Journal of Communications and Information Networks, Vol.3, No.3, Sept. 2018
DOI: 10.1007/s41650-018-0029-y

Research paper

Reinforcement Learning-Based Control for Unmanned Aerial Vehicles

Geyi Sheng, Minghui Min, Liang Xiao, Sicong Liu

***Abstract*—Estates, especially those of public security-related companies and institutes, have to protect their privacy from adversary unmanned aerial vehicles (UAVs). In this paper, we propose a reinforcement learning-based control framework to prevent unauthorized UAVs from entering a target area in a dynamic game without being aware of the UAV attack model. This UAV control scheme enables a target estate to choose the optimal control policy, such as jamming the global positioning system signals, hacking, and laser shooting, to expel nearby UAVs. A deep reinforcement learning technique, called neural episodic control, is used to accelerate the learning speed to achieve the optimal UAV control policy, especially for estates with a large area, against complicated UAV attack policies. We analyze the computational complexity for the proposed UAV control scheme and provide its performance bound, including the risk level of the estate and its utility. Our simulation results show that the proposed scheme can reduce the risk level of the target estate and improve its utility against malicious UAVs compared with the selected benchmark scheme.**

***Keywords*—unmanned aerial vehicles, security, reinforcement learning, nec, privacy**

I. Introduction

Unmanned aerial vehicles (UAVs) can perform data acquisition for civilian and commercial operations, such as weather monitoring, traffic control, and communication relaying, with high mobility and low cost[1,2]. However, camera-equipped UAVs are sometimes used to violate user privacy and security[3-5], such as in illegal surveillance and reconnaissance operations, smuggling, mid-air collisions, and eavesdropping attacks[6-8]. For example, at least two UAVs entered the restricted airspace at the White House, which can cause severe issues because criminals can carry out such violations via overflights. The continuous use of such overflight strategies can reveal political secrets, and if sensitive photographs are uploaded to the Internet, they may compromise the safety of the country[9].

Unauthorized UAV intrusions to an estate can be addressed by attacking the malicious UAVs by, for example, jamming its radar and radio signals, hijacking the global positioning system (GPS) signals, and using netguns and lasers[10-12]. For instance, the UAV capture and control strategy proposed in Ref. [13] takes over the controller of a rotorcraft UAV via a destructive GPS-spoofing attack. These UAV control policies vary in terms of UAV expel strength and costs. For example, laser shooting is more expensive and harder to operate than others policies, while jamming and GPS spoofing are cheaper and more portable. However, laser shooting can be used to intercept malicious UAVs more accurately and is more destructive[6]. Therefore, UAV control systems have to optimize their UAV control policies to expel or destroy UAVs with different attack and flight patterns, especially during security-sensitive time periods in which privacy is highly critical in the target estate.

In this paper, we propose a UAV control framework that incorporates existing UAV detection, identification, and tracking methods, such as acoustic sensing, radio frequency emission sensing, and electro-optical sensing, to protect the privacy of the target estate[6]. This framework chooses the control policy to expel the unauthorized UAV, such as command jamming, GPS spoofing, hacking electronics and shooting laser beams, without being aware of the attack model[14]. The UAV control policy is chosen based on the security level of the estate in the protected area and the distance between the unauthorized UAVs and the estate to reduce the risk level of the protected estate and improve its utility. In state-of-the-art methods, the optimal control policy depends on accurate knowledge of the attack model in each time slot, which is hard to acquire, especially in dynamic UAV control systems.

Control policy selections in a dynamic game can be approximately formulated as a Markov decision process (MDP) with finite states, in which the estate observes the states consist-

Manuscript received Jun. 10, 2018; accepted Jul. 24, 2018. This work was supported by the National Natural Science Foundation of China (Nos. 61671396 and 91638204). The associate editor coordinating the review of this paper and approving it for publication was X. Cheng.

G. Y. Sheng, M. H. Min, L. Xiao, S. C. Liu. Department of Communication Engineering, Xiamen University, Xiamen 361005, China (e-mail: 353387202@qq.com; 1175377092@qq.com; lxiao@xmu.edu.cn; liusc@xmu.edu.cn).

ing of the previous attack mode of the UAV and the current security level of the target estate. Therefore, reinforcement learning (RL) techniques can be applied to defend against an unauthorized UAV in a dynamic game and help the estate derive the optimal control policy in an MDP[15].

The reinforcement learning-based control framework proposed in Ref. [16] exploits the Q-learning algorithm, which is a model-free RL technique, to achieve the optimal control policy in a dynamic game without being aware of the attack model of the unauthorized UAV. To improve its performance in large-scale networks, we propose a neural episodic control (NEC)-based[17] unauthorized UAV control scheme, which employs a convolutional neural network (CNN) to generate the key to search the memory module called differentiate neural dictionary (DND). The outputs of the DND are the estimated long-term expected utilities of each control policy. By applying this deep reinforcement learning technique, the NEC-based UAV control algorithm is able to significantly compress the state space observed by the estate and thus reduce the time required to achieve the optimal control policy.

We prove that the proposed scheme achieves an optimal control policy selection after enough time slots have elapsed in the dynamic game. The control performance bound is provided in terms of the risk level and the utility of the estate. The proposed RL-based UAV control algorithm can improve the utility of the estate and decrease its risk level. The utility achieved by this scheme depends on the security level of the estate and has a negative linear correlation with the control policy. Simulation results show that the proposed RL-based scheme decreases the risk level of the estate and increases its utility more than the benchmark scheme as proposed in Ref. [16] and prove its convergence to optimal performance.

The main contributions of this paper are summarized as follows:

(1) We formulate a UAV control framework in which an estate applies different control methods to prevent an unauthorized UAV from stealing data.

(2) We propose a Q-learning-based UAV control scheme to achieve the optimal control policy, in which the estate chooses the control method according to the previous location of the UAV and the current security level of the target estate. This scheme enables the estate to achieve the optimal control performance without knowing the attack model in a dynamic control game.

(3) We develop an NEC-based UAV control scheme for the estate with enough computational resources to support deep learning to further accelerate the learning speed, reduce the risk level of the estate, and improve its utility.

(4) We provide the performance bound of the RL-based UAV control scheme and prove its convergence to optimal performance.

The rest of this paper is organized as follows. We review related work in section II and present the system model in section III. The Q-learning based UAV control scheme and the NEC-based UAV control scheme are developed in section IV and section V, respectively. We analyze the performance bound of the RL-based UAV control scheme in section VI. Simulation results are provided in section VII and the conclusions of this work are drawn in section VIII.

II. Related Work

The real-time UAV anomaly detection system proposed in Ref. [18] uses the recursive least squares method to estimate UAV parameters. The three-dimensional guidance law for rotary UAV interception proposed in Ref. [19] combines proportional navigation-based guidance and velocity feedback. The theory and practice of UAV capture and control via GPS signal spoofing are analyzed and demonstrated in Ref. [13]. GPS spoofing signals on autonomous UAVs was verified and assessed via experimental results in Refs. [20,21], which show that spoofing signals can affect the navigation system of UAVs so that they go off course or show abnormal operation. Several methods that can take down the malicious UAVs are described in Ref. [22], such as jamming, GPS spoofing, hacking, netguns, laser and so on.

RL techniques have been used to improve network security. The minimax Q-learning-based spectrum allocation method developed in Ref. [23] increases the spectrum efficiency in cognitive radio networks[24]. The two-dimensional Q-learning-based anti-jamming communication scheme proposed in Ref. [25] can increase the signal-to-interference-plus-noise ratio of secondary users against cooperative jamming in cognitive radio networks. The spoofing detection schemes proposed in Ref. [12] use Q-learning and Dyna-Q techniques to obtain the optimal test threshold of the physical-layer authentication in wireless networks. The deep Q-network-based transmission scheme developed in Ref. [26] achieves optimal power and node mobility control to address jamming in the underwater acoustic networks. A deep Q-learning based UAV power allocation strategy is proposed in Ref. [27] to achieve the optimal power allocation against smart attacks without knowing the attack model and the channel model. The hotbooting policy hill climbing (PHC)-based UAV relay strategy[28] helps vehicular ad-hoc networks resist jamming in a dynamic game.

A Q-learning-based malicious UAV control scheme is proposed in Ref. [16] to select the optimal control policy based on the system state, which consists of the previous UAV location and the current security level of the estate in the protected area without knowing the attack model of the malicious UAV. Compared with our previous work in Ref. [16], this work investigates unauthorized UAV control based on NEC to accelerate the learning speed of the unauthorized UAV control

scheme so as to reduce the risk level of the estate in the protected area and improve the utility of the estate. Simulation results show that the proposed schemes can achieve a better performance compared with the benchmark scheme we selected against unauthorized UAVs.

III. System Model

A smart UAV control system consisting of a target estate, an unauthorized UAV, and UAV detection devices, such as satellites, is considered in this paper. The estate applies multiple control methods to prevent the unauthorized UAV from stealing data, such as command jamming, GPS spoofing, hacking electronics for the takeover of controllers, and shooting an electromagnetic (EM) laser beam, as shown in Fig. 1. The control policies vary in terms of UAV expel strength and cost. For instance, though the laser shooting method suffers from higher energy consumption and more complicated operation than the GPS spoofing method, it can destroy the unauthorized UAV more accurately.

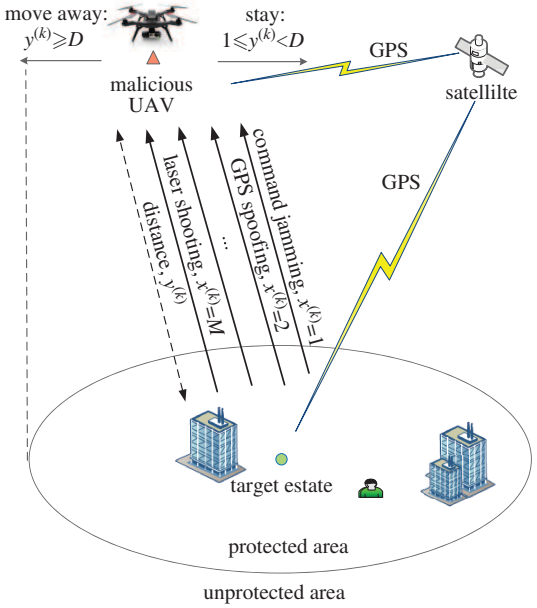


Figure 1 Smart UAV control system that uses RL to choose the control policy $x^{(k)}$, such as command jamming, GPS spoofing, and laser shooting to defend against the unauthorized UAV, which is $y^{(k)}$ meters away from the protected area of the target estate

The control policies are split into different categories according to their impact strength on the unauthorized UAV. Without loss of generality, let Level-$(i+1)$ control policy denote a stronger control policy against the unauthorized UAV than that of Level-i. For example, laser shooting can be labeled with a higher level than GPS spoofing if the former is considered to be stronger to defend against a given unauthorized UAV.

Once an unauthorized UAV has been identified, the smart UAV control system chooses a control policy at time k, denoted by $x^{(k)} \in X = \{0,1,2,\cdots,M\}$, where X is the action set of the estate. The smart unauthorized UAVs control system takes no action if $x = 0$ and defends against the UAV with the Level-x control policy if $x > 0$.

The state of the unauthorized UAV at time k observed by the estate is denoted by $y^{(k)} \in Y = \{0,1,2,\cdots,D\}$, in which the UAV is crashed if $y^{(k)} = 0$, and is $y^{(k)}$ meters away from the estate if $1 \leqslant y^{(k)} \leqslant D$. More specifically, the UAV is able to move away from the attack target area (i.e., $y^{(k)} \geqslant D$) or stay in the area of interest (i.e., $1 \leqslant y^{(k)} < D$). The next state of the UAV depends on the current control policy and the current state of the unauthorized UAV (i.e., the distance between the unauthorized UAV and the estate). The state transfer probability is denoted by $P_{x,y,y'} = \Pr(y'|x,y)$, where y' is the next state of the unauthorized UAV if the estate applies the Level-x control policy against the UAV at the present state y.

For ease of reference, the commonly used notations are summarized in Tab. 1. The time index k in the superscript is omitted if there is no possibility of confusion.

Table 1 List of notations

symbol	description
$C^{(k)}(x,y)$	control cost
$G^{(k)}(y)$	control gain
$R^{(k)}(y)$	risk level
$U^{(k)}$	utility of the estate
η	cost coefficient of the control policy
μ	cost coefficient of the distance between the unauthorized UAV and the estate
ϕ	security level of the estate
$\boldsymbol{s}$	state of the estate
$\boldsymbol{K}_x$	key vector of the NEC
$\boldsymbol{V}_x$	Q values of NEC
$\boldsymbol{\varphi}$	experience sequence
$\boldsymbol{\theta}$	weights of the CNN

IV. Q-Learning-Based UAV Control Scheme

We propose a Q-learning-based UAV control scheme to choose the control policy via trial and error. The control policy is chosen based on the state, which consists of the previous UAV location and the current security level of the target estate. The next state observed by the estate is independent of the previous states and actions for a given estate state and UAV control policy in the current time slot. Therefore, the unauthorized UAV control process can be viewed as an MDP, in which the Q-learning technique can derive the optimal policy without being aware of the attack model.

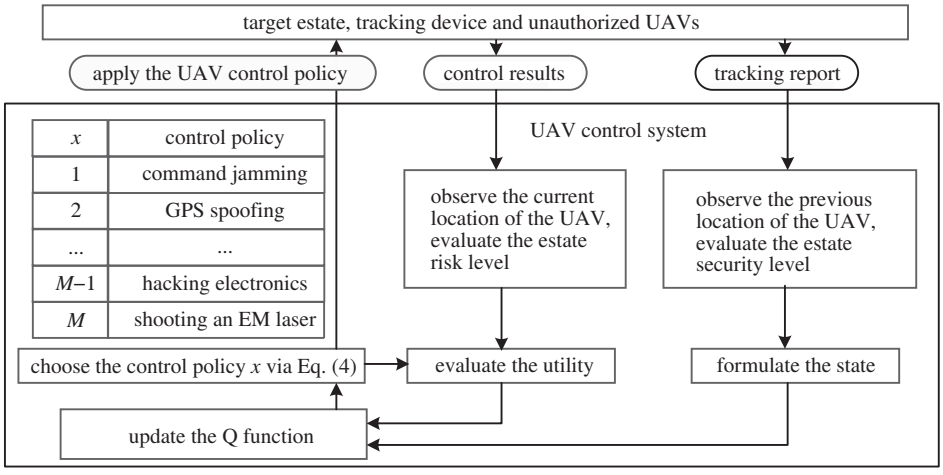


Figure 2 Reinforcement-learning-based UAV control framework

As illustrated in Fig. 2, we initialize the distance between the unauthorized UAV and the estate according to the tracking report, and the learning parameters are set to achieve a good control performance. The smart UAV control system observes the state of the estate at time k, denoted by $\boldsymbol{s}^{(k)}$, which consists of the previous location of the unauthorized UAV and the current security level of the target estate, i.e., $\boldsymbol{s}^{(k)} = [y^{(k-1)}, \phi^{(k)}] \in \Lambda$, where Λ is the state set of the estate. Based on the estate state, the UAV control system chooses control policy $x^{(k)} \in X$.

The UAV control system observes the current location of the unauthorized UAV, and evaluates the security level of the target estate in the protected area to determine its utility. Intuitively, the estate obtains more control gains if a UAV that is closer is crashed, because such a UAV is more likely to steal sensitive information. For simplicity, the control gain of the UAV control system at time k denoted by $G^{(k)}(y)$ is modeled as a linear function of the distance between the unauthorized UAV and the protected estate $y^{(k-1)}$ at the previous time slot, i.e., $G^{(k)}(y) = A - By^{(k-1)}$, where A and B are constant. The control cost of the UAV control system at time k denoted by $C^{(k)}(x,y)$ is a function of the control policy $x^{(k)}$ and the distance between the UAV and the estate $y^{(k)}$. We model the cost of the estate as $C^{(k)}(x,y) = \eta x^{(k)} + \mu y^{(k)}$, where η and μ are the cost coefficients. The risk level of the estate in the protected area is denoted by $R^{(k)}(y)$ and is given by $R^{(k)} = \nu y^{(k)}/D$, showing that the closer the UAV gets to the target estate, the easier it can be destroyed, and thus there is less risk for the estate, where ν is the risk coefficient.

The utility of the UAV control system at time k based on the control gain and the control cost is denoted by $U^{(k)}$ and given by

$$U^{(k)} = \mathrm{I}\big(y^{(k)} = 0\big)\big(A - By^{(k-1)}\big)\phi^{(k)} - \eta x^{(k)} - \mu y^{(k)}, \quad (1)$$

where $\phi^{(k)} \in (0,1]$ indicates the security level of the estate in the protected area at time k and $\mathrm{I}(\sigma)$ is the indicator function, which equals 1 if σ is true and 0 otherwise.

The proposed Q-learning-based UAV control system maintains a Q-function for each action-state pair, denoted by $Q(\boldsymbol{s},x)$, which is the expected discounted long-term reward observed by the UAV control system. The Q-function is updated at time k according to the iterative Bellman equation as follows:

$$Q\big(\boldsymbol{s}^{(k)},x^{(k)}\big) \leftarrow (1-\alpha)Q\big(\boldsymbol{s}^{(k)},x^{(k)}\big) + \alpha\big(U^{(k)} + \gamma V\big(\boldsymbol{s}'\big)\big), \quad (2)$$

where $\boldsymbol{s}'$ is the next state if the estate applies the Level-x control policy to intercept the unauthorized UAV at state $\boldsymbol{s}^{(k)}$, the learning rate $\alpha \in (0,1]$ is the weight of the current experience, the discount factor $\gamma \in [0,1]$ indicates the uncertainty of the estate on future rewards, and the value function denoted by $V(\boldsymbol{s})$ maximizes $Q(\boldsymbol{s},x)$ over the action set given by

$$V\big(\boldsymbol{s}^{(k)}\big) = \max_{x' \in X} Q\big(\boldsymbol{s}^{(k)},x'\big). \quad (3)$$

To make a tradeoff between exploitation and exploration, the UAV control policy is chosen according to the ε-greedy policy. More specifically, the UAV control policy $x^{(k)}$ that maximizes the Q-function is chosen with a high probability of $1-\varepsilon$, while other actions are selected with a low probability to avoid staying in a local maximum, i.e.,

$$\Pr\big(x^{(k)} = \hat{x}\big) = \begin{cases} 1-\varepsilon, & \hat{x} = \arg\max\limits_{x'} Q\big(\boldsymbol{s}^{(k)},x'\big), \\ \dfrac{\varepsilon}{|X|-1}, & \text{o.w.} \end{cases} \quad (4)$$

The proposed Q-learning-based UAV control scheme is summarized in Algorithm 1.

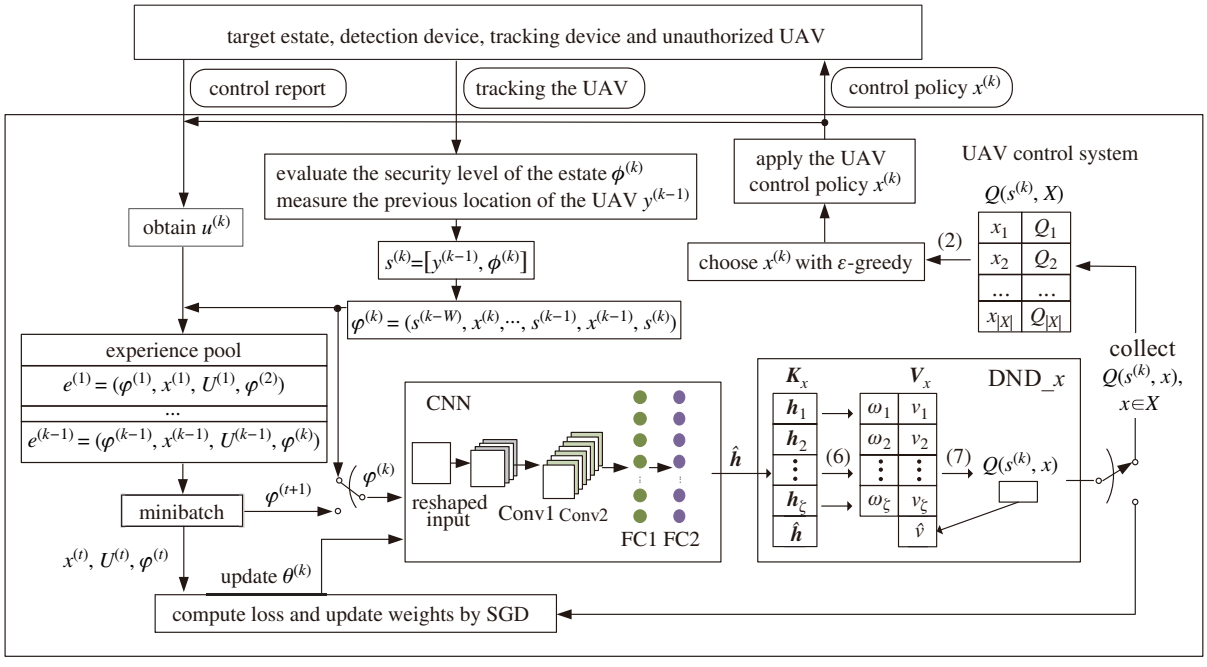


Figure 3 Illustration of the NEC-based UAV control scheme for an estate to choose its UAV control policy, such as shooting an electromagnetic laser beam, GPS spoofing, command jamming, and hacking electronics for the takeover of controllers

Algorithm 1 Q-learning-based UAV control algorithm

1: Initialize α and γ
2: $\boldsymbol{Q} = \boldsymbol{0}$
3: Randomly choose $\boldsymbol{s}^{(0)} \in \Lambda$
4: **for** $k = 1, 2, 3, \cdots$ **do**
5: Measure the previous location of the UAV $y^{(k-1)}$
6: Estimate the current security level of the estate $\phi^{(k)}$
7: $\boldsymbol{s}^{(k)} = [y^{(k-1)}, \phi^{(k)}]$
8: Choose $x^{(k)}$ via (4)
9: Apply control policy $x^{(k)}$
10: Observe the current location of the UAV $y^{(k)}$
11: Estimate the risk level of the estate $R^{(k)}$
12: Evaluate $U^{(k)}$ via Eq. (1)
13: Update $Q(\boldsymbol{s}^{(k)}, x^{(k)})$ via Eq. (2)
14: Update $V(\boldsymbol{s}^{(k)})$ via Eq. (3)
15: **end for**

V. NEC-Based UAV Control Scheme

We propose an NEC-based UAV control scheme to accelerate the learning speed of the estate with a larger state-action space. This scheme uses a CNN to compress the state space and a DND to store similar past experiences. The UAV control system chooses the control policy $x^{(k)} \in X$ based on a DND denoted by $(\boldsymbol{K}_x, \boldsymbol{V}_x)$, where $\boldsymbol{K}_x$ saves the keys produced by the CNN to look up the estimated Q values stored in the database denoted by $\boldsymbol{V}_x$.

The smart UAV control system observes the current state of the estate $\boldsymbol{s}^{(k)} = [y^{(k-1)}, \boldsymbol{\phi}^{(k)}]$. The estate state $\boldsymbol{s}^{(k)}$ is processed by the CNN to produce the keys denoted by $\hat{\boldsymbol{h}}$. The experience sequence $\boldsymbol{\varphi}^{(k)}$ is based on the current state and the previous W state-action pairs, i.e., $\boldsymbol{\varphi}^{(k)} = \{\boldsymbol{s}^{(k-W)}, x^{(k-W)}, \boldsymbol{s}^{(k-W-1)}, x^{(k-W-1)}, \cdots, x^{(k-1)}, \boldsymbol{s}^{(k)}\}$.

The estate reshapes the state sequence $\boldsymbol{\varphi}^{(k)}$ into an $n_1 \times n_1$ matrix and inputs it to the CNN, as shown in Algorithm 2. The CNN consists of two convolutional (Conv) layers and two fully connected (FC) layers. The first Conv layer has f_1 filters, each of size $n_2 \times n_2$ and stride 1, while the second Conv layer has f_2 filters, each of size $n_3 \times n_3$ and stride 1. Both layers use rectified linear units (ReLUs) as their activation function. The first FC layer involves r_1 ReLUs, and the second FC layer has r_2 ReLUs.

The estate uses the output of the CNN as the lookup key $\hat{\boldsymbol{h}}$ for the DND, which generates the weight for the jth Q-value, with $1 \leqslant j \leqslant \zeta$ and $\zeta \leqslant k$, in $\boldsymbol{V}_x$ of the DND denoted by $\boldsymbol{\omega}_j$ and given by

$$\boldsymbol{\omega}_j = \frac{k\left(\hat{\boldsymbol{h}}, \boldsymbol{h}_j\right)}{\sum_p k\left(\hat{\boldsymbol{h}}, \boldsymbol{h}_p\right)}, \tag{5}$$

where $\boldsymbol{h}_j$ is the jth corresponding key stored in $\boldsymbol{K}_x$ of the DND and $k(\hat{\boldsymbol{h}}, \boldsymbol{h}_j)$ is a kernel function measuring the distance between the lookup key and the corresponding key in memory given by

$$k\left(\hat{\boldsymbol{h}}, \boldsymbol{h}_j\right) = \exp\left(-\frac{\|\hat{\boldsymbol{h}} - \boldsymbol{h}_j\|_2^2}{2}\right). \tag{6}$$

The Q-value of the action $x^{(k)}$ under current state $\boldsymbol{s}^{(k)}$ is esti-

Algorithm 2 NEC-based UAV control algorithm

Initialize α, γ, $\boldsymbol{\theta}$, T, and W
for $k = 1, 2, 3, \cdots$ **do**
  Measure the previous UAV distance $y^{(k-1)}$
  Estimate the current security level of the estate $\phi^{(k)}$
  $\boldsymbol{s}^{(k)} = \left[y^{(k-1)}, \boldsymbol{\phi}^{(k)}\right]$
  if $k \leqslant W$ **then**
    Select $x^{(k)}$ at random
  else
    $\boldsymbol{\varphi}^{(k)} = \{\boldsymbol{s}^{(k-W)}, x^{(k-W)}, \cdots, x^{(k-1)}, \boldsymbol{s}^{(k)}\}$
    Add $\{\boldsymbol{\varphi}^{(k)}, x^{(k)}, U^{(k)}, \boldsymbol{\varphi}^{(k+1)}\}$ to the experience pool
    for $t = 1, 2, \cdots, T$ **do**
      Randomly sample $\boldsymbol{e}^{(t)}$ from the experience pool
    end for
    Formulate $\mathscr{T}$ with $\{\boldsymbol{e}^{(t)}\}_{1 \leqslant t \leqslant T}$
    Input $\boldsymbol{\varphi}^{(k)}$ to the CNN
    Get the output of the CNN as the key $\hat{\boldsymbol{h}}$
    Generate $\omega_j^{(k)}$ via Eq. (5)
    Calculate the Q values for $x^{(k)}$ via Eq. (7)
    Add $\left(\hat{\boldsymbol{h}}, Q\left(\boldsymbol{s}^{(k)}, x^{(k)}\right)\right)$ to $(\boldsymbol{K}_x, \boldsymbol{V}_x)$
    Update $\boldsymbol{\theta}^{(k)}$ using the SGD algorithm via Eq. (8)
    Select $x^{(k)}$ via Eq. (4)
    Apply the control policy $x^{(k)}$ to expel the unauthorized UAV
    Observe the UAV location $y^{(k)}$
    Estimate the risk level of the estate $R^{(k)}$
    Evaluate $U^{(k)}$ via Eq. (1)
  end if
end for

mated according to

$$Q\left(\boldsymbol{s}^{(k)}, x^{(k)}\right) = \sum_j \omega_j v_j, \tag{7}$$

where v_j is the jth element in vector $\boldsymbol{V}_x$. The new key-value pair $(\hat{\boldsymbol{h}}, Q(\boldsymbol{s}^{(k)}, x^{(k)}))$ is then added at the end of the respective vectors $(\boldsymbol{K}_x, \boldsymbol{V}_x)$. If key $\hat{\boldsymbol{h}}$ already exists in $\boldsymbol{K}_x$, the DND updates the corresponding Q value denoted by $\hat{v}$ in $\boldsymbol{V}_x$ with $Q(\boldsymbol{s}^{(k)}, x^{(k)})$.

The architecture is replicated once for each control policy, with the CNN being shared between each separate $(\boldsymbol{K}_x, \boldsymbol{V}_x)$. The estate chooses the control policy $x^{(k)}$ according to the ε-greedy strategy, measures the distance between the UAV and the estate, estimates the security level of the data in the target estate, and evaluates its reward or utility $U^{(k)}$ via Eq. (1).

The estate saves the current experience sequence $\boldsymbol{\varphi}^{(k)}$, the control policy $x^{(k)}$, and the utility $U^{(k)}$ as a control experience denoted by $\{\boldsymbol{\varphi}^{(k)}, x^{(k)}, U^{(k)}, \boldsymbol{\varphi}^{(k+1)}\}$. The estate randomly selects a control experience from the experience pool, which stores the control experiences of the previous k time slots. The weights of the CNN used to compress the state space shown in Fig. 3 and denoted by $\boldsymbol{\theta}^{(k)}$ are updated according to a stochastic gradient descent (SGD) algorithm similar to the one in Ref. [29]. More specifically, the estate randomly chooses T experience sequences from the experience pool $\mathscr{E}$ to formulate a minibatch denoted by $\mathscr{T}$ with $\{\boldsymbol{e}^{(t)}\}_{1 \leqslant t \leqslant T}$, where $\boldsymbol{e}^{(t)}$ is the tth selected control experience including the control policy, the utility, and the previous and new experience sequences. The loss function of the minibatch $\mathscr{T}$ represents the squared error of the target optimal Q-value. The estate chooses the CNN weights $\boldsymbol{\theta}^{(k)}$ that minimize the loss function as follows,

$$\boldsymbol{\theta}^{(k)} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathscr{T}}\left[\left(U - Q\left(\boldsymbol{\varphi}, x; \boldsymbol{\theta}^{(k)}\right) + \gamma \max_{x'} Q\left(\boldsymbol{\varphi}', x'; \boldsymbol{\theta}^{(k-1)}\right)\right)^2\right], \tag{8}$$

where $\boldsymbol{\varphi}'$ is the next state sequence according to the chosen experience from the experience pool, x' is the next control strategy, and γ is the discount factor of the learning process, as shown in Algorithm 2.

The NEC-based UAV control scheme does not always outperform the Q-learning based scheme. More specifically, the NEC-based scheme can decrease the risk level of the estate and increase the utility of the estate at the cost of higher computational complexity compared with the Q-learning-based scheme. Therefore, the NEC-based control policy selection scheme is more suitable for estates with sufficient computational resources. On the other hand, estates with restricted computational resources cannot afford using the complicated NEC algorithm and have to resort to the Q-learning based scheme with less complexity and computational costs to choose the control policy in time. For example, the Q-learning algorithm takes 95% less time on average to choose the control policy in a time slot compared with the NEC-based algorithm, as shown in the experiment in which the same amount of computational resources are consumed.

VI. Performance Evaluations

We analyze the performance bound of the RL-based UAV control scheme regarding the risk level and the utility of the estate and discuss the computational complexity of the scheme. Repeated control policy selection in a dynamic game against UAVs can be viewed as an MDP, because the future state observed by the estate, including the location of the UAV and the security level of the estate, is independent of the previous states for a given current state and UAV control policy. Therefore, according to Refs. [16,17], reinforcement-learning techniques, such as Q-learning in Algorithm 1 and NEC in Algorithm 2, enable the estate to achieve the optimal UAV control policy after enough time slots have elapsed with a probability of 1.

Theorem 1 Let D_0 represent the distance between the unauthorized UAV and the estate at the previous time slot. An estate using Algorithm 1 and 2 in the dynamic game can carry out policy $x^* = 1$ after enough time slots have elapsed to destroy UAVs ($y = 0$) and achieve a risk level of $R = 0$ and a utility given by

$$U = (A - BD_0)\,\phi - \eta, \tag{9}$$

if

$$\phi BD_0 < A\phi - \mu - \eta. \tag{10}$$

Proof By Eq. (1), if $y = 0$, we have

$$U(x) = (A - BD_0)\,\phi - \eta x. \tag{11}$$

It is clear that the utility of the control system has a negative linear correlation with x; thus we have, for any $x \in X$,

$$U(1) = (A - BD_0)\,\phi - \eta > (A - BD_0)\,\phi - \eta x = U(x). \tag{12}$$

Similarly, if $y \neq 0$, we have

$$U = -\eta x - \mu y. \tag{13}$$

It is also clear that the utility of the control system has a negative linear correlation with x; thus we have, for any $x \in X$,

$$U(0) = -\mu y > -\eta x - \mu y = U(x). \tag{14}$$

Therefore, $U(0)$ is maximum at $\min\limits_{1 \leqslant y \leqslant D} y$, i.e., $U(0) = -\mu$. Thus, we have $U(1) \geqslant U(0)$, if expression (10) hold.

According to Ref. [16], the RL-based control scheme can achieve the optimal control policy $x^* = 1$ in the MDP after a sufficiently long time. Therefore, the proposed algorithm can achieve $x^* = 1$. By Eq. (1), we have $y = 0$, $R = 0$, and thus Eq. (9) is proven.

Remark 1 A UAV control system applies the RL-based control scheme to achieve the optimal policy without being aware of the attack model in the dynamic control game. If the unauthorized UAV is near the protected area, as shown in Fig. 1, the estate will choose the control policy with lower cost $x = 1$ to intercept the unauthorized UAV.

The computational complexity of the NEC-based UAV control system presented in Algorithm 2 mostly depends on the CNN. Let $m_{\psi-1}$ be the number of the inputs to the CNN in Algorithm 2, f_ψ be the number of the filters in the CNN, n_ψ be the spatial size of each filter, and m_ψ be the spatial size of the output feature maps of Conv layer ψ.

Theorem 2 The computational complexity of the CNN denoted by Γ in Algorithm 2 is given by

$$\Gamma = O\left(f_1 f_2^2 n_3^2 (n_1 - n_2 + 1)^2 (n_1 - n_2 - n_3 + 2)^2\right). \tag{15}$$

Proof According to Ref. [17], the total complexity of the CNN is $O(\sum_{(\psi=1)}^{2} m_{\psi-1} f_\psi n_\psi^2 m_\psi)$. The first Conv layer includes f_1 filters each of size $n_2 \times n_2$ with an $n_1 \times n_1$ matrix as the input and $f_1(n_1 - n_2 + 1)^2$ feature maps as the output, and the second Conv layer has f_2 filters each of size $n_3 \times n_3$ with $f_2(n_1 - n_2 - n_3 + 2)$ feature maps as the output. Therefore, we have

$$\Gamma = O\big(f_1 (n_1 - n_2 + 1)^2 \big(f_1 n_1^2 n_2^2 + f_2^2 n_3^2 (n_1 - n_2 - n_3 + 2)^2\big)\big). \tag{16}$$

According to the CNN architecture in Ref. [29], we have $f_1 n_1^2 n_2^2 \ll f_2^2 n_3^2 (n_1 - n_2 - n_3 + 2)^2$, and thus the computational complexity is given by Eq. (15).

VII. Simulation Results

Simulations were carried out to evaluate the proposed RL-based UAV control scheme for an estate with a topology as shown in Fig. 4. The estate and the detection devices were distributed around residential areas to protect themselves against the unauthorized UAV. In these simulations, the cost coefficients η and μ were 0.1 and 2.0, respectively. We supposed that the estate applies four control methods to prevent the unauthorized UAV from stealing data, namely command jamming, GPS spoofing, hacking electronics, and shooting an electromagnetic laser beam.

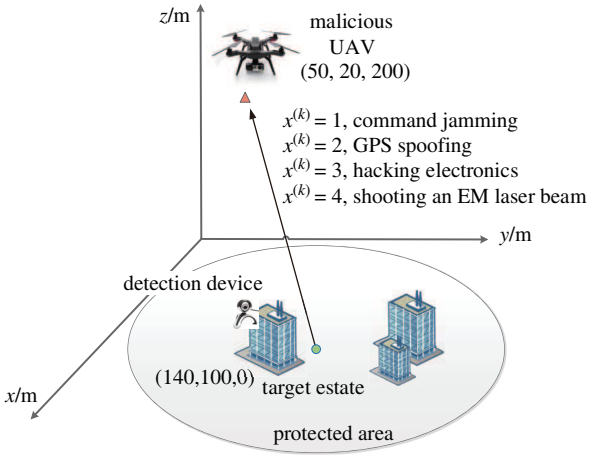


Figure 4 Initial topology of the control system in our simulation

Unless otherwise specified, the learning rate was 0.7 and the discount factor was 0.3. The estate reshapes the state sequence into a 6×6 matrix. The first Conv layer was set to have 20 filters of size 3×3 and 180 ReLUs, and the second Conv was set to have 40 filters of size 2×2 and 180 ReLUs. The discount factor in the updating process for the CNN weights was set as 1.0, the minibatch size was set as 4, and the number of previous state-action pairs was 11, according to the deep-Q network design method presented in Ref. [29].

The unauthorized UAV is more likely to be destroyed with a stronger control policy and at a shorter distance from the estate. As a special case, we set $P_{x,y,y'}$ as follows,

$$P_{x,y,y'} = \Pr(y'|x,y) = \begin{cases} \dfrac{\exp(\frac{mx}{ny})}{1 + \exp(\frac{mx}{ny})}, & y' = 0, \\ \dfrac{1}{(|Y| - 1)\,(1 + \exp(\frac{mx}{ny}))}, & \text{o.w.} \end{cases} \tag{17}$$

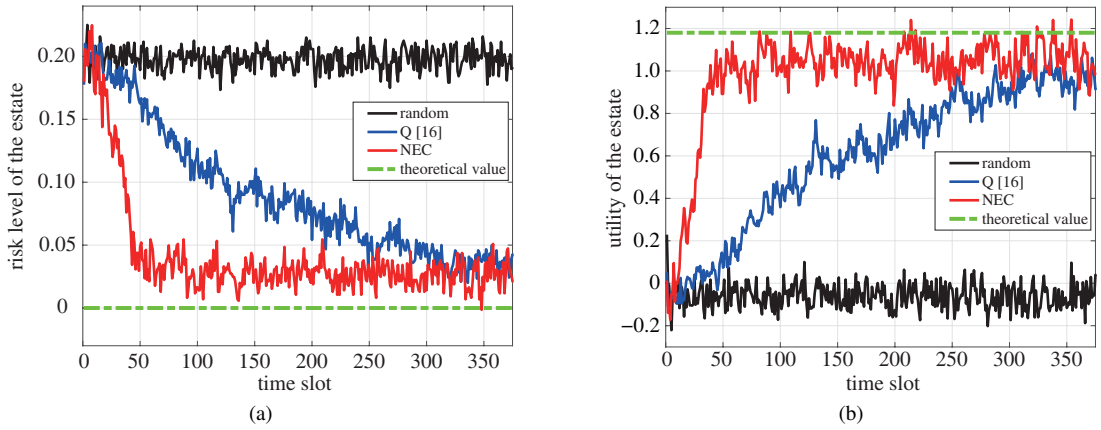


Figure 5 UAV control scheme performance in our simulations against an unauthorized UAV with its settings configured as shown in Fig. 4, with $\alpha = 0.7$, $\gamma = 0.3$, $A = 2.8$, $B = 0.5$, $\eta = 0.1$, and $\mu = 2.0$: (a) risk level of the estate; (b) utility of the estate

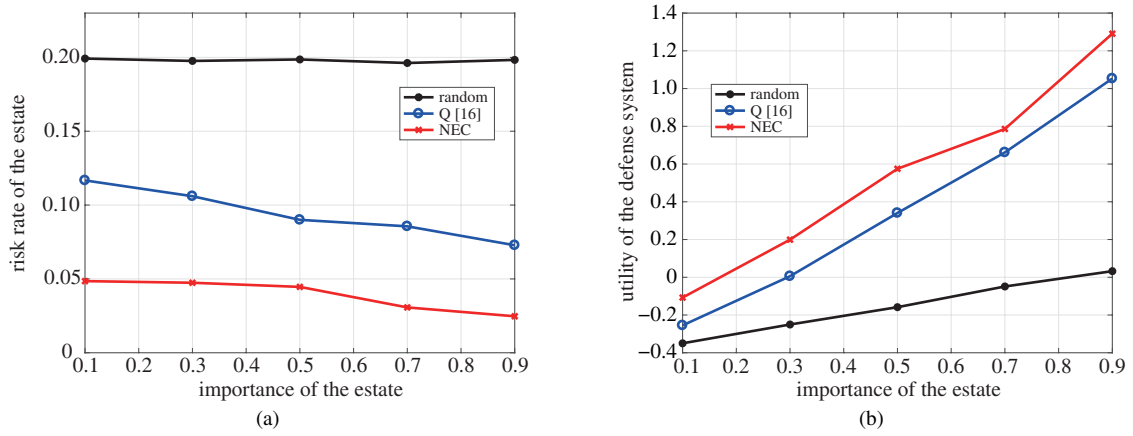


Figure 6 Performance of the UAV control system for the given security level of the estate, averaged over 400 time slots: (a) risk level of the estate; (b) utility of the estate

where m and n represent the impact weights of the control policy x and the distance y on the control results, respectively.

As shown in Fig. 5, the proposed NEC-based UAV control scheme converges to the performance bound given by Theorem 1. The NEC-based scheme outperforms the Q-learning-based scheme, yielding a lower risk level and higher utility, and both outperform the random strategy. As shown in Fig. 5(a), the risk level decreases over time with our proposed NEC-based UAV control scheme and converges to 2.0% after approximately 100 time slots, which is approximately 87.4% lower than for the Q-learning-based strategy. The risk level of the Q-learning based strategy converges after approximately 300 time slots, which is approximately 76.1% lower than the risk level of the random strategy.

As shown in Fig. 5(b), the utility of the estate when using the NEC-based UAV control scheme increases quickly after the start of the learning process and converges to a certain value that is much higher than that for the Q-learning-based strategy. For example, the utility of the estate with our proposed NEC-based scheme exceeds the Q-learning-based scheme by 65.0% at the 100th time slot. This is because the estate adjusts the control policy via the NEC technique. The Q-learning-based scheme requires less computational complexity than the NEC-based one. For example, the Q-learning-based strategy takes 94.9% less time to choose the control strategy in a time slot compared with the NEC-based scheme.

The average performance over 400 time slots, presented in Fig. 6, shows that the risk level decreases with the importance of the protected estate while the utility of the estate increases with the importance of the protected estate. For instance, if the importance of the protected estate is 0.9 instead of 0.1, the risk level of the estate using the NEC-based technique decreases by one time and the utility of the estate increases by 13 times.

If the security level of the protected estate is measured as 0.7 in each time slot as show in Fig. 6, the Q-learning-based UAV control scheme has a 1.5-times higher utility and 54.7% lower risk level compared with the random strategy. The NEC-based scheme further decreases the risk level by 50.4% and increases the utility by 40.7% compared with the Q-learning-based UAV control scheme.

VIII. Conclusion

In this paper, we have proposed an RL-based control system to defend against unauthorized UAVs exploiting Q-learning to improve the control performance without being aware of the attack policy in a dynamic game. The NEC technique was adopted to further accelerate the learning speed and improve the performance of the estate. Simulation results show that the proposed RL-based UAV control scheme can reduce the risk level of the target estate of a protected area and increase the utility of the estate more than the selected benchmark scheme. For instance, the risk level of the proposed NEC-based UAV control scheme converges to 2.0% after approximately 100 time slots, which is approximately 87.4% lower than that of the Q-learning based strategy. Moreover, the utility of the NEC-based scheme exceeds the Q-learning-based scheme by 65.0%.

References

[1] Q. Wang, Z. Chen, W. Mei, et al. Improving physical layer security using UAV-enabled mobile relaying [J]. IEEE Wireless Communication Letters, 2017, 6(3): 310-313.

[2] M. Dong, K. Ota, M. Lin, et al. UAV-assisted data gathering in wireless sensor networks [J]. The Journal of Supercomputing, 2014, 70(3): 1142-1155.

[3] D. He, S. Chan, M. Guizani. Communication security of unmanned aerial vehicles [J]. IEEE Wireless Communications, 2016, 24(4): 134-139.

[4] K. Mansfield, T. Eveleigh, T. H. Holzer, et al. Unmanned aerial vehicle smart device ground control station cyber security threat model [C]//IEEE International Conference on Technologies for Homeland Security (HST), Waltham, MA, 2012: 722-728.

[5] H. Sedjelmaci, S. M. Senouci, N. Ansari. Intrusion detection and ejection framework against lethal attacks in UAV-aided networks: A Bayesian game-theoretic methodology [J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(5): 1143-1153.

[6] D. Sathyamoorthy. A review of security threats of unmanned aerial vehicles and mitigation steps [J]. Journal of Defence and Security, 2015: 6(1): 81-97.

[7] Y. Xiao, V. K. Rayi, B. Sun, et al. A survey of key management schemes in wireless sensor networks [J]. Computer Communications, 2007, 30(11-12): 2314-2341.

[8] T. Wang, J. Tan, W. Ding, et al. Inter-community detection scheme for social Internet of things: A compressive sensing over graphs approach [J]. IEEE Internet of Things Journal, 2018.

[9] E. Vattapparamban, I. Guvenc, A. Yurekli, et al. Drones for smart cities: Issues in cybersecurity, privacy, and public safety [C]//International Wireless Communications and Mobile Computing Conference (IWCMC), Paphos, Cyprus, 2016: 216-221.

[10] L. Xiao, Y. Li, C. Dai, et al. Reinforcement learning-based NOMA power allocation in the presence of smart jamming [J]. IEEE Transactions on Vehicular Technology, 2018, 67(4): 3377-3389.

[11] S. Lv, L. Xiao, Q. Hu, et al. Anti-jamming power control game in unmanned aerial vehicle networksneural episodic control [C]//IEEE Global Communications Conference (GLOBECOM), Singapore, 2017.

[12] L. Xiao, Y. Li, G. Han, et al. PHY-layer spoofing detection with reinforcement learning in wireless networks [J]. IEEE Transactions on Vehicular Technology, 2016, 65(12): 10037-10047.

[13] A. J. Kerns, D. P. Shepard, J. A. Bhatti, et al. Unmanned aircraft capture and control via GPS spoofing [J]. Field Robotics, 2014, 31(4): 617-636.

[14] F. Yang, J. Gao. Dimming control scheme with high power and spectrum efficiency for visible light communications [J]. IEEE Photonics Journal, 2017, 9(1): 7901612.

[15] L. Xiao, Y. Li, G. Han, et al. A secure mobile crowdsensing game with deep reinforcement learning [J]. IEEE Transactions on Information Forensics and Security, 2018, 13(1): 35-47.

[16] M. Min, L. Xiao, D. Xu, et al. Learning-based defense against malicious unmanned aerial vehicles [C]//IEEE 87th Vehicular Technology Conference, Porto, Portugal, 2018: 1-5.

[17] A. Pritzel, B. Uria. Neural episodic control [J]. arXiv:1703.01988, 2017.

[18] Z. Birnbaum, A. Dolgikh, V. Skormin, et al. Unmanned aerial vehicle security using recursive parameter estimation [J]. Journal of Intelligent & Robotic Systems, 2016, 84(1-4): 107-120.

[19] B. Zhu, A. Zaini, L. Xie. Distributed guidance for interception by using multiple rotary-wing unmanned aerial vehicles [J]. IEEE Transactions on Industrial Electronics, 2017, 64(7): 5648-5656.

[20] S. Seo, B. Lee, S. Im, et al. Effect of spoofing on unmanned aerial vehicle using counterfeited GPS signal [J]. Positioning, Navigation, and Timing, 2015, 4(2): 57-65.

[21] X. Du, Y. Xiao, M. Guizani, et al. An effective key management scheme for heterogeneous sensor networks [J]. Ad Hoc Networks, 2007, 5(1): 24-34.

[22] Y. Zeng, R. Zhang, T. J. Lim. Wireless communications with unmanned aerial vehicles: opportunities and challenges [J]. IEEE Communications Magazine, 2016, 54(5): 36-42.

[23] B. Wang, Y. Wu, K. Liu, et al. An anti-jamming stochastic game for cognitive radio networks [J]. IEEE Journal on Selected Areas in Communications, 2011, 29(4): 877-889.

[24] C. Zhang, W. Zhang. Spectrum sharing for drone networks [J]. IEEE Journal on Selected Areas in Communications, 2017, 35(1): 136-144.

[25] L. Xiao, Y. Li, J. Liu, et al. Power control with reinforcement learning in cooperative cognitive radio networks against jamming [J]. The Journal of Supercomputing, 2015, 71(9): 3237-3257.

[26] L. Xiao, D. Jiang, X. Wan, et al. Anti-jamming underwater transmission with mobility and learning [J]. IEEE Communications Letters, 2018, 22(3): 542-545.

[27] L. Xiao, C. Xie, M. Min, et al. User-centric view of unmanned aerial vehicle transmission against smart attacks [J]. IEEE Transactions on Vehicular Technology, 2018, 67(4): 3420-3430.

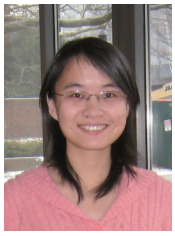
[28] L. Xiao, X. Lu, D. Xu, et al. UAV relay in VANETs against smart jamming with reinforcement learning [J]. IEEE Transactions on Vehicular Technology, 2018, 67(5): 4087-4097.

[29] V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533.

About the Authors

Geyi Sheng received her B.S. degree in communication engineering from Xiamen University, Xiamen, China, in 2017, where she is currently pursuing her M.S. degree in the Department of Communication Engineering. Her research interests include network security and wireless communications.

Minghui Min received her B.S. degree in automation from Qufu Normal University, Rizhao, China, in 2013, and her M.S. degree in control theory and control engineering from Shenyang Ligong University in joint training with Shenyang Institute of Automation, Chinese Academy of Sciences, Shengyang, China, in 2016. She is currently pursuing her Ph.D. degree in the Department of Communication Engineering, Xiamen University, Xiamen, China. Her research interests include network security and wireless communications.



Liang Xiao [corresponding author] (M'09, SM'13) is currently a Professor in the Department of Communication Engineering, Xiamen University, Fujian, China. She has served in several editorial roles, including an associate editor of IEEE Trans. Information Forensics & Security and IET Communications. Her research interests include wireless security, smart grids, and wireless communications. She won the best paper award for 2016 IEEE INFOCOM Big security WS. She received her B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, China, in 2000, her M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and her Ph.D. degree in electrical engineering from Rutgers University, NJ, in 2009. She was a visiting professor in Princeton University, Virginia Tech, and the University of Maryland, College Park. She is a senior member of the IEEE.

Sicong Liu (S15-M17) received his B.S.E. and his Ph.D. degree, both in electronic engineering, from Tsinghua University, Beijing, China in 2012 and 2017 (with the highest honor). From 2010 to 2011, he was a visiting scholar in the City University of Hong Kong, China. From 2017 to 2018, he served as a senior research engineer in Huawei Technologies Co., Ltd. Currently, he is an assistant professor in the Department of Communications Engineering, School of Information Science and Technology, Xiamen University, China. Sicong Liu has published over 35 journal and conference research papers. He owns 7 Chinese invention patents. He is one of the core members that draft the Broadband Power Line Communications Standard in China. He has won the Best Doctoral Dissertation Award of Tsinghua University. He is a reviewer of many top journals and has served as the guest editor of the Future Internet Journal and a TPC chair/member of IEEE ICC, IEEE SmartGridComm, and several international conferences. His research interests lie in sparse signal processing, interference mitigation, wireless communications, network security, and machine learning.