A DEEP REINFORCEMENT LEARNING APPROACH FOR INTEGRATED AUTOMOTIVE RADAR SENSING AND COMMUNICATION

Lifan Xu, Ruxin Zheng and Shunqiao Sun

Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL, USA

2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM) | 978-1-6654-0633-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SAM53842.2022.9827815

ABSTRACT

We present a deep reinforcement learning approach to design an automotive radar system with integrated sensing and communication. In the proposed system, sparse transmit arrays with quantized phase shifter are used to carry out transmit beamforming to enhance the performance of both radar sensing and communication. Through interaction with environment, the automotive radar learns a reward that reflects the difference between mainlobe peak and the peak sidelobe level in radar sensing mode or communication user feedback in communication mode, and intelligently adjust its beamforming vector. The Wolpertinger policy based action-critic network is introduced for beamforming vector learning, which solves the dimension curse due to huge beamforming action space.

***Index Terms*—** Deep reinforcement learning, integrated sensing and communication (ISAC), automotive radar, sparse array, beamforming

I. INTRODUCTION

Millimeter Wave (mmWave) vehicle radar, which operates at 76-81 GHz, is one of the key technologies of autonomous driving system and can sense the environment under various weather conditions [1]. Multi-input multi-output (MIMO) radar is a widely used cost-effective and scalable solution for increasing antenna aperture size by exploiting the idea of virtual sum coarray [1]. The use of sparse arrays combined with MIMO radars can further reduce costs without losing angular resolution [2]. Autonomous vehicles need to exchange information with road infrastructure and other neighbor vehicles to achieve operation coordination, especially in vehicle platooning [3]. Traditionally, the automotive radar sensing and vehicle communication functions are implemented via separated hardware. A dual-function radar communication (DFRC), or integrated sensing and communication (ISAC) system utilizes the same hardware platform to send electromagnetic waves for both environment sensing and communication with neighboring devices [4]–[10], which has found applications in autonomous vehicles [11], [12].

Different strategies have been considered to design the ISAC system. For example, array beampattern, such as the null region of the transmit pattern [13] or the unique steering vector through transmit antenna selection [14] has been used to encode the communication message. On the other hand, optimization techniques have been proposed to optimize the transmit matrix to realize sensing and communication functions simultaneously by forming multiple beams [5], [15]. A data-driven deep learning method has been proposed to realize the ISAC system by forming the transmit matrix in a quantized manner [16]. However, the generalization of the data-driven deep learning approach to the scenarios that are not considered during the training process is limited.

In practical systems, the values of antenna phase shifter are quantized instead of continuous. For example, the Texas Instruments (TI) imaging radar cascaded by four AWR2243 radar chipsets supports 6-bit phase shifter [17]. When the number of bits of phase shifter and antenna elements increase, it becomes challenging to use optimization or data-driven machine learning approaches to design transmit beamforming. Deep reinforcement learning has been introduced in beamforming design for mobile communications [18], [19] using quantized phase shifters. Value-based reinforcement learning, such as deep Q-networks (DQNs) [20], [21], has great generalization performance and has high training efficiency. However, its computation complexity is extremely high due to the dimensional curse.

In this paper, we present a Wolpertinger policy-assisted [22] reinforcement learning framework to intelligently learn a quantized transmit beamforming vector for a sparse tranmit array in the automotive radar ISAC system. The action-critic network is used to deal with the dimension curse of deep reinforcement learning techniques due to huge action space. The proposed ISAC automotive radar system not only improves signal-to-noise ratio (SNR) for both radar sensing and communication, but also is able to reduce mutual radar interference [23], multipath propagation, especially in urban streets and tunnel scenarios.

II. SYSTEM MODEL

We consider an automotive radar ISAC system exploiting frequency-modulated continuous-wave (FMCW) consisting of N_t transmit and N_r receiver antennas.

Assume there are N_s data streams to transmit, and there are N_{RFt} radio frequency (RF) beamforemers connected to N_t transmit antennas. We assume $N_s \leq N_{RFt} \leq N_t$. The

This work has been funded in part by U.S. National Science Foundation (NSF) under Grant CCF-2153386.

system contains a baseband precoder $\mathbf{F}_{BB} \in \mathbb{C}^{N_{RFt} \times N_s}$ and a RF precoder $\mathbf{F}_{RF} \in \mathbb{C}^{N_t \times N_{RFt}}$, constructed using phase shifters. The FMCW radar emits N consecutive chirps with carrier frequency f_c, bandwidth B, and pulse duration time T. The emitted waveform at the m-th transmitter during the n-th chirp is

$$x_m(n,t) = e^{j2\pi\left(f_c t + \frac{1B}{2T} t^2\right)} e^{j2\pi\Omega_n}, \tag{1}$$

where $e^{j2\pi\Omega_n}$ is the symbol along slow time that can be used to encode the communication message. Automotive radar ISAC system switches between radar sensing and communication functions. The system block diagram of integrated automotive radar sensing and communication is shown in Fig. 1.

To simply our formulation, we consider single RF chain, i.e., $N_{RFt} = 1$. All transmitters carry the same communication symbol, i.e., $N_s = 1$. Therefore $\mathbf{F}_{BB}$ is not applied, and only analog beamforming $\mathbf{F}_{RF} \in \mathbb{C}^{N_t \times 1}$ will be employed through q-bit quantized phase shifters design. The phase is selected from a quantized set $\mathcal{D}$ with 2^q possible phase values that are uniformly distributed in $(-\pi, \pi]$. During the n-th chirp, transmit signal from N_t transmit antennas forms a vector as

$$\mathbf{x}(n,t) = e^{j2\pi\left(f_c t + \frac{1B}{2T} t^2\right)} e^{j2\pi\Omega_n} \mathbf{F}_{RF}. \tag{2}$$

II-A. Radar Transmit Beamforming

The radar transmit beampattern can be expressed as

$$G(\theta) = \mathbf{b}_t^H(\theta) \mathbf{W} \mathbf{b}_t(\theta), \tag{3}$$

where the array response $\mathbf{b}_t(\theta)$ is

$$\mathbf{b}_t(\theta) = \left[e^{\frac{j2\pi d_1 \sin\theta}{\lambda}}, e^{\frac{j2\pi d_2 \sin\theta}{\lambda}}, \cdots, e^{\frac{j2\pi d_{N_t} \sin\theta}{\lambda}} \right]^T,$$

with $d_m, m = 1, \cdots, N_t$ denoting the transmit antennas location. The precoded waveform $\mathbf{W} \in \mathbb{C}^{N_t \times N_t}$ is given as

$$\mathbf{W} = \mathbb{E}\left[\mathbf{x}(n,t) \mathbf{x}^H(n,t) \right] = \beta \mathbf{F}_{RF} \mathbf{F}_{RF}^H, \tag{4}$$

where β denotes the expectation of the FMCW chirp signal. The automotive radar sensing function aims to design the analog beamforming matrix $\mathbf{F}_{RF}$ so that its main beam is steered to the target of interest through adjusting its quantized phase shifter. The analog beamformer $\mathbf{F}_{RF}$ is replaced with a beamformer $\mathbf{w}_r$, defined below

$$\mathbf{w}_r = 1/\sqrt{N_t} \left[e^{j\phi_1}, e^{j\phi_2}, \cdots, e^{j\phi_{N_t}} \right]^T, \tag{5}$$

where phase $\phi_m, m = 1, \cdots, N_t$ is chosen from a quantized set $\mathcal{D}$ with 2^q possible phase values.

Assuming there are K targets located in the radar main beam θ, the radar received signals during the n-th chirp is

$$\mathbf{y}_r(n,t) = \sum_{k=1}^{K} \alpha_k e^{j2\pi\left[f_c(t-\tau_k) + \frac{B}{2T}(t-\tau_k)^2 + \Omega_n\right]} \mathbf{1}_{N_t \times 1}^T \mathbf{w}_r \mathbf{b}_r + \mathbf{n}(n,t),$$

where α_k and τ_k are the reflection coefficient and delay of the k-th target, respectively. Here, $\mathbf{b}_r \in \mathbb{C}^{N_r \times 1}$ denotes the receive steering vector.

II-B. Communication System

Assume there are N_c antennas at a communication receiver. The mmWave channels are expected to have limited scattering [24]. We consider a geometric channel model with L independent propagation paths, where the value of L is small compared to N_t for limited scattering. Then the downlink channel matrix $\mathbf{H} \in \mathbb{C}^{N_c \times N_t}$ can be expressed as

$$\mathbf{H} = \sqrt{\frac{N_t N_c}{L}} \sum_{l=1}^{L} \alpha_l \mathbf{b}_c(\theta_{cl}) \mathbf{b}_t^H(\theta_{tl}), \tag{6}$$

where α_l is the l-th complex path gain. Here, $\mathbf{b}_c(\theta_{cl})$ and $\mathbf{b}_t(\theta_{tl})$ are communication system receive and radar transmit array steering vectors, respectively, with θ_{cl} and θ_{tl} as the angles of arrival and departure of the l-th path. The received signal at communication receiver is

$$\mathbf{y}_c(n,t) = \sqrt{\rho} e^{j2\pi\left[f_c(t-\tau_c) + \frac{1B}{2T}(t-\tau_c)^2\right]} e^{j2\pi\Omega_n} \mathbf{H} \mathbf{F}_{RF} + \mathbf{n}(n,t), \tag{7}$$

where ρ denotes the average received power, and τ_c is the delay between radar transmitter and communication receiver. In the communication mode, the analog beamformer $\mathbf{F}_{RF}$ is replaced with a beamformer $\mathbf{w}_c$, defined below

$$\mathbf{w}_c = 1/\sqrt{N_t} \left[e^{j\Phi_1}, e^{j\Phi_2}, ..., e^{j\Phi_{N_t}} \right]^T, \tag{8}$$

where phase $\Phi_m, m = 1, \cdots, N_t$ is chosen from a quantized set $\mathcal{D}$ with 2^q possible phase values.

In the following, we aim to develop a deep reinforcement learning approach to intelligently adjust the beamforming vectors $\mathbf{w}_r$ and $\mathbf{w}_c$ such that the target of interest and communication receiver are illuminated under the main beam alternatively by the automotive radar.

III. TRANSMIT BEAMFORMING DESIGN

To achieve a high angular resolution at low hardware cost, a sparse transmit antenna array is considered. The sparse antenna array geometry is selected such that the peak sidelobe level (PSL) is minimized. For N_t antennas with q-bit quantized phase shifter, the total number of phase states is $2^{q \times N_t}$. This dimension becomes exploded as the number of array elements increases and a higher resolution quantized phase shifter is adopted. Traditional optimization techniques are not suitable for optimal beamforming vector design due to the high computational complexity. We propose an action-critic architecture based deep reinforcement learning framework to adjust the quantized phase of transmit antennas to form the desired transmit beamforming.

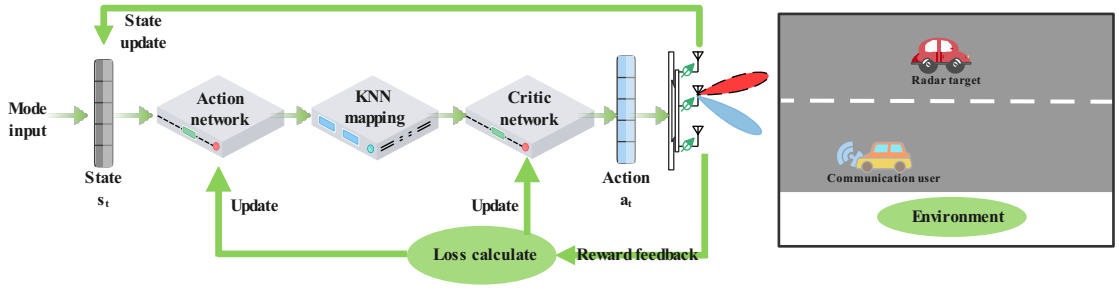


Fig. 1: The proposed automotive radar ISAC system diagram with action-critic network.

III-A. Deep Reinforcement Learning

Leveraging the trial and reward mechanism, reinforcement learning intelligently searches for strategy π [25] by mapping state s_t, action a_t and reward r_t to the action-value function $Q^*(s_t, a_t)$ as $Q^*(s_t, a_t) = \max_\pi \mathbb{E}[r_t | s_t = s, a_t = a, \pi]$. Yet, the problem under reinforcement learning exploiting DQN becomes intractable due to the extremely large action dimension. To mitigate this dimension explosion problem, we adapt the Wolpertinger policy-based reinforcement learning framework to adjust the phase shifters. The Wolpertinger policy includes an action network, k-nearest neighbor (KNN) mapping, and a critic network [22] (see Fig. 1).

1) **Action Network:** This network is an approximator parameterized by θ^π to map the input state to an output proto-action. The KNN technique is used to map the proto-action to the feasible space.

2) **Critic Network:** This network parameterized by θ^Q evaluates the Q-value of all action-state pairs from the KNN mapping by $a_t = \arg \max_{a_t \in g_t(\hat{a}_t)} Q_{\theta^Q}(s_t, a_t)$. The loss function is given as $Loss = (y - Q_{\theta^Q}(s, a))^2$, where y is the output of the target network that is used to generate independent and identically distributed (i.i.d.) data to stabilize the training process. The deep deterministic policy gradient (DDPG) is used to train networks [26].

Different from the DQN, the Wolpertinger policy introduces the KNN block to do time-wise tractable training and it has a logarithmic-time lookup complexity.

III-B. Beamforming Design with DRL

The beamforming design is implemented using the action-critic network.

1) **Action Space:** Consider a transmit antennas array with N_t elements, and each element is equipped with a q-bit quantized phase shifter. The dimension of action is $\mathbb{R}^{2^{q \times N_t}}$ and each action is mapped to a beamforming vector $\mathbf{w}_r$. One element phase change is equal to taking an action from the action space.

2) **State:** After taking an action from the action space, the state is changed. The state vector $\mathbf{s}$ is consisted of the transmit array phase shifter status. At the i-th iteration, the state $\mathbf{s}_i = [\theta_1, \theta_2, ..., \theta_{N_t}]$, where θ_{n_t} is selected from the quantized set $\mathcal{D}$.

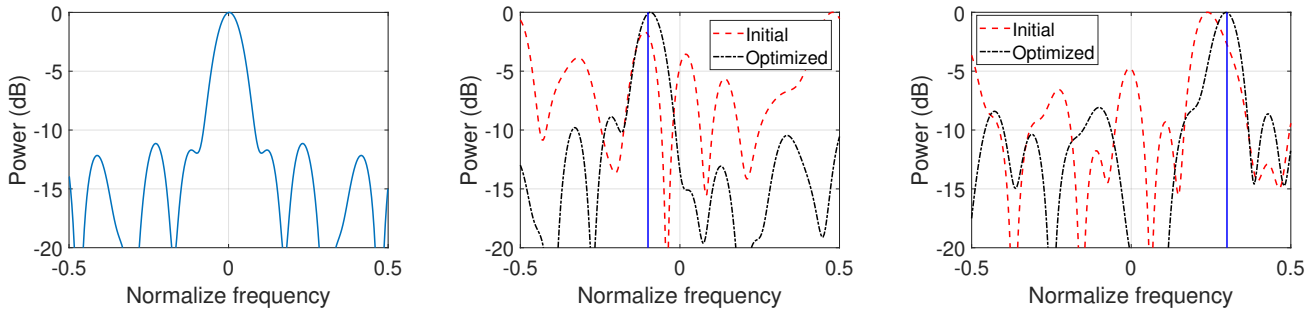
Algorithm 1 DRL based automotive radar ISAC system

```
Initialize networks with corresponding parameters.
Mode selection.
if mode = radar then
    Initialize ξ_0 = 0 and g_r0 = 0.
else
    Initialize g_c0 = 0.
end if
Initial sample a random beamforing vector w_1 as initial state s_1 and record action a_1.
for i = 1 to T do
    Receive proto-action â_i from actor network.
    Action embedding g(â_i) through KNN mapping.
    Execute action a_i passed from critic network, calculate reward and update state s_{i+1} = a_i.
    Update ξ_i and g_ri or g_ci according selected mode.
    Update all networks.
end for
```

3) **Reward:** In the radar tracking mode, the region of interest (ROI) is known in the radar search mode when targets' range, Doppler and angle are estimated. Suppose the interest main beam region is θ_{ROI}, and the 3-dB beam width is given by $\Delta_\theta = 2\arcsin\left(\frac{1.4\lambda}{\pi D}\right)$. Thus, the ROI is spanned from $[-1/2\Delta_{\theta\text{ROI}}, 1/2\Delta_{\theta\text{ROI}}]$. The areas out of the 3-dB mainlobe is defined as sidelobe regions. The difference between mainlobe peak and PSL is $\xi_i = \max(\text{P}_{\text{ROIi}}) - \max(\text{P}_{\text{SLi}})$ at i-th iteration. Here, $\max(\text{P}_{\text{ROIi}})$ and $\max(\text{P}_{\text{SLi}})$ are maximum main beam level and PSL at i-th iteration, respectively. The received power is $g_r = \left|\mathbf{y}_r^H(n,t)\mathbf{y}_r(n,t)\right|$. The term ξ and g_r are used to guarantee the main beam is steered to ROI, while the PSL is minimized. The reward is given by

$$r_i = \begin{cases} 1, & \text{if } \xi_i > \xi_{i-1} \\ 0, & \text{if } \xi_i \le \xi_{i-1} \text{ and } g_{ri} \le g_{ri-1} \\ -1, & \text{if } \xi_i \le \xi_{i-1} \text{ and } g_{ri} > g_{ri-1}. \end{cases} \quad (9)$$

For communication mode, the received gain is expressed as $g_c = |\mathbf{H}\mathbf{w}_c|^2$. Assume the channel parameters have been



(a) (b) (c)

Fig. 2: (a) Beampattern of sparse transmit array; (b) The transmit beamforming under radar mode; (c) The transmit beamforming under communication mode; Ground truth directions are indicated in blue lines.

Table I: Hyper-parameters for training

Parameter	Value	
Models	Actor-Net	Critic-Net
Relay Buffer	4096	4096
Mini-batch	64	64
Learning rate	0.01	0.01
Decay	0.01	0.001

estimated. The reward in communication mode is given by

$$r_i = \begin{cases} 1, & \text{if} \quad g_{ci} > g_{ci-1} \\ 0, & \text{if} \quad g_{ci} = g_{ci-1} \\ -1, & \text{if} \quad g_{ci} < g_{ci-1}. \end{cases} \tag{10}$$

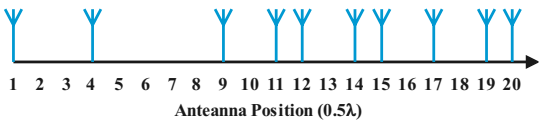
This dynamic gain will be reported to automotive radar by communication user through an uplink channel.

The pseudo code of the DRL based automotive ISAC using Wolpertinger policy is given by Algorithm 1.

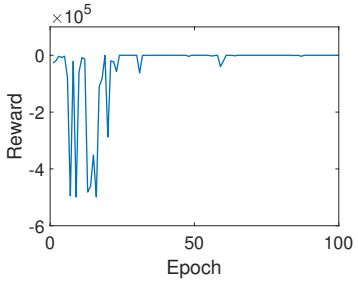
IV. NUMERICAL RESULTS

We consider a FMCW radar system with 10 transmit antennas, and field of view of 60°. The transmit sparse array has a physical aperture size of 10λ, with corresponding 3-dB beamwidth as $\Delta_f = 0.09$. The transmit array geometry is shown in Fig. 3 (a), and its beampattern is shown in Fig. 2 (a). Each antenna has a 2-bit quantized phase shifter, and therefore the total quantized action space has dimension over 1 million, i.e., $2^{2\times 10} = 1,048,576$. The hyper-parameters for model training is described in Table I. All networks are trained on a Lambda machine with an Intel Core^{TM} i9-10920X CPU and 4 Nvidia Quadro RTX 6000 GPU.

We first switch our system to radar sensing mode, and the normalized ROI frequency is set as $f = -0.1$ corresponding to $\theta = -6^\circ$. In the initial state, the transmit array was assigned a random phase. As shown in Fig. 2 (b), after sufficient iterations, an optimized beam vector is learned, and under which, the transmit beamforming is steered to the desired direction, while the sidelobe level is suppressed as well. In the communication mode, a communication user



(a)



(b)

Fig. 3: (a) The transmit sparse array geometry; (b) The reward during the training.

is located at the $f = 0.3$ corresponding to $\theta = 17^\circ$. The performance comparison between an initial randomly generated beamforming vector and the learned optimal beamvector is shown in Fig. 2 (c). Note the main beam is slightly off the ground truth, which may be due to the limited quantized resolution of phase shifters. As Fig. 3 (b) shown, after 20 epochs of training, the network can intelligently adjust the phases so that the main beam is steered to the ROI based on the current observation state.

V. CONCLUSIONS

We proposed Wolpertinger policy based reinforcement learning framework to design an automotive radar ISAC system, which can intelligently adjust its quantized phase shifters to steer its main beam to track target of interest or enhance communication capacity. The proposed approach works well for an extremely large dimension of action space while avoiding exhausted action search. The feasibility of proposed method has been validated via simulations.

VI. REFERENCES

[1] S. Sun, A. P. Petropulu, and H. V. Poor, "MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 98–117, 2020.

[2] S. Sun and Y. D. Zhang, "4D automotive radar sensing for autonomous vehicles: A sparsity-oriented approach," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 879–891, 2021.

[3] J. Axelsson, "Safety in vehicle platooning: A systematic literature review," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1033–1045, 2016.

[4] B. Li, A. P. Petropulu, and W. Trappe, "Optimum co-design for spectrum sharing between matrix completion based MIMO radars and a MIMO communication system," *IEEE Trans. Signal Process.*, vol. 64, no. 17, pp. 4562–4575, 2016.

[5] F. Liu, L. Zhou, C. Masouros, A. Li, W. Luo, and A. Petropulu, "Toward dual-functional radar-communication systems: Optimal waveform design," *IEEE Trans. Signal Process.*, vol. 66, no. 16, pp. 4264–4279, 2018.

[6] A. Hassanien, M. G. Amin, E. Aboutanios, and B. Himed, "Dual-function radar communication systems: A solution to the spectrum congestion problem," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 115–126, 2019.

[7] L. Zheng, M. Lops, Y. C. Eldar, and X. Wang, "Radar and communication coexistence: An overview: A review of recent methods," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 85–99, 2019.

[8] K. V. Mishra, M. R. B. Shankar, V. Koivunen, B. Ottersten, and S. A. Vorobyov, "Toward millimeter-wave joint radar communications: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 100–114, 2019.

[9] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3834–3862, 2020.

[10] L. G. de Oliveira, B. Nuss, M. B. Alabd, A. Diewald, M. Pauli, and T. Zwick, "Joint radar-communication systems: Modulation schemes and system design," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 3, pp. 1521–1551, 2022.

[11] D. Ma, N. Shlezinger, T. Huang, Y. Liu, and Y. C. Eldar, "Joint radar-communication strategies for autonomous vehicles: Combining two key automotive technologies," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 85–97, 2020.

[12] ——, "FRaC: FMCW-based joint radar-communications system via index modulation," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1348–1364, 2021.

[13] X. Wang, A. Hassanien, and M. G. Amin, "Dual-function MIMO radar communications system design via sparse array optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 3, pp. 1213–1226, 2019.

[14] A. Hassanien, M. G. Amin, Y. D. Zhang, and F. Ahmad, "Dual-function radar-communications: Information embedding using sidelobe control and waveform diversity," *IEEE Trans. Signal Process.*, vol. 64, no. 8, pp. 2168–2181, 2015.

[15] A. M. Elbir, K. V. Mishra, and S. Chatzinotas, "Terahertz-band joint ultra-massive MIMO radar-communications: Model-based and model-free hybrid beamforming," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1468–1483, 2021.

[16] A. M. Elbir and K. V. Mishra, "Joint antenna selection and hybrid beamformer design using unquantized and quantized deep learning networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1677–1688, 2019.

[17] Texas Instruments Inc., "Design guide: TIDEP-01012 imaging radar using cascaded mmWave sensor reference design (REV. A)," [Available Online] https://www.ti.com/lit/ug/tiduen5a/tiduen5a.pdf, Mar., 2020.

[18] Y. Zhang, M. Alrabeiah, and A. Alkhateeb, "Reinforcement learning of beam codebooks in millimeter wave and Terahertz MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 904–919, 2022.

[19] ——, "Reinforcement learning for beam pattern design in millimeter wave and massive MIMO systems," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Nov. 1-4 2020, pp. 445–449.

[20] A. M. Ahmed, A. A. Ahmad, S. Fortunati, A. Sezgin, M. S. Greco, and F. Gini, "A reinforcement learning based approach for multitarget detection in massive MIMO radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 5, pp. 2622–2636, 2021.

[21] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[22] G. Dulac-Arnold, R. Evans, H. van Hasselt, P. Sunehag, T. Lillicrap, J. Hunt, T. Mann, T. Weber, T. Degris, and B. Coppin, "Deep reinforcement learning in large discrete action spaces," *arXiv preprint arXiv:1512.07679*, 2015.

[23] S. Alland, W. Stark, M. Ali, and A. Hedge, "Interference in automotive radar systems: Characteristics, mitigation techniques, and future research," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 45–59, 2019.

[24] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, 2011.

[25] Y. Zhai, C. Baek, Z. Zhou, J. Jiao, and Y. Ma, "Computational benefits of intermediate rewards for goal-reaching policy learning," *Journal of Artificial Intelligence Research*, vol. 73, pp. 847–896, 2022.

[26] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.