSITION PROCESS

It is assumed that every state update of agent is a time step. At time step k, the WCE stays at state x^k, selects action a^k according to SA, and then reaches the state x^{k+1}. At time step $k+1$, the sensor node numbered x^{k+1} in V_c has been fully charged. Assume that the duration between time step k to time step $k+1$ is Δt^k. Next, we will discuss the state as the time step $k+1$.

As for sensor nodes, remaining energy e_i^{k+1} of n_i can be calculated by Eq.(8).

$$e_i^{k+1} = \begin{cases} E_{\max} & n_i \in V_c \text{ and } m = x^{k+1} \\ E_{\max} - p_i \Delta t^k & n_i \in V_c \text{ and } m = x^k \\ e_i^k - p_i \Delta t^k & \text{others} \end{cases} \tag{8}$$

Δt^k consists of two parts as shown in Eq.(9): 1) the driving time from the sensor node numbered x^k to the sensor node numbered x^{k+1} in V_c. 2) the charging time for the sensor node numbered x^{k+1} in V_c. And U is the charging power.

$$\Delta t^k = \frac{d_m^k}{v} + \frac{E_{\max} - \left(e_m^k - p_m \frac{d_m^k}{v}\right)}{U - p_m}, \quad \left(m = x^{k+1}\right) \tag{9}$$

As for the WCE, at time step $k+1$, the ec_m and ed_m can be calculated by Eq.(10), Eq.(11). And then the ER_c^{k+1} and ER_d^{k+1} are shown as Eq.(12) and Eq.(13) respectively.

$$ec_m = \frac{E_{\max} - \left(e_m^k - p_m \frac{d_m^k}{v}\right)}{1 - \rho}, \quad \left(m = x^{k+1}\right) \tag{10}$$

$$ed_m = \mu \frac{d_m^k}{v}, \quad \left(m = x^{k+1}\right) \tag{11}$$

$$ER_c^{k+1} = ER_c^k - ec_m \tag{12}$$

$$ER_d^{k+1} = ER_d^k - ed_m \tag{13}$$

D. REWARD MODEL

In the RL, the agent learns by reward value, so the setting of the reward function is especially important, which determines whether the algorithm can converge and the speed of convergence. The problem in this study is to maximize the energy utilization of WCE in a charging round. Therefore, after performing an action a^k, the reward r^k of a^k should consider two aspects: 1) to maximize EU_c. 2) to minimize EU_d. The reward function is as Eq.(14),

$$r^k = \alpha_1 \frac{ec_m}{E_{\max}} + \alpha_2 \frac{K}{ed_m} \tag{14}$$

As shown in Eq.(14), $0 < \alpha_1, \alpha_2 < 1, \alpha_1 + \alpha_2 = 1$. α_1 and α_2 respectively represent the proportions of the two factors. K represents the unit energy value. The higher $ec_m \left(m = x^k\right)$ is, the higher r^k is. In addition, the lower $ed_m \left(m = x^k\right)$ is, the higher r^k is.

V. PROPOSED ALGORITHM: SRL-FA

To overcome the shortcoming of the original reinforcement learning too much invalid learning, the swarm reinforcement learning (SRL) is introduced. In SRL, multi agents learn simultaneously. However, the existing SRL often falls into local optimum. Meanwhile, for optimization problems, a population-based method such as Firefly Algorithm (FA) have been recognized that they are able to find rapidly optimal solutions. Therefore, in this section, a Swarm Reinforcement Learning based on FA (SRL-FA) is proposed.

In SRL-FA, agents all learn concurrently with two stages, Individual Learning the Charging Path and Learning through Exchanging Information. The improvement of SRL in this study is shown in the latter stage. And these two stages are discussed in detail in subsection A and subsection B. And the learning framework of the WCE is shown as Fig.3. Y is the number of interaions.

A. INDIVIDUAL LEARNING THE CHARGING PATH

In this stage, each agent learns individually by using a usual RL. As for the agent ag^i, the learning process in an iteration is as follows: Simulated Annealing (SA) is used to select the action. After selecting, WCE judges whether it can reach this action and return to the CS. If it can, WCE adds this action to the charging path and calculates the reward of this action. Then update the Q-value ΔQ^i_{ag} according to Eq.(15); Otherwise, the WCE will back to the CS.

$$\Delta Q^i_{ag}(x^k, a^k) = \psi(r^k + \gamma \max_{a^{k+1}} Q^i_{ag}(x^{k+1}, a^{k+1}) - Q^i_{ag}(x^k, a^k)) \tag{15}$$

where ψ is the update factor, and γ $(0 < \gamma < 1)$ is the discount factor. Q^i_{ag} is the learned experience of ag^i, and ΔQ^i_{ag} is the new Q-value in this iteration.

After an iteration ends, the WCE will evaluate the charging path W^i_{ag} obtained by ag^i in this iteration y. The evaluation indicator V^i_{ag} can be calculated through Eq.(16). Then the Q^i_{ag} is updated as Eq.(17),

$$V^i_{ag} = \eta(W^i_{ag}) \tag{16}$$

$$Q^i_{ag}(y) = Q^i_{ag}(y-1) + V^i_{ag} \times \Delta Q^i_{ag} \tag{17}$$

Based on the above statement, the Individual Learning Algorithm is shown as Algorithm 1:

Algorithm 1 Individual Learning Algorithm

```
Input: $Y$ and the number of the agents $na$.
Output: Q-value and charging path of all agents.
for $y \leftarrow 1$ to $Y$ do
    for $ag \leftarrow 1$ to $na$ do
        for $k \leftarrow 1$ to $M+1$ do
            Select the action $a^k$ according to SA;
            Calculate $ER_c^{k+1}$, $ER_d^{k+1}$ according to Eq.(12), Eq. (13);
            if WCE can reach $a^k$ and back to the CS and $ER_c^k$ then
                Calculate $r^k$ according to Eq.(14);
                Calculate $\Delta Q^i_{ag}$ according to Eq.(16);
                $W^i_{ag}(k) \leftarrow a^k$;
            else
                WCE backs to the CS;
                Break;
            end if
        end for
    end for
end for
```

B. LEARNING THROUGH EXCHANGING INFORMATION

In this stage, each agent learns through updating its Q-value by referring to the other agent. The Q-value update method is important, it determines the performance of the algorithm. The existing SRL algorithms update Q-value directly. This update method is only suitable for the continuous problems. However, the charging path planning problem in this study is a discrete problem. Therefore, this stage should be improved. Suppose the agents have independently learned for Y iteration. Then the improvement idea is shown in Fig.4. Q_{best} is the best Q-value of all the agents at only the previous iteration. G_{best} is the best Q-value found by all the agents so far. P^i_{ag} is the best Q-value found by the agent ag^i so far. V is a so-called velocity. ω, C_1, C_2 are weight parameters. R_1, R_2 are uniform random number in the range from 0 to 1.

As figure shows, the improvement idea is that we update the Q-value by updating the path of agent. Among the existing SRL algorithms, PSO-Q can update the path. PSO-Q is the combination of PSO and SRL algorithm. However, PSO is easy to fall into local optimality. Firefly Algorithm (FA) is similar to PSO and performs better than PSO. Therefore, to maintain the advantages of PSO combined with SRL and improve the SRL, FA is introduced into Q-value Update method. And the Q-value Update Algorithm based on FA is proposed. And the details of the algorithm are described in the following sections.

1) RELATIONSHIP BETWEEN FIREFLY ALGORITHM AND OUR STUDY

Firefly Algorithm (FA) regards the value of the objective function as the absolute brightness of a firefly. The Fireflies

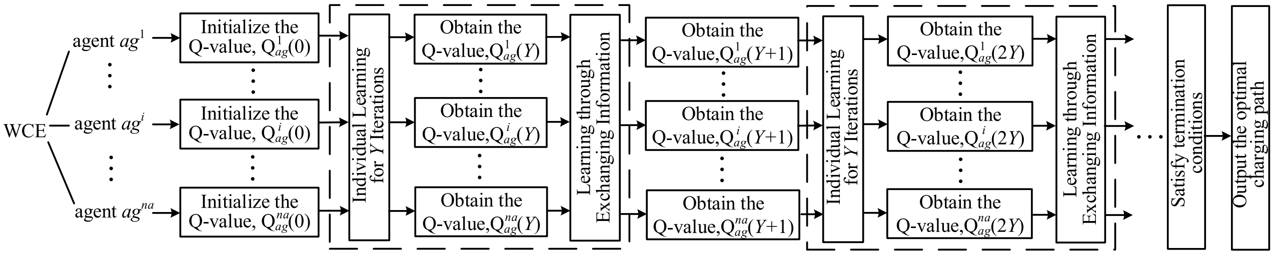

WCE
agent ag^1
agent ag^i
agent ag^{na}
Initialize the Q-value, $Q^1_{ag}(0)$
Initialize the Q-value, $Q^i_{ag}(0)$
Initialize the Q-value, $Q^{na}_{ag}(0)$
Individual Learning for Y Iterations
Obtain the Q-value,$Q^1_{ag}(Y)$
Obtain the Q-value,$Q^i_{ag}(Y)$
Obtain the Q-value,$Q^{na}_{ag}(Y)$
Learning through Exchanging Information
Obtain the Q-value,$Q^1_{ag}(Y+1)$
Obtain the Q-value,$Q^i_{ag}(Y+1)$
Obtain the Q-value,$Q^{na}_{ag}(Y+1)$
Individual Learning for Y Iterations
Obtain the Q-value,$Q^1_{ag}(2Y)$
Obtain the Q-value,$Q^i_{ag}(2Y)$
Obtain the Q-value,$Q^{na}_{ag}(2Y)$
Learning through Exchanging Information
Satisfy termination conditions
Output the optimal charging path

FIGURE 3. Learning framework of the WCE.

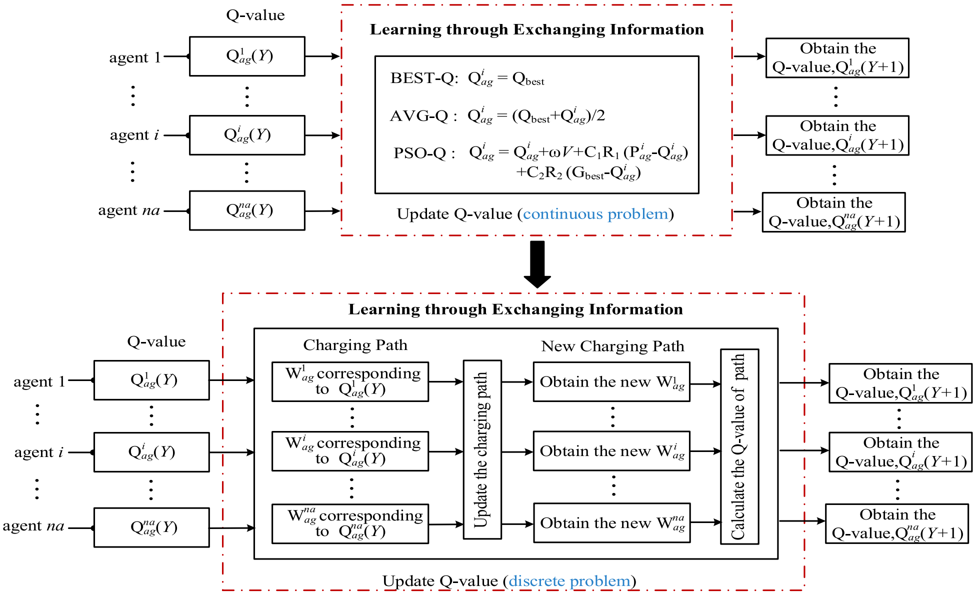

Q-value
agent 1
agent i
agent na
$Q^1_{ag}(Y)$
$Q^i_{ag}(Y)$
$Q^{na}_{ag}(Y)$
Learning through Exchanging Information
BEST-Q: $Q^i_{ag} = Q_{best}$
AVG-Q : $Q^i_{ag} = (Q_{best}+Q^i_{ag})/2$
PSO-Q : $Q^i_{ag} = Q^i_{ag}+\omega V+C_1R_1(P^i_{ag}-Q^i_{ag})+C_2R_2(G_{best}-Q^i_{ag})$
Update Q-value (continuous problem)
Obtain the Q-value,$Q^1_{ag}(Y+1)$
Obtain the Q-value,$Q^i_{ag}(Y+1)$
Obtain the Q-value,$Q^{na}_{ag}(Y+1)$
Learning through Exchanging Information
Q-value
agent 1
agent i
agent na
$Q^1_{ag}(Y)$
$Q^i_{ag}(Y)$
$Q^{na}_{ag}(Y)$
Charging Path
W^1_{ag} corresponding to $Q^1_{ag}(Y)$
W^i_{ag} corresponding to $Q^i_{ag}(Y)$
W^{na}_{ag} corresponding to $Q^{na}_{ag}(Y)$
Update the charging path
New Charging Path
Obtain the new W^1_{ag}
Obtain the new W^i_{ag}
Obtain the new W^{na}_{ag}
Calculate the Q-value of path
Obtain the Q-value,$Q^1_{ag}(Y+1)$
Obtain the Q-value,$Q^i_{ag}(Y+1)$
Obtain the Q-value,$Q^{na}_{ag}(Y+1)$
Update Q-value (discrete problem)

FIGURE 4. The improvement idea of SRL in this study.

are hermaphroditic. Therefore the main idea of FA is that a little brighter firefly will move towards the brightest one within the visible distance range. And if there is no brighter one than a particular firefly, it will move randomly.

This study considers updating the Q-value by transforming the charging path of each agent. Therefore, we regard the charging path W^i_{ag} as the solution of the firefly ffy^i, and the corresponding fitness value $\eta\left(W^i_{ag}\right)$ as the absolute brightness F^i_{ffy} of the ffy^i. And then the way of updating Q-value is as follows: For each firefly ffy^i, find the other firefly $ffy^{\max}$ with the highest fitness value in its visible distance range. If $ffy^{\max}$ exists, move ffy^i toward $ffy^{\max}$. Otherwise, ffy^i will move randomly. Then calculate the Q-value corresponding to the new charging path.

Since the solution of a firefly in this study is discrete, and the WCE has limited energy, it will back to the CS if it would use up its energy. Therefore the dimensions of the charging path for each learning process may vary. We cannot simply calculate the distance using Euclidian distance. Then we define the distance between any two fireflies as the number of different arcs between them. To be specific, we assume the solution of ffy^i and ffy^j are shown in Fig.5. The green solid lines are the same arcs between ffy^i and ffy^j, and black dotted lines are different arcs between ffy^i and ffy^j. Therefore the distance between them is 4.

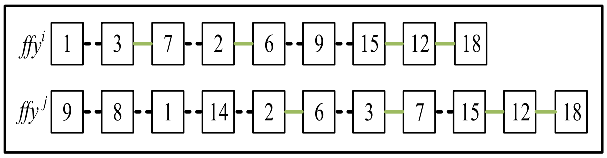
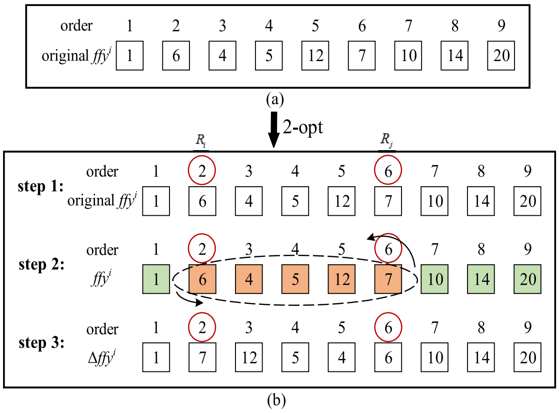


FIGURE 5. The distance between ffy^i and ffy^j.

2) THE WAY OF FIREFLY MOVING

In this section, we will introduce the way the fireflies move. And there are two situations of firefly moving. Let D_i be the visible distances range of ffy^i. Then we take firefly ffy^i as the example.

Situation 1: There are no fireflies that the fitness value is larger than ffy^i within D_i.



order 1 2 3 4 5 6 7 8 9
original ffy^i 1 6 4 5 12 7 10 14 20
(a)
2-opt
R_i
R_j
step 1:
order 1 2 3 4 5 6 7 8 9
original ffy^i 1 6 4 5 12 7 10 14 20
step 2:
order 1 2 3 4 5 6 7 8 9
ffy^i 1 6 4 5 12 7 10 14 20
step 3:
order 1 2 3 4 5 6 7 8 9
Δffy^i 1 7 12 5 4 6 10 14 20
(b)

FIGURE 6. The random movement process of ffy^i.

The Way of Firefly Moving 1: In situation 1, ffy^i moves randomly. In this study, the 2-opt operation is performed on ffy^i. The main idea of 2-opt is: Firstly, generate two random numbers R_i, R_j $\left(R_i < R_j < P_L^i\right)$ as two positions in W_{ag}^i, where P_L^i is the length of the W_{ag}^i. Then, to obtain the new charging path ΔW_{ag}^i, the path before R_i and after R_j in W_{ag}^i is added to the ΔW_{ag}^i. The path between R_i and R_j in W_{ag}^i is reversed, and then added it to the ΔW_{ag}^i. To be specific, we assume the solution of ffy^i is shown in Fig.6 (a). And the random movement process of ffy^i and the new solution Δffy^i are shows in Fig.6.

Situation 2: There is a $ffy^{\max}$ with the highest fitness value within D_i. In this Situation, if ffy^i moves directly to $ffy^{\max}$, it is easy to fall into local optimality. To enhance the global search capability, we accept the randomly generated new firefly $\overline{ffy^i}$ with a certain probability P. The probability P can be calculated by Eq.(18).

$$P = \begin{cases} exp\left(-F_{ffy}^{max}/\overline{F_{ffy}^i}\right), & F_{ffy}^{max} > \overline{F_{ffy}^i} \\ 1, & F_{ffy}^{max} \leqslant \overline{F_{ffy}^i} \end{cases} \tag{18}$$

Based on the above statement, there are two ways of moving. The rand() function randomly generates a real number in range [0, 1]. Then, the details are as follows.

The Way of Firefly Moving 2: If rand() $>P$, ffy^i moves directly to $ffy^{\max}$. Then the solution of $ffy^{\max}$ is assigned to ffy^i. Therefore, the update formula is shown as Eq.(19),

$$\begin{aligned} W_{ag}^i &\leftarrow W_{ag}^{\max} \\ Q_{ag}^i &\leftarrow Q_{ag}^{\max} \end{aligned} \tag{19}$$

The Way of Firefly Moving 3: If rand () $\leqslant P, ffy^i$ is updated by $\overline{ffy^i}$. Then the solution of $\overline{ffy^i}$ is assigned to ffy^i. Therefore, the update formula is shown as Eq.(20),

$$\begin{aligned} W_{ag}^i &\leftarrow \overline{W_{ag}^i} \\ Q_{ag}^i &\leftarrow \overline{Q_{ag}^i} \end{aligned} \tag{20}$$

Inspired by the idea of Firefly Algorithm, the Q-value Update Algorithm based on FA is shown as Algorithm 2:

Algorithm 2 Q-Value Update Algorithm Based on FA

Input: na, firefly fly^i and absolute brightness F_{fly}^i.

Output: new charging path of ag^i and new ΔQ_{ag}^i.

```
repeat
    $count \leftarrow 0, k \leftarrow 1$;
    while $k \leq na$ do
        Calculate the distance $d_{i,k}$ between $fly^i$ and $fly^k$;
        if $d_{i,k} < D_i$ then
            if $F_{fly}^i < F_{fly}^k$ then
                $count \leftarrow count + 1$;
                Save the number $k$ and its corresponding path;
            end if
        end if
    end while
    if $count = 0$ then
        Move $fly^i$ around randomly;
    else
        Calculate $P$ according to Eq.(18);
        if $P < rand()$ then
            Find the one with the highest fitness value in the saved path and move $fly^i$ to it;
        else
            Randomly generate a new solution and move $fly^i$ to it;
        end if
    end if
    Get a new $\Delta fly^i$ and calculate the new $\Delta Q_{ag}^i$;
until $i = na$
```

VI. PERFORMANCE EVALUATION

In this section, experimental simulations are carried out to demonstrate the advantages of the proposed algorithm Swarm Reinforcement Learning based on Firefly Algorithm (SRL-FA). We compare our algorithm with other reinforcement learning algorithms and charging scheduling algorithms in Performance Comparison. Moreover, we investigate the impact of several important parameters on algorithm performance in Properties Analysis.

A. SIMULATION ENVIRONMENT

As listed in TABLE 2, 100 to 200 sensor nodes are deployed in a 2000m × 2000m square. To analyze the performance of SRL-FA in different kinds of networks, the sensor nodes in this study has two ways of distribution, random and uniform. The corresponding networks named C1 and C2 respectively. Uniform distribution means the spacing of each sensor node is the same. The other parameters listed in TABLE 2 have been chosen mostly based on [30].

B. PERFORMANCE COMPARISON

In this section, the Swarm Reinforcement Learning based on Firefly Algorithm (SRL-FA) is compared with other

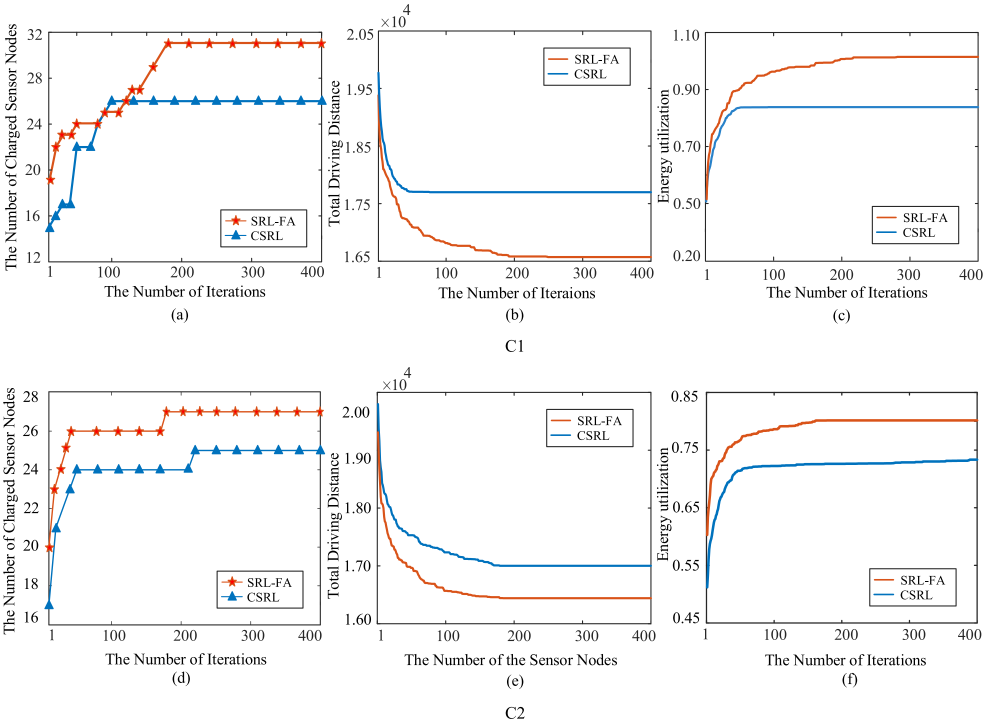


FIGURE 7. Performance comparison between SRL-FA and CSRL [30] in terms of (a)(d) the number of charged sensor nodes, (b)(e) total driving distance and (c)(f) energy utilization in two network scenarios respectively.

TABLE 2. Parameters used in this study.

Parameters	Symbols	Values
Number of nodes	N	100-200
Location of CS	-	0m, 0m
Energy consumption rate of sensor node	pi	$[0w, 1w]$
Energy consumption rate of WCE's driving	μ	$(8v + 5)$ J/s
Charging power of WCE	U	10J/s
Charging loss rate	ρ	0.2
Energy threshold for sending a charging request	R	$0.4E_{\max}$
Energy capacity of sensor node	$E_{\max}$	10800J
Minimum energy of sensor node	$E_{\min}$	540J
Energy capacity of WCE for driving	$E_d^{\max}$	210KJ
Energy capacity of WCE for charging	$E_c^{\max}$	210KJ
Factor in Eq.(14)	α_1	0.4
Factor in Eq.(14)	α_2	0.6
Unit distance value in Eq.(14)	K	3500

Reinforcement Learning (RL) algorithms and two classic charging algorithms, FCFS and NJNP under the on-demand charging architecture to analyze its performance.

1) COMPARISONS WITH RL ALGORITHMS

Firstly, we perform an analysis concerning the performance of the redesigned reinforcement learning algorithm in this study. Under the same two networks setting where 200 sensor nodes are deployed randomly(C1) and uniformly(C2) respectively, we compare SRL-FA with 1) the original RL algorithm with an agent [30] and 2) the existing three Swarm Reinforcement Learning (SRL) algorithm named BEST-Q, AVG-Q and PSO-Q [31], [32]. Different performance metrics are considered, including energy utilization η of WCE, the number *nc* of sensor nodes that have been charged and the driving distance of WCE. The results are shown below and they are average of 30 runs. The red line in Fig.7 and Fig.8 represent our algorithm.

TABLE 3. Simulation results of five reinforcement learning algorithms in two network scenarios.

Scenario		Energy Utilization	*nc*	Total Driving Distance(m)
C1	**SRL-FA**	**0.9974**	**31**	**16545**
	CSRL	0.8412	26	17699
	BEST-Q	0.8508	26	17136
	AVG-Q	0.8668	27	16851
	PSO-Q	0.9862	29	16702
C2	**SRL-FA**	**0.8351**	**27**	**16451**
	CSRL	0.7767	25	17002
	BEST-Q	0.7422	25	16899
	AVG-Q	0.7491	25	16679
	PSO-Q	0.7720	26	16494

In Fig.7, we compare the SRL-FA with CSRL. The figure shows that the optimization accuracy of SRL-FA is better than CSRL. Since CSRL is based on original RL. There is only one agent in CSRL to explore. SRL-FA is based on SRL. There are multiple agents in SRL-FA. Moreover, agents learn through exchanging information. Therefore, the ability to explore is increased. As shown in TABLE3, in two network scenarios, 1) the energy utilization obtained by SRL-FA is 19% and 7% higher than CSRL; 2) the number of charged sensor nodes obtained by SRL-FA is 19% and 8% higher than CSRL respectively. Therefore, the result confirms that SRL-FA based on SRL is superior to CSRL based on original RL.

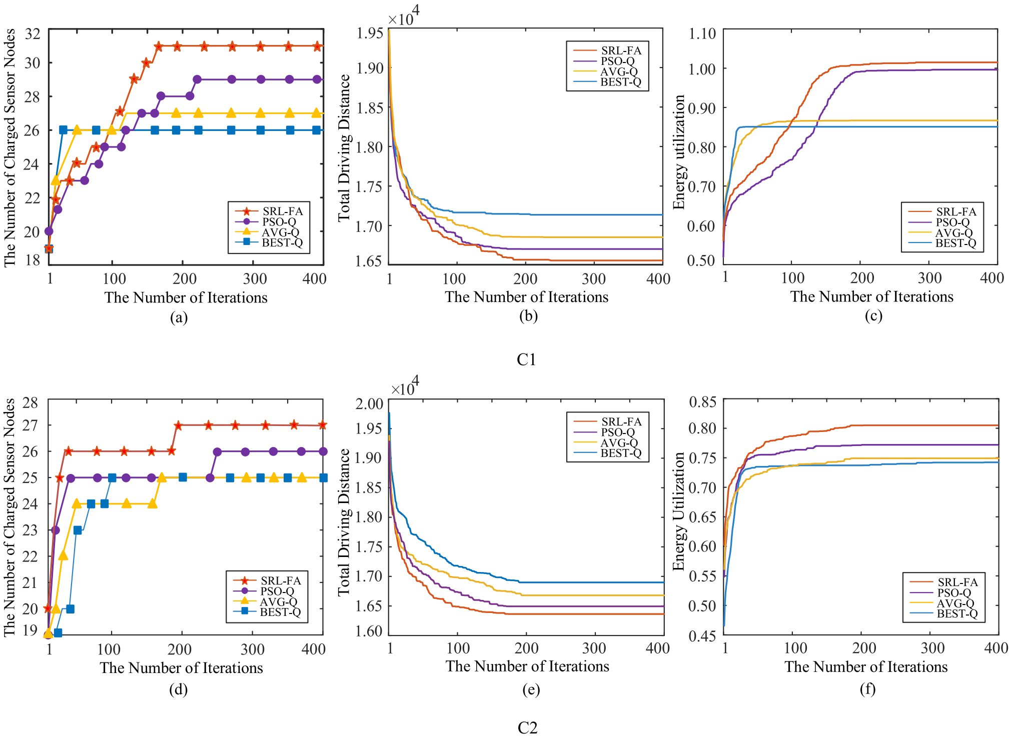


FIGURE 8. Performance comparison between SRL-FA and other swarm reinforcement learning algorithm [31], [32] in terms of (a)(d) the number of charged sensor nodes, (b)(e) total driving distance and (c)(f) energy utilization in two network scenarios respectively.

SRL-FA is improved by SRL algorithm. To verify the superiority of SRL-FA, we compare SRL-FA with the existing SRL algorithms. The SRL algorithms being compared are BEST-Q, AVG-Q and PSO-Q. From Fig.8, it can be observed that, SRL-FA performs better than the other three algorithms. Moreover, as listed in TABLE3, in two network scenarios, 1) the energy utilization obtained by SRL-FA is 17% and 12% higher than BEST-Q, 15% and 11% higher than AVG-Q and 2% and 7% higher than PSO-Q respectively. 2) the number of charged sensor nodes obtained by SRL-FA is 16% and 7% higher than BEST-Q, 13% and 7% higher than AVG-Q and 6%and 3% higher than PSO-Q respectively. Due to the performance of the SRL algorithm highly depends on the method of exchanging information, therefore the result confirms that SRL-FA is well designed.

2) COMPARISONS WITH ON-DEMAND CHARGING SCHEDULING ALGORITHMS

To measure the performance of SRL-FA on on-demand charging scheduling algorithms, we compare our algorithm with two classic on-demand charging scheduling algorithms, FCFS and NJNP in two network scenarios.

As demonstrated in Fig.9(b)(e), the energy utilization of SRL-FA is always higher than FCFS and NJNP. This is because we use TSP solutions to formulate the charging path, which can achieve global optimization. FCFS schedules the incoming charging requests based on their temporal property and ignores the driving distance, therefore it has the least energy utilization. Although NJNP overcomes the drawback of FCFS, it always selects the nearest sensor node and ignores the residual energy of the sensor node. And the charging energy used by NJNP may not be high, resulting in less energy utilization. Next Fig.9(c)(f) compares charging success rate, which is defined as the ratio of the number of sensor nodes which have been successfully charged to the number of sensor nodes sending the charging request. As Figure shows, the SRL-FA performs well. And the energy of the WCE is limited, the sensor nodes that the corresponding charging request does not be responded may not be dead. Thus the charging success rate does not reflect the survival rate of the sensor nodes. And then we compare the number of dead sensor nodes of three algorithms to evaluate system stability, the result is shown in Fig.9(a)(d). From the results, it can be observed that, with the growth of the number of the sensor nodes, the charging success rate decreases and the number of dead sensor nodes increases. It is because the charging requests will increase with the growth of the number of the sensor nodes and the energy of the WCE is limited. WCE cannot serve such a large number of sensor nodes. But the simulation results show that SRL-FA performs better than FCFS and NJNP.

C. PARAMETERS ANALYSIS

In this section, we will study the impact of different parameters such as the number of the agents, the speed of WCE and the update factor on the performance of SRL-FA. And we fix the network scale at 200 sensor nodes in a 2000m × 2000m field.

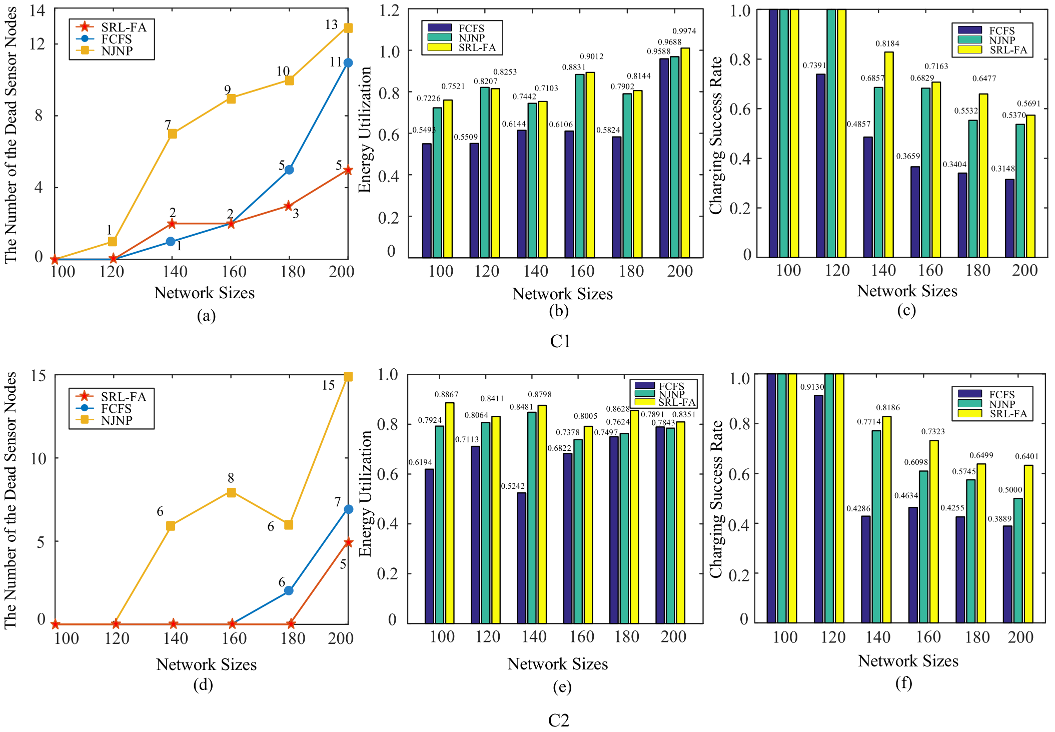


FIGURE 9. Performance comparison between SRL-FA, FCFS [16] and NJNP [17] in terms of (a)(d) the number of the dead sensor nodes, (b)(e) energy utilization and (c)(f) charging success rate in two network scenarios respectively.

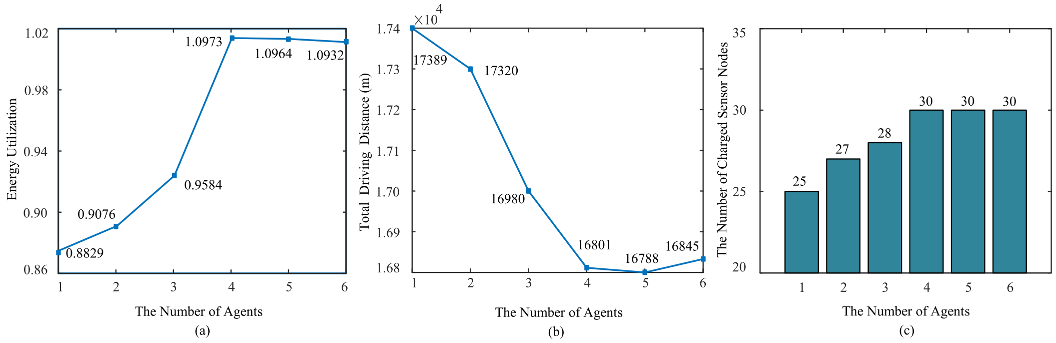


FIGURE 10. Impact of the number of agents on the charging process (a) energy utilization, (b) total driving distance and (c) the number of the charged sensor nodes.

1) IMPACT OF THE NUMBER OF AGENTS

SRL-FA is based on SRL. The number of agents may influence the algorithm performance. Therefore, we study the impact of it by varying its value from 1 to 6. Fig.10 shows that with the growth of the number of the agents, the energy utilization as well as the number of charged sensor nodes increases and the driving distance decreases. The reason is that multiple agents learn simultaneously to make exploration more full. However, when the number of agents more than 4, the growth is not obvious, which implies that the performance of SRL-FA is near optimal.

2) IMPACT OF THE SPEED OF WCE

An important factor that determines the mobile charger's ability in performing charging tasks is its driving speed v. We explore the performance of SRL-FA with varying v from 5 to 10 m/s. The results are shown in Fig.11. It can be clearly observed that with the speed increases, the SRL-FA performs better. It is because WCE can accomplish charging faster with larger speed. However, the driving energy consumption of WCE is related to speed. The higher the speed is, the higher the consumption is. Therefore, as shown in Fig.11, when the speed of WCE more than 8m/s, there is no obvious improvement in performance.

3) IMPACT OF THE VALUE OF UPDATE FACTOR

According to Eq.(15). ψ is the update factor and it can be seen as the equilibrium between the new Q-value and the learned Q-value during the learning process. If ψ set at 1, it means the learned Q-value has no effect during the learning process. To study the impact of the update factor on the performance of SRL-FA, we set ψ as 01.-0.9, and the result as shown

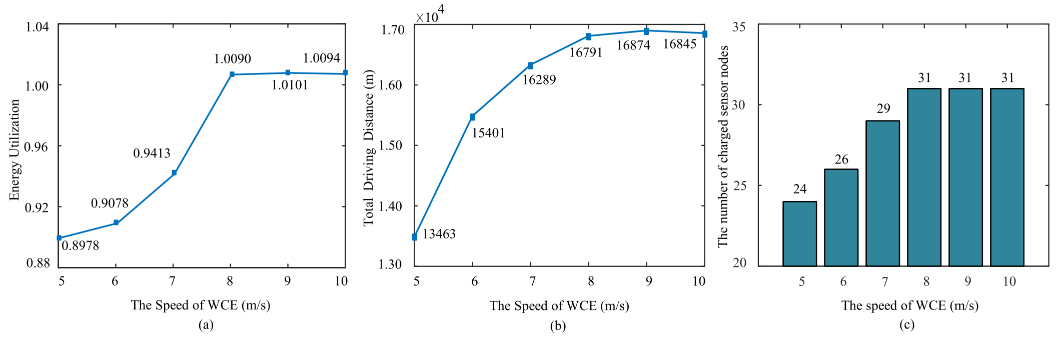


FIGURE 11. **Impact of the speed of WCE on the charging process (a) energy utilization, (b) total driving distance and (c) the number of the charged sensor nodes.**

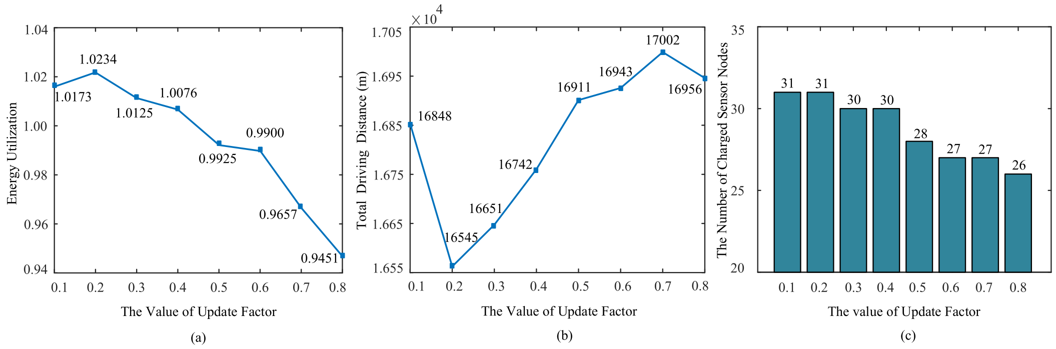


FIGURE 12. **Impact of the value of update factor on the charging process (a) energy utilization, (b) total driving distance and (c) the number of the charged sensor nodes.**

in Fig.12. We can see that there is a significant improvement in the value of charged sensor nodes and energy utilization between $\psi = 0.1$ and $\psi = 0.2$. And the performance of the algorithm turns to be worse with the increase of ψ. Therefore, the algorithm performs better when $\psi = 0.2$.

VII. CONCLUSION AND FUTURE WORK

In this study, an on-demand charging algorithm based on Swarm Reinforcement Learning is proposed, named SRL-FA. With the application of reinforcement learning algorithm, SRL-FA can help WCE achieve autonomous path planning. Moreover, SRL-FA totally consider the performance of the WCE with limited energy and the response to the charging requests. Therefore, SRL-FA can improve the performance of WCE and sensor networks.

And then a large number of experiments are conducted to verify the performance of SRL-FA, which is compared with the existing swarm reinforcement learning algorithms and classic on-demand charging algorithms. The simulation results demonstrate that SRL-FA is well designed and can effectively prolong the lifespan of networks as well as WCE's energy utilization under the limited energy of the WCE. We further analyze how the parameters affect SRL-FA, such as the number of agents, the speed of the WCE and update factor.

In the future, we are planning to extend this work by using multiple WCEs and considering the energy consumption of the sensor nodes are dynamic. It may lead to more cooperative works among them to address more practical problem in WRSNs.

REFERENCES

[1] M. Ekman and H. Palsson, "Ground target tracking of vehicles in a wireless ground sensor network," in *Proc. 15th Int. Conf. Inf. Fusion*, Singapore, 2012, pp. 2392–2399.

[2] E. H. Houssein, M. R. Saad, K. Hussain, W. Zhu, H. Shaban, and M. Hassaballah, "Optimal sink node placement in large scale wireless sensor networks based on Harris' hawk optimization algorithm," *IEEE Access*, vol. 8, pp. 19381–19397, 2020.

[3] K. Romer and F. Mattern, "The design space of wireless sensor networks," *IEEE Wireless Commun.*, vol. 11, no. 6, pp. 54–61, Dec. 2004.

[4] M. M. Ahmed, E. H. Houssein, A. E. Hassanien, A. Taha, and E. Hassanien, "Maximizing lifetime of large-scale wireless sensor networks using multi-objective whale optimization algorithm," *Telecommun. Syst.*, vol. 72, no. 2, pp. 243–259, Oct. 2019.

[5] D. Dondi, A. Bertacchini, D. Brunelli, L. Larcher, and L. Benini, "Modeling and optimization of a solar energy harvester system for self-powered wireless sensor networks," *IEEE Trans. Ind. Electron.*, vol. 55, no. 7, pp. 2759–2766, Jul. 2008.

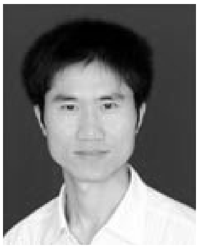
[6] A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulos, P. Fisher, and M. Soljacic, "Wireless power transfer via strongly coupled magnetic resonances," *Science*, vol. 55, no. 7, pp. 2759–2766, 2008.

[7] Z. Lyu, Z. Wei, Y. Lu, X. Wang, M. Li, C. Xia, and J. Han, "Multi-node charging planning algorithm with an energy-limited WCE in WRSNs," *IEEE Access*, vol. 7, pp. 47154–47170, 2019.

[8] W. Xu, W. Liang, H. Kan, Y. Xu, and X. Zhang, ''Minimizing the longest charge delay of multiple mobile chargers for wireless rechargeable sensor networks by charging multiple sensors simultaneously,'' in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Dallas, TX, USA, Jul. 2019, pp. 881–890.
[9] Y. Shi, L. Xie, Y. T. Hou, and H. D. Sherali, ''On renewable sensor networks with wireless energy transfer,'' in *Proc. IEEE Int. Conf. Comput. Commun.*, Shanghai, China, Apr. 2011, pp. 1350–1358.
[10] L. Fu, P. Cheng, Y. Gu, J. Chen, and T. He, ''Minimizing charging delay in wireless rechargeable sensor network,'' in *Proc. IEEE Int. Conf. Comput. Commun.*, Turin, Italy, Apr. 2013, pp. 2922–2930.
[11] L. Xie, Y. Shi, Y. T. Hou, W. Lou, and H. D. Sherali, ''On traveling path and related problems for a mobile station in a rechargeable sensor network,'' in *Proc. 14th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Bengaluru, India, 2013, pp. 109–118.
[12] X. Rao, Y. Yan, M. Zhang, W. Xu, X. Fan, H. Zhou, and P. Yang, ''You can recharge with detouring: Optimizing placement for roadside wireless charger,'' *IEEE Access*, vol. 6, pp. 47–59, 2018.
[13] L. He, Y. Gu, J. Pan, and T. Zhu, ''On-demand charging in wireless sensor networks: Theories and applications,'' in *Proc. 10th Int. Conf. Mobile Ad-Hoc Sensor Syst.*, Hangzhou, China, 2013, pp. 28–36.
[14] X. S. Yang, *Nature-Inspired Metaheuristic Algorithm*. Beckington, U.K.: Luniver Press, 2008.
[15] G. Sun, Y. Liu, M. Yang, A. Wang, and Y. Zhang, ''Charging nodes deployment optimization in wireless rechargeable sensor network,'' in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–6.
[16] L. He, Y. Zhuang, J. Pan, and J. Xu, ''Evaluating on-demand data collection with mobile elements in wireless sensor networks,'' in *Proc. IEEE 72nd Veh. Technol. Conf.*, Ottawa, ON, Canada, Fall 2010, pp. 1–5.
[17] L. He, L. Kong, Y. Gu, J. Pan, and T. Zhu, ''Evaluating the on-demand mobile charging in wireless sensor networks,'' *IEEE Trans. Mobile Comput.*, vol. 14, no. 9, pp. 1861–1875, Sep. 2015.
[18] A. Kaswan, A. Tomar, and P. K. Jana, ''An efficient scheduling scheme for mobile charger in on-demand wireless rechargeable sensor networks,'' *J. Netw. Comput. Appl.*, vol. 114, pp. 123–134, Jul. 2018.
[19] J. Zhu, Y. Feng, M. Liu, Z. Zhang, and C. Ma, ''Node failure avoidance mobile charging in wireless rechargeable sensor networks,'' in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–6.
[20] C. Lin, D. Han, J. Deng, and G. Wu, ''P^{2}S: A primary and passer-by scheduling algorithm for on-demand charging architecture in wireless rechargeable sensor networks,'' *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8047–8058, Sep. 2017.
[21] L. Fu, L. He, P. Cheng, Y. Gu, J. Pan, and J. Chen, ''ESync: Energy synchronized mobile charging in rechargeable wireless sensor networks,'' *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7415–7431, Sep. 2016.
[22] C. Zhao, H. Zhang, F. Chen, S. Chen, C. Wu, and T. Wang, ''Spatiotemporal charging scheduling in wireless rechargeable sensor networks,'' *Comput. Commun.*, vol. 152, pp. 155–170, Feb. 2020.
[23] A. Tomar, L. Muduli, and P. K. Jana, ''An efficient scheduling scheme for on-demand mobile charging in wireless rechargeable sensor networks,'' *Pervasive Mobile Comput.*, vol. 59, pp. 1574–1192, Oct. 2019.
[24] L. Khelladi, D. Djenouri, M. Rossi, and N. Badache, ''Efficient on-demand multi-node charging techniques for wireless sensor networks,'' *Comput. Commun.*, vol. 101, pp. 44–56, Mar. 2017.
[25] W. Xu, W. Liang, X. Jia, Z. Xu, Z. Li, and Y. Liu, ''Maximizing sensor lifetime with the minimal service cost of a mobile charger in wireless sensor networks,'' *IEEE Trans. Mobile Comput.*, vol. 17, no. 11, pp. 2564–2577, Nov. 2018.
[26] K.-L.-A. Yau, H. G. Goh, D. Chieng, and K. H. Kwong, ''Application of reinforcement learning to wireless sensor networks: Models and algorithms,'' *Computing*, vol. 97, no. 11, pp. 1045–1075, Nov. 2015.
[27] R. S. Sutton and A. G. Barto, ''Reinforcement learning: An introduction,'' *IEEE Trans. Neural Netw.*, vol. 9, no. 5, pp. 1054–1054, Sep. 1998.
[28] M. Nisio, ''Optimal control for stochastic partial differential equations and viscosity solutions of bellman equations,'' *Nagoya Math. J.*, vol. 123, pp. 13–37, Sep. 1991.
[29] H. Wu, G.-H. Tian, Y. Li, F.-Y. Zhou, and P. Duan, ''Spatial semantic hybrid map building and application of mobile service robot,'' *Robot. Auto. Syst.*, vol. 62, no. 6, pp. 923–941, Jun. 2014.
[30] Z. Wei, L. Fei, Z. Lyu, X. Ding, L. Shi, and C. Xia, ''Reinforcement learning for a novel mobile charging strategy in wireless rechargeable sensor networks,'' in *Proc. Int. Conf. Wireless Algorithm*, Tianjin, China, 2018, pp. 485–496.
[31] H. Iima and Y. Kuroe, ''Swarm reinforcement learning algorithms based on particle swarm optimization,'' in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Singapore, Oct. 2008, pp. 1110–1115.
[32] H. Iima and Y. Kuroe, ''Reinforcement learning through interaction among multiple agents,'' in *Proc. SICE-ICASE Int. Joint Conf.*, Busan, South Korea, 2006, pp. 1–6.

ZHEN WEI was born in 1965. He received the Ph.D. degree from the Hefei University of Technology, in 2005. He is currently a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology. His research interests include wireless communications and wireless sensor networks.

MENG LI was born in 1995. She received the B.S. degree from Anhui Polytechnic University, in 2017. She is currently pursuing the M.S. degree with the School of Computer Science and Information Engineering, Hefei University of Technology. Her main research interest includes wireless rechargeable sensor networks.



ZHENCHUN WEI was born in 1978. He received the Ph.D. degree from the Hefei University of Technology, in 2007. He is currently an Associate Professor with the School of Computer Science and Information Engineering, Hefei University of Technology. His research interests include wireless communications and wireless sensor networks.

LEI CHENG was born in 1971. He received the Ph.D. degree from the Hefei University of Technology, in 2003. He is currently an Associate Professor with the School of Computer Science and Information Engineering, Hefei University of Technology. His research interests include wireless communications and wireless sensor networks.

ZENGWEI LYU was born in 1989. He received the B.S., M.S., and Ph.D. degrees from the School of Computer Science and Information Engineering, Hefei University of Technology, in 2012, 2015, and 2019, respectively. His main research interest includes wireless rechargeable sensor networks.

FEI LIU was born in 1994. She received the M.S. degree from the Hefei University of Technology, in 2019, where she is currently pursuing the Ph.D. degree with the School of Computer Science and Information Engineering. Her main research interest includes wireless rechargeable sensor networks.