IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 27, 2022, accepted May 27, 2022, date of publication June 8, 2022, date of current version June 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3181152

Beamforming Optimization for IRS-Assisted mmWave V2I Communication Systems via Reinforcement Learning

YEONGROK LEE[1], (Student Member, IEEE), JU-HYUNG LEE[2], (Member, IEEE), AND YOUNG-CHAI KO[1], (Senior Member, IEEE)

[1]School of Electrical Engineering, Korea University, Seoul 02841, South Korea
[2]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90007, USA

Corresponding authors: Young-Chai Ko (koyc@korea.ac.kr) and Ju-Hyung Lee (juhyung.lee@usc.edu)

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2021-0-01810) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

ABSTRACT Intelligent reflecting surface (IRS), which can provide a propagation path where non-line-of-sight (NLOS) link exists, is a promising technology to enable beyond fifth-generation (B5G) mobile communication systems. In this paper, we jointly optimize the base station (BS) and IRS beamforming to enhance network performance in the mmWave vehicle-to-infrastructure (V2I) communication system. However, the joint optimization of the beamforming matrix for BS and IRS is challenging due to non-convex and time-varying issues. To tackle those issues, we propose a novel reinforcement learning algorithm based on deep deterministic policy gradient (DDPG) method. Simulation results corroborate that the proposed algorithm converges in both systems *with* and *without* IRS, and the case *with* IRS improves the network performance from as little as about 5% to as much as about 100% depending on the environments such as the number of vehicles or deployment. Simulation results also show that in the IRS-assisted communication, up to 10% higher network throughput can be achieved in *Dense* V2I network scenario compared to *Sparse* case.

INDEX TERMS Intelligent reflecting surface (IRS), deep reinforcement learning (DRL), vehicle-to-infrastructure communications (V2I), mmWave.

I. INTRODUCTION

Witnessing an exponential growth of the number of connected machines, unprecedented requirements are expected across wireless communications [1]–[3]. Examples of connected machines include not only new form-factors, such as augmented reality (AR), virtual reality (VR), and hologram devices but also autonomous mobile devices, such as an unmanned aerial vehicle (UAV) and autonomous driving. Each requires different categories of service: enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-type communications (mMTC) as standardized in the fifth-generation (5G) [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Hassan Omar.

However, the existing resource (such as, sub-6 GHz) band may not be enough to satisfy all the services and requirements of beyond 5G (B5G) due to its resource scarcity. In this regard, the standardization work and academic research toward B5G or sixth-generation (6G) are already actively underway [4]. In particular, many studies have investigated the potential of other higher frequency bands such as millimeter wave (mmWave) [5], [6] and Terahertz (THz) bands [7]–[9].

This high-frequency band communication poses other challenges: severe path loss and extra signal attenuation. Particularly, the signal in such a high frequency band is attenuated by atmospheric conditions, e.g., water vapor and oxygen [10]. Moreover, due to its high directivity, the signal is severely attenuated in a non-line-of-sight (NLOS) environment [10]. Nevertheless, it can be overcome by the additional deployment of a base station (BS) or by utilizing a massive

multiple input multiple output (MIMO) at the expense of implementation complexity and hardware cost.

Recently, as a promising alternative solution for the challenges in the high-frequency band, an intelligent reflecting surface (IRS) has been introduced with an emerging material called metasurface [11]–[16]. IRS is also called a reconfigurable intelligent surface (RIS). Each IRS element is programmable such that the amplitude and the phase of a signal can be reconfigured as desired. Another noteworthy advantage is that IRS requires fewer RF chains than conventional multiple antennas systems; thereby, the power consumption and the hardware cost can be reduced. Thus, the low-cost and reconfigurable alternative can deal with the limitations of high-frequency communications by installing more of it instead of conventional BS and by intelligently reconfiguring it to obtain favorable channel conditions [13], [17]–[19].

A. RELATED WORKS

1) IRS-ASSISTED COMMUNICATION SYSTEMS

Numerous studies have been introduced to utilize the various advantages of IRS. IRS has been mainly considered for supporting the downlink multi-user multiple input single output (MU-MISO) case in fixed terrestrial networks and its performance and optimization have been widely studied [20]–[23]. These studies mainly focused on optimizing well-known communication techniques such as beamforming to improve network performance such as throughput and secrecy rate in various situations.

In [24] and [25], the authors introduced the operation method and the cost of the existing technologies with which IRS can replace such as backscatter, MIMO relay, and massive MIMO and compared them with IRS. Moreover, the IRS-assisted communication has been considered for mobile networks as well as fixed networks, such as UAV-enabled wireless networks and vehicular networks. In particular, in dynamic network scenarios [26]–[28], not only the beamforming optimization for IRS but also trajectory design for UAV is addressed to fully utilize the advantage of dynamic networks. However, since its dynamics make the optimization problem to be non-convex, it could be inefficient to optimize the transmit beamforming matrix with conventional methods [29]–[31]. In order to efficiently solve optimization problems that are difficult to solve with conventional methods, a recent study introduced a technique for optimizing the beamforming matrices in the IRS-assisted network to which deep reinforcement learning is applied [32]–[34].

In some studies, feasibility studies of IRS technology have been conducted in various network scenarios, such as optimizing the transmit beamforming or IRS reflecting matrix to increase network performance such as coverage and spectral efficiency. However, few studies have contributed to careful consideration of the channel characteristics of IRS-assisted communications, which directly affect coverage and spectral efficiency.

2) mmWave VEHICLE-TO-INFRASTRUCTURE COMMUNICATION SYSTEMS

The growth of automation technologies leads to the openness of mobile networks advancing toward 6G. To satisfy the requirements of enhanced throughput and reduced latency, the high-frequency bands, mmWave or THz band, are also considered in vehicle-to-infrastructure (V2I) or vehicle-to-everything (V2X) communications. As pointed out in [35] and [36], the vehicle communication using mmWave depends on LOS and focused-reflected paths, not on scattering and diffracting paths. Some studies have analyzed the vehicle communications with mmWave band to tackle the issues and introduced some challenges [37]–[40].

IRS can also be leveraged for the mmWave-V2X networks, particularly in urban environments which suffer from securing LOS channel conditions and from limited coverage. Using IRS instead, costs can be reduced compared with conventional BS to guarantee coverage and LOS channels in the mobile mmWave-V2X network scenario; thereby, it is of great importance to optimize the beamforming architecture to provide more wide coverage. However, there are only a few studies that have addressed the issue of beamforming design of IRS-assisted networks in the mobile networks [41], [42].

B. CONTRIBUTIONS AND PAPER ORGANIZATION

In this paper, we maximize the network throughput by jointly optimizing the beamforming matrix of BS and reflecting matrices of IRS in mmWave V2I network with IRS-assisted communication systems. To do so, the deep reinforcement learning (DRL) method is proposed by taking into account the characteristic of the non-stationary V2I network environment [43]. The main contributions of this work are summarized as follows:

- We jointly optimize the BS beamforming matrix and the IRS reflecting matrices to maximize the throughput of the IRS-assisted mmWave V2I network.
- We propose a novel DRL algorithm based on deep deterministic policy gradient (DDPG) method [44], to address the non-convex and time-varying optimization while considering the mmWave V2I network channel and environment.
- Simulation results demonstrate that the proposed DDPG-based algorithm converges and IRS helps improve mmWave V2I network performance. Moreover, the comparison results are presented over the number of vehicles and the network density.

The remainder of this paper is organized as follows. In Section II, IRS-assisted mmWave V2I communication system model is presented. In Section III, the rate maximization problem is proposed. In Section V, simulation results are provided, followed by concluding remarks in Section VI.

Notation: The boldface capital letters and lower case letters denote matrices and column vectors, respectively. Capital calligraphic letters denotes finite discrete set and $|\cdot|$ denotes cardinality of the set if applied to the finite discrete set, or absolute value if applied to the complex

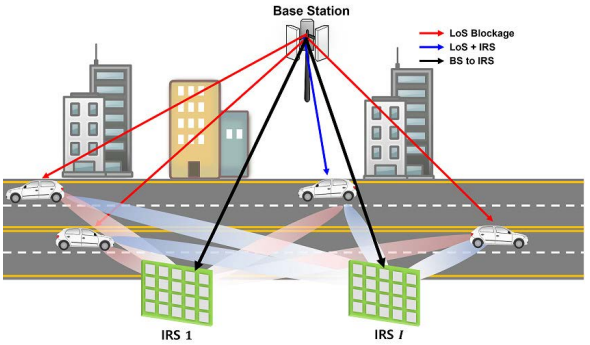


FIGURE 1. An illustration of IRS-assisted mmWave V2I network model.

number. For example, $|\mathcal{M}|$ is the finite discrete set of the BS antennas. $(\cdot)^{\mathrm{H}}$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{-1}$ denote the Hermitian transpose, the transpose and matrix inversion of matrices or vectors, respectively. tr$(\cdot)$ and diag$(\cdot)$ denote trace and diagonal matrix with elements in vector. $\mathbb{C}^{A\times B}$ and $\mathbb{R}^{A\times B}$ denote the space of $A \times B$ complex-values matrices and real-values matrices, respectively. $\mathbb{E}[\cdot]$ denotes the statistical expectation. Re$(\cdot)$ and Im$(\cdot)$ denote the real and imaginary part of a complex number, respectively.

II. SYSTEM MODEL

As illustrated in Fig. 1, we consider the multi-IRS-assisted mmWave V2I network scenario. In particular, to consider a practical environment, we here consider an urban case with UMi Street Canyon model [37] in which a base station (BS) and multiple IRSs exist around the street. Table 1 gives the description of the symbols and notations used throughout this paper.

A. CHANNEL MODEL

In this paper, we assume that the V2I network operates in the mmWave band. In particular, throughout this paper, we consider the Saleh-Valenzuela (SV) channel model [5] with slow fading, which is a conventional channel model of mmWave MIMO case. In addition, Doppler effect [37] is further considered by taking account of the characteristics of mobile V2I network. This channel model can be expressed by baseband equivalent channel matrix $\mathbf{H}$ given as

$$\mathbf{H}(t) = \mathbf{H}_{\mathrm{NLOS}}(t) + \mathbf{H}_{\mathrm{LOS}}(t), \tag{1}$$

where $\mathbf{H}_{\mathrm{NLOS}}(t)$ and $\mathbf{H}_{\mathrm{LOS}}(t)$ are denoted as (2) and (3), shown at the bottom of the page, respectively. In (2) and (3), C is the number of clusters, L_i is the number of paths for the i-th cluster, $\beta_{i,j}$ is the path loss, and

TABLE 1. Description of symbols and notations.

Symbol/Notation	Description
M	Number of the BS antennas
I	Number of the IRSs
R	Number of reflecting elements of an IRS
K	Number of vehicles
δ	Discretized time length
$\boldsymbol{q}[n] \in \mathbb{R}^{K\times 2}$	Position of vehicles at time $t = n\delta$
$\boldsymbol{q}_k[0] \in \mathbb{R}^{1\times 2}$	Iniitial position of the k-th vehicle
$\boldsymbol{v}[n] \in \mathbb{R}^{K\times 2}$	Velocity of vehicles at time $t = n\delta$
$\boldsymbol{a}[n] \in \mathbb{R}^{K\times 2}$	Acceleration of vehicles at time $t = n\delta$
$\mathbf{W} \in \mathbb{C}^{M\times K}$	BS's transmit beamforming matrix
$\mathbf{\Phi}_i \in \mathbb{C}^{R\times R}$	Reflecting elements array of i-th IRS
$\mathbf{H}_i \in \mathbb{C}^{M\times R}$	Channel from BS to i-th IRS
$\mathbf{h}_{d,k} \in \mathbb{C}^{M\times 1}$	Direct channel from BS to i-th IRS
$\mathbf{g}_{k,i} \in \mathbb{C}^{R\times 1}$	Channel from i-th IRS to k-th vehicle
$\alpha_{r,i}$	r-th Reflecting elements of i-th IRS
y_k	Multi-hop received signal for k-th vehicle
γ_k	Multi-hop SINR for k-th vehicle
C	Multi-hip network throughput
P_t	Transmission power budget

$f_{i,j} = v^{\mathrm{R}} \cos(\varphi^{\mathrm{R}}_{i,j}) \sin(\theta^{\mathrm{R}}_{i,j})/\lambda$ is Doppler frequency shift, $\varphi^{\mathrm{R}}_{i,j}$ and $\varphi^{\mathrm{T}}_{i,j}$ denote azimuth angle of arrival and departure, $\theta^{\mathrm{R}}_{i,j}$ and $\theta^{\mathrm{T}}_{i,j}$ are the elevation angle of arrival and departure, respectively, all for the j-th ray of the i-th cluster. With a slight abuse of notation, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{R}}$ for scalar value represents the value for transmitter and the receiver, respectively. An array response vector of uniform linear array can be expressed as

$$\boldsymbol{a}(\varphi_{i,j}) = \frac{1}{\sqrt{M}}[1,\ e^{j2\pi\frac{d}{\lambda}\sin\varphi_{i,j}},\ \ldots,\ e^{j2\pi\frac{d}{\lambda}(M-1)\sin\varphi_{i,j}}]. \tag{4}$$

Unlike $\boldsymbol{a}^{\mathrm{T}}$ and $\boldsymbol{a}^{\mathrm{R}}$, the array response vector of IRS, $\boldsymbol{a}^{\mathrm{I}}(\varphi^{I}_{i,j}, \theta^{I}_{i,j})$ is based on a uniform planar array (UPA), not a uniform linear array (ULA), which can be expressed as following

$$\boldsymbol{a}^{\mathrm{I}}(\varphi^{\mathrm{I}}_{i,j}, \theta^{\mathrm{I}}_{i,j}) = \frac{1}{\sqrt{N}}[1,\ e^{j2\pi\frac{d}{\lambda}\sin\varphi^{\mathrm{I}}_{i,j}\cos\theta^{\mathrm{I}}_{i,j}},\ \ldots, \times e^{j2\pi\frac{d}{\lambda}(M-1)\sin\varphi^{\mathrm{I}}_{i,j}\cos\theta^{\mathrm{I}}_{i,j}}], \tag{5}$$

where φ^{I} and θ^{I} denote azimuth and elevation angle of the j-th ray of the i-th cluster for the IRS elements, respectively. Note that in UPA vector, not only the azimuth angle but also the elevation angle are considered.

$$\mathbf{H}_{\mathrm{NLOS}}(t) = \sqrt{\frac{MN}{\sum_{i=1}^{N_{cl}} L_i}} \sum_{j=1}^{N_{cl}} \sum_{l=1}^{L_i} \beta_{i,j} e^{-j2\pi f_{i,j} t} \boldsymbol{a}^{\mathrm{R}}(\varphi^{\mathrm{R}}_{i,j}, \theta^{\mathrm{R}}_{i,j})(\boldsymbol{a}^{\mathrm{T}}(\varphi^{\mathrm{T}}_{i,j}, \theta^{\mathrm{T}}_{i,j}))^{\mathrm{H}}), \tag{2}$$

$$\mathbf{H}_{\mathrm{LOS}}(t) = I_{\mathrm{L}}(d_0)\sqrt{MN} e^{j\eta[n]} \beta_0 e^{-j2\pi f_0 t} \boldsymbol{a}^{\mathrm{R}}(\varphi^{\mathrm{R}}_0, \theta^{\mathrm{R}}_0)(\boldsymbol{a}^{\mathrm{T}}(\varphi^{\mathrm{T}}_0, \theta^{\mathrm{T}}_0))^{\mathrm{H}}, \tag{3}$$

For the detail of channel, the parameters of LOS in (3) are expressed, similarly to parameters of NLOS, in (2) by using the subscript 0. In (3), $\eta[n] \sim \mathcal{U}(0, 2\pi)$ denotes a random variable that changes the phase according to the environment, and $I_{\mathrm{L}}(d_0)$ is a function for LOS probability at the distance d_0 between the transceiver. As considered in [12], we assume that the channel state can be perfectly estimated by using channel estimation techniques for various mmWave communication systems.

B. V2I NETWORK SCENARIO

Consider that a BS is equipped with M antennas and communicating with K vehicles each equipped with a single antenna ($M \geq K$). In addition, I IRSs assist the network between BS and vehicles to enhance communication performance. We assume that all IRSs are equipped with R passive reflective elements. For mathematical convenience, throughout this paper, we denote the set of BS antenna, IRS, elements of IRS, and vehicle as $\mathcal{M} \in \{m = 1, 2, \ldots, M\}$, $\mathcal{I} \in \{i = 1, 2, \ldots, I\}$, $\mathcal{R} \in \{r = 1, 2, \ldots, R\}$, and $\mathcal{K} \in \{k = 1, 2, \ldots, K\}$, respectively.

We also denote $\mathbf{H}_i \in \mathbb{C}^{R \times M}$ as the channel matrix between BS and the i-th IRS, $\boldsymbol{g}_{i,k} \in \mathbb{C}^{R \times 1}$ as the channel vector between the i-th IRS to the k-th vehicle, and $\boldsymbol{h}_k \in \mathbb{C}^{M \times 1}$ as the channel matrix from the BS to the k-th vehicle.

By following the discrete-time state-space model [45], [46], we consider that time is discretized in slots of length δ. Then, the position of vehicles, $\boldsymbol{q}[n] = \left[q_1[n], q_2[n], \ldots, q_K[n]\right]^{\mathrm{T}}$, at time n can be expressed as

$$\boldsymbol{q}[n+1] = \boldsymbol{q}[n] + \boldsymbol{v}[n]\delta + \frac{1}{2}\boldsymbol{a}[n]\delta^2, \quad \forall n, \tag{6}$$

where $q_k[0]$ is the initial position of the vehicle k at time $n = 0$, assuming that each vehicle has a different initial position. This position of vehicles constraint can be written as

$$\boldsymbol{q}_{\min} \leq \boldsymbol{q}_k[n] \leq \boldsymbol{q}_{\max}, \quad \forall k, n, \tag{7}$$

where $\boldsymbol{q}_{\min} = [q_{\min,x}, q_{\min,y}]^{\mathrm{T}}$ is the minimum value and $\boldsymbol{q}_{\max} = [q_{\max,x}, q_{\max,y}]^{\mathrm{T}}$ is the maximum value of coordinate in 2-D Cartesian coordinate plane, respectively.

Similar to (6) and (7), the velocity of the vehicles, $\boldsymbol{v}[n] = [v_1[n]v_2[n] \ldots v_K[n]]^T$, at time n can be also expressed as

$$\boldsymbol{v}[n+1] = \boldsymbol{v}[n] + \boldsymbol{a}[n]\delta, \quad \forall n, \tag{8}$$

where $v_k[0]$ is the initial velocity of the k-th vehicle.

In our V2I network model, there are two types of communications link; *i)* the direct link from BS-to-vehicle, and *ii)* the reflected link from BS-to-IRSs-to-vehicle, as depicted in Fig. 1. For ease of analysis, the following conditions are assumed; the signals of both links can be transmitted to the receiver synchronously, there is no reflection between IRSs, and there is a central controller between BS and IRSs which coordinates them to synchronize and for beamforming.

C. NETWORK MODEL

Let us denote $\boldsymbol{w}_k \in \mathbb{C}^{M \times 1}$ and $\mathbf{W} \in \mathbb{C}^{M \times K}$ for the transmit beamforming vector and matrix at BS for the vehicle k, respectively. Then we can write the transmit signal, x, at BS as

$$\boldsymbol{x} = \sum_{k=1}^{K} \boldsymbol{w}_k s_k, \tag{9}$$

where $s_k \sim \mathcal{CN}(0, 1)$ is the transmitted symbol for the vehicle k at BS. In addition, there is a maximum transmission output limit in BS. We consider the following power constraint at BS as

$$\mathbb{E}\left[\|x\|^2\right] = \mathrm{tr}\left(\mathbf{W}\mathbf{W}^{\mathrm{H}}\right) \leq P_t, \tag{10}$$

where P_t is the total transmit power at BS.

The i-th IRS reflecting elements can be expressed as $\alpha_{r,i} = \beta_{r,i}e^{j\theta_{n,i}}$, where $\beta_{r,i} \in [0, 1]$ and $\theta_{r,i} \in [0, 2\pi]$ are the amplitude and the phase of the i-th IRS's the r-th elements, respectively. Additionally, the i-th IRS reflecting matrix is a diagonal matrix denoted by $\Phi_i = \mathrm{diag}\left(\alpha_{1,i}, \alpha_{2,i}, \ldots, \alpha_{R,i}\right) \in \mathbb{C}^{R \times R}$. We assume that ideal IRS reflecting elements mounted on all the IRSs that do not affect the power of the signal like a mirror. This assumption can be denoted by $\left|\alpha_{r,i}\right| = 1$ for all the values of n and i. In other words, we suppose that $\beta_{r,i} = 1$ for all r and i for the remainder of the paper. We also assume that the IRS reflecting elements are configured in a square shape. That is, when the total number of IRS elements is R, the number of horizontal and vertical elements can be considered as $\sqrt{R}$.

In our network configuration, there are two network model, i.e., *single-hop* and *multi-hop*. For single-hop case, BS can transmit a signal directly to vehicles, i.e., BS-vehicles, which represents the conventional V2I network model. While, for multi-hop case, IRS and BS support the V2I network, i.e., BS-IRS-vehicles and BS-vehicles; That is, there is not only a direct link from BS but also a multi-hop link from BS-IRS.

1) SINGLE-HOP

For single-hop case, the received signal at the vehicle k, y_k, can be expressed as

$$y_k^s = \underbrace{\sum_{k=1}^{K} \boldsymbol{h}_{d,k}^{\mathrm{H}} \boldsymbol{w}_k s_k}_{\text{desired signal}} + \underbrace{\sum_{l=1, l\neq k}^{K} \boldsymbol{h}_{d,k}^{\mathrm{H}} \boldsymbol{w}_l s_l}_{\text{interference signal}} + n_k, \tag{11}$$

where $\boldsymbol{h}_{d,k} \in \mathbb{C}^{M \times 1}$ is the baseband equivalent channel from BS to vehicle k, s_k denotes the transmit signal for the vehicle k and $n_k \sim \mathcal{CN}\left(0, \sigma_k^2\right)$ is the independent and identically distributed (i.i.d.) Gaussian noise at the vehicle k. Then, SINR for the single-hop link at vehicle k, γ_k^s, is given by

$$\gamma_k^s = \frac{\left(\boldsymbol{h}_{d,k}^{\mathrm{H}} \boldsymbol{w}_k\right)^2}{\left(\sum_{l=1, l\neq k}^{K} \boldsymbol{h}_{d,k}^{\mathrm{H}} \boldsymbol{w}_l\right)^2 + \sigma_k^2}, \tag{12}$$

and the network throughput of the single-hop network model is given as

$$C^s = \sum_{k=1}^{K} \log_2(1 + \gamma_k^s). \tag{13}$$

2) MULTI-HOP

Single BS and multiple IRSs are connected to multiple vehicles in the multi-hop network model. We consider that the received signal y_k at the vehicle k is the sum of all signals from the IRSs and BS, which can be expressed as

$$y_k^m = (\boldsymbol{h}_{d,k}^{\mathrm{H}} + \underbrace{\sum_{i=1}^{I} \boldsymbol{g}_{k,i}^{\mathrm{H}} \Phi_i \mathbf{H}_i}_{\text{reflected link channel}})\boldsymbol{w}_k s_k + n_k, \tag{14}$$

and as in single-hop case, SINR for the multi-hop link at vehicle k, γ_k^m, is given by

$$\gamma_k^m = \frac{\left|(\boldsymbol{h}_{d,k}^{\mathrm{H}} + \sum_{i=1}^{I} \boldsymbol{g}_{k,i}^{\mathrm{H}} \boldsymbol{\Phi}_i \mathbf{H}_i)\boldsymbol{w}_k\right|^2}{\left|\sum_{l=1, l\neq k}^{K} (\boldsymbol{h}_{d,k}^{\mathrm{H}} + \sum_{i=1}^{I} \boldsymbol{g}_{k,i}^{\mathrm{H}} \Phi_i \mathbf{H}_i)\boldsymbol{w}_l\right|^2 + \sigma_k^2}. \tag{15}$$

Thus, the network throughput of the multi-hop network model is given as

$$C = \sum_{k=1}^{K} \log_2(1 + \gamma_k^m). \tag{16}$$

III. NETWORK THROUGHPUT MAXIMIZATION FOR IRS-ASSISTED mmWave V2I NETWORKS

This section addresses the network throughput maximization problem for the multi-IRS-assisted mmWave V2I network by optimizing the beamforming matrices. In this paper, we define the overall performance of the system for one time slot as the *network throughput*. We also define the average value of network throughput over the entire time as *average network throughput (ANT)*.

A. PROBLEM FORMULATION

Throughout this paper, we aim to jointly optimize the BS transmit beamforming matrix $\mathbf{W}$ and the IRSs reflecting beamforming matrices Φ_i for maximizing the network throughput. The following problem, (P1), corresponds to the network throughput under the constraints related to the characteristics of IRSs and the actual conditions of vehicles given as

$$\begin{aligned} * \ \text{(P1)} \quad & \max_{\mathbf{W}, \Phi_i} C \\ \text{s.t. } & (6) - (8) \\ & \mathrm{tr}\left(\mathbf{W}\mathbf{W}^{\mathrm{H}}\right) \leq P_t, && \text{(17a)} \\ & \left|\alpha_{r,i}\right| = 1, \quad \forall r, i && \text{(17b)} \\ & \Phi_i = \mathrm{diag}(\alpha_{1,i}, \ \alpha_{2,i}, \ \ldots, \ \alpha_{R,i}), \quad \forall i, && \text{(17c)} \end{aligned}$$

where P_t denotes the transmission power of BS. In (P1), (17a) is the power contraint at BS while the constraints of (17b) and (17c) represent the characteristics of the IRSs reflecting beamforming matrices. Each of these represents a form of aan IRS reflecting beamforming matrix that an IRS reflecting element reflects all the transmitted signals without power loss.

However, since the problem (P1) is non-convex with non-convex objective functions and constraints, it is challenging to solve it with general convex optimization techniques. Although there are some methods to approach the non-convex optimization problem that provide sub-optimal solutions such as the successive convex approximation (SCA) method [47], [48], but that is still difficult to apply to this system model, as the optimization problem is composed of some entities with stochastic channel. Furthermore, the variables considered in this system are too many to utilize the conventional exhaustive search-based method. Thus, we propose a joint beamforming method via a DRL method, which is elaborated in the following section.

IV. BEAMFORMING OPTIMIZATION FOR IRS-ASSISTED V2I NETWORKS VIA DEEP REINFORCEMENT LEARNING

In this section, we introduce the DRL-based beamforming optimization method. Firstly, the MDP model is designed, which casts the optimization problem (P1). Next, our proposed algorithm, based on DDPG [44], is introduced.

A. MARKOV DECISION PROCESS MODELING

We design the problem (P1) into environment, state, behavior, and reward of the MDP model.

1) ENVIRONMENT

Our environment consists of the proposed communication systems, in which there is an agent that interacts with this environment to find the optimal actions and policies that maximize cumulative rewards. The environment includes all information related to the networks such as the BS, vehicle and IRS. Specifically, the transmission power of BS, the characteristics of IRS elements, the state of vehicles, and the channel information are included in the environment. At each time step n, an agent observes a state $s[n]$ from the state space $\mathcal{S}$, accordingly takes an action $a[n]$ from the action space $\mathcal{A}$ based on a policy $\pi(s, a)$, which is a mapping from the state space to the action space. Let us define the cardinalities of the state space and action space as $|\mathcal{S}|$ and $|\mathcal{A}|$, respectively. After performing the action, the current state $s[n]$ of the environment changes to the next state $s[n+1]$. In addition, the agent receives current reward $r[n]$.

We point out that V2I network environment has a periodic pattern such that the movement of vehicles in this scenario follows a similar pattern for a certain period. Thus, we set one episode of the environment as $n = 0, \cdots, T-1$ and consider the initial time slot $n = 0$ and the final time slot $n = T-1$.

2) STATE

In this system, the agent obtains a state of the system by observing the environment. We aim to optimize the

BS beamforming matrix and the IRS reflecting matrix to maximize the network throughput of the network scenario. Therefore, observable information related to this network throughput is included in the state. In particular, the state includes the followings: the BS transmit beamforming matrix $\mathbf{W}[n]$, the IRSs reflecting matrices $\Phi_i[n]$, and the channel information ($\mathbf{H}_i[n]$, $\boldsymbol{g}_{k,i}[n]$, and $\boldsymbol{h}_{k,i}[n]$).

Then we can write the state $s[n]$ as

$$s[n] = \{\mathbf{W}[n], \Phi_i[n], \mathbf{H}_i[n], \times \boldsymbol{g}_{k,i}[n], \boldsymbol{h}_{k,i}[n], i \in \mathcal{I}, k \in \mathcal{K}\}, \quad \forall n. \tag{18}$$

3) ACTION

In our problem (P1), the BS transmit beamforming matrix $\mathbf{W}$ and the IRSs reflecting beamforming matrices Φ_i are jointly optimized to maximize the total throughput of the system. Accordingly, the action space of the system includes those matrices. Thus, the action $a[n]$ is given by

$$a[n] = \{\mathbf{W}[n], \Phi_i[n], i \in \mathcal{I}\}, \quad \forall n. \tag{19}$$

Note that the beamforming vectors and reflecting elements are continuous values rather than discrete values; accordingly, the action is also determined in the continuous action space.

4) REWARD

The aim of optimization problem (P1) is to maximize the total network throughput of IRS-assisted V2I networks. We set the reward function as the network throughput of the multi-hop network in (16). Therefore, for the time slot n, the instantaneous reward $r[n]$ is given by

$$r[n] = C[n], \ \forall n, \tag{20}$$

where $C[n]$ denotes the sum rate of the system at the time step n.

B. BEAMFORMING OPTIMIZATION VIA DRL

Under the designed MDP model, we employ a DDPG-based DRL algorithm for beamforming optimization. Before describing the considered DDPG algorithm, we firstly introduce a deep Q-Network (DQN), which is the basis of our algorithm. For the convenience of expression, state $s[n]$, action $a[n]$, and reward $r[n]$ are shortened as s_n, a_n and r_n, respectively.

1) DEEP Q-NETWORK

Deep Q-Network (DQN) is one of the most widely used reinforcement learning algorithms and is based on model-free, value-based and off-policy. The method learns the optimal policy to maximize cumulative future rewards. The cumulative reward R at time n is expressed as

$$R_n = \sum_{t=n}^{\infty} \gamma^{t-n} r_t, \tag{21}$$

where γ is the discount rate, which distinguish between present and future rewards by setting a higher weight to the present reward, e.g., $0 \leq \gamma \leq 1$.

We define the expected sum of future reward as the action-value function, $Q(s, a)$, when the action, a, is performed in a state s with a policy π. The action-value function, $Q(s, a)$, is also called Q-value. The agent needs to find the optimal Q-value to maximize the reward and the optimal action-value function $Q^*(s, a)$ to find this optimal value is defined as

$$Q^*(s, a) = \max_{\pi} \mathbb{E}_{\pi}[R_t | s_t, a_t, \pi]. \tag{22}$$

This optimal Q-value is a function that can obtain the best reward when action a is taken in state s. The role of policy π is to calculate the Q-value by mapping state s and action a. The model-free DRL trains the policy by using a Bellman equation [49] to find the optimal Q-value, which can be expressed as

$$Q^*(s, a) = \mathbb{E}_{s'}[r_t + \gamma \max_{a'} Q^*(s', a') | s, a]. \tag{23}$$

However, in reality, the optimal Q-value cannot be found due to the lack of information, so a function that converges to the optimal Q-value is found by updating the Q-value through the policy π. This operation of repeatedly updating the Q-value is expressed as an equation as follows

$$Q^{\pi}_{n+1}(s, a) = \mathbb{E}[r + \gamma \mathbb{E}_{r,a'}[Q^{\pi}_n(s', a')], \tag{24}$$

and Q^{π}_n will converge to Q^* as n goes to infinity.

This iterative process, called value iteration, enables the problem in (24) to find the optimal Q-value. However, note that the iterative training and learning becomes challenging as the dimension of state and action increases since cause a severe complexity in calculation of Q-value and storing data. To solve this issue, the authors of [50] proposed a method to find the approximate Q-value through DNN instead of using the Q-table created by finding the deterministic Q-value, of which algorithm is called a deep Q-Network (DQN). Here, the loss function L_i is given as

$$L(\theta_i) = \mathbb{E}_{s,a,r,s'}[(y_i - Q(s, a; \theta_i))^2], \tag{25}$$

where

$$y_i = \mathbb{E}_{s,a,r,s'}[r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) | s, a], \tag{26}$$

and θ denotes DNN parameters for finding the optimal policy by stochastic gradient descent (SGD) method. Therefore, we can calculate the optimal weight θ through SGD of the loss function as follows

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}[(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a, ; \theta_i)) \nabla_{\theta_i} Q(s, a; \theta_i)]. \tag{27}$$

However, the trajectory data used in (27) and the value function are temporally correlated while learning the policy, which degrades the performance. In particular, if samples are correlated, this method does not perform well because the SGD method assumes that each sample is independent and evenly distributed. Note that DQN method uses two tricks to address this issue: *i)* the experience is not used for learning immediately, but stored in the replay buffer; and *ii)* when

data accumulates more than a certain amount, it is randomly extracted and used for learning. This idea is called *experience and replay*, and it makes the samples independent.

2) DEEP DETERMINISTIC POLICY GRADIENT

For the DQN method, there are some hurdles for the application. The amount of computation rapidly increases as the number of actions or states increases due to the burden of Q-value calculation. Besides, it can only deal with discrete actions, not with continuous ones. In value-based methods such as DQN, values must be discretized to the handle continuous action, but this method has some limitations. When the action space is discretized, the action space increases exponentially, making learning almost impossible, and since the optimal action can also be removed in the discretization process, it is difficult to find the desired action.

Most of the algorithms used to solve this problem are based on policy gradient (PG), especially, actor-critic method. In particular, compared to other stochastic policy gradient methods, the deterministic policy gradient (DPG) method is learned through a deterministic action space rather than considering the probability distribution of the action space, so the amount of computation is small and convergence is fast [51].

In this paper, we consider the deep deterministic policy gradient (DDPG) that combines the advantages of DQN and DPG for our system model. DDPG was introduced in [44] by improving the conventional DPG, a model-free, off-policy, actor-critic learning method. We make some modification of the original DDPG by using the experiance replay such as in DQN. Unlike many actor-critic methods that are on-policy methods, DDPG is also applicable because it is an off-policy method. There are two more problems, one of which is the problem of updating the actor network and the critic network using the gradient obtained from the time difference (TD) error.

The main critic network $Q(s, a|\theta^Q)$ and the main actor network $\mu(s|\theta^\mu)$ can be expressed as

$$Q(s, a) = \mathbb{E}[r_1^\gamma]S_1 = s, A_1 = a; \pi], \tag{28}$$
$$\mu(s) = \mathbb{E}[r_1^\gamma]S_1 = s; \pi]. \tag{29}$$

In addition, time delayed copy of the critic and the actor network are defined as $Q'(s, a|\theta^{Q'})$ and $\mu'(s|\theta^{\mu'})$, respectively. Those networks are also called target critic network and target actor network, respectively. When selecting the action for the next time step through an actor network $\mu'(s|\theta^{\mu'})$ in DDPG, a random action is selected for exploration. In the paper that first proposed DDPG [44], Ornsten-Uhlenbeck (OU) noise derived from OU process $\mathcal{N}$ is used. Random action is selected by adding this noise to the output value of the network. The formula for selecting a random action in DDPG in this way can be written as

$$a_l = \mu(s_l|\theta^\mu) + \mathcal{N}_l. \tag{30}$$

The time difference target y_i and the loss function L to be used in the critic network can be written respectively as

$$y_i = r_i + \gamma Q'(s', u'(s'; \theta^{u'})|\theta^{Q'}), \tag{31}$$
$$L = \frac{1}{N_b}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2. \tag{32}$$

The gradient of objective function can be calculated as

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \int_S \rho^\pi(s) \int_A \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) da\, ds \\ &= \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)],\end{aligned} \tag{33}$$

where J is the objective function, in the form of a discounted cumulative reward, which is given as

$$J(\theta) = \mathbb{E}[Q(s, a)|\pi_\theta(a|s) Q^\pi(s, a)]. \tag{34}$$

For the next learning step, the parameters of the main actor network θ^μ are updated as $\theta^\mu \leftarrow \theta^\mu - lr_\mu \nabla_{\theta^\mu}$. Finally, the target critic and actor network are updated through a soft update target parameter τ, which controls the learning frequency of the target networks. This parameter update process is summarized as

$$\begin{aligned}\theta^{Q'} &\leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \\ \theta^{\mu'} &\leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}.\end{aligned} \tag{35}$$

The following subsection introduces our proposed reinforcement learning algorithm for beamforming optimization based on the DDPG algorithm with some modifications.

3) PROPOSED DDPG-BASED ALGORITHM

Fig. 2 shows the training process of our proposed algorithm for IRS-assisted mmWave V2I communication systems. As described in Section II, we aim to jointly optimize the BS transmit beamforming matrix $\mathbf{W}$ and the IRSs reflecting matrices Φ_i. The real and imaginary elements, $\Re(w_{m,k})$ and $\Im(w_{m,k})$, of $\mathbf{W}$ are continuous in the range $[-1, 1]$, respectively. Similarly, the amplitude and the phase of IRS elements $\beta_{r,i}$ and $\theta_{r,i}$ of Φ_i are also continuous in the range $[0, 1]$ and $[0, 2\pi]$, respectively. Also, all the channel matrices have continuous complex-values. Note that our MDP model consists of continuous values for both states and actions.

In DRL, a non-linear activation function is used to prevent a situation in which gradient vanishes or explodes when learning a neural network. Since most of the non-linear activation functions have very limited domains, our system model does not handle a wide range of values such as elements of a channel matrix. Therefore, to solve the gradient vanishing or exploding problem, *ReLU6* is used for an activation function, as well as the operation of the batch normalization [52].

First, the activation function, ReLU6, is a modified form of the widely used ReLU [53]. The ReLU function has the advantage of efficiently solving the gradient vanishing or exploding problem, and ReLU6 has an additional advantage that it can make a quick learning when the feature is sparse like our system. Also, batch normalization is a method of

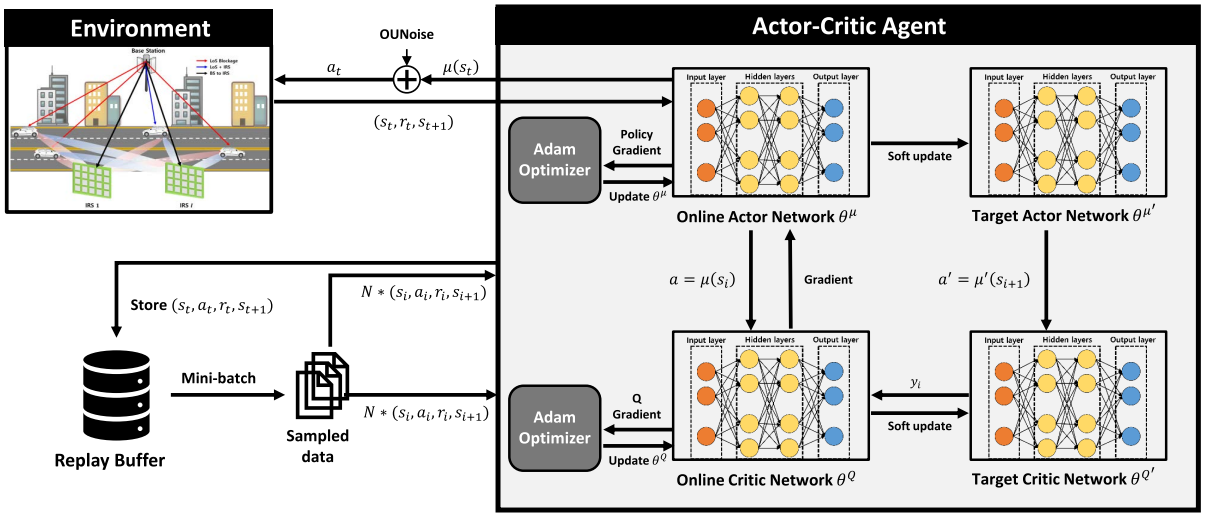


FIGURE 2. Learning diagram of the proposed DDPG-based algorithm.

normalizing the mean and variance of input values for each layer in the neural network so that the distribution is not deformed [52].

V. NUMERICAL EVALUATIONS

This section presents the numerical results of the proposed joint beamforming optimization. We, here, compare and evaluate the proposed scheme by configuring various network scenarios by changing the number of BS antennas, vehicles, and IRSs.

A. SIMULATION SETUP

1) ENVIRONMENT

In this section, we consider a special case in Section II where all the vehicles are moving at a constant velocity in the x direction. The initial position of vehicles is as $q_{\min,x} = 0$ in the x-coordinate. Therefore, the discrete position of vehicles in this environment can be expressed as

$$0 \le q_k[n] \le q_{\max,x}, \quad \forall k, n, \tag{36}$$

where $q_{\max,x}$ is the maximum in the x coordinate that vehicle can reach and the velocity of vehicles is assumed to be a constant, v_0, given as

$$v[n] = v_0, \quad \forall n. \tag{37}$$

In our simulation situation, we assume that the episode end after a certain period of time T, regardless of the location of the car. Even though $q_{\max,x}$ can be any value, we set $q_{\max,x}$ to $q_{\max,x} = q_K[T-1]$ which is the largest value of position of any vehicles. We include the position and speed of the vehicle defined in this section as additional constraints in (P1). The detail of considered environment configuration is illustrated in Fig. 3. In this environment, we set $T = 50$ with a time step size of 0.1 [sec] as an example. Unless otherwise stated, the parameters related to BS, IRS, and vehicle are

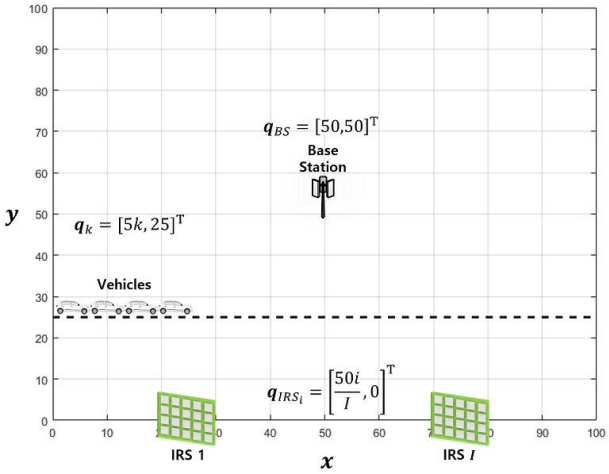


FIGURE 3. System model of the simulation environment.

used as summarized in Table 2. Throughout this section, *ANT* is regarded as the main performance metric of our network scenario defined as

$$ANT = \frac{\sum_l^L r[n]}{T} \text{ [bps/Hz]}. \tag{38}$$

2) DRL NETWORK

The network structure of the proposed DDPG-based algorithm is shown in Table 3. In the case of online and target actor networks, since the entire action is received as input, the number of nodes is the same as the size of the action space $\mathcal{A}$. Since online and target critic networks receive both state and action as inputs and determine the Q value, the number of nodes is equal to the sum of the state space $\mathcal{S}$ and the action space $\mathcal{A}$.

We employ *Tensorflow 2* with some modifications to handle the complex values to implement the proposed algorithm. In our networks considered for the simulations, there are two

Algorithm 1 Proposed DDPG-Based Algorithm

Input: **Learning parameters** : $E, T, \gamma, \tau, B, N_b, \mu_a, \mu_c, \tau_a$ and τ_c

Randomly initialize critic network $Q(s, a|\theta^Q)$ and the actor network $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ;
Initialize the target critic network Q' and the target actor network μ' with weight $\theta^{Q'} \leftarrow \theta^Q$ and $\theta^{\mu'} \leftarrow \theta^\mu$;
Initialize replay buffer R;
for episode $= 1$ to E **do**
    Initialize the environment and a random OUnoise process $\mathcal{N}$ for action exploration;
    Receive initial observation state s_0;
    for $l = 0$ to $T - 1$ **do**
        Executes the beamforming design based on the state s_l and the policy μ, and $a_l = \mu(s_l|\theta^\mu) + \mathcal{N}_l$;
        Perform the action a_l and records reward r_l and the next state s';
        Store the transition (s_l, a_l, r_l, s') in R;
    end for
    Sample a random mini-batch of N_b transitions from R;
    Set $y_i = r_i + \gamma Q'(s', \mu'(s'; \theta^{\mu'})|\theta^{Q'})$;
    Minimize the loss function to update the critic network:
    $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$,
    Update the online critic network weights θ^Q as:
    $\theta^Q \leftarrow \theta^Q - lr_Q \nabla_{\theta^Q}$;
    Update the online actor network by sampled stochastic policy gradient ascent as:
    $\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s)|_{s=s_i}$;
    Update θ^u can be expressed as:
    $\theta^\mu \leftarrow \theta^\mu - lr_\mu \nabla_{\theta^\mu}$;
    Soft update the target critic network and the target actor network:
    $\theta^{Q'} \leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'}$,
    $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'}$;
end for
return P

hidden layers with the number of each node given as 400 and 300, respectively. Note that all the layers are fully connected. The activation function of the hidden layer uses the ReLU6 function, which can solve the gradient vanishing problem that occurs during training in DRL [53]. The activation function of the output layer of the critic network is used only to determine the Q value. Therefore, the output value is used by using a linear function as the activation function. Unlike the critic network, in the actor network, since the behavior is the beamforming matrix of the BS and the reflecting element of the IRS, the range of the network output value must be in $[-1, 1]$ according to the constraint of the optimization problem. Since the most used activation function in this situation is the *tanh* function, it is used as the activation function of the actor network [54], [55]. In the learning process, the parameters specified in Table 2 are used unless otherwise stated.

TABLE 2. Environment and learning parameters.

Parameter	Value
Transmission power of the BS	$P_t = 1$ [W]
Number of time slots per episode	$T = 50$
Total observation time	$T/10 = 5$ [sec]
Location of BS	$q_{BS} = [50, 50]^{\mathrm{T}}$ [m]
Location of IRSs	$\boldsymbol{q}_{IRS_i} = [\frac{50i}{I}, 0]^{\mathrm{T}}$ [m]
Initial location of Vehicles	$\boldsymbol{q}_k = [5k, 25]^{\mathrm{T}}$
Velocity of vehicles	$\boldsymbol{v}[n] = v_0 = 30$ [km/h]
Number of episodes	$E = 15000$
Discount rate of reward	$\gamma = 0.99$
Soft target update parameter	$\tau = 0.005$
Buffer size experience replay	$B = 100000$
Mini-batch size	$N_b = 100$
Starting training step (Warming up)	100
Learning rate of main actor network	$\mu_a = 0.0001$
Learning rate of main critic network	$\mu_c = 0.001$
Learning rate of target actor network	$\tau_a = 0.0001$
Learning rate of target critic network	$\tau_c = 0.001$
Actor target update step	2

B. CONVERGENCE

Firstly, we present the convergence behavior of the proposed DDPG-based algorithm in Fig. 4. Particularly in Fig. 4(a) shows convergence curve as iterations go. Note that the variance and the mean of each result are drawn in bold and shaded forms, respectively, by conducting three different simulations. Here, three cases are considered, 1) *No-IRS*, 2) *Single-IRS* and 3) *Multi-IRS*. In this section, We consider that *Multi-IRS* case has two IRSs.

Fig. 4(b) shows the result for the last 5000 episodes in Fig. 4(a) to show and compare the convergence and variance of each. It is shown that *Multi-IRS* obtains the best performance on average, although *Single-IRS* and *Multi-IRS* achieve a comparable ANT; while the variance of *Multi-IRS* is relatively small compared to *Single-IRS* and *No-IRS*. On the other hand, Fig. 4(c) shows the average result for the last 5000 episodes in Fig. 4(a) to explicitly compare the converged policy of each simulation. The achievable throughput after convergence in each case is 14.7 for Multi-IRS, 14.25 for Single-IRS, and 9.75 for No-IRS. Multi-IRS converges around 5500 episodes, while Single-IRS and No-IRS take around 4000 and 2000 episodes, respectively. Those results suggest that Multi-IRS achieves the best throughput performance at the expense of training complexity.

C. AVERAGE NETWORK THROUGHPUT

We compare the ANT performance of our proposed schemes denote as *w/ IRS-DRL*, with two different schemes, denoted as *Random* and *w/o IRS-DRL*.

TABLE 3. Network structures of the proposed DDPG-based algorithm.

Network	Connectivity	Layer[1]	Number of Nodes	Activation function
Policy network $\theta^{\mu}, \theta^{\mu'}$	Fully connected	Input Layer	$\|2 \cdot \mathcal{S}\|$	-
		1st Hidden Layer	400	ReLU6
		2nd Hidden Layer	300	ReLU6
		Output Layer	$\|2 \cdot \mathcal{A}\|$	tanh
Q network $\theta^{Q}, \theta^{Q'}$	Fully connected	Input Layer	$\|2 \cdot \mathcal{S}\| + \|2 \cdot \mathcal{A}\|$	-
		1st Hidden Layer	400	ReLU6
		2nd Hidden Layer	300	ReLU6
		Output Layer	1	Linear

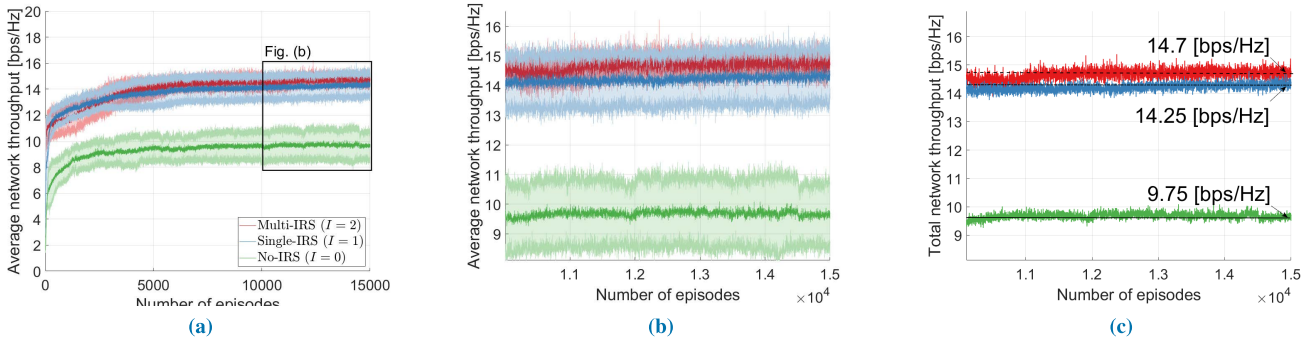


FIGURE 4. (a) Convergence curves of the proposed-DDPG algorithm, (b) Enlarged figure of last 5000 episodes in Fig. (a) and (c) Converged ANT curve obtained through the optimal policy.

1) W/WWWWW/IRS-DRL (Proposed)

In a situation that BS and IRS are beamformed together using the proposed DDPG-based DRL, the optimal BS beamforming matrix and IRS reflecting matrix are selected to maximize the ANT.

2) W/O IRS-DRL

It uses the same learning method as *w/ IRS-DRL*, that is, DRL based on DDPG, but considers the situation where there is no IRS in the environment. That is, it is a method of maximizing the ANT by selecting only the BS beamforming matrix.

3) RANDOM

In the same environment as *w/ IRS-DRL*, this is a method of randomly selecting the BS beamforming matrix and IRS reflecting matrices. Because the matrices are chosen randomly, the performance of this scheme is said to be the actual lower bound. The channel is generated by *Matlab* simulation and the simulation results are averaged over 10,000 times.

Fig. 5 and Table 4 show the convergence curve and the convergence values of each algorithms, respectively. Fig. 5(a) and 5(b) shows the baseline convergence curves in $M = 4$, $K = 2$ and $M = 4, K = 4$, respectively. First, *Random* selects the elements of BS beamforming matrix and IRS reflecting matrix completely randomly, so the performance is very poor in both cases.

TABLE 4. Comparison of ANT.

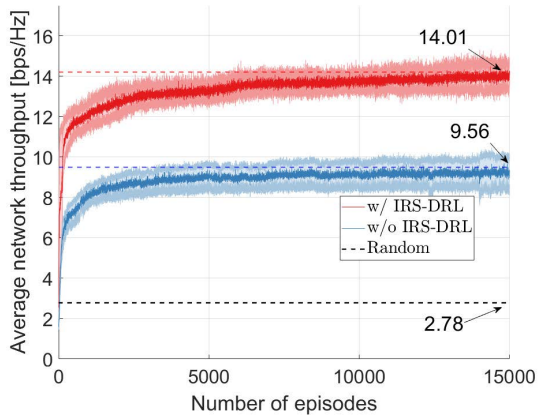
Algorithms	Avg. Net. Thr. [bps/Hz]	
	$M = 4, K = 2$	$M = 4, K = 4$
Random	2.78	1.81
w/o IRS-DRL	9.56	14.94
w/ IRS-DRL	14.01	22.77

In Fig. 5(a), we compare the ANT between our proposed scheme *w/ IRS-DRL*, and *w/o IRS-DRL* and show that *w/ IRS-DRL* scheme provides the throughput about 46.55% higher. Similarly, in Fig. 5(b), the ANT at *w/ IRS-DRL* is about 52.41% higher. This means that the use of IRS is a good way to improve communication performance in our proposed simulation environment.

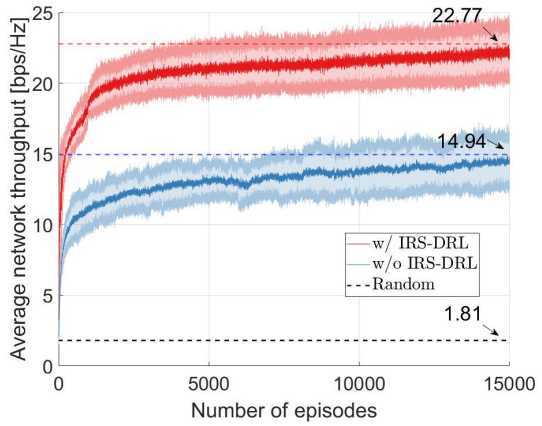
D. IMPACT OF THE NUMBER OF VEHICLES

We analyze and compare the effects of IRS and the number of vehicles on the ANT. For convenience of comparison, when using IRS, we consider that only one IRS is used. In addition, the number of BS antenna is fixed to $M = 8$, and the number of IRS reflecting elements is fixed to $R = 16$ as an example.

Fig. 6, shows the ANT according to the number of vehicles. In both *Single-IRS* and *No-IRS* cases, the ANT increases as



(a) Baseline convergence curve in $M = 4, K = 2$



(b) Baseline convergence curve in $M = 4, K = 4$

FIGURE 5. Convergence curves for comparison between out method and baselines.

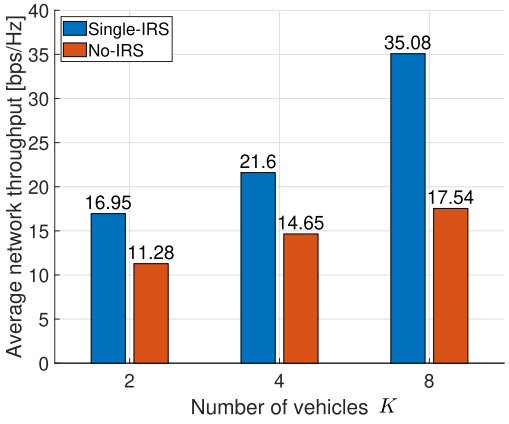


FIGURE 6. Learning diagram of the proposed DDPG-based algorithm.

the number of vehicles increases. In particular, we can see that the ANT of *Single-IRS* enhances significantly compared to the case of *No-IRS* at the number of vehicles increases. Those trends are related to the channel rank [56]. The rank in the MIMO channel is sufficient when the number of vehicles is relatively small compared to the number of BS antennas. However, when the difference between the number of BS antennas and the number of vehicles is small, the gain through the channel cannot be sufficiently obtained because the channel rank is low. Remarkably, the IRS improves overall channel condition by providing an additional channel rank and reducing correlation between different channels, as in the case of $K = 8$.

TABLE 5. ANT comparison for network density.

Algorithms	Avg. Net. Thr. [bps/Hz]	
	Sparse	*Dense*
No-IRS $(I = 0)$	20.86	22.52
Single-IRS $(I = 1)$	22.43	24.64
Multi-IRS $(I = 2)$	22.50	25.80

E. IMPACT OF NETWORK DENSITY

Table 5 shows the ANT over the network density. Note that it is assumed the initial location of vehicles in *Dense* and *Sparse* deployments are $\boldsymbol{q}_k[0] = [35 + 10\ k, 25]^{\mathrm{T}}$ and $\boldsymbol{q}_k[0] = [33.3\ k, 25]^{\mathrm{T}}$, respectively, while the period of this result is considered with a short time for $T = 5$, for convenience.

For *Sparse* network scenario, the network throughput improves by about 7.53% in *Single-IRS* and about 7.86% in *Multi-IRS* compared with *No-IRS*. Besides, for *Dense* case, the network throughput improves about 9.41% in *Single-IRS* and about 14.56% in *Multi-IRS* compared with *No-IRS*. As shown in this table, IRS can enhance the network throughput and multiple IRS further improve it. It is worth noting that in a dense network environment, in general, the interference power may significantly degrade the network performance. Nevertheless, in IRS-assisted communications, the reflective elements mitigate interference power well, improving overall network performance even in the dense network environment.

VI. CONCLUSION

This paper investigated a system in which the BS and the IRS perform beamforming jointly for mmWave V2I communications network. We proposed a novel DDPG-based DRL algorithm that optimizes the BS beamforming matrix and the IRS reflecting matrices to maximize network performance. Simulation results showed that IRS could improve the network performance in mmWave V2I communications network in dense as well as sparse network environments. Improving reinforcement learning structures and algorithms to support more base station antennas, IRS reflective elements, and vehicles could be an interesting future research. It is also interesting to consider generalized situations, such as those in which there are vehicles moving in random directions rather than following a road. Additionally, extending the optimization and considering beam tracking may be of interest for future studies. It is also worth investigating flexible MADRL frameworks to adapt quickly to a new environment with meta and split learning [57].

REFERENCES

[1] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and C. J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Sahin, and A. Kousaridas, "A tutorial on 5G NR V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1972–2026, 2021.

[4] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, 2020.

[5] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter wave communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1616–1653, 3rd Quart., 2018.

[6] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[7] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[8] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62367–62414, 2020.

[9] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 136–143, Dec. 2014.

[10] T. Bai, A. Alkhateeb, and R. W. Heath, "Coverage and capacity of millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 70–77, Sep. 2014.

[11] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[12] W. Qingqing and Z. Rui, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[13] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[14] W. Tang, M. Z. Chen, J. Y. Dai, Y. Zeng, X. Zhao, S. Jin, Q. Cheng, and T. J. Cui, "Wireless communications with programmable metasurface: New paradigms, opportunities, and challenges on transceiver design," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 180–187, Apr. 2020.

[15] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410–1414, Oct. 2019.

[16] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[17] L. Dai, B. Wang, M. Wang, X. Yang, J. Tan, S. Bi, S. Xu, F. Yang, Z. Chen, M. D. Renzo, C.-B. Chae, and L. Hanzo, "Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45913–45923, 2020.

[18] J. Hu, H. Zhang, B. Di, L. Li, K. Bian, L. Song, Y. Li, Z. Han, and H. V. Poor, "Reconfigurable intelligent surface based RF sensing: Design, optimization, and implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2700–2716, Nov. 2020.

[19] O. Özdogan, E. Björnson, and E. G. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 581–585, May 2020.

[20] J. Chen, Y.-C. Liang, Y. Pei, and H. Guo, "Intelligent reflecting surface: A programmable wireless environment for physical layer security," *IEEE Access*, vol. 7, pp. 82599–82612, 2019.

[21] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1488–1492, Sep. 2019.

[22] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.

[23] Q. Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, May 2020.

[24] M. Di Renzo, K. Ntontin, J. Song, F. H. Danufane, X. Qian, F. Lazarakis, J. De Rosny, D.-T. Phan-Huy, O. Simeone, R. Zhang, M. Debbah, G. Lerosey, M. Fink, S. Tretyakov, and S. Shamai, "Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 798–807, 2020.

[25] E. Björnson, O. Özdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020.

[26] H. Lu, Y. Zeng, S. Jin, and R. Zhang, "Aerial intelligent reflecting surface: Joint placement and passive beamforming design with 3D beam flattening," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4128–4143, Jul. 2021.

[27] Q. Zhang, W. Saad, and M. Bennis, "Reflections in the sky: Millimeter wave communication with UAV-carried intelligent reflectors," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[28] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 716–720, Jan. 2020.

[29] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[30] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064–3076, May 2020.

[31] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809–1822, Aug. 2020.

[32] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663–1677, Jun. 2021.

[33] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[34] C. Huang, R. Mo, and Y. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Jun. 2020.

[35] M. Giordani, A. Zanella, and M. Zorzi, "Millimeter wave communication in vehicular networks: Challenges and opportunities," in *Proc. 6th Int. Conf. Modern Circuits Syst. Technol. (MOCAST)*, May 2017, pp. 1–6.

[36] F. Jameel, S. Wyne, S. J. Nawaz, and Z. Chang, "Propagation channels for mmWave vehicular communications: State-of-the-art and future research directions," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 144–150, Feb. 2019.

[37] S. Buzzi and C. D'Andrea, "On clustered statistical MIMO millimeter wave channel simulation," 2016, *arXiv:1604.00648*.

[38] D. He, L. Wang, K. Guan, B. Ai, J. Kim, and Z. Zhong, "Channel characterization for mmWave vehicle-to-infrastructure communications in urban street environment," in *Proc. Eur. Conf. Antennas Propag. (EuCAP)*, 2019, pp. 1–5.

[39] Y. Wang, K. Venugopal, R. W. Heath, Jr., and A. F. Molisch, "MmWave vehicle-to-infrastructure communication: Analysis of urban microcellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7086–7100, Aug. 2018.

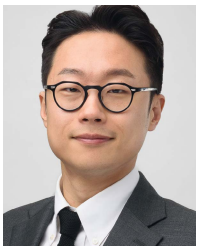
[40] T. Zugno, M. Drago, M. Giordani, M. Polese, and M. Zorzi, "NR V2X communications at millimeter waves: An End-to-End performance evaluation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[41] Y. Chen, Y. Wang, J. Zhang, and Z. Li, "Resource allocation for intelligent reflecting surface aided vehicular communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12321–12326, Oct. 2020.

[42] Y. U. Ozcan, O. Ozdemir, and G. K. Kurt, "Reconfigurable intelligent surfaces for the connectivity of autonomous vehicles," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2508–2513, Mar. 2021.

[43] M. A. A. Careem and A. Dutta, "Spatio-temporal recommender for V2X channels," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–7.
[44] 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," 2015, *arXiv:1509.02971*.
[45] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.
[46] J.-H. Lee, J. Park, M. Bennis, and Y.-C. Ko, "Integrating LEO satellite and UAV relaying via reinforcement learning for non-terrestrial networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020, pp. 1–6.
[47] J.-H. Lee, K.-H. Park, Y.-C. Ko, and M.-S. Alouini, "A UAV-mounted free space optical communication: Trajectory optimization for flight time," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1610–1621, Mar. 2020.
[48] J.-H. Lee, K.-H. Park, Y.-C. Ko, and M.-S. Alouini, "Spectral-efficient network design for high-altitude platform station networks with mixed RF/FSO systems," *IEEE Trans. Wireless Commun.*, early access, Mar. 4, 2022, doi: 10.1109/TWC.2022.3154401.
[49] A. K. Dixit, *Optimization in Economic Theory*. Oxford, U.K.: Oxford Univ. Press, 1990.
[50] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 33–529, Feb. 2015.
[51] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, vol. 32, 2014, pp. 387–395.
[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456.
[53] P. Panda, S. Venkataramani, A. Sengupta, A. Raghunathan, and K. Roy, "Energy-efficient object detection using semantic decomposition," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 9, pp. 2673–2677, Sep. 2017.
[54] B. L. Kalman and S. C. Kwasny, "Why tanh: Choosing a sigmoidal function," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 4, Jun. 1992, pp. 578–581.
[55] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018, *arXiv:1811.03378*.
[56] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[57] J. Kim, Y. Park, G. Kim, and S. J. Hwang, "SplitNet: Learning to semantically split deep networks for parameter reduction and model parallelization," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, Aug. 2017, pp. 1866–1874.

YEONGROK LEE (Student Member, IEEE) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2016, where he is currently pursuing the Ph.D. degree with the School of Electrical Engineering. His current research interests include software defined radio (SDR), intelligent reflecting surface (IRS), vehicular communication, and machine learning (ML) in communications and networks.



JU-HYUNG LEE (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Electronic Engineering, Korea University, Seoul, South Korea, in 2016 and 2021, respectively. He is currently a Postdoctoral Researcher at the University of Southern California, Los Angeles, USA. His research interests include optimization and algorithm design for non-terrestrial networks, machine learning for wireless communications, free-space optical communications, and signal processing techniques. He has received awards including the Best Paper Awards in IEEE ICTC, in 2021, the Travel Grant in IEEE GLOBECOM, in 2020, the Bronze Prize in IEEE Seoul Section Student Paper Contest, in 2020, and the Graduate Research Excellence Award at Korea University, in 2021.

YOUNG-CHAI KO (Senior Member, IEEE) received the B.Sc. degree in electrical and telecommunication engineering from Hanyang University, Seoul, South Korea, and the M.S.E.E. and Ph.D. degrees in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1999 and 2001, respectively. From January 2001 to March 2001, he was a Research Scientist with Novatel Wireless. In March 2001, he joined the Wireless Center, Texas Instruments Inc., San Diego, CA, USA, as a Senior Engineer. He is currently with the School of Electrical Engineering, Korea University, as a Professor. His current research interests include the design and evaluations of multi-user cellular systems, MODEM architecture, mm-wave, and tera Hz wireless systems.