REAL-TIME LEARNING FOR THz RADAR MAPPING AND UAV CONTROL

Anna Guerra [†], *Francesco Guidi* [⋆], *Davide Dardari* [†], *Petar M. Djurić* [⋄]

[†] DEI, University of Bologna, Italy. E-mail: {anna.guerra3, davide.dardari}@unibo.it
[⋆] CNR-IEIIT, National Council Research of Italy, Italy. E-mail: francesco.guidi@ieiit.cnr.it
[⋄] ECE, Stony Brook University, New York. E-mail: petar.djuric@stonybrook.edu

ABSTRACT

In this paper we consider a joint detection, mapping and navigation problem by an unmanned aerial vehicle (UAV) with real-time learning capabilities. We formulate this problem as a Markov decision process (MDP), where the UAV is equipped with a THz radar capable to electronically scan the environment with high accuracy and to infer its probabilistic occupancy map. The navigation task amounts to maximizing the desired mapping accuracy and coverage and to decide whether targets (e.g., people carrying radio devices) are present or not. With the numerical results, we analyze the robustness of the considered Q-learning algorithm, and we discuss practical applications.

Index Terms**—** Autonomous Navigation, Reinforcement Learning, Q-learning, Unmanned Aerial Vehicles.

1. INTRODUCTION

Perception and cognition are two essential features for next generation radar systems. A cognitive radar (CR) is able to learn from the environment and to adjust its behaviour based on the received rewards or penalties that represent a feedback on the CR actions [1].

More recently, in [2,3] a massive multiple-input multiple-output (MIMO) CR has been investigated for multi-target detection using a reinforcement learning (RL) algorithm. In these papers, no prior information about the statistical model of the disturbance, or of the number of targets, was assumed for the proper functioning of the radar. Following a similar research direction, [4] showed the optimization of the trajectory of a unmanned aerial vehicle (UAV)-radar for environment mapping and detection using a RL approach where rewards were predicted within a finite temporal horizon. Indeed, *time* is a key aspect for UAV networks because of their limited energy autonomy [5–7] and, thus, it should be properly accounted for when designing the UAV control for time-critical applications (e.g., search–and–rescue). In [5], an information-seeking algorithm is developed for extraterrestrial exploration and return-to-base application, whereas in [8, 9] a similar problem is solved using RL for source localization. Algorithms for UAVs formation, navigation and self-localization have been proposed in [10–14], and RL for enhancing communications has been studied in [15–18].

The advent of sixth generation (6G) cellular systems fosters the exploitation of new frequency bands, which suggests the importance to investigate indoor detection and mapping using Terahertz (THz) radar technologies, as they are expected to guarantee unprecedented levels of radio localization accuracy [19]. The advantage of operating at THz rather than microwaves is that the surface illuminated by the interrogation signal reflects back in different directions (*diffuse scattering*) and not just specularly [20]. Beyond 100 GHz, the diffuse scattering is comparable with the specular component, allowing to improve the reconstruction of the surrounding thanks to the richer backscattered signal.

In this paper, our aim is to explore this technology in the context of CR-UAV. To be successful in indoor detection and mapping, the CR-UAV has to autonomously decide where to go to improve the detection task within a limited available time. Increasing the ambient awareness through mapping can also accelerate the overall learning process and the completion of the UAV primary task. Thus, starting from [4], valid when empirical models are available, we consider a THz radar exploiting a Q-learning algorithm with a combination of intrinsic (mapping) and extrinsic (detection) rewards. Finally, we show the impact of the THz radar parameters on the attainable performance through a simulation analysis.

2. PROBLEM FORMULATION

The UAV trajectory is designed to maximize the target detection, mapping accuracy and coverage subject to the mission time T_{M} and collision avoidance. We formulate the optimization problem as a Markov decision process (MDP). This problem can be solved using a model-free RL method. An example of an indoor environment is shown in Fig. 1.

Markov Decision Processes: Following the same notation as in [21], a MDP is defined by a tuple containing the state space $\mathcal{S}$, the action space $\mathcal{A}$, the reward space $\mathcal{R}$, and the probability of transitioning from one state s_k, at time instant k, to the next state s_{k+1}. Notably, the random state at time instant k, indicated with S_k, represents the knowledge about the environment available to the agent at time instant k,

978-1-7281-7289-7/20/$31.00 ©2020 IEEE

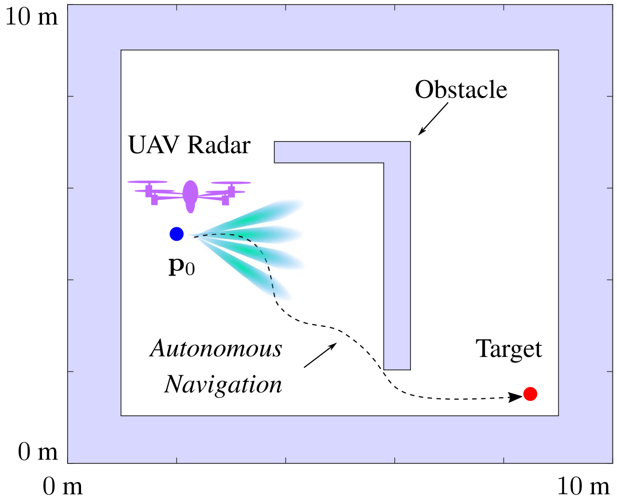

Fig. 1: Considered UAV scenario and reference map.

and it can take values $s_k \in \mathcal{S}$. The actions are chosen according to a specific policy $\pi\left(a_k|s_k\right)$, which is referred to as a probability density function (pdf) of an action.[1] The optimal policy selects actions that maximize a value function by

$$\pi^*\left(a_k|s_k\right) = \arg\max_{a_k} Q_\pi\left(s_k, a_k\right), \tag{1}$$

where the Q-function, $Q_\pi\left(\cdot\right)$, is the expected sum of discounted rewards over all possible policies and is given by

$$Q_\pi\left(s_k, a_k\right) = \mathbb{E}_\pi\left\{\sum_{l=0}^{\infty} \gamma^l R_{k+l+1}\middle| S_k = s_k, A_k = a_k\right\}, \tag{2}$$

with $0 \leq \gamma \leq 1$ being the discount rate and where the expected reward at time instant $k+1$, is $r_{k+1}(s_k, a_k) = \mathbb{E}\left[R_{k+1}|S_k = s_k, A_k = a_k\right]$. Optimal policies share the same *optimal action-value function* defined as

$$Q^*\left(s_k, a_k\right) = \arg\max_{\pi} Q_\pi\left(s_k, a_k\right), \qquad \forall s_k, \forall a_k. \tag{3}$$

State: The state vector $\mathbf{s}_k$ at time k contains the UAV location, the map of the environment and a detection variable, i.e., $\mathbf{s}_k = \left[\mathbf{p}_k, \mathbf{m}_k, \mathrm{t}_k\right]^\mathsf{T}$, where $\mathbf{p}_k = \left[x_k, y_k\right]^\mathsf{T} \in \mathbb{R}^2$ is the true UAV position, $\mathbf{m}_k \in \mathbb{B}^{N_\mathsf{cell}}$ is the true map at time k described as a vector of N_cell cells in which the map is discretized, and $\mathrm{t}_k \in \mathbb{B}$ is the target variable (equal to one if the target is present and zero otherwise). As the environment is considered stationary, it is $\mathrm{t}_k = \mathrm{t}$ and $\mathbf{m}_k = \mathbf{m}, \forall k$, with $\mathbf{m} = \left[m_1, \ldots, m_i, \ldots, m_{N_\mathsf{cell}}\right]^\mathsf{T}$, containing the occupancy value of each cell, i.e., $m_i \in \mathbb{B}$, and N_cell being the total number of cells. The state space is[2]

$$\mathcal{S} = \underbrace{\mathbb{R}^2}_{\text{UAV position}} \times \underbrace{\mathbb{B}^{N_\mathsf{cell}}}_{\text{Map}} \times \underbrace{\mathbb{B}}_{\text{Target}}. \tag{4}$$

Actions: The UAV navigation actions can be defined as $\mathbf{a}_k = \Delta\mathbf{p}_k = \left[\Delta x_k, \Delta y_k\right]^\mathsf{T} \in \mathbb{R}^2$ in terms of position displacement $\Delta\mathbf{p}_k$ according to $N_\mathrm{a} = 4$ actions, where the action space, for steps of Δ, is

$$\mathcal{A} = \Big\{\underbrace{[\Delta,\, 0]}_{\text{Right}}, \underbrace{[-\Delta,\, 0]}_{\text{Left}}, \underbrace{[0,\, \Delta]}_{\text{Up}}, \underbrace{[0,\, -\Delta]}_{\text{Down}}\Big\}. \tag{5}$$

Rewards: Following the *information foraging* philosophy [23, 24], we consider an extrinsic reward that is task-specific (detection) and it maps state-action pairs to a real-valued reward, and an intrinsic reward that only indirectly depends on the world state via the UAV internal belief of the state [23]. Intrinsic rewards are usually used for *reward shaping*, for example in situations with sparse rewards. The combination of intrinsic and extrinsic rewards allows to speed up the learning process and to get better policies. According to this formulation, the reward is defined as [23]

$$r_{k+1} = r_{\mathrm{i},\, k+1} + \eta\, r_{\mathrm{e},\, k+1}, \tag{6}$$

where we omitted the state and action dependence, η is a normalizing factor, $r_{\mathrm{i},\, k+1} = r_{\mathrm{c},k+1} + r_{\mathrm{m},k+1}$ is an intrinsic reward used for obtaining a sufficient knowledge of the surrounding environment, and $r_{\mathrm{e},\, k+1} = r_{\mathrm{d},k+1}$ is a reward for the considered UAV task. More specifically, $r_{\mathrm{d},k+1}$ is defined as the reward accounting for the detection rate that is

$$r_{\mathrm{d},k+1} = \mathcal{Q}_h(\sqrt{\lambda_k},\, \sqrt{\xi}), \tag{7}$$

where $\mathcal{Q}_h$ is the Marcum's $\mathcal{Q}$-function of order h, λ_k is the measured signal-to-noise ratio (SNR) at time instant k and ξ is the considered signal detection threshold [4, (37)], [25,26].

For each radar position, we also define a mapping reward both in terms of coverage ($r_{\mathrm{c},k+1}$) and accuracy ($r_{\mathrm{m},k+1}$) as

$$r_{\mathrm{c},k+1} \triangleq \frac{\sum_{i\in\mathcal{I}_k} \mathbf{1}(i \in \mathcal{D}_k)}{N_\mathsf{cell}}, \quad r_{\mathrm{m},k+1} \triangleq \frac{H_{k+1|k}(\mathbf{m})}{|\mathcal{I}_k|}, \tag{8}$$

where $\mathcal{D}_k \subseteq \mathcal{I}_k$ represents the subset of the indices of the intercepted cells that are discovered for the first time, and $\mathcal{I}_k$ is the set of the indices of all the cells illuminated by the radar at the kth time slot, and $\mathbf{1}(x) = 1$ if the logical condition x is verified, otherwise it is 0. Considering $r_{\mathrm{m},k+1}$, it holds

$$H_{k+1|k}(\mathbf{m}) = -\sum_{i\in\mathcal{I}_k} b_{k+1|k}(m_i)\, \log_2\left(b_{k+1|k}(m_i)\right), \tag{9}$$

where $H_{k+1|k}(\mathbf{m})$ represents the entropy indicating the level of lack of information about $\mathbf{m}$, $|\mathcal{I}_k|$ is the cardinality of $\mathcal{I}_k$ [4, (35)], and $b_{k+1|k}(m_i)$ is the predicted belief of occupancy state of the ith cell at time slot k. Note that such reward is designed in a way to favor actions that reduce the uncertainty about the environment in the shortest possible time. Finally we consider a numerical penalty for avoiding crashes with obstacles and targets.

[1] Note that π for a discrete state-action is a probability mass function.

[2] When the dimension of the state space is large (e.g., for large outdoors), policy iteration might suffer for the "curse of dimensionality" [22].

3. STATE ESTIMATION AND CONTROL

The CR on UAV is a system comprising two estimation processes. The first is a "*State Estimator*" that implements an occupancy grid (OG) for mapping and a detection module that determines if a target is present. The second step is a "*Policy Estimator*" for the UAV navigation.

3.1. State Estimator: Mapping with OG

The map of the environment is estimated using an OG algorithm [4], and energy measurements collected by the radar from each steering direction and different tested distances, according to the model described in [4, (13)] and [27, (35-37)].

Let $b_k(m_i)$ be the belief of the occupancy state of the ith cell at time instant k. Given the binary nature of m_i and to avoid numerical instability, the OG uses log-odds, defined as $\ell_k(m_i) \triangleq \log\left(\frac{b_k(m_i)}{1-b_k(m_i)}\right)$. The major steps are summarized as follows.

Initialization: The belief of each cell composing the map is initialized as $b_0(m_i) = 0.5$ (complete uncertainty).

Measurement Update: A new energy matrix is collected for each steering direction and time bin and it is compared with the expected received power, evaluated according to the THz scattering model of [20] and the actual knowledge of the map. More specifically, it accounts for the scattering term

$$\rho = 8\,\pi \frac{S^2\, L\, \cos(\theta_i)}{F_{\alpha_\mathrm{r}}} \left(\frac{1+\cos(\Psi)}{2}\right)^{\alpha_\mathrm{r}}, \tag{10}$$

where S is the scattering coefficient, θ_i is the incident angle with respect to the normal of the obstacle, $\Psi = \theta_s - \theta_r$ is the difference between the reflected (θ_r) and the scattered (θ_s) angles, and L is the length of the scattering object. F_{α_r} is a scaling factor, and α_r is the width of the scattering lobe.

Hence, the likelihood functions for the case of occupied/free cells (i.e., $p(\mathbf{o}_k|m_i)$) are computed as in [4, (22-23)], where $\mathbf{o}_k$ is the observation collected at the kth instant.

Log-Odd Update: Finally, for each time instant, the log-odd update is

$$\ell_k(m_i) = \log\left(\frac{p(\mathbf{o}_k|m_i)}{1-p(\mathbf{o}_k|m_i)}\right) + \ell_{k-1}(m_i). \tag{11}$$

3.2. Policy Estimator: Control with Q-learning

Q-learning is an off-policy temporal-difference (TD) control algorithm approach where the policy is learnt run-time while the UAV is navigating the environment. It is a model-free tabular algorithm whose main steps are reported in Alg. 1, where we included the possibility of choosing a random action with probability ϵ (ϵ-greedy approach). TD methods use a generalized policy iteration (GPI) mechanism to alternatively estimate the optimal policy in (1) and the optimal Q-value in (3).

Algorithm 1 Q-Learning Navigation for a Single Episode

```
Parameters: Set $(\gamma, \alpha, \epsilon)$ and the mission time $T_\mathsf{M}$;
Initialization: Initialize the $Q$-table to zeros, and $\mathbf{s}_0$ ;
while $k < T_\mathsf{M}$ do
    Generate a random value $\epsilon_k$;
    if $\epsilon_k < \epsilon$ then
        Choose a random action $\mathbf{a}_k \in \mathcal{A}$; (exploration)
    else
        Choose a greedy action $\mathbf{a}_k \in \mathcal{A}$ that corresponds to
        the maximum $Q$-value in $Q(\mathbf{s}_k, :)$; (exploitation)
    end
    UAV moves to the new state, collects the reward $r_{k+1}$ and
    updates the $Q$-table according to (12).
end
```

The advantages of using TD methods instead of Monte Carlo or dynamic programming is that there is no need of a model for the environment's dynamics and an update of the return is made at each time step.

Moreover, a sample return is considered instead of the expected return in (2) by the use of sample episodes. For discrete states and actions, the Q-value in (2) can be represented by a Q-table that, at each time instant, is updated as [21]

$$\begin{aligned} Q(\mathbf{s}_k, \mathbf{a}_k) \leftarrow & Q(\mathbf{s}_k, \mathbf{a}_k) + \\ & + \alpha\left[r_{k+1} + \gamma \max_{\mathbf{a}} Q(\mathbf{s}_{k+1}, \mathbf{a}) - Q(\mathbf{s}_k, \mathbf{a}_k)\right], \end{aligned} \tag{12}$$

where α is the learning rate, and the max operator is used to have a greedy policy. In this case, the learned action-value function directly approximates the optimal action-value function in (3), independently from the policy being followed.

4. CASE STUDY

We now assess the navigation and mapping performance by accounting for a realistic propagation environment and different radar parameters. For the THz scattering model, we set $S = 0.5$ (rough surface), $L = 0.5$ and $\alpha_\mathrm{r} = 1$ [20]. Then, we considered an effective radiated isotropic power (EIRP) of 30 dBm, a receiver noise figure of 4 dB, a transmitted signal with central frequency of 140 GHz, and 1 GHz bandwidth. The mapping is performed by a radar equipped with an antenna array of 100 antennas such that 10 steering directions are required for scanning the environment, and with a reading range (RR) that is alternatively set to 3 m and 7 m. The radar is initially assumed to be in $\mathbf{p}_0 = (2, 5)$ m and it moves with steps of $\Delta = 0.5$ m, equal to the cell width. For mapping parameters, we refer to [4]. For the detection module, we considered an antenna with 0 dBi gain, a RR of 7 m and a target always present and located alternatively in $(8.5, 1.5)$ m, and $(8.5, 8.5)$ m. We set ξ in (7) by considering a desired false alarm probability of 10^{-3}. We fixed $T_\mathsf{M} = 400$, $N_\mathsf{ep} = 20$ episodes, $\gamma = 0.99$, $\alpha = 0.9$, and

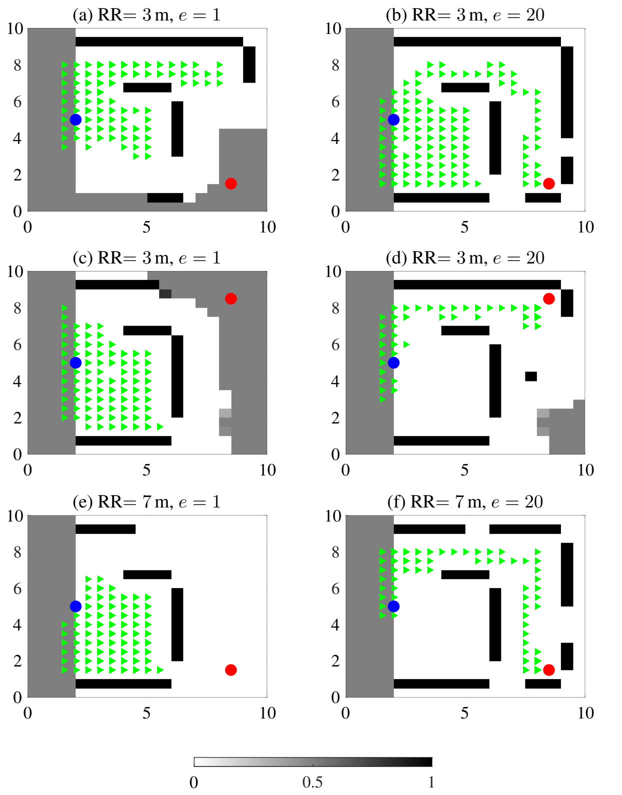

Fig. 2: Examples of estimated trajectories and maps for $e = 1$ (left) and $e = 20$ (right). Blue and red markers indicate $\mathbf{p}_0$ and the target position, respectively.

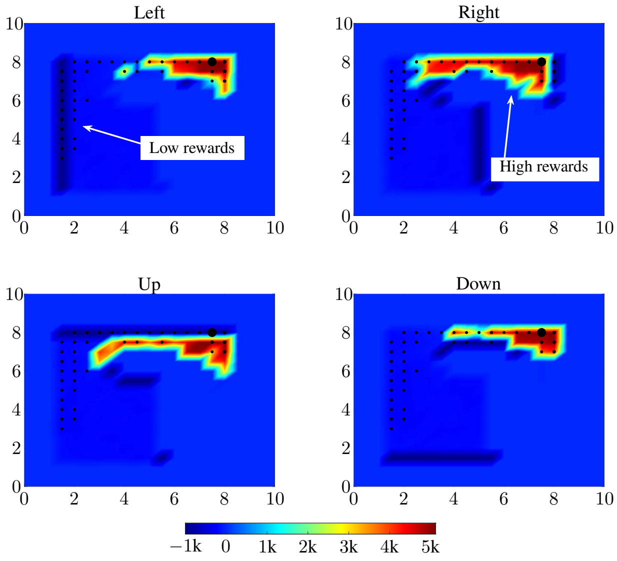

Fig. 3: Q-table related to Fig. 2-(e) with $k = T_{\mathsf{M}}$.

$\epsilon = \epsilon_{k,e}$, with $\epsilon_{k,e} = 0.2$, $\forall e > N_{\mathsf{ep}}/2$, otherwise it holds $\epsilon_{k,e} = [0.8, 0.6, 0.5, 0.3]$ for $k < [T_{\mathsf{M}}/4, T_{\mathsf{M}}/2, 3T_{\mathsf{M}}/4, T_{\mathsf{M}}]$.

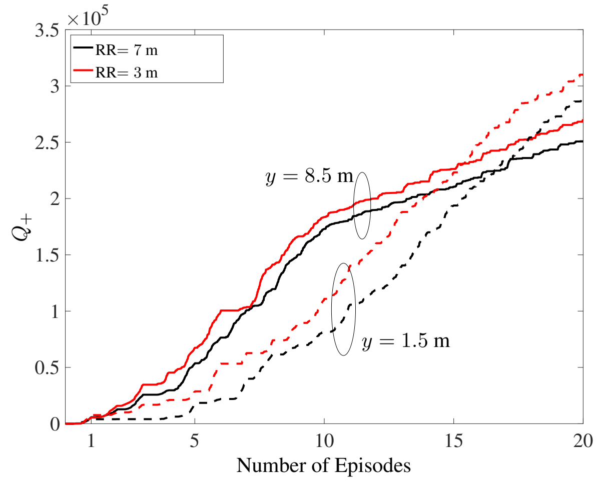

Fig. 4: Positive Q-values within a window of $N_{\mathsf{ep}} = 20$.

4.1. Results

Figure 2 shows the UAV trajectory with green markers for two different episodes, that is, $e = 1$ (left) and $e = 20$ (right), and for different radar RRs, that are RR $= 3$ m (top, middle) and RR $= 7$ m (bottom). According to the results, the UAV is capable of reconstructing a reliable copy of the map (see the reference map in Fig. 1) and of finding a good trajectory after some training episodes. In fact, during the first episode, i.e., for $e = 1$, the radar is still in an exploratory phase, as evidenced by the scarce map reconstruction, and by the followed non–optimized trajectory. This can be explained by the fact that the detection reward is sparse in the environment and mapping rewards tend faster to zero, especially for high RR. Fig. 3 reports the Q-table related to the last instant of Fig. 2-(e) for each possible action. For example, a UAV located in $(4, 7)$ will receive the highest reward by choosing the right action. By contrast, a UAV in $(2, 5)$ will receive the lowest reward by choosing the left action. Finally, Fig. 4 reports the behavior of the positive Q-values as a function of the number of episodes. Notably, for shorter RR, the UAV, driven by curiosity, is pushed to explore more, thus increasing the amount of received rewards.

5. CONCLUSION

In this paper we showed the UAV capability for autonomous navigation of an environment to accomplish the goal of detecting a target and of reconstructing a map of the indoors. We considered a Q-learning approach with a combination of intrinsic and extrinsic rewards. Our results show the possibility of attaining the objective by means of a THz radar, which augments its ambient awareness at each episode and improves its capability of accomplishing the assigned task of target detection.

6. REFERENCES

[1] S. Haykin, "Cognitive radar: a way of the future," *IEEE Signal Process. Mag.*, vol. 23, no. 1, pp. 30–40, 2006.

[2] A. M. Ahmed et al., "A reinforcement learning based approach for multi-target detection in massive MIMO radar," *IEEE Trans. Aerosp. Electron. Syst.*, pp. 1–1, 2021.

[3] P. Liu et al., "Decentralized automotive radar spectrum allocation to avoid mutual interference using reinforcement learning," *IEEE Trans. Aerosp. Electron. Syst.*, 2020.

[4] A. Guerra et al., "Reinforcement learning for UAV autonomous navigation, mapping and target detection," in *Proc. IEEE/ION Pos. Loc. Nav. Symp.*, 2020, pp. 1004–1013.

[5] S. Zhang, R. Raulefs, and A. Dammann, "Location information driven formation control for swarm return-to-base application," in *Proc. European Signal Process. Conf.* IEEE, 2016, pp. 758–763.

[6] F. Koohifar et al., "Autonomous tracking of intermittent RF source using a UAV swarm," *IEEE Access*, vol. 6, pp. 15884–15897, 2018.

[7] Emanuel Staudinger et al., "The role of time in a robotic swarm: A joint view on communications, localization, and sensing," *IEEE Commun. Mag.*, 2021.

[8] A. Guerra, D. Dardari, and P. M. Djurić, "Dynamic radar network of UAVs: A joint navigation and tracking approach," *IEEE Access*, vol. 8, pp. 116454–116469, 2020.

[9] E. Testi, E. Favarelli, and A. Giorgetti, "Reinforcement learning for connected autonomous vehicle localization via UAVs," in *Proc. IEEE Int. Workshop Metrology Agriculture Forestry*, 2020, pp. 13–17.

[10] K. Gu, Y. Wang, and Y. Shen, "Cooperative detection by multi-agent networks in the presence of position uncertainty," *IEEE Trans. Signal Process.*, vol. 68, pp. 5411–5426, 2020.

[11] A. Guerra, D. Dardari, and P. M. Djurić, "Dynamic radar networks of UAVs: A tutorial overview and tracking performance comparison with terrestrial radar networks," *IEEE Veh. Technol. Mag.*, vol. 15, no. 2, pp. 113–120, 2020.

[12] L. Wielandner, E. Leitinger, and K. Witrisal, "Information-criterion-based agent selection for cooperative localization in static networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2020, pp. 1–7.

[13] C. Wang et al., "Autonomous navigation of uavs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2124–2136, 2019.

[14] S. Zhang et al., "Self-aware swarm navigation in autonomous exploration missions," *Proc. IEEE*, vol. 108, no. 7, pp. 1168–1195, 2020.

[15] H. Bayerlein et al., "Multi-UAV path planning for wireless data harvesting with deep reinforcement learning," *arXiv preprint arXiv:2010.12461*, 2020.

[16] M. Theile et al., "UAV path planning using global and local map information with deep reinforcement learning," *arXiv preprint arXiv:2010.06917*, 2020.

[17] O. Esrafilian, R. Gangula, and D. Gesbert, "3D Map-based trajectory design in UAV-aided wireless localization systems," *IEEE Internet of Things J.*, 2020.

[18] H. Bayerlein et al., "UAV path planning for wireless data harvesting: A deep reinforcement learning approach," *arXiv preprint arXiv:2007.00544*, 2020.

[19] M. Lotti et al., "Radio simultaneous localization and mapping in the terahertz band," in *Proc. 25th Int. ITG Workshop on Smart Antennas*, 2021.

[20] S. Ju et al., "Scattering mechanisms and modeling for terahertz wireless communications," in *Proc. IEEE Int. Conf. Commun.*, 2019, pp. 1–7.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[22] R. S. Sutton et al., *Introduction to reinforcement learning*, vol. 135, MIT press Cambridge, 1998.

[23] N. Mafi, F. Abtahi, and I. Fasel, "Information theoretic reward shaping for curiosity driven learning in POMDPs," in *Proc. IEEE Int. Conf. Develop. Learning (ICDL)*, 2011, vol. 2, pp. 1–7.

[24] I. Fasel et al., "Intrinsically motivated information foraging," in *Proc. IEEE 9th Int. Conf. Develop. Learning*, 2010, pp. 101–107.

[25] A. Mariani, A. Giorgetti, and M. Chiani, "Effects of noise power estimation on energy detection for cognitive radio applications," *IEEE Trans. Commun.*, vol. 59, no. 12, pp. 3410–3420, Dec. 2011.

[26] M. Chiani, "Integral representation and bounds for Marcum Q-function," *Electronics Lett.*, vol. 35, no. 6, pp. 445–446, 1999.

[27] F. Guidi, A. Guerra, and D. Dardari, "Personal mobile radars with millimeter-wave massive arrays for indoor mapping," *IEEE Trans. Mobile Comput.*, vol. 15, no. 6, pp. 1471–1484, 2015.