Federated Deep Reinforcement Learning-based Spectrum Access Algorithm with Warranty Contract in Intelligent Transportation Systems

Rongbo Zhu, *Member, IEEE*, Mengyao Li, Hao Liu, Lu Liu, and Maode Ma, *Senior Member, IEEE*

***Abstract*—Cognitive radio (CR) provides an effective solution to meet the huge bandwidth requirements in intelligent transportation systems (ITS), which enables secondary users (SUs) to access the idle spectrum of the primary users (PUs). However, the high mobility of users and real-time service requirements resulting in the additional transmission collisions and interference, which degrades the spectrum access rate and the quality of service (QoS) of users in ITS. This paper proposes a spectrum access algorithm (Feilin) based on federated deep reinforcement learning (FDRL) to improve spectrum access rate, which maximizes the QoS reward function with considering the hybrid benefits of delay, transmission power and utility of SUs. To guarantees the utility of SUs, the warranty contract is designed for SUs to obtain compensation for data transmission failure, which promotes SUs to compete for more spectrum resources. To meet the real-time requirements and improve QoS in ITS, a spectrum access model called FDQN-W is proposed based on federated deep Q-network (DQN), which adopts the asynchronous federated weighted learning algorithm (AFWLA) to share and update the weights of DQN in multiple agents to decrease time cost and accelerate the convergence. Detailed simulation results show that, in the multiuser scenario, compared with the existing methods, the proposed algorithm Feilin increases the spectrum access success rate by 15.1%, and reduces the collision rate with SUs and the collision rate with PUs by 46.4% and 6.8%, respectively.**

***Index Terms*—Intelligent transportation systems, federated deep reinforcement learning, spectrum access, warranty contract, quality of service.**

I. Introduction

With the advent of the 5G era and massive smart devices, various kinds of real-time communication and services are enabled in intelligent transportation systems (ITS) [1]. The significant increase of on-board units (OBU) and edge devices will generate a huge number of applications and spectrum requirements in ITS [2]. It is a challenge to support the substantial number of in-vehicle users and massive connections with high quality of service (QoS) including ultra-low latency and high reliability by the limited wireless spectrum resources [2]. Meanwhile, the existing schemes show that the licensed spectrums are idle in most time [3], which promoted the development of cognitive radio (CR) systems to effectively utilize the underutilized spectrum. CR enables secondary users (SUs) access the idle spectrum resources of the primary users (PUs) on the premise of uninterfering with PUs as much as possible, which perceives the available idle channels and avoids collisions and interference among users, thereby improving QoS of users in ITS.

To perceive the spectrum state accurately, optimization-based spectrum access schemes were addressed comprehensively [4, 5, 6]. A joint channel and power allocation scheme was proposed to consider the spatial-temporal change of vehicular mobility [4]. The combined impact of unlicensed vehicular user mobility and licensed user activity was analyzed and evaluated in [6]. To improve the spectrum efficiency, wireless power transmission was applied to roadside units (RSUs) in small devices deployed in vehicle-to-everything communications [7]. Due to the similarity between spectrum resource allocation and market economic behavior, the contract theory-based schemes were studied to improve spectrum efficiency. To realize a short-term spectrum sharing mechanism, a prototype of an online auction platform was designed [8]. A user-centric distributed spectrum sharing model was presented [9], which enables PUs to share their spectrums with users. Focusing on the spectrum allocation in multi-channel vehicle-to-vehicle communication, a game model based on the generalized Nash equilibrium was proposed to reduce the interference [10].

To predict spectrum state accurately, learning-based methods were explored to capture the dynamic mobility of vehicles in ITS. A multi-agent model-free reinforcement learning scheme called SARSA was proposed to allocate spectrum resources [11]. Q-learning scheme was adopted to solve the channel and power allocation problem [12]. Deep Q network (DQN) was utilized to determine an access policy from the observed states of channels [13]. To design a global optimization algorithm with dynamic spectrum access, a group-based multihop broadcast protocol was designed based on deep reinforcement learning (DRL) [14]. The Q-learning-based methods [11, 12, 13, 14] are effective for discrete action spaces. To handle continuous actions,

Manuscript received September 25, 2021; revised April 10, 2022; accepted May 22, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61772562, and the Fundamental Research Funds for the Central Universities under Grant 2662022JC004. *(Corresponding author: Rongbo Zhu.)*

Rongbo Zhu is with the College of Informatics, Huazhong Agricultural University, 430070 Wuhan, China (e-mail: rbzhu@mail.hzau.edu.cn).

Mengyao Li and Hao Liu are with the College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China (e-mail: 2020110260@mail.scuec.edu.cn; 2019110249@mail.scuec.edu.cn).

Lu Liu is with the School of Informatics, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: l.liu@leicester.ac.uk).

Maode Ma is with the College of Engineering, Qatar University, Doha, Qatar. (e-mail: mamaode@qu.edu.qa).

reinforcement learning models based on the policy optimization were studied. Based on the deterministic policy [15], the deep deterministic policy gradient (DDPG) was developed for continuous control [16]. To obtain the solution of the minimization problem by learning stochastic and deterministic approximate optimal policies, a regularized dual-averaging policy gradient (RDA-PG) scheme was proposed [17]. However, the learning-based methods mentioned above depend on the centralized model training, which increases transmission overhead and degrades the real-time performance.

To learn optimal spectrum access strategies in a distributed manner, the reservoir computing recurrent neural network (RNN) was utilized to realize DRL by taking advantage of the underlying temporal correlation [18]. To achieve low latency, a task offloading scheme based on federated learning (FL) was designed [19], where the vehicles and RSU can share a common learning model to reduce the learning cost. To speed up the convergence of FL caused by the unbalanced data [20], the data compression methods were utilized in Internet of vehicles (IoV) [21, 22]. To address the complex and dynamic control issues, a decentralized framework based on federated deep reinforcement learning (FDRL) was proposed with cooperative edge caching [23].

Although there is a large body of work on spectrum access in ITS, most of them are limited to spectrum resource allocation. It remains open how to develop an efficient spectrum access scheme that can degrade transmission collision and interference and maximize the spectrum utilization to meet the mobility of vehicles and real-time service requirements in ITS. This stagnation underscores the technical challenges in the exploration of cooperative and game design from a SU's perspective, which we describe as follows. First, with the significant increase of OBUs and edge devices in ITS, the requirements for spectrum resources aggravate the contradiction of limited spectrum between SUs and PUs. Especially for SUs, severe competition of spectrum resources leads to additional transmission collisions and interferences. Most of the existing methods do not consider how to ensure SU's utility while improving spectrum access, exacerbating the phenomenon of spectrum scarcity. Second, the mobility of vehicles and real-time service requirements raise the QoS level in ITS. It is difficult to cope with complex dynamic networks using traditional spectrum allocation methods. The autonomous learning capacity of reinforcement learning can accurately predict the state transition probability in complex dynamic ITS. Hence it is very suitable for spectrum holes discovery and spectrum access in CR. However, most existing DRL-based schemes rely on the centralized models with long-time training, which are hard to meet the real-time requirements in ITS. Therefore, a decentralized real-time learning-based spectrum access scheme is needed. However, developing a distributed learning-based spectrum access algorithm together with guaranteeing the utility of SUs and QoS can easily become intractable and is a challenging task.

Focusing on improving the spectrum access rate and QoS in ITS, this paper fills the gap by developing an efficient spectrum access mechanism with the warranty contract by utilizing FDRL to provide real-time services. The main contributions of this paper are summarized as follows:

(1) A FDRL-based spectrum access algorithm (Feilin) is proposed, which effectively integrates the warranty contract and the spectrum access model FDQN-W to improve the spectrum access rate in ITS.

(2) A warranty contract is designed to promote SUs competing for more spectrum resources and guarantee the utility of SUs. In the warranty contract, the RSU obtains the transmission failure probability of all users in the cluster in real-time, and calculates the rewards and losses of transactions between the PU and SUs according to the transmission failure probability. The PU designs warranty contracts for all SUs, and each SU maximizes its own utility by selecting the optimal warranty contract.

(3) An efficient spectrum access model called FDQN-W is proposed based on federated DQN (FDQN) to meet the real-time requirements and improve QoS in ITS. FDQN-W takes the delay, transmission power and the utility of SU as the reward function, which adopts an asynchronous federated weighted learning algorithm (AFWLA) to share and update the weights of DQN in multiple agents to speed up the convergence. To reduce the waiting time of weights uploading in FDQN, AFWLA adaptive selects the number of aggregated local models and updates the parameters of the global model according to the percentage of accuracy of each aggregated local model.

The rest of this paper is organized as follows. Section II reviews the related work. In Section III, the system model and the spectrum access algorithm Feilin are presented. Sections IV and V present the warranty contract and FDQN-W, respectively. The simulation results and analysis are presented in Section VI, and finally, Section VII concludes the paper.

II. Related Work

The scarcity of spectrum resources and real-time service requirements in ITS require spectrum sharing with existing wireless communication systems. Existing work focuses on optimization-based spectrum resource allocation and sharing schemes, and learning-based spectrum access algorithms, which provides potential feasible methods to improve the QoS of users in ITS.

A. Optimization-based Spectrum Allocation and Sharing Schemes

Focusing on the mobility in high speed ITS, an emotionally inspired cognitive agent was introduced [24], which adopts a probabilistic and deterministic finite automaton using a fear factor. To improve the security of information transmission, a channel sensing scheme was proposed [25], which evokes the trust of its cognitive users (CUs) by analyzing predefined attributes. To relax the constraints of software defined radio (SDR) deployments, a resource allocation approach was developed and evaluated [26]. To determine the idle spectrum and estimate the channel quality, a system called V-Scope was presented [27], which utilizes spectrum sensors on public vehicles to collect and report measurements on the road and builds various models. Considering the spectrum sensing and access of multi-channel optimal opportunities for full-duplex radio, a joint learning and spectrum access scheme was proposed to maximize the throughput [28].

However, in practical scenarios of ITS, the huge number of edge devices aggravate the scarcity of spectrum. It is necessary to ensure that SUs try to access the idle spectrum without interfering with PUs. Hence, efficient game and incentive mechanisms are good ways to promote SUs and PUs to share spectrums. A content sharing framework was proposed [29], which combines contract theory and Lyapunov optimization to design a new random incentive method. To ensure PUs are compensated for sharing their licensed bands, a blockchain-based platform was proposed [30]. To solve the collisions and unfair channel allocation, a contract theory-based bargaining approach was proposed in a centralized manner [31]. To maximize the operation efficiency of devices, a communication channel allocation and resource optimization scheme was proposed based on spectrum clustering and non-cooperative game [32]. To effectively utilize the limited spectrum resource, a heterogeneous spectrum allocation scheme based on three-phase bargaining game was proposed [33]. The game and contract-based schemes mentioned above provide efficient solutions to promote PUs and SUs to share idle spectrums.

B. *Learning-based Spectrum Access Algorithms*

Due to the mobility of vehicles and users in ITS, it is difficult to acquire accurate channel state information (CSI). In addition, it is an intractable and challenging task to establish an accurate model to depict the system with massive devices in ITS. Due to the powerful prediction capacity, learning-based schemes were explored to improve spectrum access performance [34, 35]. To solve the complex dynamic resource allocation problem, a virtualized framework was proposed [36], which adopted a high-performance asynchronous advantage actor-critic learning algorithm. A Q-learning-based spectrum access scheme was proposed to adaptively distribute multimedia data on free spectrum holes [37]. To reduce the overestimation of action-value function, multi-pseudo Q-learning was adopted for the continuous action space [38], which utilizes sub greedy policy to replace the greedy policy in Q-learning. A DQN-based spectrum sensing strategy was designed to overcome the challenges of unknown dynamics and prohibitive computation and maximize the expected long-term successful transmissions [39]. The schemes mentioned above provide efficient ways to solve the problem of single-user spectrum access by the centralized learning models in ITS.

Considering the high-reliability services in multiusers scenario, a federated edge caching framework was proposed to solve complex dynamic control and caching problems [40]. To improve the convergence speed of the FL algorithm [41], a momentum federated learning algorithm was proposed [42], which integrates the momentum gradient descent (MGD) method on a central server. To select a subset of clients with significant weight updates, the optimal sampling strategy of FL was proposed with an Ornstein-Uhlenbeck process [43]. A vertical FL-based cooperative sensing scheme was proposed to improve spectrum sensing and data privacy-preserving capability [44]. To overcome spectrum scarcity, a deep learning approach was proposed for modeling the resource allocation problem [45], which addresses a non-cooperative spectrum access problem in different environments.

Although there are many effective spectrum allocation and access methods in ITS, most schemes ignore the utility of SUs and the distributed dynamic mobility of users and real-time service requirements in ITS [46]. This paper fills this gap by developing a FDRL-based spectrum access algorithm to overcome the limitations mentioned above in ITS.

III. The Proposed Algorithm

A. *The Problem Description*

Consider a hierarchical system consisting of one base station (BS), K RSUs, M PUs and J SUs in ITS. Let N denote the number of spectrum channels. The system architecture of the proposed spectrum access algorithm is shown in Fig. 1, which includes multiple clusters. Table I lists the notations that we use in this paper.

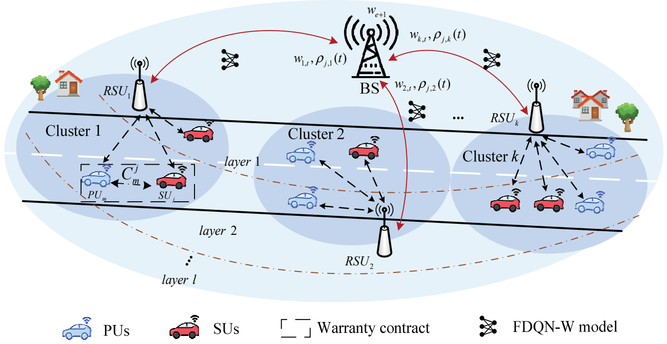


Fig. 1. The system architecture.

TABLE I. Notations

Symbol	Description
$C_{m,t}^j$	The warranty contract of SU_j provided by PU_m at t
$\rho_j^l(t)$	The transmission failure ratio of layer l
$\rho_{j,k}(t)$	The packets transmission failure of RSU_k
T_0	The maximum tolerance delay
$p(DT_{n,t} \leq T_0)$	The probability of transmission delay no more than T_0
$q_{n,t}$	The percentage of loaded data of SU_j
$R_{n,t}$	The transmission rate on the channel n at time t
$P_{n,t}$	The transmission power on the channel n at time t
$P_{\lim}$	The maximum transmission power of SU_j
$R_{\lim}^n$	The capacity of the channel n
$Pr_{m,t}$	The channel selling price of PU_m
$L_{j,t}$	The loss of SU_j with the failure transmission
$I_{m,t}^j$	The warranty fee that SU_j pays to PU_m
$F_{m,t}^j$	The net compensation of SU_j paid by PU_m
$H_{m,t}^j$	The total compensation of SU_j paid by PU_m
$U_{S,t}$	The transmission success reward of SU_j
$U_{F,t}$	The transmission failure reward of SU_j
$Pr_{m,t}$	The best-selling price for available channels
$U_{SU}(C_{m,t}^j)$	The utility of SU_j with signing a warranty contract
$U_{SU}(C_{0,t})$	The utility of SU_j without signing a warranty contract
R_1	The negative reward of a collision with any PU
R_2	The negative reward of a collision with any SU
N_m	The number of samples in the transition memory

The BS hierarchizes users according to the system status, aggregates, updates and shares parameters of the proposed FDQN-W model with each RSU. The cluster is formed by the communication range of the RSU. In each cluster, there is a cluster head RSU_k ($k\in[1, K]$) that collects data packets from SUs and PUs. The PU designs the warranty contacts for all SUs in the same cluster, and the SU will choose the optimal warranty contact to access the idle channel without interfering with the communication of the PUs.

The proposed algorithm Feilin includes two phases: spectrum sensing and spectrum access. In the spectrum sensing phase, the BS classifies SUs and PUs into different levels according to the signal-to-noise ratio (SNR) periodically. RSU_k senses and evaluates the CSI, including delay, power, and data transmission failure ratio. Then PU_m designs the warranty contract $C^j_{m,t}$ for SU_j based on the idle spectrum information and the current CSI in the corresponding cluster. In the spectrum access phase, the transmission collision and interference are considered in FDQN-W. If and only if the SU successfully receives the packet, it sends an acknowledgement (ACK) frame to the corresponding transmitter. Based on the ACK information, SUs count transmission failure ratio $\rho^l_j(t)$ and predict the channel state. The weight w^t_k of FDQN-W and $\rho_{j,k}(t)$ in RSU_k is shared and updated by the BS. RSU_k adopts the proposed model FDQN-W to share the idle spectrum resources with SUs according to the QoS requirements of different traffics, and adjusts the transmission power and other parameters of the corresponding SU_j in ITS.

To improve the spectrum access rate and QoS in ITS, Feilin considers the comprehensive benefits of delay, power consumption and SU utility. Define the reward function $QoS(t)$ at time t as:

$$QoS(t)=\lambda\sum_{n=1}^{N}p(DT_{n,t}\le T_0)\cdot q_{n,t}+\beta\frac{\sum_{n=1}^{N}R_{n,t}}{\sum_{n=1}^{N}P_{n,t}}+\theta U_{SU}(C^j_{m,t}) \quad (1)$$

where λ , β and θ are three weight parameters, and satisfy the sum is 1, $p(DT_{n,t} \le T_0)$ is the probability that the transmission delay $DT_{n,t}$ of SU_j on the channel n is less than maximum tolerance delay T_0, $R_{n,t}$ is the transmission rate on the channel n, $P_{n,t}$ is the transmission power on the channel n, $C^j_{m,t}$ denotes the warranty contract of SU_j designed by PU_m, and $q_{n,t}$ is the percentage of loaded data of SU_j on the channel n:

$$q_{n,t}=\frac{R_{n,t}}{\sum_{n=1}^{N}R_{n,t}} \quad (2)$$

The goal of the proposed spectrum access algorithm Feilin is to maximize the reward function $QoS(t)$ under the transmission power constraint $C1$ and transmission rate constraint $C2$ as:

$$\begin{aligned}P:&\quad \max QoS(t)\quad s.t.\\ C1:&\quad \sum_{n=1}^{N}P_{n,t}\le P_{\text{lim}}\\ C2:&\quad \sum_{j=1}^{J}R_{n,t,j}\le R^n_{\text{lim}}\end{aligned} \quad (3)$$

where P_{lim} denotes the maximum transmission power of SU_j, R^n_{lim} is the maximum transmission rate on the channel n, and $R_{n,t,j}$ represents the transmission rate of SU_j on the channel n.

B. Algorithm Description

To solve the optimization problem mentioned above, the proposed spectrum access algorithm Feilin maximizes the QoS reward with considering the hybrid benefits of delay, transmission power and utility in ITS. The pseudo-code and flowchart of Feilin are shown in Algorithm 1 and Fig. 2, respectively.

```
Algorithm 1. The spectrum access algorithm Feilin in ITS
Input: $P_{\text{lim}}$, $R^n_{\text{lim}}$
Output: Spectrum access selection set $A_{k,t}=\{a^1_{k,t},a^2_{k,t},...,a^j_{k,t}\}$, $QoS(t)$
begin
  $RSU_k$ senses and acquires the channel state information $s^j_{k,t}$;
  for each $PU_m$ in $RSU_k$
      $PU_m$ designs the warranty contract $C^j_{m,t}$ for $SU_j$;
      $SU_j$ calculates the utility $U_{SU}(C^j_{m,t})$
  end for
  for each $SU_j$ in $RSU_k$
      $SU_j$ selects the optimal warranty contract $C^j_{m,t}$;
  end for
  for each $RSU_k$
      $RSU_k$ utilizes FDQN-W ( $s^j_{k,t}$, $P_{\text{lim}}$, $R^n_{\text{lim}}$, $U_{SU}(C^j_{m,t})$ ) to obtain $A_{k,t}$
      $=\{a^1_{k,t},a^2_{k,t},...,a^j_{k,t}\}$, and dispatches $A_{k,t}$ to SUs;
  end for
  SUs access the channel according to $A_{k,t}$, and $QoS(t)$ is obtained.
end
```

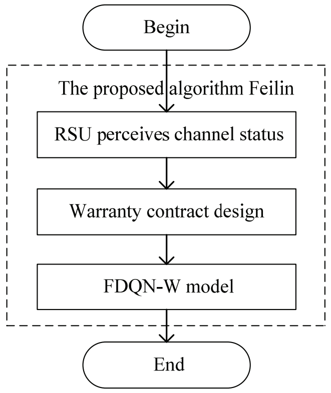


Fig. 2. The flowchart of the proposed algorithm Feilin.

In the proposed algorithm Feilin, each RSU senses and collects channel state information $s^j_{k,t}$, aggregates data packets from SUs and PUs, and obtains the delay and transmission failure ratio according to the ACK and SNR. Then the BS classifies SUs and PUs into different layers according to the SNR and determines the range of transmission failure ratio of each layer. RSU_k calculates the reward and loss of spectrum sharing between PUs and SUs in the same cluster. PU_m in RSU_k designs the warranty contract $C^j_{m,t}$ for SU_j that tries to access the idle spectrum. SU_j calculates the utility $U_{SU}(C^j_{m,t})$ and determines the optimal warranty contract $C^j_{m,t}$. Finally, RSU_k adopts the spectrum access model FDQN-W, whose weights are downloaded from the BS periodically, and performs local training to update the parameters w^t_k by minimizing the loss function and adjusts the $QoS(t)$ according to the CSI $s^j_{k,t}$. In the local training, FDQN-W takes the delay, transmission power and $U_{SU}(C_{m,t}^{j})$ as the reward function. In the federated aggregation, AFWLA is proposed to adaptive select the number of aggregated local models to reduce time cost. We can obtain the final spectrum selection set $A_{k,t}=\{a_{k,t}^{1},a_{k,t}^{2},...,a_{k,t}^{j}\}$ for each SU in RSU_k, where $a_{k,t}^{j}$ ($a_{k,t}^{j}\in[0,N]$) is the action that SU_j chooses at time t in RSU_k. Finally, we can get the optimal reward function $QoS(t)$ according to the formula (1).

IV. Warranty Contract Design

To design the optimal warranty contract for SUs with considering the transmission success ratio and interference in ITS, the BS layers SUs and PUs according to the distance, transmission rate and signal-to-noise ratio (SNR). Then PU_m designs warranty contracts for the SUs in the same cluster. The procedure of the warranty contract design is shown in Fig. 3.

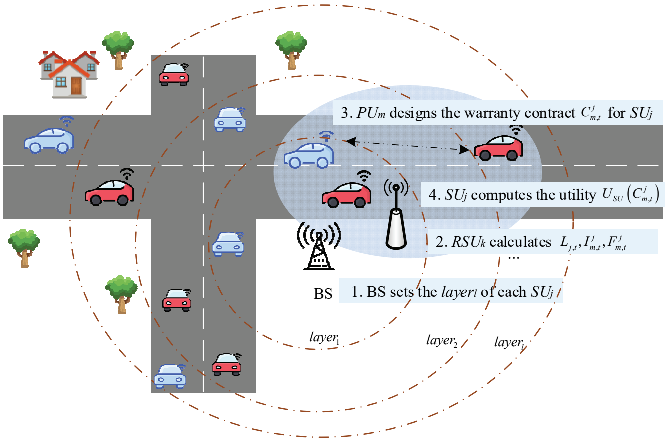

3. PU_m designs the warranty contract $C_{m,t}^{j}$ for SU_j
4. SU_j computes the utility $U_{SU}\left(C_{m,t}^{j}\right)$
2. RSU_k calculates $L_{j,t},I_{m,t}^{j},F_{m,t}^{j}$
...
BS
1. BS sets the $layer_l$ of each SU_j
$layer_1$
$layer_2$
$layer_l$

Fig. 3. The procedure of the warranty contract design in ITS.

1) The BS sets the layer of each SU_j

To depict the transmission interference in ITS, we calculate SNR_j^n of the channel n for SU_j:

$$SNR_j^n=\frac{g_j^nP_j^n}{\sigma^2+I_j^n} \tag{4}$$

where g_j^n represents the channel power gain between SU_j and the BS assigned to the channel n, P_j^n is the transmission power allocated to the channel n between SU_j and the BS, σ^2 is noise power, and I_j^n indicates the interference as:

$$I_j^n=\sum_{m=1}^{M}P_m^ng_m^n+\sum_{y=1,y\neq j}^{J}P_y^ng_y^n \tag{5}$$

where M is the number of PUs, the first term on the right side is the total interference of PUs, the second term is the interference of other SUs, P_m^n represents the transmission power allocated to the channel n between PU_m and the BS, g_m^n denotes the channel power gain between PU_m and the BS assigned to the channel n, P_y^n is the transmission power allocated to the channel n between other SUs and the BS, and g_y^n represents the channel power gain between other SUs and the BS assigned to the channel n.

Each RSU_k counts the number $N_{f,k}(t)$ of packet transmission failure of all SUs during the current period in the cluster k, and calculates the probability of packets transmission failure $\rho_{j,k}(t)$ as:

$$\rho_{j,k}(t)=N_{f,k}(t)/N_{p,k}(t) \tag{6}$$

where $N_{p,k}(t)$ is the total number of transmitted packets during the current period in the cluster k.

RSU_k uploads the transmission failure ratio $\rho_{j,k}(t)$ to the BS, and the BS calculates the global transmission failure probability $\rho_j(t)$ as:

$$\rho_j(t)=\frac{1}{K}\sum\nolimits_{k=1}^{K}\rho_{j,k}(t)\,. \tag{7}$$

Then the BS sets the transmission failure ratio $\rho_j^l(t)\in(\rho_{j,\min}^l(t),\rho_{j,\max}^l(t)]$ of the corresponding layer l according to the transmission failure ratio as:

$$\rho_{j,\max}^l(t)=\rho_j(t)*(L-l+1)/L \tag{8}$$

$$\rho_{j,\min}^l(t)=\rho_j(t)*(L-l)/L \tag{9}$$

where L denotes the total number of layers in the system.

According to SNR_j^n, we can get the corresponding transmission rate by the Shannon formula. To guarantee the spectrum access, the current channel quality is quantified by l of users and transmission failure ratio. We can get l of each user, transmission rate $R_{n,t}$, and the range of transmission failure ratio $\rho_j^l(t)$ under different channel condition according to the SNR_j^n. With formulas (4), (5), (8) and (9), we can obtain the layer l of each user as shown in Table II.

TABLE II
Layer Under Different Channel Conditions in ITS

l	SNR_j^n (dB)	$R_{n,t}$ (Mbps)	The transmission failure ratio
1	-10 ~ -6	25 ~ 31	$\rho_j^l(t)$
2	-6 ~ -0.5	31 ~ 36	
3	-0.5 ~ 5	36 ~ 41	
4	5 ~ 7	41 ~ 44	
5	7~ 9	44 ~ 48	
6	9 ~ 10.5	48 ~ 54	
7	10.5 ~ 14.5	54 ~ 59	
8	14.5 ~ 18	59 ~ 66	
9	>18	66 ~ 70	

2) Calculation of the rewards and the loss

Let $L_{j,t}$ denote the loss of SU_j with the failure transmission, we have:

$$L_{j,t}=|(U_{S,t}^j-Pr_{m,t})-(U_{F,t}^j-Pr_{m,t})|=|U_{S,t}^j-U_{F,t}^j| \tag{10}$$

where $U_{S,t}^j$ and $U_{F,t}^j$ are the transmission rewards with successful and failure transmission respectively, which are linearly related to the successful transmission ratio $\rho_j^l(t)$ as:

$$U_{S,t}^j=R_{n,t}(1-\rho_j^l(t))\Delta t \tag{11}$$

$$U_{F,t}^j=R_{n,t}\rho_j^l(t)\Delta t \tag{12}$$

where Δt is the transmission slot.

Let $I_{m,t}^j$ represent the warranty fee, $F_{m,t}^j$ denote the net compensation when the transmission of SU_j fails, we have:

$$I_{m,t}^{j}=\rho_{j}^{l}(t)L_{j,t} \tag{13}$$

$$F_{m,t}^{j}=(1-\rho_{j}^{l}(t))L_{j,t} \tag{14}$$

3) Design the warranty contract

To promote SUs access the idle spectrum of PUs, we consider the spectrum leasing from the perspective of SUs in ITS. PU_m designs a warranty contract $C_{m,t}^{j}$ for SU_j with l in the same cluster. Then we can get the set $C_{m,t}=(C_{m,t}^{1},C_{m,t}^{2},\cdots,C_{m,t}^{j},\cdots,C_{m,t}^{J}),\forall m\in\{1,\cdots,M\}$, and the total compensation $H_{m,t}^{j}$ is:

$$H_{m,t}^{j}=I_{m,t}^{j}+F_{m,t}^{j} \tag{15}$$

4) Calculation of the utility of SU_j

The utility U_{PU_m} of PU_m selling the channel to SU_j can be expressed as:

$$U_{PU_m}=Pr_{m,t}b_{m}+(1-\rho_{j}^{l}(t))I_{m,t}^{j}-\rho_{j}^{l}(t)F_{m,t}^{j}-C(b_{m}) \tag{16}$$

where b_m is the bandwidth shared by PU_m and $C(b_m)$ is the channel selling cost of PU_m:

$$C(b_{m})=(b_{m})^{\tau} \tag{17}$$

where τ is a natural number.

Let $b_{\min}$ and $b_{\max}$ denote the lower and upper bounds of the spectrum sharing capabilities of all PUs, respectively. Take the partial derivative of U_{PU_m} with respect to the spectral bandwidth b_m, we have:

$$\frac{\partial U_{PU_m}}{\partial b_{m}}=Pr_{m,t}-\tau\cdot(b_{m})^{\tau-1} \tag{18}$$

Let the derivative $\frac{\partial U_{PU_m}}{\partial b_{m}}$ =0, we can get the best-selling price $Pr_{m,t}$ for available channels:

$$Pr_{m,t}=\tau(b_{m})^{\tau-1} \tag{19}$$

$Pr_{m,t}$ is convex when its exponent is between 0 and 1, namely $1<\tau<2$ to ensure that $C(b_m)$ is convex.

The utility $U_{SU}(C_{m,t}^{j})$ of SU_j buying the channel of PU_m can be expressed as:

$$\begin{aligned}U_{SU}(C_{m,t}^{j})&=(1-\rho_{j}^{l}(t))\ln(U_{S,t}^{j}-Pr_{m,t}b_{m}-I_{m,t}^{j})\\&+\rho_{j}^{l}(t)\ln(U_{F,t}^{j}-Pr_{m,t}b_{m}+F_{m,t}^{j})\end{aligned} \tag{20}$$

The expected utility $U_{SU}(C_{0,t})$ when SU_j does not sign the insurance contract is:

$$U_{SU}(C_{0,t})=(1-\rho_{j}^{l}(t))\ln(U_{S,t}^{j}-Pr_{m,t}b_{m})+\rho_{j}^{l}(t)\ln(U_{F,t}^{j}-Pr_{m,t}b_{m}) \tag{21}$$

V. The Proposed Model FDQN-W

A. Model Building

Due to the mobility of vehicles and distributed features of ITS, it is difficult to acquire the accurate CSI. To meet the real-time requirements and improve QoS in ITS, FDQN-W takes delay, transmission power and the utility of each SU as the reward function, which adopts AFWLA to share and update the weights of DQN in RSUs to speed up the convergence. The proposed model is shown in Fig. 4.

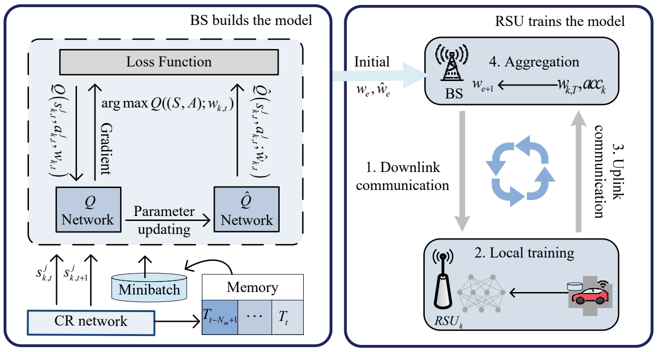


Fig. 4. The proposed model FDQN-W.

As shown in Fig. 4, FDQN-W is made up of two components: Q-network (MainNet) and $\hat{Q}$ -network (TargetNet). The former is used to select a spectrum selection action, and the latter is utilized for performance evaluation. The parameters $\hat{w}_t$ in the $\hat{Q}$ -network periodically are updated along with w_t in the Q-network. There is a transition memory that maintains N_m training samples in FDQN-W.

In the model building phase, the BS random initializes w , and sets $w_t=w$ and $\hat{w}_t=w$, which are the weights of multilayer perception (MLP) in Q-network and $\hat{Q}$ -network, respectively. MainNet is used to select the system with the information tuple $T_t=(s_{k,t}^{j},a_{k,t}^{j},r(s_{k,t}^{j},a_{k,t}^{j}),s_{k,t+1}^{j})$, which represents the current training sample. Then the BS dispatches w_t and $\hat{w}_t$ to each RSU.

B. Model Training

Algorithm 2. Model training and the optimal action algorithm

```
Input: $s_{k,t}^{j}$ , $P_{\lim}$ , $R_{\lim}^{n}$ , $U_{SU}(C_{m,t}^{j})$
Output: Spectrum access selection set $A_{k,t}=\{a_{k,t}^{1},a_{k,t}^{2},...,a_{k,t}^{j}\}$
begin
RSU_k obtains $w_e,\hat{w}_e$ from the BS.     //Downlink communication
RSU_k obtains $a_{k,t}^{j}$ according to $s_{k,t}^{j}$ with the expression (22), and
  $r(s_{k,t}^{j},a_{k,t}^{j})$ with the expression (23).  //Local training
RSU_k saves the four-tuple $T_t=(s_{k,t}^{j},a_{k,t}^{j},r(s_{k,t}^{j},a_{k,t}^{j}),s_{k,t+1}^{j})$ in $N_m$.
 if epoch % $N_m=0$
      for $t=1,2\cdots T$ do // The number $T$ of iteration
          RSU_k randomly select some samples from $M$ for the batch
           training.
          Update $w_{k,t}$ in the $Q$ network and $\hat{w}_{k,t}$ in the $\hat{Q}$ network
           according to $\nabla_{w_{k,t}}f(w_{k,t})$ .
          Local training according to the expression (27);
      end for
      RSU_k uploads $w_{k,T}$ to the BS.     //Uplink communication
      After receiving $w_{e+1}$ according to AFWLA, RSU_k updates its
       weight with $w_{e+1}$ .               //Aggregation
end if
end
```

Each RSU trains the model FDQN-W and obtains the optimal spectrum selection action set. The process of model training and the optimal spectrum selection action in FDQN-W includes four phases as shown in algorithm 2: downlink communication phase, local training phase, uplink communication phase and aggregation phase.

1) Downlink communication phase

In the downlink communication phase, RSUs download the global parameters from the BS. *RSU*$_k$ obtains w_e and $\hat{w}_e$ from the BS in *e*-th communication round $e \in [1, E]$, where *E* represents the number of total aggregation.

2) Local training phase

To determine the optimal action $a_{k,t}^j$, *RSU*$_k$ balances the short-term gains and the long-term gains to get the most benefit in the local training process of DQN. Considering that the average reward of actions is unknown in DQN, to avoid getting stuck in a local optimal solution, the ε-greedy algorithm is adopted to optimize the action selection process, and we have:

$$a_{k,t}^j = \begin{cases} randn(N), & rand < \varepsilon \\ \arg\max Q(s_{k,t}^j, a_{k,t}^j), & rand \geq \varepsilon \end{cases} \tag{22}$$

where *rand* is a random number (*rand*∈ [0,1]), and $Q(s_{k,t}^j, a_{k,t}^j)$ refers to the reward when the agent takes the action $a_{k,t}^j$ under the state $s_{k,t}^j$.

When *RSU*$_k$ takes action $a_{k,t}^j$ upon $s_{k,t}^j$, it can obtain the feedback reward $r(s_{k,t}^j, a_{k,t}^j)$:

$$r(s_{k,t}^j, a_{k,t}^j) = \begin{cases} QoS(t), & \text{Without collision,} \\ R_1, & \text{Collision with any PU,} \\ R_2, & \text{Collision with any SU,} \end{cases} \tag{23}$$

where R_1 and R_2 denote the negative rewards when SU collides with any PU and SU, respectively.

After obtaining $r(s_{k,t}^j, a_{k,t}^j)$, *RSU*$_k$ saves the transition T_t in the transition memory with the size N_m in each training agent, where $T_t = (s_{k,t}^j, a_{k,t}^j, r(s_{k,t}^j, a_{k,t}^j), s_{k,t+1}^j)$, and N_m is updated by the most recent transitions.

In the iterative process, the task is to find the objective parameters $w_{k,t}$ with the samples in the transition memory to minimize the overall loss of *RSU*$_k$.

RSU$_k$ randomly selects a minibatch samples (S_n <=N_m) from the transition memory for batch training, and gets $Q(s_{k,t}^j, a_{k,t}^j; w_{k,t})$ of each sample by T_t:

$$\begin{aligned} Q(s_{k,t}^j, a_{k,t}^j; w_{k,t}) &= (1-\alpha) Q(s_{k,t}^j, a_{k,t}^j; w_{k,t}) \\ &+ \alpha[r(s_{k,t}^j, a_{k,t}^j) + \gamma \max_{a_{t+1\in A}} Q(s_{k,t+1}^j, a_{k,t+1}^j; w_{k,t+1})] \end{aligned} \tag{24}$$

where $\alpha \in [0,1)$ denotes the learning rate, γ represents the discount factor which determines the influence of the future feedback on the current decision, and we have $\gamma \in (0,1)$.

To minimize the overall loss of *RSU*$_k$, we get the loss function $f(w_{k,t})$ as:

$$\begin{aligned} f(w_{k,t}) &= (r(s_{k,t}^j, a_{k,t}^j) - Q(s_{k,t}^j, a_{k,t}^j; w_{k,t}) \\ &+ \gamma \cdot \hat{Q}(s_{k,t}^j, \arg\max_{a_{k,t+1}^j} Q(s_{k,t+1}^j, a_{k,t+1}^j; w_{k,t}); \hat{w}_{k,t})^2 \end{aligned} \tag{25}$$

Then we can get the gradient $\nabla_{w_{k,t}} f(w_{k,t})$ of $w_{k,t}$ to be updated as:

$$\begin{aligned} &\nabla_{w_{k,t}} f(w_{k,t}) = [(r(s_{k,t}^j, a_{k,t}^j) - Q(s_{k,t}^j, a_{k,t}^j; w_{k,t}) + \\ &\gamma \cdot \hat{Q}(s_{k,t}^j, \arg\max_{a_{k,t+1}^j} Q(s_{k,t+1}^j, a_{k,t+1}^j; w_{k,t}); \hat{w}_{k,t}) \cdot \nabla_{w_{k,t}} Q(s_{k,t}^j, a_{k,t}^j; w_{k,t})] \end{aligned} \tag{26}$$

RSU$_k$ updates the local model parameters $w_{k,t+1}$ as:

$$w_{k,t+1} = w_{k,t} - \eta \cdot \nabla_{w_{k,t}} f(w_{k,t}) \tag{27}$$

where $\eta \geq 0$ is the step size.

RSU$_k$ repeats the steps mentioned above to get an optimal learning model. By updating the parameters $w_{k,t}$ of MLP, we can get the approximate optimal *Q* value $Q(S, A)$ as:

$$Q(S, A) \approx Q((S, A); w_t) \tag{28}$$

where *S* and *A* denote the set of $s_{k,t}^j$ and $a_{k,t}^j$. $Q(s_{k,t}^j, a_{k,t}^j; w_{k,t})$ and $\hat{Q}(s_{k,t}^j, a_{k,t}^j; \hat{w}_{k,t})$ are saved by the *RSU*$_k$, and the parameters $\hat{w}_{k,t}$ are periodically updated in the $\hat{Q}$ network with $w_{k,t}$ in the Q network. *RSU*$_k$ records local model accuracy acc_k every *T* iterations.

3) Uplink communication phase

In the uplink communication phase, each RSU not only trains the local model, but also uploads the local parameters to the BS for the federated aggregation, as shown in Fig. 4. Due to the different computing and communication overhead of each RSU, the number nup_e of aggregated local models has a crucial effect on time cost in this phase. Therefore, it is necessary to consider the number of participants in aggregation. After *T* iterations of local model training, *RSU*$_k$ uploads parameters $w_{k,T}$ and acc_k to the BS.

4) Aggregation phase

In the aggregation phase, AFWLA is proposed to aggregate and update the global parameters w_{e+1} in BS, which utilizes the average accuracy $Acc_{avg,e}$ of the uploaded local training model as the benchmark for each aggregation.

For the (*e*+1)-th round federated aggregation, the BS updates the number nup_{e+1} of uploaded $w_{k,T}$ and acc_k until there is n_e models whose accuracies are all higher than the last round $Acc_{avg,e}$ (for the first round aggregation, we set $nup_1 = n_1 = K$), and we can get:

$$Acc_{e+1} = \sum_{k=1}^{nup_{e+1}} acc_k \tag{29}$$

$$w_{e+1} = \sum\nolimits_{k=1}^{nup_{e+1}} \frac{acc_k}{Acc_{e+1}} w_{k,T} \tag{30}$$

Calculate the average accuracy $Acc_{avg,e+1}$ of the updated models as:

$$Acc_{avg,e+1} = \frac{1}{nup_{e+1}} \sum_{k=1}^{nup_{e+1}} acc_k \tag{31}$$

Let n_{e+1} denote the number of the updated model's accuracy is higher than $Acc_{avg,e+1}$, we can obtain:

$$n_{e+1} = \sum_{k=1}^{nup_{e+1}} count(acc_k > Acc_{avg,e+1}) \tag{32}$$

where $count(\cdot)$ denotes the binary function, and the value is 1 if $acc_k > Acc_{avg,e+1}$, otherwise it is 0.

In AFWLA, only the RSU that has reached *T* rounds local training will upload the parameters $w_{k,T}$ and acc_k. All RSUs update the global parameters w_{e+1} and begin the next *T* rounds local training.

C. Performance Analysis

Let w_* denote the optimal solution, we have the following assumptions [47]:

Assumption 1: For all k, we assume the following:

1) $f(w_k)$ is convex;

2) $f(w_k)$ is L-smooth, i.e., for any w_k and w_k', $f(w_k') \le f(w_k) + \nabla f(w_k) \cdot (w_k' - w_k) + (L/2)\|w_k - w_k'\|^2$.

The feasibility of the linear regression and the update rule of FDQN-W are guaranteed by Assumption 1. We then have the lemma as follow:

Lemma 1: $f(w)$ is μ-strongly convex and L-smooth.

Proof: Straightforward from Assumption 1, according to the definition of convex, $f(w)$ is the finite-sum structure of $f(w_k)$ and triangle inequality.

Theorem 1: Considering that $f(w)$ is L-smooth and μ-strongly, let $\eta_t = 1/L$ and $w_* = \arg\min f(w)$, we have

$$\|w_t - w_*\| \le (1 - \frac{\mu}{L})^t \|w_1 - w_*\| \tag{33}$$

hence the gradient dispersion can be derived as $O(\varpi) = \dfrac{L}{\mu}\log(\|w_1 - w_*\| / \varpi)$, which is used to illustrate how the parameters w_t are distributed in each participant.

Proof: According to the μ-strongly convexity of $f(w)$, we have

$$\nabla f(w)(w - w_*) \ge f(w) - f(w_*) + \frac{\mu}{2}\|w - w_*\|^2 \tag{34}$$

Thus, we can obtain the following:

$$\begin{aligned} \|w_{t+1} - w_*\|^2 &= \|w_t - \eta \nabla f(w_t) - w_*\|^2 \\ &= \|w_t - w_*\|^2 - 2\eta \nabla f(w_t)(w_t - w_*) + \eta^2 \|\nabla f(w_t)\|^2 \\ &\le \|w_t - w_*\|^2 - 2\eta(f(w) - f(w_*) \\ &+ \frac{\mu}{2}\|w_t - w_*\|^2) + \eta^2 \|\nabla f(w_t)\| \end{aligned} \tag{35}$$

By smoothing $f(w)$, we can obtain the gradient bound:

$$\begin{aligned} f(w_*) &\le f(w - \frac{1}{L}\nabla f(w)) \\ &\le f(w) - \frac{1}{2L}\|\nabla f(w)\|^2 \end{aligned} \tag{36}$$

By incorporating, (35) and (36) can be transformed as:

$$\begin{aligned} \|w_{t+1} - w_*\|^2 &= \|w_t - \eta \nabla f(w_t) - w_*\|^2 \\ &\le \|w_{t+1} - w_*\|^2 - \eta\mu\|w_t - w_*\|^2 + 2\eta(\eta L - 1)(f(w) - f(w_*)) \\ &\le (1 - \frac{\mu}{L})\|w_t - w_*\|^2 \le (1 - \frac{\mu}{L})\|\Delta^* w\|^2 \end{aligned} \tag{37}$$

where η is set as the learning rate of the last step.

The convergence in expectations can be formulated as:

$$[f(w_t) - f(w_*)] \le \varpi^t [\Delta^t (f(w))] \tag{38}$$

Thus, $f(w)$ is proven to be bounded where $\Delta^t(f(w)) = f(w_1) - f(w_*)$.

The proposed FDQN-W can obtain the optimal spectrum access strategy by using the aforementioned theoretical analysis and results, as shown in Algorithm 2.

For the time complexity of the proposed algorithm Feilin, in the incentive mechanism based on warranty contracts, the time complexity of the PU designing warranty contracts and uploading transaction information is *O*(*n*). In the model FDQN-W, the main execution time is model training in each RSU. Each RSU maintains its policy and performs decisions independently during decision making phase. Let episode *E* denote the aggregation round of AFWLA. The complexity is *O*(*ET*) = $O(n^2)$, where *T* refer to the training steps in each episode, respectively. Therefore, the time complexity of the proposed algorithm Feilin is $O(n^2)$.

VI. Simulation Results and Analysis

A. Experimental Setup

To verify the effectiveness of the proposed mechanism in ITS, the simulation results and analysis are presented. We consider the system with 1 BS, the number *M* of PUs, the number *J* of SUs and the number *K* of RSUs vary in the range of [1, 11], [1, 20] and [2, 5], respectively. The transmission rate of SUs varies from 25Mbps to 70Mbps, and the transmission power of SUs is in the range of 45mW to 55mW. The negative rewards R_1 and R_2 are set to -2 and -1, respectively. The experimental learning rate α is 0.01 and the discount factor γ is 0.9 [48]. The values of the parameters are shown in Table III.

TABLE III
Simulation Setup

Parameter	Value
The transmission rate *R* of SU	25~70Mbps
The transmission power *P* of SU	35~45mW
The negative feedback R_1	-2
The negative feedback R_2	-1
The bandwidth b_m	5~20Mbps
The learning rate α	0.01
The discount factor γ	0.9
The exploration rate ε	[0.3, 0.9, 1]
Training steps *T*	1500
Episode *E*	150

The performance of the proposed spectrum access algorithm Feilin is evaluated in terms of the average spectrum access success rate, the average collision rate and the average reward. First, the impact of λ, β, θ, the number of SUs and the number of channels on the performance of the proposed scheme is analyzed. Then the performance of the proposed spectrum access algorithm Feilin is compared with DQN+RC [18], Q-learning [11], PG+RDA [17], and MPQ-L+DPG [38]. All results in the following scenarios are the average of 1000 independent experiments.

B. Performance Analysis of the Proposed Scheme

In this scenario, the impact of λ, β and θ on the performance of Feilin is analyzed. In the experiment, the number of SUs, PUs, and channels are set to 2, 6, and 6, respectively. Since the goal of Feilin is to maximize the reward in ITS, the values of $QoS(t)$ are shown in Fig. 5 with varying λ, β and θ.

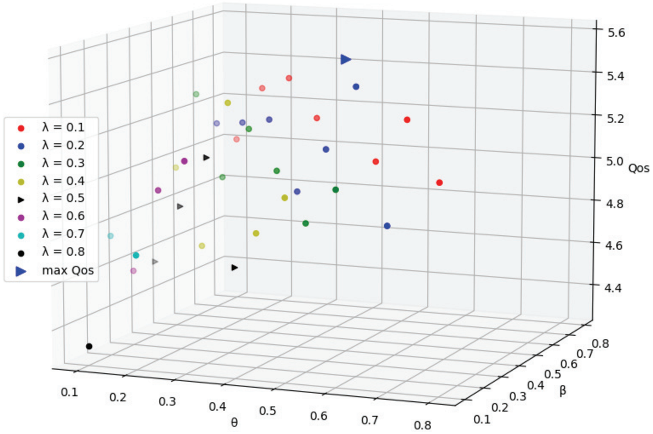


Fig. 5. The reward with varying λ, β and θ.

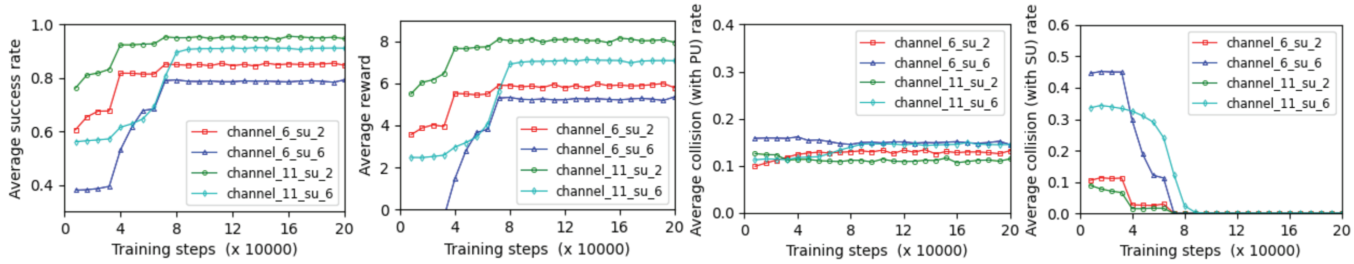
As shown in Fig. 5, the reward QoS will vary randomly with different λ, β and θ. To obtain the maximal QoS in convenience, λ is initialized to 0.1 and increases to 0.8 with step 0.1. We can get the optimized maximal reward 5.4 when λ, β and θ are 0.1, 0.5 and 0.4 respectively. Hence, the optimized parameters will be configured in the following experiments.

The average access success rate, average reward, average collision rates with PUs and SUs of the proposed scheme with a varying number of channels and SUs are shown in Fig. 6.

As shown in Fig. 6(a), the average successful access rate converges to a stable value with the number of iterations increasing. For the system with 6 SUs and 6 channels, the initial successful access rate is 0.39, and continues to increase during the training process; when the training steps increase to 72000, the access rate reaches the stable state to 0.782. If the number of SUs decreases to 2, the initial average access rate reaches 0.602, and upgrades to 0.814 at 48000 steps; as the training steps increase to 75000, the average access rate becomes steady to 0.845, which is higher than that of 6 SUs by 8.1%. As the number of channels increases to 11 (with 2 SUs), the average access rate is higher than that of 6 channels with 2 SUs by 10.8%. It can be seen that with the same number of channels, the smaller the number of SUs, the higher the spectrum successful access rate. The reason is that fewer SUs decrease the channel competition and increase the spectrum successful access rate. Meanwhile, the fewer SUs will also decrease the additional transmissions of weights of the FDQN-W model and save computation cost. The more available channels, the higher the average successful access rate is. Hence, we can see that the average access rate with 11 channels and 2 SUs is higher than those of the others clearly.

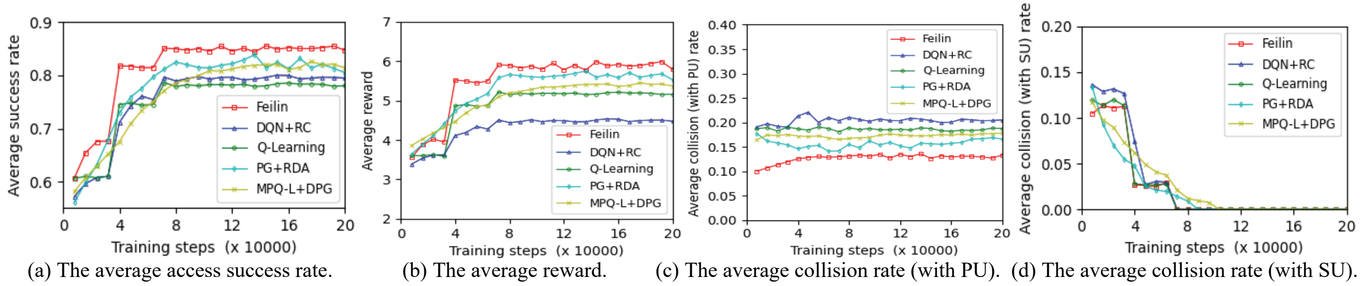
For the average reward, we can see the similar results as shown in Fig. 6(b). For the system with 6 channels and 6 SUs, when the number of training steps is less than 35000, the negative rewards are continuously obtained due to the high collision rate with SUs, resulting in the reward value less than 0; the average reward increases to 5.4 when the training steps reach 71000. As the number of SUs decreases to 2, the average reward will increase to 6.01, which is higher than that of the system with 6 SUs by 11.3%. As the number of channels increases to 11, the average reward is higher than the other schemes obviously, and the maximal reward reaches 8.1. The results also demonstrate the number of available channels has a great influence on the average reward in ITS.

From Fig. 6(c), for the proposed algorithm, we can see that the varying number of channels and SUs have a small impact on the average collision rate with the PU. The average collision rates of four cases fluctuate between 0.1 and 0.17. In general, more channels and fewer SUs will result in a lower collision rate with the PU. For the system with 11 channels and 2 SUs, the collision rate is 0.1, which is the lowest. Because once there is a collision with any SU during the transmission, the negative reward will be obtained, which decreases the reward and makes the model to adjust the spectrum strategy adaptively.



(a) The average access success rate. (b) The average reward. (c) The average collision rate (with PU). (d) The average collision rate (with SU).

Fig. 6. The performance of the proposed scheme with varying number of channels and SUs.



(a) The average access success rate. (b) The average reward. (c) The average collision rate (with PU). (d) The average collision rate (with SU).

Fig. 7. Performance of different schemes with varying training steps.

As shown in Fig. 6(d), the average collision rate with SU for different cases converges to 0 after continuous learning. For the system with 11 channels and 6 SUs, the collision rate decreases from 0.334 to 0 with 80,000 steps. Meanwhile, for the system with 11 channels and 2 SUs, the initial collision rate is 0.1, and it will reach 0 after 54,000 training steps. It's clear that the less number of SUs decrease the collisions and speed up the convergence of the training model. On the contrary, the less channels and larger number of SUs lead to higher collision probability in ITS.

C. Performance Comparisons

1) Performance Analysis with Varying Training Steps

In this scenario, the effectiveness of Feilin is evaluated and compared with DQN+RC, Q-Learning, PG+RDA, and MPQ-L+DPG in terms of average access success rate, average reward, and the average collision rates with PUs and SUs. In the experiments, the number of SUs, PUs, and channels are 2, 6, and 6, respectively. For all models, there is a hidden layer composed of 64 neurons. The average access success rate, average reward, and average collision rates of Feilin with varying training steps are shown in Fig. 7.

As shown in Fig. 7(a), the average success rates of the five models all are increasing until they reach a stable value. For DQN+RC, the average access rate increases from 0.57 to 0.77. While the average access rate of Q-Learning rises from 0.61 to 0.79. For DQN+RC and Q-Learning, the average access rates are almost the same. For PG+RDA and MPQ-L+DPG, the average access rates increases from 0.56 to 0.82, and 0.58 to 0.81, respectively. For the proposed Feilin, the average access rate rises to 0.85, which is higher than those of DQN+RC, Q-Learning, PG+RDA and MPQ-L+DPG by 10.4%, 7.6%, 3.2% and 4.1% respectively. The reason is that the incentives of the proposed warranty contract and the sharing of training parameters promote SUs access the idle spectrum in ITS, which guarantees the reward of SUs and increases the spectrum successful access rate. For PG+RDA and MPQ-L+DPG based on the policy optimization reinforcement learning, the access success rates rise steadily by utilizing the softmax algorithm. Meanwhile, Feilin, DQN+RC and Q-Learning are based on the Q-Learning reinforcement learning, which are suitable to solve the discrete action space problem. The varing ε in the greedy algorithm helps to find the optimal learning strategy in Q function. As shown in Fig. 7(a), the spectrum successful access rates of Feilin increase greatly when the training steps are 32000 and 64000 respectively. The results also demonstrate Feilin can achieve convergence faster than PG+RDA and MPQ-L+DPG.

From Fig. 7(b), we can see that the average rewards of the five schemes rise as the training steps increasing. For DQN+RC, Q-learning, PG+RDA, and MPQ-L+DPG, the average rewards increase from 3.4 to 4.3, 3.6 to 5.1, 3.6 to 5.6, and 3.8 to 5.3 respectively. For the proposed Feilin, the average reward increases to 6.21 as training steps increase to 70000, which is higher than those of DQN+RC, Q-Learning, PG+RDA and MPQ-L+DPG by 44.4%, 21.8%, 10.9% and 17.2%, respectively. The results in Fig. 7(b) are consistent with those in Fig. 7(a), which demonstrate the proposed Feilin can provide more reward for SUs.

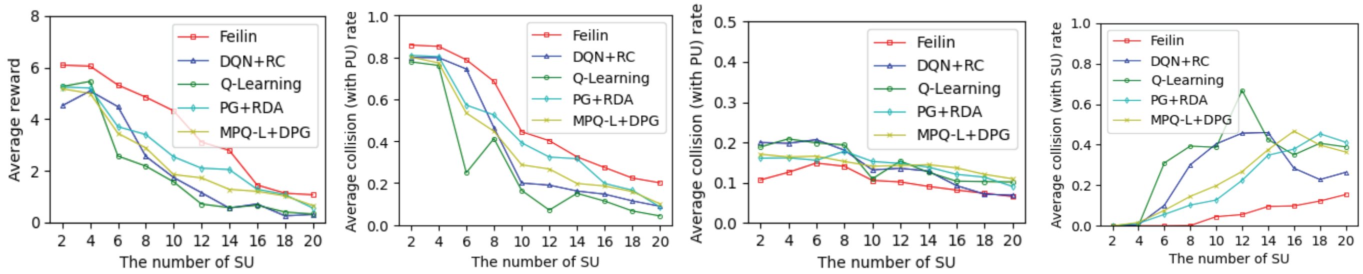
As shown in Fig. 7(c), for DQN+RC, the average collision rate with PU fluctuates between 0.19 and 0.24 as the training steps increasing, which is higher than those of other schemes. For Q-Learning, PG+RDA and MPQ-L+DPG, the average collision rates oscillate around 0.18, 0.15 and 0.17, respectively. For Feilin, the collision rate with PU is 0.13, which is lower than others greatly. The reason is that Feilin adopts the warranty contract to coordinate the spectrum access behaviors between SUs and PUs and guarantee the reward of SUs, which reduces the collisions between the SUs and PUs. As shown in Fig. 7(d), the average collision (with SU) rates of the five schemes are almost identical within 0.15 and 0.

The results in Fig. 7 show that the proposed spectrum algorithm Feilin can continuously optimize the model in time, and effectively learn the correlation between channel state and the spectrum access behaviors, which improves the spectrum successful access rate greatly.

2) Performance Analysis with Varying Number of SUs

In this scenario, the effectiveness of Feilin is evaluated and compared with DQN+RC, Q-Learning, PG+RDA and MPQ-L+DPG in terms of average access success rate, average reward, and the average collision rates with varying number of SUs. In the experiments, the number of PUs and the number of channels are all 6. The results of different schemes are shown in Fig. 8.

From Fig. 8(a), we can see that the average success rate of the five schemes decrease gradually with the number of SUs. For DQN+RC, MPQ-L+DPG, PG+RDA, the average success rates decrease from 0.8 to 0.1, 0.79 to 0.11 and 0.81 to 0.1, respectively. For Feilin, the average success rate is the highest among the five schemes, and it decreases from 0.86 to 0.21, which is higher than those of Q-Learning, DQN+RC, PG+RDA and MPQ-L+DPG by 15.1%, 11.4%, 10.1% and 9.3%, respectively. The reason is that Feilin adopts AFWLA to speed up the parameters sharing to depict the channel state and requirements of SUs accurately, which helps SUs to take accurate actions to access the idle spectrum successfully.



(a) The average access success rate. (b) The average reward. (c) The average collision rate (with PU). (d) The average collision rate (with SU).

Fig. 8. Performance of different schemes with varying number of SUs.

For the average reward, we can see the similar results as shown in Fig. 8(b). It's clear the average reward of Feilin is always higher than others. As the number of SUs increasing, the average rewards of DQN+RC, Q-learning, PG+RDA and MPQ-L+DPG decrease from 5.1 to 0.3, 5.4 to 0.3, 5.2 to 0.5 and 5.1 to 0.6, respectively. For Feilin, the average reward decreases from 6.1 to 1.1. The reason is that Feilin considers the hybrid reward and tries to maximize the rewards by the warranty contract and FDRL, which promotes SUs access the idle spectrum with additional rewards.

As shown in Fig. 8(c), when the number of SUs increases, the average collision rate with PUs degrades. For DQN+RC, Q-Learning, PG+RDA and MPQ-L+DPG, the average collision rates with PUs decrease from 0.21 to 0.08, 0.22 and 0.11, 0.18 and 0.1, and 0.17 and 0.12, respectively. For the proposed Feilin, the average collision rate is always lower than those of the four schemes, especially when the number of SUs is less than 10. As the number of SUs increasing to 20, the average collision rate of Feilin is lower than those of Q-Learning, DQN+RC, PG+RDA and MPQ-L+DPG by 6.8%, 7.0%, 5.7% and 6.7%, respectively. It seems that we get contradictory results on the average collision rate with the increasing number of SUs. However, the essential reason is that the collisions among SUs increase greatly with the larger number of SUs.

From Fig. 8(d), the average collision rates of the five schemes increase significantly as the number of SUs larger than 4, and fluctuate greatly as the number of SUs increasing. When the number of SUs is 12, the average collision rate of Q-Learning reaches 0.67, which is higher than that of Feilin by 0.61. For Feilin, the average collision rate with SU is the lowest. As the number of SUs increases to 20, the collision rate of Feilin increases to 0.15, which is lower than those of DQN+RC , Q-Learning, PG+RDA and MPQ-L+DPG by 46.4%, 62.5%, 35.6% and 39.1%, respectively. The results show that Feilin can guarantee the rewards of SUs to avoid additional collisions among SUs in multiple SUs scenarios, which provides more spectrum access opportunities to mobile users in ITS.

VII. Conclusion

To improve the spectrum successful access rate and guarantee the real-time requirements and QoS in ITS, this paper proposes a spectrum access algorithm Feilin based on FDRL, which adopts the warranty contract to guarantee the utilities of the SUs, and the asynchronous federated weighted learning algorithm (AFWLA) to decrease time cost. Meanwhile, Feilin promotes SUs access the idle spectrum by maximizing the hybrid reward including the delay and transmission power, which considers the transmission collisions and interference between SUs and PUs in ITS. Detailed experiments and analysis validate the performance of our proposed scheme, which improves the average spectrum successful access rate and degrades the average transmission collision rate of users. By signing a warranty contract, SUs become more motivated to access the idle channels. Compared with Q-Learning, DQN+RC, PG+RDA and MPQ-L+DPG, the proposed Feilin speeds up the model training process with less training time, increases the spectrum successful access rate and can provide high QoS for users in ITS.

References

[1] Y. Zhang, Y. Li, R. Wang, M. Shamim Hossain, and H. Lu, "Multi-aspect aware session-based recommendation for intelligent transportation services," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4696-4705, May 2020.

[2] J. He, K. Yang, H.-H. Chen, "6G cellular networks and connected autonomous vehicles," *IEEE Network*, vol. 35, no. 4, pp. 255-261, 2021.

[3] Y. Zhang, X. Ma, J. Zhang, M. Shamim Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive Internet of things: Improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58 - 64, May 2019.

[4] Y. Liu, Q. Zhang, and L. Ni, "A general framework for spectrum sensing using dedicated spectrum sensor networks," *ACM Trans. Sen. Netw.*, vol. 15, no. 1, pp. 1-23, 2019.

[5] S. Midya, A. Roy, K. Majumder, S. Phadikar, "QoS aware distributed dynamic channel allocation for V2V communication in TVWS spectrum," *Computer Networks*, vol. 171, pp. 1-17, 2020.

[6] D. B. Rawat, R. Alsabet, C. Bajracharya, M. Song, "On the performance of cognitive internet-of-vehicles with unlicensed user-mobility and licensed user-activity," *Computer Networks*, vol. 137, pp. 98-106, 2018.

[7] D. -T. Do, M. -S. Van Nguyen, M. Voznak, A. Kwasinski and J. N. de Souza, "Performance analysis of clustering car-following V2X system with wireless power transfer and massive connections," *IEEE Internet of Things Journal*, 2021, doi: 10.1109/JIOT.2021.3070744.

[8] S. Zhan, S. Chang, C. Chou and Z. Tsai, "Spectrum sharing auction platform for short-term licensed shared access," *in Proceedings of Wireless Days*, 2017, pp. 184-187.

[9] A. S. Shafigh, S. Glisic, E. Hossain, B. Lorenzo and L. A. DaSilva, "User-centric distributed spectrum sharing in dynamic network architectures," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 15-28, Feb. 2019.

[10] S. Chouikhi, L. Khoukhi, M. Esseghir, L. Merghem-Boulahia, "Generalized Nash equilibrium approach for radio resource sharing and power allocation in vehicular networks," *Computer Networks*, vol. 182, 2020.

[11] A. Kaur and K. Kumar, "Energy-efficient resource allocation in cognitive radio networks under cooperative multi-agent model-free reinforcement learning schemes," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1337-1348, Sept. 2020.

[12] A. Omidkar, A. Khalili, H. H. Nguyen and H. Shafiei, "Reinforcement learning based resource allocation for energy-harvesting-aided D2D communications in IoT networks," *IEEE Internet of Things Journal*, 2022, doi: 10.1109/JIOT.2022.3151001.

[13] H. Q. Nguyen, B. T. Nguyen, T. Q. Dong, D. T. Ngo and T. A. Nguyen, "Deep Q-learning with multiband sensing for dynamic spectrum access," *in Proceeding of IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2018, pp. 1-5.

[14] Y. Wang, X. Li, P. Wan and R. Shao, "Intelligent dynamic spectrum access using deep reinforcement learning for VANETs," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 15554-15563, 15 July15, 2021.

[15] Y. Wang, J. Sun, H. He and C. Sun, "Deterministic policy gradient with integral compensator for robust quadrotor control," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3713-3725, Oct. 2020.

[16] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, Daan Wierstra, "Continuous control with deep reinforcement learning," *in Proceedings of 4th International Conference on Learning Representations (ICLR)*, 2016.

[17] L. Li, D. Li, T. Song and X. Xu, "Actor-critic learning control with regularization and feature selection in policy gradient estimation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1217-1227, March 2021.

[18] H. Chang, H. Song, Y. Yi, J. Zhang, H. He and L. Liu, "Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1938-1948, April 2019.

[19] J. Cao, K. Zhang, F. Wu, and S. Leng, "Learning cooperation schemes for mobile edge computing empowered Internet of Vehicles," *in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC),* Seoul, South Korea, May 2020, pp. 1-6.

[20] A. Nilsson, S. Smith, G. Ulm, E. Gustavsson, and M. Jirstrand, "A

performance evaluation of federated learning algorithms," *in Proc. 2^{nd} Workshop Distrib. Infrastruct. Deep Learn. (DIDL)*, Dec. 2018, pp. 1-8.

[21] F. Sattler, S. Wiedemann, K.-R. Muller, and W. Samek,"Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[22] D. Rothchild et al., "FetchSGD: Communication-efficient federated learning with sketching," *in Proc. Int. Conf. Mach. Learn.*, Nov. 2020, pp. 8253-8265.

[23] X. Wang, C. Wang, X. Li, V. C. M. Leung and T. Taleb, "Federated deep reinforcement learning for Internet of things with decentralized cooperative edge caching," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9441-9455, Oct. 2020.

[24] F. Riaz, M. M. Rathore, A. Sohail, N. I. Ratyal, S. Abid, S. Khalid, T. Shehryar, A. Waheed, "Emotion-controlled spectrum mobility scheme for efficient syntactic interoperability in cognitive radio-based unmanned vehicles," *Computer Communications*, vol. 160, pp. 1-13, 2020.

[25] G. Rathe, D. Ahmad, F. Kurugollu, et al., "CRT-BIoV: A cognitive radio technique for blockchain-enabled Internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4005-4015, July 2021.

[26] G. Kakkavas, K. Tsitseklis, V. Karyotis and S. Papavassiliou, "A software defined radio cross-layer resource allocation approach for cognitive radio networks: From theory to practice," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 740-755, June 2020.

[27] T. Zhang, N. Leng, and S. Banerjee, "A vehicle-based measurement framework for enhancing whitespace spectrum databases," *in Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom'14)*, 2014, pp. 17-28.

[28] M. Hammouda, R. Zheng and T. N. Davidson, "Learning-theoretic multi-channel spectrum sensing and access in full-duplex cognitive radio networks with unknown primary user activities," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 4, pp. 885-897, 2019.

[29] A. Asheralieva and D. Niyato, "Combining contract theory and Lyapunov optimization for content sharing with edge caching and device-to-device communications," *IEEE/ACM Transactions on Networking*, vol. 28, no. 3, pp. 1213-1226, June 2020.

[30] T. Ariyarathna, P. Harankahadeniya, S. Isthikar, N. Pathirana, H. M. N. D. Bandara and A. Madanayake, "Dynamic spectrum access via smart contracts on blockchain," *in Proceeding of IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, pp. 1-6.

[31] S. K. Dhurandher, N. Gupta and P. Nicopolitidis, "Contract theory based medium access contention resolution in TDMA cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8026-8035, Aug. 2019.

[32] S. Zhao, Y. Feng, G. Yu, "D2D communication channel allocation and resource optimization in 5G network based on game theory," *Computer Communications*, vol. 169, pp. 26-32, 2021.

[33] S. Kim, "Heterogeneous network spectrum allocation scheme based on three-phase bargaining game," *Computer Networks*, vol. 177, 2020.

[34] M. S. Frikha, S. M. Gammar, A. Lahmadi, L. Andrey, "Reinforcement and deep reinforcement learning for wireless Internet of things: A survey," *Computer Communications*, vol. 178, pp. 98-113, 2021.

[35] F. Obite, A. D. Usman, E. Okafor, "An overview of deep reinforcement learning for spectrum sensing in cognitive radio networks," *Digital Signal Processing: A Review Journal,* vol. 113, 2021.

[36] M. Chen, T. Wang, K. Ota, M. Dong, M. Zhao, A. Liu, "Intelligent resource allocation management for vehicles network: An A3C learning approach," *Computer Communications*, vol. 151, pp. 485-494, 2020.

[37] X.-L. Huang, Y.-X. Li, Y. Gao, and X.-W. Tang, "Q-learning-based spectrum access for multimedia transmission over cognitive radio networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 110-119, March 2021.

[38] W. Shi, S. Song, C. Wu and C. L. P. Chen, "Multi pseudo Q-learning-based deterministic policy gradient for tracking control of autonomous underwater vehicles," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3534-3546, Dec. 2019.

[39] S. Wang, H. Liu, P. H. Gomes and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 2, pp. 257-265, June 2018.

[40] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316-4336, 2021.

[41] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, Virginia Smith, "Federated optimization in heterogeneous networks," arXiv:1812.06127, 2020.

[42] W. Liu, L. Chen, Y. Chen, and W. Zhang, "Accelerating federated learning via momentum gradient descent," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 8, pp. 1754-1766, Aug. 2020.

[43] Mónica Ribero, Haris Vikalo, "Communication-efficient federated learning via optimal client sampling," arXiv:2007.15197, 2020.

[44] Y. Zhang, Q. Wu and M. Shikh-Bahaei, "Vertical federated learning based privacy-preserving cooperative sensing in cognitive radio networks," *in Proceeding of IEEE Globecom Workshops (GC Wkshps)*, 2020, pp. 1-6.

[45] U. Challita, L. Dong, W. Saad, "Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4674-4689, July 2018.

[46] M. Veres, M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3152-3168, 2020.

[47] S. Wang et al., "Adaptive Federated Learning in Resource Constrained Edge Computing Systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205-1221, June 2019.

[48] Vektor Dewanto, Marcus Gallagher, "Examining average and discounted reward optimality criteria in reinforcement learning," arXiv:2107.01348, 2021.

Rongbo Zhu (Member, IEEE) received the B.S. and M.S. degrees from Wuhan University of Technology, China, in 2000 and 2003, respectively; the Ph.D. degree from Shanghai Jiao Tong University, China, in 2006. From 2011 to 2012, Dr. Zhu was a visiting scholar at Virginia Tech, USA. He is currently a Professor with the College of Informatics, Huazhong Agricultural University, Wuhan, China. He has published over 60 papers in the areas of wireless networks and mobile computing.

Mengyao Li received the B.S. degree in network engineering from the South-Central University for Nationalities, China in 2020. She is currently pursuing the M.S. degree in computer system architecture with the College of Computer Science, South-Central University for Nationalities, China. Her research interests are in the areas of cognitive radio and ITS.

Hao Liu received the B.S. degree in network engineering from the South-Central University for Nationalities, China in 2019. He is currently pursuing the M.S. degree in computer system architecture with the College of Computer Science, South-Central University for Nationalities, China. His research interests are in the areas of cognitive radio networks and mobile computing.

Lu Liu received the M.S. degree from Brunel University, Uxbridge, U.K., in 2003 and the Ph.D. degree from the University of Surrey, Guildford, U.K., in 2008. He is the Head of School of Informatics and a Professor of informatics, University of Leicester, Leicester, U.K.. His research interests include data analytics, service computing, artificial intelligence, and the Internet of Things. He is a Fellow of British Computer Society.

Maode Ma (Senior Member, IEEE) received his Ph.D. degree in computer science from Hong Kong University of Science and Technology, China, in 1999. Dr. Ma is currently a Professor with the College of Engineering, Qatar University, Qatar. From 2001 to 2020, He was a Professor in the School of Electrical and Electronic Engineering at Nanyang Technological University in Singapore. Dr. Ma has more than 200 international academic publications. He serves as a Senior Editor for IEEE Communications Surveys and Tutorials, and Guest Editor of IEEE Communications Magazine, Computer Communications. Dr. Ma is a Distinguished Lecturer of IEEE Communication Society. He is a Fellow of the IET.