# Reconfigurable Beamforming for Automotive Radar Sensing and Communication: A Deep Reinforcement Learning Approach

Lifan Xu, *Graduate Student Member, IEEE*, Shunqiao Sun, *Senior Member, IEEE*, Yimin D. Zhang, *Fellow, IEEE*, and Athina P. Petropulu, *Fellow, IEEE*

***Abstract*— In this article, we present a novel low-cost, dual-function radar-communication system that addresses dynamic environments such as those arising in automotive applications. The low cost is achieved by using a sparse phased arrays equipped with quantized double-phase shifters. The operation in dynamic environments is achieved via a deep reinforcement learning (DRL) approach that adaptively selects a small subset of transmit antennas and adjusts the phase shifters such that the transmitted energy is concentrated on the communication user and the target of interest, while the interference to other radars is reduced. The action space in the DRL approach increases fast with the number of antennas and the number of bits used in quantization, and as a result the complexity of the design problem grows exponentially. To tackle the resulting curse of dimensionality in the action space, we adopt the Wolpertinger strategy, which incorporates the nearest neighborhood component to project the vast action space into a smaller, more manageable space while maintaining the desired performance. Numerical results demonstrate the feasibility of our proposed method.**

***Index Terms*— Adaptive beamforming, automotive radar, deep reinforcement learning (DRL), dual-function radar-communication (DFRC), sparse array**

## I. Introduction

As self-driving technology advances, the integration of vehicle-to-everything (V2X) communication into automotive radar can help enhance road safety, alleviate traffic congestion, and improve the driving experience. By enabling vehicles to communicate with other vehicles, infrastructure, pedestrians, and networks, V2X can facilitate the creation of a more connected and intelligent transportation system [3], [4]. However, as radar becomes an integral part of automobiles and the demand for communication functions increases, limited spectrum resources may become a challenge. Thus, finding ways for radar and communication functions to coexist in the same band has been drawing a lot of attention [5]–[10]. Dual-function radar-communication (DFRC) systems are a new type of technology that offers the radar and communication functionalities out of a single platform and via the same waveform, thereby offering efficient spectrum use, reduced hardware complexity, improved safety, higher accuracy, and new application possibilities [11]–[23]. DFRC systems are ideally suited for autonomous driving vehicles. Unlike communication systems at cellular base stations, automotive radars use a relatively small number of antennas, and thus greatly benefit from sharing antennas and waveform for both radar and communication functions.

Several key factors must be considered in the design of automotive DFRC systems. First, for the possibility of mass production, it is essential that these systems are cost-effective while providing high angular resolutions. This requirement makes sparse arrays with large apertures particularly attractive because they offer a balanced trade-off between performance and cost [24]. To achieve low-cost hardware solutions, phase shifters are designed to select discrete values from a predefined set, limiting the design flexibility of transmit signals due to the constraints imposed by the phase shifters' degrees of freedom. For instance, the Texas Instruments AWR2243 radar chipset utilizes 6-bit phase shifters [25]. Second, current automotive multi-input multi-output (MIMO) radars use orthogonal waveforms, thus spreading the transmit energy over the entire field of view (FOV) [26], [27]. This may create problems, such as mutual interference [27], [28] and ghost targets due to multipath wave propagation. Third, automotive DFRC systems operate in highly dynamic environments and must be capable of detecting targets with low radar cross-sections (RCSs), such as pedestrians and cyclists, amidst stronger reflective objects like vehicles and light poles. Current state-of-the-art systems heavily rely on MIMO radar technology with static transmit parameters, including fixed array geometries and non-adaptive orthogonal waveforms [26], [27], which may not always provide optimal sensing performance in dynamic automotive conditions.

Part of this work was presented in the IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2022) [1] and the 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023) [2]. This work was supported in part by U.S. National Science Foundation (NSF) under Grants CCF-2153386, ECCS-2340029, ECCS-2033433, ECCS-2236023, ECCS-2320568, ONR under W911NF2320103, and Alabama Transportation Institute (ATI).

L. Xu and S. Sun are with the Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL, USA (emails: lxu36@crimson.ua.edu, shunqiao.sun@ua.edu).

Y. D. Zhang is with the Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA, USA (email: ydzhang@temple.edu).

A. P. Petropulu is with the Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ, USA (email: athinap@rutgers.edu).

To address the above automotive DFRC system challenges, in this paper, we propose a low-cost, reconfigurable adaptive beamforming scheme in automotive DFRC system to enhance both sensing and communication performance while suppressing mutual interference. Low-cost and low power implementation is afforded via the use of sparse phased arrays equipped with quantized double-phased shifters (DPSs). Interference suppression and detection of weak targets is achieved through beamforming. Unlike the high cost of digital hardware associated with MIMO radar, our approach uses a phased array equipped with DPSs. The DPSs enable the design of flexible beams that can suppress automotive radar mutual interference, minimize the occurrence of ghost targets due to multipath, and enhance the signal-to-noise ratio (SNR) of targets with low RCS in the tracking phase. The introduction of DPSs opens up the possibility of forming richer beams towards both communication user and target of interest [29], [30]. The use of sparse phased arrays reduces hardware cost and power consumption. By optimally designing the transmit array along with the beamforming weights, one can closely approximate the beampattern performance of a full array. However, the high complexity of the design problem remains a challenging bottleneck. Commonly used optimization methods for beamformer optimization, such as linear programming and alternative optimization [31]–[36], lead to daunting computational costs due to the NP-hard nature of the problem given the selection of antennas and quantized phase shifters. We propose a deep reinforcement learning (DRL) approach as a feasible solution to adaptively select the antennas and adjust the phase shifters to achieve beamforming performance in the automotive DFRC system. DRL allows an agent to learn its optimal action through interaction with its environment via trial-and-error [37]. However, the dimensionality of the action space of the proposed automotive DFRC system is huge, which may require a significant large training process for DRL. To address this issue, we adopt the Wolpertinger's strategy that encompasses the nearest neighborhood component to project the daunting action space to a small size space, thus significantly reducing the complexity of the training process while maintaining the desired good performance.

We begin with a brief literature review of automotive DFRC, sparse arrays, adaptive beamforming and DRL.

### A. Related Work on Automotive DFRC

Generally, DFRC systems can be categorized as radar-centric, communication-centric, or based on a joint design approach. In radar-centric systems, communication information is embedded in radar waveforms. For example, when using amplitude-modulated signaling, the communication information can be associated with the sidelobe amplitude of the transmitted waveform [38]. In MIMO radar, communication symbols are encoded into each of the orthogonal waveforms. Phase modulation can also be used to embed communication symbols into the phase of transmitted chirps [39]. Radar-centric systems achieve good sensing performance but their communication rate is low. However, with frequency hopping [40], antenna index modulation [41], [42], and sparse array configuration techniques [32], their communication data rate can be further increased. Communication-centric systems are optimized for the communication function and use typical communication signals, such as orthogonal frequency-division multiplexing (OFDM), for sensing [43]–[45]. Adaptive systems, capable of dynamically adjusting their parameters to achieve a desired trade-off between the radar and communication subsystems, and tailored to the specific requirements of the application have also been proposed [41], [46]–[48].

DFRC systems have significant advantages in autonomous vehicle applications [17]. Deep neural networks have been introduced in [49] for communication multi-user demodulation and target tracking. In [18], frequency-modulated continuous-wave (FMCW) signals from a selected subset of antennas are transmitted in a randomized fashion. Index modulation is utilized to embed message through the selection of carrier frequencies and antennas.

Sparse arrays have been widely adopted in automotive radar systems to achieve a larger antenna array aperture by deploying antenna elements placed in a non-uniform spacing pattern [24]. Beamforming has been proposed to achieve dual functions for vehicle networks [50] with the fast beam tracking capability of a road-side unit, and thus it is not applicable for systems on moving vehicles. Recently, an integrated MIMO system is proposed in [51] that performs target tracking and downlink communications, while also receiving uplink signals from other communication nodes to facilitate bi-directional communications. Integration of sparse array configuration and waveform permutation was introduced in [32] to embed communication symbols effectively. Beamforming with sparse arrays was considered in [52] through selective antenna positioning, ensuring both radar and communication functionalities are maintained. In [53], sparse array-based beamformers for DFRC systems were designed by utilizing both amplitude and phase modulations to encode information into the transmit beam patterns. In [34], an alternating direction method of multipliers (ADMM) solver was developed to design sparse array beamforming that accommodates multiple downlink users. More recently, a new scheme is proposed in [54] to jointly develop multiple beamformers that deliver communication information via a common sparse array for integrated radar and communication (IRC) systems. This approach enhances the applicability and simplifies the hardware implementation by eliminating the need for continual antenna switching. Nevertheless, the antenna selection problem is NP-hard, and the computational cost of relaxed optimization algorithms is still extremely high.

### B. Deep Reinforcement Learning

Machine learning techniques, including deep learning (DL) and reinforcement learning (RL), have become powerful tools for solving beamforming matrix and antenna selection problems [55]. Among them, DL networks can be trained to select antennas and determine phase shifters with a low computational complexity, provided that a large amount of training data is available. However, the applicability of DL algorithms in real-life scenarios, such as for tracking noise and channel

changes during rapid vehicle movements, may still be limited due to inherent biases in the dataset. RL algorithms, on the other hand, can be trained by interacting with the environment and receiving feedback in the form of rewards or penalties. This makes them well suited for decision-making tasks in complex and dynamic environments, such as target tracking using a massive MIMO radar. However, RL faces a scaling dilemma when the goal is to obtain an optimal beamforming matrix. To address this issue, we adopt deep RL (DRL) to learn complex policies and representations directly from raw data, such as online beamforming learning. Although it is more computationally expensive than traditional RL, DRL offers significant advantages in scalability, performance, and data requirements [56], [57].

### C. Motivation and Contributions

In this paper, our system operates in tracking and searching modes. In the searching mode, the system performs a quick target perception. During the tracking mode, based on the perception results, the sparse transmit array is optimized to carry out adaptive transmit beamforming to enhance both radar sensing and communication functionalities. This approach reduces system costs and optimizes the target tracking performance, while simultaneously suppressing interference towards victim radars. Furthermore, by applying DRL, the system dynamically adjusts the transmit parameters to achieve an optimal solution through interactions with the environment. Our contributions are as follows:

- We propose a low-cost reconfigurable transmit beamforming scheme, incorporating joint antenna selection and quantized double-phase shifters, to implement DFRC systems for automotive applications. We optimize the array configuration and the beamforming weights in order to simultaneously form two beams, one towards the communication receiver and the other one towards the target searching direction, while creating nulling towards other radars, allowing multiple radars to coexist and operate simultaneously within the same frequency band.
- We propose a DRL approach to adaptively select a sparse subset of transmit antennas and adjust quantized double-phase shifters to achieve optimal sparse transmit beamforming in the highly dynamic automotive scenarios.
- We propose a DRL framework based on Wolpertinger's strategy to tackle the dimensionality curse within the action space, a challenge linked to dynamic antenna selection and adaptive beamforming when employing quantized double-phase shifters. The proposed approach integrates the strengths of both deep Q-network and deep deterministic policy gradient within the actor-critic networks.

The basic concept of DRL applied to automotive radar transmit beamforming is described in [1]. An enhanced DRL algorithm along with preliminary simulation results are presented in [2]. The substantive novel contributions of this paper beyond [1], [2] include the development of a comprehensive analysis of communication signal loading, the development of improved transmit beamforming algorithms based on optimization, a thorough numerical investigation including the comparison with a baseline optimization approach, and an analysis of radar sensing outcomes.

### D. Article Organization

The rest of the paper is organized as follows. In Section II, we describe the signal model of automotive radar for sensing and communication and formulate the transmit beamforming design exploiting sparse arrays and quantized double-phase shifters. In Section III, we present a DRL-based optimization method for of transmit beamforming. We validate our models and methods with extensive numerical simulations in Section IV. Finally, Section V concludes the paper.

Throughout this paper, upper-case and lower-case bold characters denote matrices and vectors, respectively. Matrix vectorization operation is denoted by $\mathrm{vec}(\cdot)$. The conjugate transpose and transpose are denoted by $(\cdot)^H$ and $(\cdot)^T$, respectively. The complex values set is $\mathbb{C}$. The notations $\odot$ and $\otimes$ denote the Hadamard product and Kronecker product, respectively.

## II. System model

We consider a colocated phase-controlled FMCW automotive DFRC system consisting of a reconfigurable unit uniform linear transmit array, which means that the transmit array can be adaptive corresponding to the scenario by enabling or disabling antennas. The transmit array is used for both radar sensing and communication functions (see Fig. 1). The communication information is embedded in the transmitted waveform through slow-time encoding or a hybrid of fast-time and slow-time encoding.

### A. Automotive Radar

Consider an FMCW automotive radar operating at a central frequency $f_c$ with a bandwidth of $B$ and a pulse duration time of $T$. This radar transmits a linear frequency ramp with a pulse repetition interval $T_p$. At the $m$-th chirp, the corresponding uncoded transmitted signal is given by [27]

$$x(m,t) = \mathrm{rect}\left(\frac{t - mT_p}{T}\right) e^{\mathrm{j}2\pi\left(f_c t + \frac{1}{2}\cdot\frac{B}{T}t^2\right)}, \qquad (1)$$

where $t$ is the fast-time with $0 \le t \le T$.

In MIMO radar, the transmitting antennas transmit mutually orthogonal waveforms; this enables the receiving antenna to extract the contribution of each transmitting antenna. Consider a MIMO radar with $N_T$ transmit antennas that transmit orthogonal FMCW waveforms. There are many ways to realize waveform orthogonality in an FMCW radar, such as time division, frequency division, and code division [27], [58].

By exploiting orthogonality, each receive antenna can extract the received signal corresponding to each transmitting antenna. After applying fast Fourier transform (FFT) to the fast-time samples of each signal, a decoded data matrix is obtained, which is then subjected to a slow-time FFT. This process results in a range-Doppler spectrum for a specific

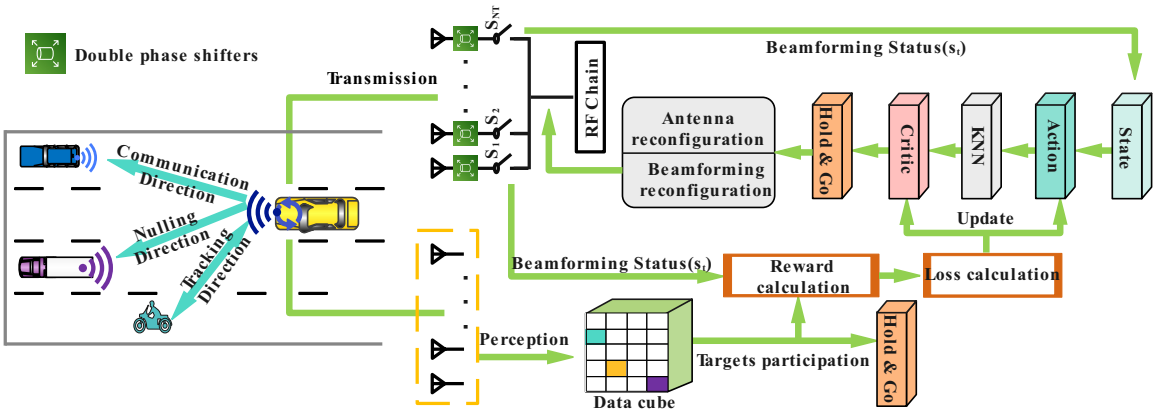


Fig. 1. The proposed automotive radar DFRC system diagram in the tracking mode.

transmit-receiver antenna pair. To extract the angle information corresponding to the pair, thresholding techniques such as the constant false alarm rate (CFAR) algorithm can be used. By following this process, the virtual array steering vector $\mathbf{a} \in \mathbb{C}^{N_T N_R}$ can be given by

$$\mathbf{a} = \sum_{k=1}^{K} \alpha_k \left[\mathbf{a}_t(\theta_k) \otimes \mathbf{a}_r(\theta_k)\right], \tag{2}$$

where $\mathbf{a}_t(\theta_k) = [1, e^{\mathrm{j}(2\pi/\lambda)\tilde{d}_{21}\sin(\theta_k)}, \cdots, e^{\mathrm{j}(2\pi/\lambda)\tilde{d}_{N_T1}\sin(\theta_k)}]^T$ and $\mathbf{a}_r(\theta_k) = [1, e^{\mathrm{j}(2\pi/\lambda)d_{21}\sin(\theta_k)}, \cdots, e^{\mathrm{j}(2\pi/\lambda)d_{N_R1}\sin(\theta_k)}]^T$ denote the transmit and receiver steering vectors, respectively. To establish a common reference point, we define the first element of each vector as the reference. Several classical techniques exist for obtaining the angle spectrum from the steering vectors, including digital beamforming and compressive sensing [27].

### B. Automotive Radar Sensing and Communication Design

A moving target-oriented DFRC system with a radar-centered reconfigurable phased transmitting array isolates moving targets from static targets such as buildings and trees through the range-Doppler spectrum of the echo and ego vehicle speed. The change of the estimated target parameters can guide the system to automatically adjust the activation and deactivation of the reconfigurable array and the corresponding beamforming vector.

*1) Radar Transmit Beamforming:* The radar transmit beampattern is given by [59]

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

where $\mathbf{W} \in \mathbb{C}^{N_T \times N_T}$ is the beamforming weight matrix composed with quantized phase terms, expressed as

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

$\mathbb{E}[\cdot]$ denotes the statistical expectation, $\mathbf{x}(m,t)$ is the transmit waveform vector for the $N_T$ transmit antennas at the $m$-th chirp and is given by

$$\mathbf{x}(m,t) = x(m,t)e^{j\phi_m}\mathbf{f}_{\mathrm{RF}}. \tag{5}$$

Here, $\phi_m$ is the slow-time code at the $m$-th chirp. To perform the radar sensing function, the analog precoder $\mathbf{f}_{\mathrm{RF}}$ is designed to steer the mainlobe to the region of interest of the radar. $\mathbf{f}_{\mathrm{RF}}$ is controlled by the phase shifters, which can be replaced by the radar sensing beamformer $\mathbf{w}_r$, defined as

$$\mathbf{w}_r = \frac{1}{\sqrt{N_T}}\left[e^{j\omega_1}, e^{j\omega_2}, \cdots, e^{j\omega_{N_T}}\right]^T, \tag{6}$$

where $\omega_i \in \mathcal{D}$ for all $i \in \{1, \cdots, N_T\}$.

*2) Communication Model:* Assume that the communication receiver has an array consisting of $N_c$ elements, and the number of independent propagation paths ($L$) is less than $N_T$ because millimeter wave (mmWave) channels have limited scattering [60]. In such scenarios, the downlink channel matrix is denoted by $\mathbf{H}_d \in \mathbb{C}^{N_c \times N_T}$ and can be expressed as follows:

$$\mathbf{H}_d = \sqrt{\frac{N_T N_c}{L}} \sum_{l=1}^{L} \beta_l \mathbf{b}_c(\theta_{cl})\mathbf{a}_t^H(\theta_{tl}), \tag{7}$$

where $\beta_l$ represents the complex path gain for the $l$-th path. Additionally, $\mathbf{b}_c(\theta_{cl})$ and $\mathbf{a}_t(\theta_{tl})$ denote the receive and transmit array steering vectors of the $l$-th path, respectively, for the communication system. The angles of arrival and departure for the $l$-th path are denoted by $\theta_{cl}$ and $\theta_{tl}$, respectively. The received signal at the communication receiver can be expressed as:

$$\mathbf{y}_c(m,t) = \sqrt{\rho}x(m, t-\tau_c)e^{j\phi_m}\mathbf{H}_d\mathbf{f}_{\mathrm{RF}} + \mathbf{n}(n,t), \tag{8}$$

where $\rho$ denotes the average received power and $\tau_c$ is the delay between the radar transmitter and the communication receiver.

In the communication mode, $\mathbf{f}_{\mathrm{RF}}$ is replaced with a beamformer $\mathbf{w}_c$, defined as

$$\mathbf{w}_c = \frac{1}{\sqrt{N_T}}\left[e^{j\Omega_1}, e^{j\Omega_2}, \cdots, e^{j\Omega_{N_T}}\right]^T, \tag{9}$$

where $\Omega_i \in \mathcal{D}$ for all $i \in \{1, \cdots, N_T\}$.

On the receive end of the communication channel, the received signal can also be represented as:

$$\mathbf{y}_c(m,t) = \sqrt{\rho}x(m, t-\tau_c)e^{j\phi_m}\mathbf{h}_{\mathrm{lump}} + \mathbf{n}(n,t). \tag{10}$$

In this equation, the lump channel information vector $\mathbf{h}_{\mathrm{lump}} = \mathbf{H}_d\mathbf{f}_{\mathrm{RF}}$ is present. To estimate the channel information $\mathbf{h}_{\mathrm{lump}}$ for the full phase array, beam sounding techniques

can be employed, as described in the prior work [61], [62]. Accurate synchronization of the signal delay between the sender and the receiver is necessary to extract communication code from the received. To achieve time synchronization, Global Positioning System (GPS) technology or atomic clock technology can be used [63]–[65]. For example, the Pulse-Per-Second (PPS) signals from two GPS modules can achieve a synchronization accuracy of 60 ns [66]. If the estimated channel $\overline{\mathbf{h}}_{\text{lump}}$ matches to $\mathbf{h}_{\text{lump}}$, after passing through the mixer and the low-pass filter, the received communication signal becomes

$$\mathbf{y}_c = e^{j\phi_m}\mathbf{1} + \mathbf{n}. \tag{11}$$

The received reconstructed signal $\mathbf{y}_c$ can be used to evaluate the communication performance.

*3) Communication Information Embedding:* In the phased-array beamforming mode, each chirp carries a communication symbol from the binary phase-shift keying (BPSK) constellation, also known as the slow-time coding method in this paper. At the receive end, the transmitted signal can be extracted using equation (11). However, in automotive radar, the number of chirps is typically limited to several hundreds in order to enable coherent processing of the target object echo, which sets an upper bound on the communication rate. To increase the communication capacity, a hybrid-coded method can be adopted, which combines fast-time and slow-time coding. In this coding scheme, the duration of each chirp is divided into $L$ short time periods, which are then encoded as follows: The code sequence for $N_T$ transmit antennas at the $m$-th chirp is given by

$$x_C(m,t) = \sum_{l=1}^{L} e^{j\phi_l}\text{rect}\left(\frac{t - m(l-1/2)T_l}{T_l}\right) e^{j2\pi\left(f_c t + \frac{1}{2}\cdot\frac{B}{T}t^2 + \phi_{N_T m}\right)}, \tag{12}$$

where $T_l = T/L$ represents the duration of a fast-time chirp.

### C. Transmit Beamforming Exploiting Sparse Arrays and Quantized Double-Phase Shifters

We present the transmit beamforming using sparse arrays with quantized phase shifters. The combined application of transmit beamforming and sparse arrays allows for a cost-efficient and effective system design. As shown in Fig. 1, the transmit antennas are sparsely selected to form two distinct beams: one directed towards the radar target and the other towards the communication user. Concurrently, a null is created in the directions of the victim radars, functioning to significantly minimize mutual interference.

*1) Antenna Selection:* In order to adapt different sparse transmitting arrays in different scenarios, we designed a reconfigurable sparse array based on the current radar scenario by activating a subset of available transmit antennas, and keeping the rest of the antennas inactive to reduce the power consumption of the radar system and mutual coupling between the antennas. A selection matrix denoted as $\mathbf{S} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{N_T}]$ is defined, where each column vector $\mathbf{u}_i$ represents the status of $i$-th antenna. The $i$-th entry of $\mathbf{u}_i$ is set to "1" if the antenna is activated, and "0" otherwise. The length of the transmit array determines the transmit beam's width, meaning that a larger aperture size results in a narrower beamwidth [67]. Therefore, we fix the first and the last antennas of the array to maintain a consistent array aperture and select $M$ other antennas in-between, resulting in a total of $M+2$ antennas being activated. As a result, the trace of $\mathbf{S}$, denoted as $\text{tr}(\mathbf{S})$, is $M+2$. The first element of $\mathbf{u}_1$ and the last element of $\mathbf{u}_{N_T}$ are set to "1", while all other entries in $\mathbf{u}_1$ and $\mathbf{u}_{N_T}$ are set to "0".

*2) Beam Synthesis via Double-Phase Shifters:* In a full-phase array DFRC system, radar sensing and communication will share the same transmitter array. Therefore, in order to achieve both functions, the radar sensing beamformer $\mathbf{w}_r$ and the communication beamformer $\mathbf{w}_c$ need to be merged as a single beamformer $\mathbf{w}$, and this is accomplished using double-phase shifters. Its working principle relies on each antenna being connected to a unique RF chain via two phase shifters. The beamformer can synthesize two main beams, focusing the energy of the array on the respective directions of interest for the two functions.

By connecting a pair of phase shifters to each antenna, we introduce more degrees of freedom to shape the beampattern [29]. The hybrid transmit beamformer $\mathbf{w}$ is given as

$$\begin{aligned}\mathbf{w} &= c_1\mathbf{w}_r + c_2\mathbf{w}_c \\ &= \frac{1}{\sqrt{N_t}}[c_1 e^{j\omega_1} + c_2 e^{j\Omega_1}, c_1 e^{j\omega_2} + c_2 e^{j\Omega_2}, \cdots, \\ &\quad c_1 e^{j\omega_{N_T}} + c_2 e^{j\Omega_{N_T}}]^T,\end{aligned} \tag{13}$$

where $c_1 \in [0,1]$ and $c_2 \in [0,1]$ with $c_1^2 + c_2^2 = 1$ are weighting factors that balance radar sensing and communication capabilities.

*3) Beamforming Optimization Problem Formulation:* In practical applications, however, phase shifters usually apply a limited number of phase shift angles due to complex implementation and overhead challenges. At the same time, in order to reduce the interference of transmitting antennas to uninteresting targets, additional constraints need to be introduced to improve the above optimization problem.

In this section, we delve into the intricate process of crafting a highly optimized transmit beamformer, denoted by $\mathbf{w}$, as well as an antenna selection matrix, represented by $\mathbf{S}$. To achieve both radar sensing and communication functions simultaneously, a common sparse array can be designed using two separate beamformers, $\mathbf{w}_r$ for radar and $\mathbf{w}_c$ for communication [52]. The objective of this design is to ensure that a certain level of power is maintained towards both the radar targets and the communication destination. Furthermore, the beamformer must generate minimal interference towards other directions while maintaining low peak sidelobe levels.

In essence, the problem at hand can be formulated as

$$\begin{aligned}\min_{\{\mathbf{w},\mathbf{S},\alpha_1,\alpha_2,\alpha_3\}} \ & \gamma_1\alpha_1+\gamma_2\alpha_2+\gamma_3\alpha_3\\ \text{s.t.} \quad & \left|\mathbf{w}^H\mathbf{S}\mathbf{a}(\theta_r)\right| = p_1,\\ & \left|\mathbf{w}^H\mathbf{S}\mathbf{a}(\theta_c)\right| = p_2,\\ & \left|\mathbf{w}^H\mathbf{S}\mathbf{a}(\theta_l)\right| \le \rho_1+\alpha_1, \theta_l\in\overline{\Theta},\\ & \left|\mathbf{w}^H\mathbf{S}\mathbf{a}(\theta_i)\right| \le \rho_2+\alpha_2,\\ & \mathbf{w} = c_1\mathbf{w}_r + c_2\mathbf{w}_c,\\ & \mathbf{w}_r,\mathbf{w}_c\in\mathcal{D},\\ & \left|\theta_r-\hat{\theta}_r\right| \le \alpha_3,\\ & \mathrm{tr}(\mathbf{S}) = M+2.\end{aligned} \tag{14}$$

The coefficient of balance $\gamma_i$ denotes the proportion of each part in the cost function. The sensing direction, departure angle to the communication user, discretized angle in the sidelobe region, and direction of an uninterested target are represented by the variables $\theta_r$, $\theta_c$, $\theta_l$, and $\theta_i$, respectively. Here, $p_1$ and $p_2$ quantify the power assigned to the radar target and the communication user, respectively. With $\overline{\Theta}$ denoting the sidelobe region, $\rho_1$ and $\rho_2$ quantify the peak sidelobe level of beampattern and interference attenuation, respectively. To make the constraints feasible, we introduce auxiliary variables $\alpha_1$ and $\alpha_2$ to relax the constraints on the specific level of peak sidelobe and interference attenuation, and their minimum values are 0. Constraints $\mathbf{w}_r$ and $\mathbf{w}_c$ force the value of phase shifters within the range of quantitative values. The parameter $\alpha_3$ is used to measure the deviation between the ground truth $\theta_r$ and the actual main beam direction $\hat{\theta}$. The last constraint ensures that a total of $M+2$ antennas are selected. The joint optimization problem of transmit antenna selection and beamforming with quantized phase shifters is NP hard.

*4) Optimization Based Solution:* A simplified optimization problem is developed to determine the optimal beamforming weights by assuming that the selection matrix $\mathbf{S}$ of transmit antennas is already obtained via DRL. Let $\mathbf{f}_d$ denote the desired reference beampattern including the mainlobe, sidelobe, and null-space region constraints specified in problem (14). The new beampattern synthesis problem is formulated as

$$\begin{aligned}\min_{\mathbf{w}} \ & \left|\mathbf{A}\mathbf{S}\mathbf{w}-\mathbf{f}_d\right|\\ \text{s.t.} \quad & \mathbf{w} = c_1\mathbf{w}_r + c_2\mathbf{w}_c,\\ & \mathbf{w}_r,\mathbf{w}_c\in\mathcal{D}.\end{aligned} \tag{15}$$

In this problem, $\mathbf{A} = [\mathbf{a}_t(\theta_1),\mathbf{a}_t(\theta_2),\cdots,\mathbf{a}_t(\theta_K)] \in \mathbb{C}^{N_T\times K}$ is a dictionary matrix consisting of discretizing the entire region of interest (ROI) into $K$ discretized angles with a certain step. The weight vector $\mathbf{w}$ comes from the quantized set $\mathcal{D}$. The fixed sparsity antenna distribution group sparse optimization (GSO) method is used to solve the optimum beam vector $\mathbf{w}$. The values of the obtained phase shifters through GSO can achieve arbitrary high precision. These values are then quantized to the closest values in the quantized phase shifter set $\mathcal{D}$. The quantization may result in a small performance loss that will be seen in the numerical part. Further, when compared to the RL method, the GSO optimization method lacks the ability to dynamically adjust the beamforming weights when changes the target's position changes due to the absence of a feedback link.

*5) Angle Finding Under Sparse Antenna Selection:* In the tracking mode, the transmit beamforming using sparse arrays through antenna selection enhances the output SNR of both sensing targets and communication users, while suppressing interference to other automotive radars. For angle finding of radar targets in the tracking mode, a separate receive antenna array is required, which is decoupled from the transmit beamforming design. However, in the sensing mode, the virtual array is synthesized with MIMO radar technology and, therefore, the dynamic sparse transmit antenna selection would impact the virtual array beampattern. The challenge of sparse arrays is the high sidelobes or potential grating lobes. High-resolution algorithms such as compressive sensing (CS) [68] and iterative adaptive approach (IAA) [69], [70] are of great interest for angle finding, as they help to suppress the high sidelobes. Additionally, CS and IAA work with single snapshot. To deal with potential grating lobes, usually angle unfolding technique can be applied with overlapped subarrays [71].

## III. Transmit Beamforming Design Using Deep Reinforcement Learning

The transmit beamforming optimization problem is combinatorial, and thus NP hard, requiring an exhaustive search through a vast number of possible solutions. This means that the time required to find a solution grows exponentially with the size of the problem. To alleviate the optimization difficulties caused by the size explosion, and considering the limited phase tunability characteristics of practical phase shifters, we adopt a framework based on DRL to dynamically activate or deactivate antennas and tune the phase of each activated antennas. The incorporation of DRL paves the way for intelligent, dynamic decision-making that is responsive to changes in the system's environment.

### A. Deep Reinforcement Learning

The problem of sparse array beamforming is to find the optimal subarray set and the corresponding beamforming matrix policy. RL utilizes the trail and reward loop to guide the agency to an optimum solution, making it a powerful tool to solve our beamforming problem in an intelligent way. RL is a type of machine learning where an agent learns to make a sequence of decisions in an environment in order to maximize a reward. The goal of the agent is to learn a policy, which is a mapping from states to actions, that maximizes the expected cumulative reward. The mathematical framework of RL can be defined by a Markov decision process (MDP) using a tuple $(S,A,P,R,\gamma)$, where $S$ is the set of possible states, $A$ is the set of possible actions, $P$ is the state transition function, which specifies the probability of transitioning from the current state $s_t$ to the next state $s_{t+1}$ when taking action $a$, and $R$ is the reward function, which specifies the immediate reward obtained during the transition from state $s_t$ to state $s_{t+1}$ when taking action $a$. In addition, $\gamma$ is the discount factor, which determines the importance of future rewards. The goal of the

agent is to learn a policy $\pi(a|s)$ that maximizes the expected cumulative reward [72], [73]:

$$G_t = \sum_{k=1}^{\infty} \gamma^k R_{t+k+1}, \tag{16}$$

where $G_t$ is the discounted cumulative reward at time step $t$. The policy is learned by updating the estimate of the value function $V_\pi(s)$, which is the expected cumulative reward starting from state $s$ and following policy $\pi$. This can be done using the Bellman equation [73]:

$$V_\pi(s_t) = \sum_{a\in A} \pi(a|s_t)_{s_{t+1}\in S} P(s_{t+1}|s_t,a)\left[R(s_t,a,s_{t+1}) + \gamma V_\pi(s_{t+1})\right]. \tag{17}$$

DRL is a type of RL that uses deep neural networks to approximate the optimal policy or value function. This allows the agent to learn more complex and abstract representations of the state space, and enables it to generalize to unseen situations. In DRL, the agent's policy or value function is represented by a deep neural network with weights $\theta$. The network takes the state $s$ as input and outputs the action probabilities or value estimates. The weights are updated using stochastic gradient descent to minimize the loss function, which is usually the mean squared error between the predicted and the actual target values. Popularly used algorithms used in DRL include Q-learning, State-Action-Reward-State-Action (SARSA), and actor-critic methods, which are all based on the idea of using deep neural networks to approximate the Q-function or policy.

The Q-function is the expected cumulative reward for taking a certain action $a$ in a certain state $s$, and following the optimal policy thereafter [72], [73]. It is expressed mathematically as:

$$Q^*(s_t,a_t) = \mathbb{E}[R_{t+1} + \gamma \max_{a_{t+1}} Q^*(s_{t+1},a_{t+1})|s_t,a_t], \tag{18}$$

where $a_{t+1}$ is the next action, and $\max_{a_{t+1}} Q^*(s_{t+1},a_{t+1})$ is the maximum expected cumulative reward under the optimal policy in the next state.

The Q-learning algorithm updates the Q-function iteratively using the following equation:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[R_t + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right], \tag{19}$$

where $\alpha$ is the learning rate, and the update is performed after every action is taken by the agent.

SARSA is a similar algorithm that updates the Q-function using the following equation:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[R_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right]. \tag{20}$$

The update is performed after every action-state pair.

Actor-critic methods combine both the policy-based and value-based approaches by training two neural networks: one to approximate the policy and the other to approximate the value function. The policy network is trained using the policy gradient method, while the value network is trained using temporal difference learning or a variant thereof.

The deep Q-network (DQN) algorithm is a DRL algorithm that combines Q-learning with a deep neural network to learn the Q-function in high-dimensional state spaces. The key idea behind DQN is to use a neural network to approximate the Q-function, which allows the agent to learn a more accurate representation of the optimal policy than traditional tabular methods. However, when the action dimension is high, it becomes difficult to use DQN RL to find the desired mapping policy. To overcome this challenge, we use the Wolpertinger policy-based RL framework, which enables us to implement training that is manageable on a time-wise basis. The Wolpertinger policy comprises three basic elements: an action network, a K-nearest neighbor (KNN) map, and a critic network. Together, these elements make up the Wolpertinger policy. The deep deterministic policy gradient (DDPG) is used to train the networks [74].

The actor network in DDPG selects an action by mapping the current state of the environment to a continuous action space. The output of the actor network is a vector of continuous action values that can be scaled to fall within a specified range. Mathematically, the actor network is represented as:

$$\hat{a} = \mathrm{Actor}(s|\theta^\mu), \tag{21}$$

where $\theta^\mu$ contains the parameters of the actor network, and $\mathrm{Actor}(\cdot)$ is the function that maps the state to the action. The action network in DDPG outputs a continuous action value, which needs to be discretized before it can be used by the critic network to estimate the Q value. To achieve this, the KNN algorithm is used to select the $K$ actions $(a)$ in the quantization interval $\mathcal{D}$ that are closest to the continuous action output of the actor network.

In DDPG, the KNN network is used to select the best action from the set of actions generated by the actor network. The K-nearest neighbors and their associated actions are stored in the KNN map for use during training and testing. The distance metric $d(\hat{a},a)$ can be defined as the Euclidean distance:

$$d(\hat{a},a) = ||\hat{a} - a||^2. \tag{22}$$

The $K$ actions can be selected as [75]

$$a = \underset{a\in\mathcal{D}}{\overset{K}{\mathrm{argmin}}}\, d(\hat{a},a). \tag{23}$$

These $K$ actions, along with the current state, are then used as state-action pairs by the critic network to calculate the Q value. The goal of critic network is to choose the corresponding action to the maximum Q value, which can be given by

$$a = \underset{a\in\mathcal{D}}{\mathrm{argmax}}\, Q(s,a|\theta^Q), \tag{24}$$

where $\theta^Q$ is the parameter set of the critic network.

During training, the actor network learns to maximize the expected reward obtained by the agent. This is achieved by adjusting the parameters of the actor network using the policy gradient method. The policy gradient is computed using the estimated value of the state-action pair, which is provided by the critic network. The gradient of the expected reward with

respect to the parameters of the actor network, denoted by $\nabla_{\theta^\mu} J$, can be written as [74]:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q(s_i, a|\theta^Q) \nabla_{\theta^\mu} \text{Actor}(s|\theta^\mu), \tag{25}$$

where $N$ is the batch size, $s_i$ is the $i$-th state in the batch, $\mathbf{a}$ is the action selected by the actor network, and $Q(s_i, a|\theta^Q)$ is the estimated value of the state-action pair provided by the critic network. The policy gradient is computed by taking the gradient of the expected reward with respect to the parameters of the actor network, $\theta^\mu$, and is used to update the actor network during training.

In the DDPG algorithm, the target actor and target critic networks are copies of the original actor and critic networks, respectively, but with separate sets of parameters [74]. These target networks are used to generate the target actions and target Q-values that are needed for updating the original networks. The update rule for the target actor network parameters is given by:

$$\theta^{\mu'} = (1 - \eta)\theta^{\mu'} + \eta\theta^\mu, \tag{26}$$

where $\theta^{\mu'}$ is the set of parameters for the target actor network and $\eta$ is the update rate parameter that controls the rate at which the target network parameters are updated. Typically, $\eta$ is set to a small value such as 0.001. Similarly, the update rule for the target critic network parameters is given by:

$$\theta^{Q'} = (1 - \eta)\theta^{Q'} + \eta\theta^Q, \tag{27}$$

where $\theta^{Q'}$ is the set of parameters for the target critic network. During training, the original actor and critic networks are updated using the gradient descent algorithm based on the loss function. However, the target networks are not updated directly. Rather, their parameters are updated slowly to match the parameters of the original networks. This introduces a lagging between the generation of the target actions and Q-values and the update of the original networks, which can help to improve the stability and convergence of the learning process.

The critic network is trained using a temporal difference method [74]. The objective is to minimize the difference between the estimated value and the actual value of the state-action pair:

$$L(\theta^Q) = \mathbb{E}\left[(r_{i+1}+1+\gamma Q(s_{i+1}, a_{i+1}|\theta^Q) - Q(s_i, a_i|\theta^Q))^2\right], \tag{28}$$

where $s_{t+1}$ is the state at time $t+1$ and $a_{t+1}$ is the action taken at time $t+1$.

### B. Beamforming Design with DRL

In this section, we describe in detail how to use Wolpertinger policy-based reinforcement learning framework to optimize transmission beamforming for both radar sensing and communication functions and to avoid interference to specific targets, as shown in equation (14), and specific functions of function blocks in Fig. 1.

1) **Action Space:** We consider a scenario where we must select $M+2$ antennas from a pool of $N_t$ antennas, and both ends have fixed antennas. The number of potential solutions for this scenario is $Q = C_{N_t-2}^{M}$. Phased array antenna systems utilize quantized phase shifters, which enable multiple antennas to direct the transmission or reception of a signal by adjusting each antenna's output phase. This technique allows the signal to be steered in a desired direction without requiring physical movement of the antenna. Each antenna in a subphase array is connected to two $q$-bit quantized phase shifters, which have a value range of $(-\pi, \pi)$. To obtain the optimal value of $\mathbf{w}$, we must optimize the phase of the phase shifters and the sparse transmit array geometry. The dimension of the phase adjustment is $\mathbb{R}^{Q \cdot 2^{q \times (M-2)}}$.

2) **State:** Once an action is taken from the action space, the state vector $\mathbf{s}$ changes and includes the current status of the transmit array phase shifters. At the $i$-th iteration, the state is represented as $\mathbf{s}_i^T = [w_1, w_2, \cdots, w_{M+2}]_i$, where each element corresponds to the status of a specific phase shifter. The activation or deactivation of a phase shifter represents a change of one element and is considered as an action taken from the action space.

3) **Hold and Go:** After the transmit antenna array emits power, the receive array processes the received echo signal. Subsequently, the moving targets within the ROI are analyzed, and relevant parameters such as distance, Doppler, and angle, are extracted. These parameter values are then transmitted to the module, where they are dynamically adjusted based on the parameter estimation results obtained from the receive array. In the hold stage, only one set of phase shifters is explored instead of two, and a pre-beamforming check is performed before inputting two phase shifters with the desired phase. The columns of the beamformer recorder matrix $\mathbf{W}_{rc} = [\mathbf{w}_r(\theta_r), \mathbf{w}_c(\theta_c)]$ form the respective beams in the target and communication receiver directions. The fused beamformer $\mathbf{w} = 0.5\mathbf{w}_r(\theta_r) + 0.5\mathbf{w}_c(\theta_c)$ is obtained, assuming $c_1 = c_2 = \sqrt{2}/2$. To determine whether the phase of double-phase shifters should be changed, a set of flags $\mathbf{f}_d$ is used. It consists of two flag bits. The one-time trigger flag bit $f_{d1}$ detects the dimension of matrix $\mathbf{W}_{rc}$, and once the dimension satisfies the two columns, the holding phase ends and external environment interaction begins. Another flag bit $f_{d2}$ indicates whether $\mathbf{W}_{rc}$ changes and, if so, updates the reward.

4) **Environment Interaction:** The feedback component of RL is critical, making it superior to other machine learning methods for control applications. In the design of beamforming, radar sensing beam feedback consists of two components. First, the self-detection of beam directivity using the fusion beamforming vector $\mathbf{w}$ allows the agent's behavior to be adjusted by observing deviations between the obtained target and the set target through beampattern transformation. Second, the range-Doppler spectrum constructed from radar sensing echoes provides feedback to compare deviations with expectations and adjust the agent's actions. Similarly, interactive feedback for communication beamforming has two components. The first component is feedback for deviations in fusion weighting in the direction of the target, and the second component adjusts the agent's actions based on feedback from the communication channel. For the interaction of targets that need to avoid interference, the corresponding position of null

in the transmit beampattern is fed back simultaneously and immediately.

5) **Reward:** We design a comprehensive reward system that dynamically evaluates the actions chosen by the agent. This system guides the agent towards selecting the most appropriate action to maximize its rewards. The rewards are based on three evaluations: radar sensing, communication, and interference reduction. Together, these evaluations enable the agent to make informed decisions. Assume that the ROI for radar sensing covers an angle of $-\theta_{\mathrm{ROI}}/2$ to $\theta_{\mathrm{ROI}}/2$, and the 3-dB beamwidth is determined by the formula $\Delta_\theta = 2\arcsin(1.4\lambda/(\pi D))$, where $D$ is the physical aperture size of the transmit antenna array. The area beyond the first nulls of the mainlobe is referred to as the sidelobe region. At each update, the difference between the maximum level of the mainlobe peak $\max(P_{\mathrm{ROI},i})$ and the peak sidelobe level $\max(\mathrm{PSL}_i)$ is denoted by $\xi_i = \max(P_{\mathrm{ROI},i}) - \max(\mathrm{PSL}_i)$. The main beam deviation is calculated using the values of $\xi$ and $d_r$ to ensure that the main beam is directed towards directions while minimizing the peak sidelobe level. The reward is given by

$$r_{ri} = \begin{cases} 1, & \text{if } \xi_i > \xi_{i-1} \text{ and } d_{ri} \le d_{ri-1}, \\ -1, & \text{if } \xi_i \le \xi_{i-1} \text{ and } d_{ri} > d_{ri-1}, \\ 0, & \text{other cases.} \end{cases} \tag{29}$$

To evaluate communication performance, the received gain is usually represented by the expression $g_c = |\mathbf{Hw}|^2$. Assuming that the channel parameters have been estimated, the communication reward can be expressed as follows:

$$r_{ci} = \begin{cases} 1, & \text{if } g_{ci} > g_{ci-1}, \\ 0, & \text{if } g_{ci} = g_{ci-1}, \\ -1, & \text{if } g_{ci} < g_{ci-1}. \end{cases} \tag{30}$$

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

To prevent interference with other automotive radar systems, it is important that the synthesized beamformer produces a null in the direction of departure $\theta_i$. The level of attenuation can be calculated using the formula $p = |\mathbf{w}^H \mathbf{a}(\theta_i)|$, where $\mathbf{w}$ is the weight vector and $\mathbf{a}(\theta_i)$ is the array response vector in the direction $\theta_i$. The reward for minimizing interference at the $i$-th update can be given by

$$r_{pi} = \begin{cases} 1, & \text{if } p_i < p_{i-1} \\ 0, & \text{if } p_i = p_{i-1} \\ -1, & \text{if } p_i > p_{i-1}. \end{cases} \tag{31}$$

The final triple reward $r_i$ at the $i$-th update is expressed as

$$r_i = \lambda_1 r_{ri} + \lambda_2 r_{ci} + \lambda_3 r_{pi}, \tag{32}$$

where $\lambda_1, \lambda_2$, and $\lambda_3$ represent the respective weights trading off between the radar and communication functions, and interference attenuation.

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

**Algorithm 1** DRL-based automotive radar DFRC system

Initialize networks with corresponding parameters.
HOLD = TRUE, $f_{d1}$ = 0; $f_{d2}$ = 0.
Initialize $\xi_0$ = 0, $d_{r0}$ = 1, $g_{c0}$ = 0, $p_0$ = 1.
Initial sample a random beamforming vector $\mathbf{w}_{rc1}$ as initial state $s_1$ and record action $a_1$.
**for** $i = 1$ to $T$ **do**
  Receive proto-action $\widehat{a}_i$ from actor network.
  Action embedding $g(\hat{a}_i)$ through KNN mapping.
  **while** HOLD **do**
    Update $\mathbf{W}_{rc1}$.
    $f_{d1}$ = column($\mathbf{W}_{rc}$).
    **if** $f_{d1}$ == 2 **then**
      HOLD = FALSE.
      Execute action $\mathbf{w}_1$ passed from critic network.
      Calculate reward and update state $s_{i+1} = a_i$.
      Update $\xi_1$ , $d_{r1}$ , $g_{c1}$ and $p_1$.
    **end if**
  **end while**
  Update $\mathbf{W}_{rci}$.
  **if** $\mathbf{W}_{rci} \neq \mathbf{W}_{rci-1}$ **then**
    Execute action $\mathbf{w}_i$ passed from critic network.
    Calculate reward and update state $s_{i+1} = a_i$.
    Update $\xi_i$ , $d_{ri}$ , $g_{ci}$ and $p_i$.
    Update all networks.
  **end if**
**end for**

### C. Computational Complexity Analysis

The action space in the DRL system grows exponentially with the increasing number of antennas and quantized phase values, which can make the computational cost of exploring and learning the optimal policy prohibitively high. To address this challenge, the KNN block is introduced to ensure that the computational cost of the entire system remains tractable at each state. The Wolpertinger policy scales linearly with the number of selected actions, $k$. According to Lemma 1 in [76], the expected value of the maximum of the $k$ closest actions is:

$$\mathbb{E}\left[\max_{i\in 1,\ldots,k} Q(s,\hat{a})|s,\hat{a}\right] = Q(s,a) + b - p^k(c-b) - \frac{2b}{k+1}\frac{1-p^{k+1}}{1-p}.$$

Given $\hat{a}$, each nearby action has a probability $p$ of being suboptimal or faulty, resulting in a value lower than $Q(s,\hat{a})-c$. The values of the other actions are uniformly distributed within the interval $[Q(s,\hat{a})-b, Q(s,\hat{a})+b]$, where $b \le c$. The minus part $-p^k(c-b) - \frac{2b}{k+1}\frac{1-p^{k+1}}{1-p}$ lowers the expected value to below $Q(s,a)+b$, with changes governed by $\mathcal{O}(p^k)$ and $\mathcal{O}(\frac{1}{k+1})$. This significantly reduces the expected value, but the diminishing returns become apparent as $k$ increases. For applications with low dimensionality, using $5\%$ or $10\%$ of the maximum number of actions performs similarly to that using the full action set $\mathcal{A}$. However, even when considering $5\%$ of a large set $\mathcal{A}$, the number of actions to evaluate at

each step remains considerable. Fortunately, when the action space dimension is large, a single look-up is sufficient. For example, in [76], with an action space dimension of $n = 20$, using $k = 1$ ensures the convergence of the DRL. In our system, we use a single look-up $k = 1$ to expedite the action-critic process. Specifically, given an $\hat{a}$, we only evaluate the nearest action in $\mathcal{A}$. This approach is low-cost, efficient, and effectively addresses the curse of dimensionality in the action space.

The computational complexity of the proposal DRL based beamforming approach can be analyzed in terms of the number of neural networks parameters that need to be stored $\mathcal{C}^P$, additions $\mathcal{C}^A$ and multiplications $\mathcal{C}^M$. The actor and critical networks have a similar structure which takes the input dimension $N_s$ and passes it through two hidden layers with the number of neurons denoting as $N_1$ and $N_2$. The softplus and tanh functions are used for the first and second hidden layers separately, and a batch norm is implemented from layer to layer. Bias are assigned to the hidden layers and the output layer. The computation complexity of the proposed DRL is considered from the parameters to be stored $\mathcal{C}^P$, the addition operation $\mathcal{C}^A$, and the multiplication $\mathcal{C}^M$. The activation function cost is considered as addition, and the cost of the batch norm is counted to multiplication. Taking the actor-network as an example, the bundle of computation complexity is given by

$$\begin{aligned}\mathcal{C}^P &= N_sN_1 + N_1N_2 + N_2N_s + N_1 + N_2 + N_s,\\ \mathcal{C}^M &= N_sN_1 + N_1N_2 + N_2N_s + N_1 + N_2,\\ \mathcal{C}^A &= N_sN_1 + N_1N_2 + N_2N_s + 2(N_1 + N_2 + N_s).\end{aligned}$$

Considering the target and evaluation networks of actor-critic networks, the total complexity is $\mathcal{C} = 4(\mathcal{C}^P + \mathcal{C}^M + \mathcal{C}^A)$.

## IV. Numerical Results

We carry out numerical simulations to evaluate the performance of the proposed DRL assisted automotive DFRC system. A FMCW MIMO radar with phase-modulated slow-time waveforms features 15 transmit and 15 receive antennas with a half wavelength inter element spacing and its setting is given in Table I. The normalized spatial frequency of the half FOV of the array is set to 0.7, which corresponds to a half angle of the region of interest, denoted by $\theta_{\text{ROI}}/2$, of 44.2°. The 3-dB beamwidth, denoted by $\Delta_f$, is 0.119, which corresponds to a half angle, denoted by $\Delta_\theta$, of 6.81°.

TABLE I
Radar Parameters

| Parameters | Values |
|---|---|
| Carrier frequency, $f_c$ | 77 GHz |
| Maximum detection range, $R_{\max}$ | 200 m |
| Maximum detection velocity, $V_{\max}$ | 230 km/h |
| Bandwidth, $B$ | 150 MHz |
| Pulse duration, $T_d$ | 7.3 $\mu$s |
| Sampling frequency, $f_s$ | 54.648 MHz |
| Number of chirps, $M$ | 512 |

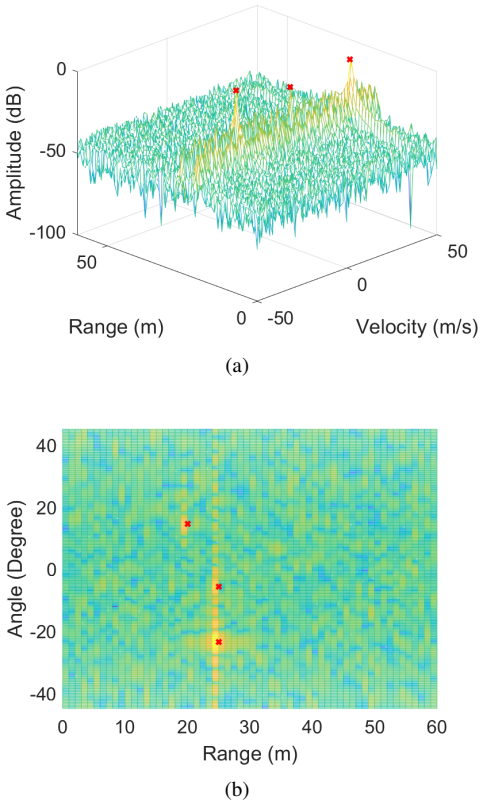


Fig. 2. Estimated target parameters in the searching mode: (a) Range-Doppler spectrum; (b) Range-angle spectrum. The red × denotes the actual positions of the targets.

To enable radar sensing and communication for specific purposes, the first crucial step is to estimate the target parameters within the radar's FOV using the searching mode. In this mode, following data processing, the resulting data spectrum can be visualized in Fig. 2. There are three objects in the FOV of the radar, including a tracking target with range $r_1 = 25$ m, $v_1 = -20$ m/s, $\theta_1 = -23.6°$, a communication user with $r_2 = 25$ m, $v_2 = 10$ m/s, $\theta_2 = -5.7°$, and a potential interference radar with $r_3 = 20$ m, $v_3 = 35$ m/s, and $\theta_3 = 14.4°$. The reflection coefficients of the three objects are normalized to $\alpha_1 = \alpha_3 = 1.0$, and $\alpha_2 = 0.3$ and are assumed to be unchanged during the processing interval. The input SNR is set to 0 dB.

For the learning model, we use the hyper-parameters as described in Table II. All networks are trained on a Lambda machine with an Intel Core i9-10920X CPU and four Nvidia Quadro RTX 6000 GPUs.

TABLE II
Hyper-parameters for Training

| Parameter | Value | |
|---|---|---|
| Models | Actor-Net | Critic-Net |
| Replay Buffer | 4096 | 4096 |
| Mini-batch | 128 | 128 |
| Learning rate | 0.001 | 0.001 |
| Decay | 0.001 | 0.001 |

### A. Performance Under Antenna Selection

We choose 12 antennas from 15 antennas to form the final transmission array, as shown in Fig. 3. There are 455 possible selection schemes. Each antenna is connected to a 3-bit quantization double-phase shifter.

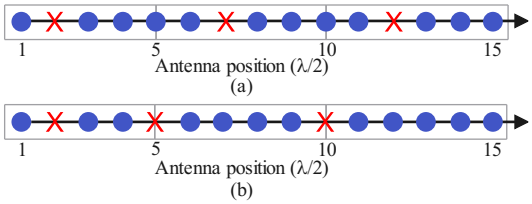


Fig. 3. (a) The transmit array configuration in the initial phase; (b) The transmit array configuration after optimization.

The antenna selection process is depicted in Figs. 3(a) and 3(b). Additionally, Fig. 4 shows the average reward attained during the training process. After approximately 60 epochs, the network intelligently adjusts the phases to steer the main beam to the ROI based on the current observation state. In the 100-th episode, the tracking target, the communication target, and the direction that requires nulling change. At this time, the reward drops sharply due to the loss of target. However, the perception information updates the hold and go module, and, therefore, RL is able to quickly adjust the transmit beamforming and reconfigure antenna locations. Fig. 5 (a) illustrates that at the outset of the iterative optimization, two beams are generated in the directions of the radar target and the communication receiver, but with a high sidelobe level in the undesired direction. After optimization, the sidelobe level is substantially reduced, effectively attenuating the interference to the specific automotive radar. Compared with DRL-optimized and GSO-optimized transmit beamforming in Fig. 5, it can be seen that the energy of communication and tracking direction optimized by DRL is more balanced and the sidelobes are reduced. Fig. 6 explores the impact of power allocation coefficients. Initially, the radar radiation power coefficient $c_1$ for the ROI was set to 0.1. Subsequently, the power was increased to $c_1 = 0.5$. Notably, power coefficients serve as effective tools for regulating the radiation pattern through the assignment of power to distinct functions.

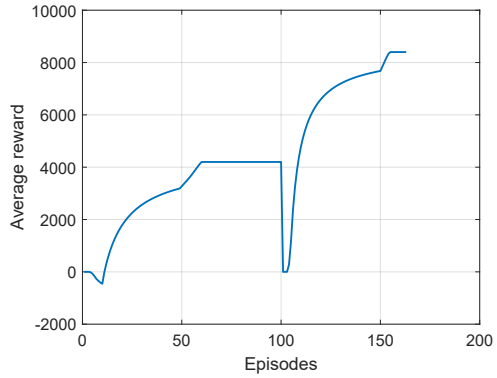


Fig. 4. The reward during training.

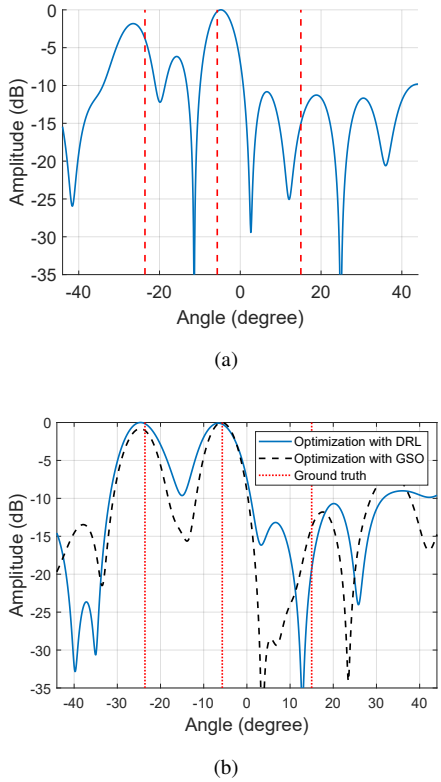


Fig. 5. (a) Transmit beamforming in the initial phase; (b) Transmit beamforming after optimization with DRL and GSO, Ground truth directions are indicated in red dash lines.

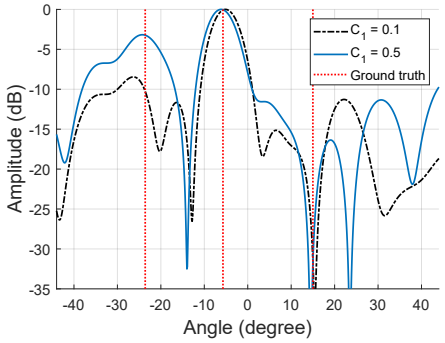


Fig. 6. DRL optimized transmit beamforming in sparse phased array with quantized phase shifters under different assigned power coefficients to radar radiation.

### B. Sensing Performance Evaluation

In phase array mode, once the emission beampattern of the fully phased array has been determined by the DRL, further adjustments can be made based on the interaction of the radar echo spectrum with the DRL agent. As depicted in the Fig. 7, after receiving the energy radiated by the antenna array, the radar's receiving antenna array processes the echo from the target. Because of the directivity of the transmitted beampattern, the echo of the target of no interest with a higher reflection coefficient is reduced and the communication target with a lower reflection coefficient is enhanced. Simultaneously, the range-angle spectrum shows that the target in the null position can no longer stand out in the spectrum, unlike in

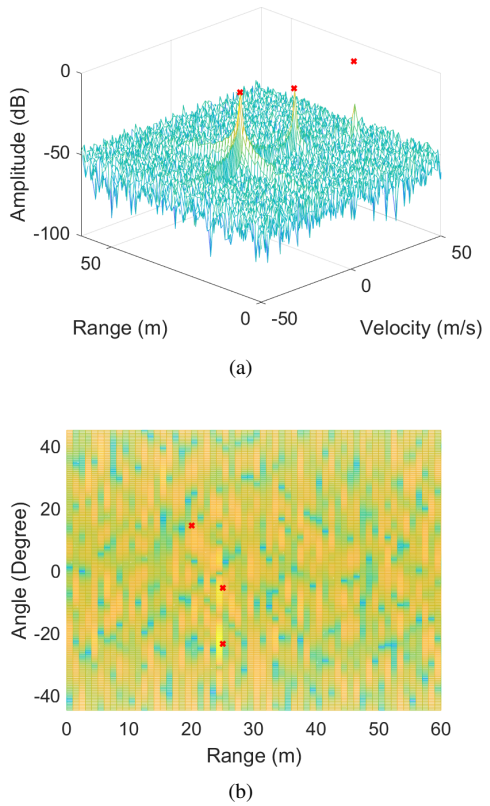


Fig. 7. Target parameters estimation in phased array sensing mode: (a) Range-Doppler spectrum; (b) Range-Angle spectrum. The red × denotes the actual positions of the targets.

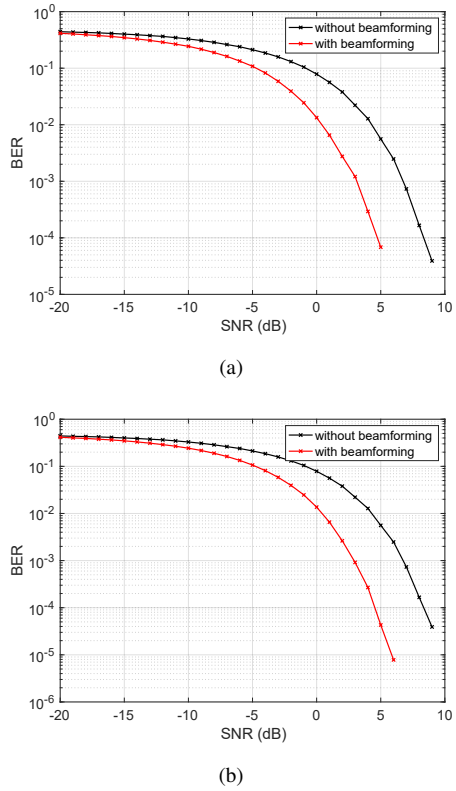


Fig. 8. BER versus SNR with binary phase-coded communication symbol along slow-time (a) and hybrid (b).

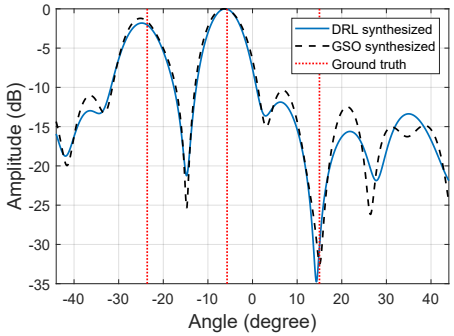


Fig. 9. Transmit beamforming in uniform phased array status with quantized phase shifters.

Fig. 2(b). In addition, due to the characteristics of the phase-controlled array, the angle resolution is also compromised relative to the MIMO radar mode.

### C. Communication Performance Evaluation

We assess the performance of the communication function by evaluating the bit error rate (BER).

Fig. 8 illustrates the BER variation trend with respect to the input SNR for two coding modes, where the number of chirps is set to $M = 512$, the number of receive antennas at the user end is set to 1, and the hybrid coding mode employs $L = 400$ time samples. Monte Carlo simulations were run for $10,000$ rounds at each SNR scenario. As seen in the plot, when the input SNR exceeds $10$ dB, the BER drops to $10^{-6}$. Hybrid coding exhibits an enhanced transmission efficiency compared to slow-time coding.

### D. Comparison With Optimization Methods

The comparison between the proposed and traditional optimization methods is of a significant value, and we select the modified beampattern synthesis method of relaxation optimization as the baseline method [77].

In the case of optimizing the beam vector of a ULA array, as shown in Fig. 9, both the DRL and the traditional optimization methods can achieve the desired beampattern, and the positions of the two mainlobes appear in the preset directions. However, due to the quantization requirements, the mainlobe may have a slight deviation, which falls within the control range of $\alpha_3$. While the deepest position of the DRL-optimized null may slightly differ from the expected direction, it still ensures the lowest possible transmission power in the direction of the desired target. The optimal phase values yielded by the GSO algorithm have arbitrary high precision. These high precision phase values are quantized to the nearest discretized phase values. The phase quantization results in a small performance loss compared with the DRL solution. In the optimization of the sparse transmission array $\mathbf{w}$, as illustrated in Fig. 5, both the DRL method and the traditional optimization method can accurately align the transmission mainlobe's direction with the radar sensing target and communication receiver. However, the DRL method has

a better beampattern synthesis performance compared to the optimization method, while both approaches can effectively regulate the radiation power in the null direction. Overall, the results indicate that the DRL method outperforms the relaxed optimization method in terms of sidelobe control.

## V. Conclusions

We presented an innovative DRL framework, inspired by the Wolpertinger's strategy, for the development of intelligent automotive radar DFRC systems. This framework is designed to optimize antenna distribution and accurately calibrate the quantized phase of low-bit double-phase shifters. Unlike traditional single phase shifters, the use of double-phase shifters in the proposed system allows for concurrent tracking of targets, enhancing communication capabilities, and reducing interference in undesired directions. This approach is particularly adept at navigating high-dimensional action spaces without requiring exhaustive action searches. In terms of communication with objects, both slow-time coding and hybrid coding methods have shown promising results. However, hybrid coding stands out in improving communication reliability. The proposed method surpasses the conventional relaxed optimization technique in effectively tuning the transmission matrix. It aligns the mainlobes precisely with the directions of radar sensing and communication users, while concurrently pointing nulls to the victim radars. Simulation results validated the feasibility and efficiency of our proposed approach, marking a significant advancement in automotive radar DFRC systems. While DRL-based reconfigurable beamforming for automotive radar sensing and communication is promising, several challenges remain. One critical area for improvement is enhancing the robustness of DRL-based beamforming systems against environmental variations, such as multipath reflections caused by surrounding objects. Another key challenge lies in achieving real-time processing and computational efficiency when implementing DRL-based beamforming on automotive radar hardware with limited computational resources.

## References


[1] L. Xu, R. Zheng, and S. Sun, "A deep reinforcement learning approach for integrated automotive radar sensing and communication," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Trondheim, Norway, June 2022, pp. 316–320.

[2] L. Xu, S. Sun, Y. D. Zhang, and A. P. Petropulu, "Joint antenna selection and beamforming in integrated automotive radar sensing-communications with quantized double phase shifters," in *Proc. IEEE 48th Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, June 2023.

[3] M. Hasan, S. Mohan, T. Shimizu, and H. Lu, "Securing vehicle-to-everything (V2X) communication platforms," *IEEE Trans. Intell. Veh.*, vol. 5, no. 4, pp. 693–713, 2020.

[4] S. Chen, J. Hu, Y. Shi, Y. Peng, J. Fang, R. Zhao, and L. Zhao, "Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G," *IEEE Commun. Stand. Mag.*, vol. 1, no. 2, pp. 70–76, 2017.

[5] S. D. Blunt, P. Yatham, and J. Stiles, "Intrapulse radar-embedded communications," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 3, pp. 1185–1200, 2010.

[6] A. R. Chiriyath, B. Paul, G. M. Jacyna, and D. W. Bliss, "Inner bounds on performance of radar and communications co-existence," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 464–474, 2016.

[7] 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.

[8] B. Li and A. P. Petropulu, "Joint transmit designs for coexistence of MIMO wireless communications and sparse sensing radars in clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 6, pp. 2846–2864, 2017.

[9] A. R. Chiriyath, B. Paul, and D. W. Bliss, "Radar-communications convergence: Coexistence, cooperation, and co-design," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 1, pp. 1–12, 2017.

[10] 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.

[11] A. Hassanien, M. G. Amin, Y. D. Zhang, and F. Ahmad, "Signaling strategies for dual-function radar communications: an overview," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 31, no. 10, pp. 36–45, 2016.

[12] P. M. McCormick, S. D. Blunt, and J. G. Metcalf, "Simultaneous radar and communications emissions from a common aperture, part I: Theory," in *IEEE Radar Conference (RadarConf)*, May 2017, pp. 1685–1690.

[13] P. Kumari, J. Choi, N. Gonzalez-Prelcic, and R. W. Heath, "IEEE 802.11ad-based radar: An approach to joint vehicular communication-radar system," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3012–3027, 2018.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] D. Ma, T. Huang, N. Shlezinger, Y. Liu, X. Wang, and Y. C. Eldar, "A DFRC system based on multi-carrier agile FMCW MIMO radar for vehicular applications," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, Dublin, Ireland, June 2020, pp. 1–7.

[19] J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, and A. Petropulu, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1295–1315, 2021.

[20] M. F. Keskin, H. Wymeersch, and V. Koivunen, "MIMO-OFDM joint radar-communications: Is ICI friend or foe?" *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1393–1408, 2021.

[21] J. Pritzker, J. Ward, and Y. C. Eldar, "Transmit precoder design approaches for dual-function radar-communication systems," *arXiv preprint arXiv:2203.09571*, 2022.

[22] J. A. Zhang, K. Wu, X. Huang, Y. J. Guo, D. Zhang, and R. W. Heath, "Integration of radar sensing into communications with asynchronous transceivers," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 106–112, 2022.

[23] X. Wang, W. Zhai, X. Zhang, X. Wang, and M. G. Amin, "Enhanced automotive sensing assisted by joint communication and cognitive sparse MIMO radar," *IEEE Trans Aerosp Electron Syst*, vol. 59, no. 5, pp. 4782–4799, 2023.

[24] 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.

[25] 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.

[26] J. Li and P. Stoica, "MIMO radar with colocated antennas," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 106–114, 2007.

[27] 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.

[28] 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.

[29] Z. Xu and A. P. Petropulu, "Phased array with improved beamforming capability via use of double phase shifters," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Trondheim, Norway, June 20-23, 2022, pp. 66–70.

[30] Z. Cheng, L. Wu, B. Wang, M. B. Shankar, and B. Ottersten, "Double-phase-shifter based hybrid beamforming for mmWave DFRC in the

presence of extended target and clutters," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 6, 2022.
[31] X. Wang, E. Aboutanios, M. Trinkle, and M. G. Amin, "Reconfigurable adaptive array beamforming by antenna selection," *IEEE Trans. Signal Process.*, vol. 62, no. 9, pp. 2385–2396, 2014.
[32] 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, 2018.
[33] A. Ahmed, S. Zhang, and Y. D. Zhang, "Antenna selection strategy for transmit beamforming-based joint radar-communication system," *Digital Signal Processing*, vol. 105, no. 102768, pp. 1–11, 2020.
[34] H. Huang, L. Wu, B. Shankar, and A. M. Zoubir, "Sparse array design for dual-function radar-communications system," *arXiv preprint arXiv:2301.00786*, 2023.
[35] S. Evmorfos, Z. Xu, and A. Petropulu, "Gflownets for sensor selection," in *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2023, pp. 1–6.
[36] ——, "Sensor selection via gflownets: A deep generative modeling framework to navigate combinatorial complexity," *arXiv preprint arXiv:2407.19736*, 2024.
[37] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process Mag*, vol. 34, no. 6, pp. 26–38, 2017.
[38] 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.
[39] ——, "Phase-modulation based dual-function radar-communications," *IET Radar, Sonar & Navigation*, vol. 10, no. 8, pp. 1411–1421, 2016.
[40] W. Baxter, E. Aboutanios, and A. Hassanien, "Joint radar and communications for frequency-hopped MIMO systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 729–742, 2022.
[41] Z. Xu, A. Petropulu, and S. Sun, "A joint design of MIMO-OFDM dual-function radar communication system using generalized spatial modulation," in *IEEE Radar Conference (RadarConf)*, 2020, pp. 1–6.
[42] T. Huang, N. Shlezinger, X. Xu, Y. Liu, and Y. C. Eldar, "Majorcom: A dual-function radar communication system using index modulation," *IEEE Trans. Signal Process.*, vol. 68, pp. 3423–3438, 2020.
[43] C. Baquero Barneto, T. Riihonen, M. Turunen, L. Anttila, M. Fleischer, K. Stadius, J. Ryynanen, and M. Valkama, "Full-duplex OFDM radar with LTE and 5G NR waveforms: Challenges, solutions, and measurements," *IEEE Trans. Microw. Theory Tech.*, vol. 67, no. 10, pp. 4042–4054, 2019.
[44] S. D. Liyanaarachchi, T. Riihonen, C. B. Barneto, and M. Valkama, "Optimized waveforms for 5G–6G communication with sensing: Theory, simulations and experiments," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 12, pp. 8301–8315, 2021.
[45] M. F. Keskin, V. Koivunen, and H. Wymeersch, "Limited feedforward waveform design for OFDM dual-functional radar-communications," *IEEE Trans. Signal Process.*, vol. 69, pp. 2955–2970, 2021.
[46] P. Kumari, S. A. Vorobyov, and R. W. Heath, "Adaptive virtual waveform design for millimeter-wave joint communication–radar," *IEEE Trans. Signal Process.*, vol. 68, pp. 715–730, 2019.
[47] S. H. Dokhanchi, M. B. Shankar, M. Alaee-Kerahroodi, and B. Ottersten, "Adaptive waveform design for automotive joint radar-communication systems," *IEEE Trans Veh Technol*, vol. 70, no. 5, pp. 4273–4290, 2021.
[48] Z. Xu and A. Petropulu, "A bandwidth efficient dual-function radar communication system based on a MIMO radar using OFDM waveforms," *IEEE Trans. Signal Process.*, vol. 71, pp. 401–416, 2023.
[49] Z. Zhang, Q. Chang, J. Xing, and L. Chen, "Deep-learning methods for integrated sensing and communication in vehicular networks," *Vehicular Communications*, vol. 40, p. 100574, 2023.
[50] W. Yuan, F. Liu, C. Masouros, J. Yuan, D. W. K. Ng, and N. Gonzalez-Prelcic, "Bayesian predictive beamforming for vehicular networks: A low-overhead joint radar-communication approach," *IEEE Trans Wirel Commun*, vol. 20, no. 3, pp. 1442–1456, 2021.
[51] X. Zhang, X. Wang, H. So, A. M. Zoubir, J. A. Zhang, and Y. J. Guo, "Transmit waveform design for integrated wideband mimo radar and bi-directional communications," *IEEE Transactions on Vehicular Technology*, 2024.
[52] X. Wang, A. Hassanien, and M. G. Amin, "Sparse transmit array design for dual-function radar communications by antenna selection," *Digital Signal Processing*, vol. 83, pp. 223–234, 2018.
[53] J. Huang, X. Zhang, X. Wang, and A. M. Zoubir, "Transmit sparse array beamformer design for dual-function radar communication systems," in *IEEE International Radar Conference (RADAR)*. IEEE, 2023, pp. 1–6.
[54] X. Zhang, X. Wang, J. Huang, and H. C. So, "Joint design of antenna selection and transmit linear array beamformer for integrated radar and communications," *Digital Signal Processing*, vol. 144, p. 104281, 2024.
[55] A. M. Elbir, K. V. Mishra, S. Chatzinotas, and M. Bennis, "Terahertz-band integrated sensing and communications: Challenges and opportunities," *arXiv preprint arXiv:2208.01235*, 2022.
[56] Y. Zhang, T. Osman, and A. Alkhateeb, "Online beam learning with interference nulling for millimeter wave MIMO systems," *arXiv preprint arXiv:2209.04509*, 2022.
[57] 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.
[58] J. De Wit, W. Van Rossum, and A. De Jong, "Orthogonal waveforms for FMCW MIMO radar," in *Proc. 2011 IEEE Radar Conference (RadarConf)*, Kansas City, MO, 2011, pp. 686–691.
[59] R. Rajamäki, S. P. Chepuri, and V. Koivunen, "Hybrid beamforming for active sensing using sparse arrays," *IEEE Trans. Signal Process.*, vol. 68, pp. 6402–6417, 2020.
[60] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, 2011.
[61] K. Wu, J. A. Zhang, X. Huang, and Y. J. Guo, "Frequency-hopping mimo radar-based communications: An overview," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 4, pp. 42–54, 2021.
[62] L. Huang, Y. Zhang, Q. Li, and J. Song, "Phased array radar-based channel modeling and sparse channel estimation for an integrated radar and communication system," *IEEE Access*, vol. 5, pp. 15 468–15 477, 2017.
[63] F. Lyu, H. Zhu, N. Cheng, H. Zhou, W. Xu, M. Li, and X. Shen, "Characterizing urban vehicle-to-vehicle communications for reliable safety applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2586–2602, 2019.
[64] O. Simeone, U. Spagnolini, Y. Bar-Ness, and S. H. Strogatz, "Distributed synchronization in wireless networks," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 81–97, 2008.
[65] S. Sedighi, K. V. Mishra, M. B. Shankar, and B. Ottersten, "Localization with one-bit passive radars in narrowband internet-of-things using multivariate polynomial optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 2525–2540, 2021.
[66] F. Lampel, F. Uysal, F. Tigrek, S. Orru, A. Alvarado, F. Willems, and A. Yarovoy, "System level synchronization of phase-coded fmcw automotive radars for radcom," in *Proc. European Conference on Antennas and Propagation (EuCAP)*, Copenhagen, Denmark, 2020, pp. 1–5.
[67] R. H. DuHamel, "Optimum patterns for endfire arrays," *Proceedings of the IRE*, vol. 41, no. 5, pp. 652–659, 1953.
[68] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when $p$ is much larger than $n$," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.
[69] W. Roberts, P. Stoica, J. Li, T. Yardibi, and F. Sadjadi, "Iterative adaptive approaches to MIMO radar imaging," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 1, pp. 5–20, 2010.
[70] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. Baggeroer, "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 425–443, 2010.
[71] L. Xu and S. Sun, "Coprime visible regions assisted angle unfolding for sparse esprit," in *2023 IEEE Radar Conference (RadarConf23)*, 2023, pp. 1–6.
[72] M. A. Wiering and M. Van Otterlo, "Reinforcement learning," *Adaptation, learning, and optimization*, vol. 12, no. 3, p. 729, 2012.
[73] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.
[74] 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.
[75] M. Alrabeiah, Y. Zhang, and A. Alkhateeb, "Neural networks based beam codebooks: Learning mmWave massive MIMO beams that adapt to deployment and hardware," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3818–3833, 2022.
[76] 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.
[77] X. Wang, E. Aboutanios, and M. G. Amin, "Thinned array beampattern synthesis by iterative soft-thresholding-based optimization algorithms," *IEEE Trans. Antennas Propag.*, vol. 62, no. 12, pp. 6102–6113, 2014.