Meta-Reinforcement Learning for Immersive Virtual Reality over THz/VLC Wireless Networks

Yining Wang*, Mingzhe Chen† ‡, Zhaohui Yang§, Walid Saad¶, Tao Luo*, Shuguang Cui‡, and H. Vincent Poor†
*Beijing Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, Beijing, China.
†Department of Electrical Engineering, Princeton University, Princeton, NJ, USA.
‡Shenzhen Research Institute of Big Data and Future Network of Intelligence Institute, the Chinese University of Hong Kong, Shenzhen, China.
§Centre for Telecommunications Research, Department of Engineering, Kings College London, WC2B 4BG, UK.
¶Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA.
Emails: wyy0206@bupt.edu.cn, mingzhec@princeton.edu, yang.zhaohui@kcl.ac.uk, walids@vt.edu,
tluo@bupt.edu.cn, shuguangcui@cuhk.edu.cn, poor@princeton.edu.

***Abstract*—In this paper, the problem of enhancing the quality of virtual reality (VR) services is studied for an indoor terahertz (THz)/visible light communication (VLC) wireless network. In the studied model, small base stations (SBSs) transmit high-quality VR images to users over THz bands and light-emitting diodes (LEDs) provide accurate indoor positioning services for VR users using VLC. Here, VR users move in real time and their movement patterns change over time according to their application. Both THz and VLC links can be blocked by the bodies of VR users. To control the energy consumption of the studied THz/VLC wireless VR network, VLC access points (VAPs) must be selectively turned on so as to ensure accurate and extensive positioning for VR users. Based on the user positions, each SBS must generate corresponding VR images and build THz links without body blockage to transmit the VR content. The problem is formulated as an optimization problem whose goal is to maximize the sum successful transmission probability of all VR users by selecting the appropriate VAPs to be turned on and controlling the user association with SBSs. To solve this problem, a policy gradient-based reinforcement learning (RL) algorithm using meta-learning framework is proposed. The proposed algorithm can effectively solve the formulated problem and enable the trained policy to quickly adapt to new user movement patterns. Simulation results demonstrate that, compared to a baseline trust region policy optimization algorithm (TRPO), the proposed meta-learning solution yields a 78% improvement in the convergence speed and about 16.4% improvement in the sum successful transmission probabilities of all VR users.**

I. Introduction

Deploying virtual reality (VR) applications over wireless networks provides new opportunities for VR to offer seamless user experience [1]. However, the scarce bandwidth of sub-6 GHz limits the ability of wireless networks to satisfy the stringent quality-of-service (QoS) requirements of VR applications in terms of delivering high data rates, low latency, and high reliability. A promising solution is to integrate VR services over high frequency terahertz (THz) bands with abundant bandwidth. However, propagation at THz covers short range and is highly prone to blockage [2]–[4]. In indoor VR scenarios, although short distances enable high-rate VR image transmission at THz frequencies, the mobile users' bodies may lead to dynamic blockages over the THz links, thus negatively affecting the immersive VR experience. In addition, to ensure a seamless interaction between the users and the virtual world, it is necessary to accurately locate VR users in real time for VR image generation and transmission. Therefore, deploying THz-enabled wireless networks to offer high-reliability VR services faces many challenges such as user positioning, reduction of link blockage, user association, and reliability assurance.

Recently, several works such as in [4]–[8] studied a number of problems related to wireless VR networks. In [5], the authors studied the use of both edge fog computing and caching to satisfy the low latency requirement of VR users. The work in [6] proposed a mobile edge computing-based VR delivery framework that can cache parts of the field of views (FOVs) so as to minimize the required transmission rate. However, the works in [5] and [6] sacrificed the quality of delivered VR videos (e.g., by reducing the resolution of VR videos or only displaying the FOV of 360° VR images) to meet the low latency constraints. The authors in [7] investigated the use of the millimeter wave (mmWave) bands to maximize the quality of the delivered video chunks in a wireless VR network. However, the works in [5]–[7] ignored the mobility of users that can significantly affect VR network performance. In [8], the authors predict the orientation and locations of VR users to minimize the occurrence of breaks in presence. However, most of the existing works such as [5]–[8] did not analyze the potential of using THz bands to provide immersive VR services. Moreover, the works in [5]–[8] ignored the need for accurate user localization in VR so as to enhance the virtual world experience. In [4], the authors derived the reliability of a THz-enabled VR system based on probabilistic line-of-sight (LoS) and non-line-of-sight (NLoS) THz links. However, the existing works on THz-enabled VR networks such as [4] did not consider the time-varying user positions that are used to generate VR images and avoid dynamic blockages of THz links. Meanwhile, in dense indoor VR scenarios, THz bands require the very narrow pencil beamforming. Therefore, although the work in [9] showed that THz has the potential for indoor positioning, it can only passively adjust the beam direction or user association after the user moves, which can detach the users from their virtual world. Visible light communication (VLC) based on light-emitting diodes (LEDs) can provide an alternative and more accurate positioning service [9]. To this end, a THz/VLC-enabled wireless VR network is needed in order to provide reliable positioning services to VR users as well as generate and transmit corresponding VR content based on the users' positions.

This work was supported in part by the U.S. National Science Foundation under Grants CCF-1908308 and CNS-1909372.

The main contribution of this work is a novel framework that jointly uses VLC and THz to service VR users. In particular, we study a dynamic THz/VLC-enabled VR network that can accurately locate VR users in real time using VLC and build THz links to transmit high-quality VR images based on the users' positions. In the studied network, only a subset of the VLC access points (VAPs) can be turned on to locate VR users to control the energy consumption of the studied wireless VR network. Based on the obtained user positions, each small base station (SBS) must determine the user association to generate corresponding VR images and build THz links to avoid blockages caused by the user bodies. The problem is formulated as a reliability maximization problem that jointly considers the VAP selection, user association with THz SBSs, and time varying user movement patterns. The reliability of VR networks is defined as the sum successful transmission probabilities of all VR users. To solve this problem, a meta-policy gradient (MPG) algorithm is proposed to find the optimal policy for VAP selection and user association. Compared to traditional reinforcement learning (RL) algorithms [10] trained for fixed environment in which each user has a fixed movement pattern, the proposed algorithm enables the trained policy to quickly adapt to new user movement patterns. Simulation results show that, compared to a baseline trust region policy optimization algorithm (TRPO), the proposed meta-learning solution yields a 78% improvement in the convergence speed and about 16.4% improvement in the sum successful transmission probability. To the best of our knowledge, *this paper is the first to study the joint use of THz and VLC for reliability maximization while considering dynamic VR user movement patterns*.

II. System Model and Problem Formulation

Consider an indoor wireless network that consists of a set $\mathcal{B}$ of B SBSs and a set $\mathcal{V}$ of V VAPs. All the VAPs and SBSs are managed by a central controller. The SBSs are evenly distributed in an indoor area $\mathcal{A}$ to serve a set $\mathcal{U}$ of U VR users over THz frequencies, as shown in Fig. 1. In the studied model, accurate locations of users are required for SBSs to build LoS THz links and generate the VR images requested by users. Each VAP provides accurate indoor positioning and tracking services for VR users using VLC. Here, we consider dual-mode user equipment (UE) that are able to access both THz and VLC bands. In the studied multi-user VR network, at each time slot t, each SBS can only serve one user with a narrow beam while each VAP can locate all the users that are not blocked in its FOV. To control the system energy consumption, the central controller selects a group of VAPs at the beginning of each time slot to locate VR users [11]. Here, not all users can be accurately located due to the user body blockage over the VLC links. Based on the obtained user positions, the central controller determines the SBSs associated with the successfully located users, and then SBSs transmit the corresponding VR images to the users using THz band. In our model, each time period n consists of T time slots. A successful transmission implies that a request of a VR user is successfully completed within a time period.

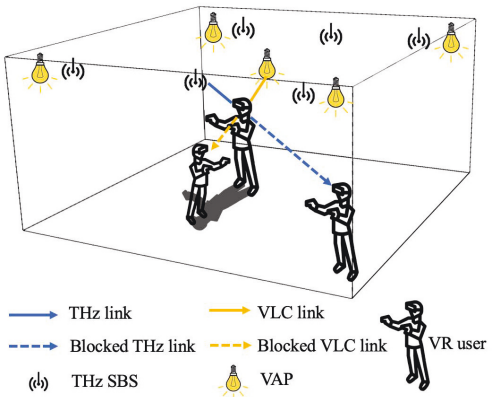

Fig. 1: A THz/VLC-enabled wireless VR network.

A. User Blockage Model

In the studied model, the LoS links (VLC or THz links) between user j and a transmitter (a VAP or an SBS) can be blocked by other VR users' bodies [4]. For a given a user j located at $\boldsymbol{v}_{j,t}^n = (x_{j,t}^n, y_{j,t}^n, z_{j,t}^n)$ at time slot t in time period n and a transmitter k located at (x_k, y_k, Z), we define a binary variable $b_{kj,t}^n$ that represents whether a blockage exist between user j and transmitter k, as follows:

$$b_{kj,t}^n = \begin{cases} 0, & \frac{\boldsymbol{q}_{kj,t}^n}{\|\boldsymbol{q}_{kj,t}^n\|} = \frac{\boldsymbol{q}_{km,t}^n}{\|\boldsymbol{q}_{km,t}^n\|} \text{and} \left\|\boldsymbol{q}_{kj,t}^n\right\| \geqslant \left\|\boldsymbol{q}_{km,t}^n\right\|, \forall m \neq j \in \mathcal{U}, \\ 1, & \text{otherwise}, \end{cases} \tag{1}$$

where $\boldsymbol{q}_{kj,t}^n = (x_k - x_{j,t}^n, y_k - y_{j,t}^n, Z - z_{j,t}^n)$, $\frac{\boldsymbol{q}_{kj,t}^n}{\|\boldsymbol{q}_{kj,t}^n\|} = \frac{\boldsymbol{q}_{km,t}^n}{\|\boldsymbol{q}_{km,t}^n\|}$ indicates that transmitter k, user j, and user m are on a straight line, and $\left\|\boldsymbol{q}_{kj,t}^n\right\| \geqslant \left\|\boldsymbol{q}_{km,t}^n\right\|$ implies that user m is located between transmitter k and user j. Hence, when a user m satisfies both $\frac{\boldsymbol{q}_{kj,t}^n}{\|\boldsymbol{q}_{kj,t}^n\|} = \frac{\boldsymbol{q}_{km,t}^n}{\|\boldsymbol{q}_{km,t}^n\|}$ and $\left\|\boldsymbol{q}_{kj,t}^n\right\| \geqslant \left\|\boldsymbol{q}_{km,t}^n\right\|$, the transmission link between transmitter k and user j will be blocked, as shown in Fig. 1. In (1), $b_{kj,t}^n = 1$ implies that the link between transmitter k and user j is LoS at time slot t in period n; otherwise, we have $b_{kj,t}^n = 0$. Here, we assume that the positions of the VR users remain unchanged during each time slot t.

B. VLC Indoor Positioning

We assume that the three-dimensional (3D) location $\boldsymbol{v}_{j,t}^n = (x_{j,t}^n, y_{j,t}^n, z_{j,t}^n)$ of each user j is determined by three VAPs from three different orientations [9], where $x_{j,t}^n$ and $y_{j,t}^n$ are coordinates of user j in the room and $z_{j,t}^n$ is the height of user j. VAPs are selectively turned on to provide stable and acceptable brightness as well as accurate user positioning services simultaneously. At each time slot t in time period n, a set $\mathcal{L}_t^n = \{l_1, l_2, l_3\}$ of three VAPs is turned on to broadcast their location information to users. When each user j receives the location information of VAP l_k, $k = 1, 2$, and 3, it can calculate the incidence angle $\psi_{l_k j,t}^n$ [12]. Note that, user j can receive the location information sent by VAP l_k at time slot t only when the following conditions are satisfied: a) VAP l_k is in the FOV of user j, as shown in Fig. 1, and b) the VLC link between VAP l_k and user j is LoS (i.e. $b_{l_k j,t}^n = 1$). Then,

the set of VAPs available for providing positioning service for user j can be given as

$$\mathcal{L}_{j,t}^n = \left\{ l_k \,\middle|\, 0 \leqslant \psi_{l_k j,t}^n \leqslant \Psi_{\frac{1}{2}}, b_{l_k j,t}^n = 1, l_k \in \mathcal{L}_t^n \right\}, \quad (2)$$

where $\Psi_{\frac{1}{2}}$ is the receiver FOV semi-angle.

Based on three different incidence angles and the corresponding VAP locations, each user j can calculate its own location $\boldsymbol{v}_{j,t}^n$ at time slot t in period n using a triangulation algorithm [12]. Then, the positioning state of user j at time slot t in period n will be

$$p_{j,t}^n(\mathcal{L}_t^n) = \begin{cases} 1, |\mathcal{L}_{j,t}^n| = 3, \\ 0, |\mathcal{L}_{j,t}^n| < 3, \end{cases} \quad (3)$$

where $|\mathcal{L}_{j,t}^n|$ represents the number of VAPs that can serve user j. Once user j is successfully located at time slot t in period n (i.e. $p_{j,t}^n(\mathcal{L}_t^n) = 1$), the VAPs transmit the position of user j to the central controller over VLC links. Based on the obtained user positions, the central controller can determines the user-SBS association and, then, the SBSs can cooperatively serve the associated users over THz. Here, we ignore the delay of the VAPs transmitting user positions to the central controller due to the small data size.

C. Transmission Model

Due to the extremely narrow pencil beamforming (narrower than mmWave) for THz [2], we assume that each user can only associate with one SBS and each SBS can only serve one user at each time slot. In time period n, let $u_{ij,t}^n \in \{0, 1\}$ be the index of the link between SBS i and user j at time slot t, i.e., $u_{ij,t}^n = 1$ implies that user j is associated with SBS i; otherwise, we have $u_{ij,t}^n = 0$. Then, we have

$$0 \leqslant \sum_{i=1}^{B} u_{ij,t}^n \leqslant 1, \forall j \in \mathcal{U}, \;\; 0 \leqslant \sum_{j=1}^{U} u_{ij,t}^n \leqslant 1, \forall i \in \mathcal{B}. \quad (4)$$

At time slot t in period n, given an SBS $i \in \mathcal{B}$ located at (x_i, y_i, Z) and its associated user $j \in \mathcal{U}$ located at $(x_{j,t}^n, y_{j,t}^n, z_{j,t}^n)$, the path loss of the THz link between SBS i and user j can be given by [4]

$$g_{ij,t}^n = \begin{cases} \left(\frac{c}{4\pi f r_{ij,t}^n}\right)^2 \delta(r_{ij,t}^n), & b_{ij,t}^n = 1, \\ 0, & b_{ij,t}^n = 0, \end{cases} \quad (5)$$

where c is the speed of light, f is the operating frequency, $r_{ij,t}^n = \sqrt{(x_i - x_{j,t}^n)^2 + (y_i - y_{j,t}^n)^2 + (Z - z_{j,t}^n)^2}$ is the distance between SBS i and user j, and $\delta(r_{ij,t}^n) \approx e^{(-K(f) r_{ij,t}^n)}$ represents the transmittance of the medium following the Beer-Lambert law with $K(f)$ being the overall absorption coefficient of the medium at THz frequency f [3]. The total noise power at each UE j that generates by thermal agitation of electrons and molecular absorption can be given by [3]

$$N_{j,t}^n = N_0 + \sum_{l \in \mathcal{B}} P \left(\frac{c}{4\pi f r_{lj,t}^n}\right)^2 \left(1 - \delta(r_{lj,t}^n)\right), \quad (6)$$

where P is the transmit power of each SBS, $N_0 = K_B T_e$ represents the Johnson-Nyquist noise generated by thermal agitation of electrons in conductors with K_B and T_e being Boltzmann constant and the temperature in Kelvin, respectively, and $\sum_{l \in \mathcal{B}} P \left(\frac{c}{4\pi f r_{lj,t}^n}\right)^2 \left(1 - \delta(r_{lj,t}^n)\right)$ is the sum of molecular absorption noise caused by the transmit power of any SBS $l \in \mathcal{B}$. Here, we assume that each user will not be interfered by other SBSs due to the narrow beam. The data rate of VR image transmission from SBS i to its associated user j at time slot t in period n can be given as

$$C_{ij,t}^n(u_{ij,t}^n) = u_{ij,t}^n W \log_2 \left(1 + \frac{P g_{ij,t}^n}{N_{j,t}^n}\right), \quad (7)$$

where W is the bandwidth of the THz band.

Given the data size S of the VR image requested by user j at time slot t in period n, the transmission delay will be

$$d_{j,t}^n(\boldsymbol{u}_{j,t}^n) = \frac{S}{\sum_{i=1}^{B} C_{ij,t}^n(u_{ij,t}^n)}, \quad (8)$$

where $\boldsymbol{u}_{j,t}^n = [u_{1j,t}^n, u_{2j,t}^n, \cdots, u_{Bj,t}^n]$. Note that the data size S of a VR image only depends on the image resolution which remains unchanged during service. Since the user position will change at next time slot, the VR image requested by user j can be successfully transmitted only when the transmission delay is within the time duration Δt of a time slot t. Then, in time period n, the transmission state of user j at time slot t can be given as

$$h_{j,t}^n(\boldsymbol{u}_{j,t}^n) = \begin{cases} 1, \; d_{j,t}^n(\boldsymbol{u}_{j,t}^n) \leqslant \Delta t, \\ 0, \; \text{otherwise}. \end{cases} \quad (9)$$

From (9), we can see that, whether the requested VR image of user j is successfully transmitted at time slot t depends on the user's locations, user association, and blockages between SBS i and user j.

D. Reliability Model

At each time slot t, a successfully served user j must satisfy two conditions: a) user j is successfully positioned and b) the VR image requested by user j is transmitted within Δt. In order to enable a seamless and immersive wireless VR experience, we assume that the waiting delay is limited to a time period that consist of T time slot. This means that each user should be successfully served at least once in a time period. Therefore, in time period n, the service state of user j until time slot t based on the selected $\mathcal{L}_t^n$ and $\boldsymbol{u}_{j,t}^n$ will be

$$w_{j,t}^n(\mathcal{L}_t^n, \boldsymbol{u}_{j,t}^n) = \left(p_{j,t}^n(\mathcal{L}_t^n) h_{j,t}^n(\boldsymbol{u}_{j,t}^n)\right) \vee w_{j,t-1}^n(\mathcal{L}_{t-1}^n, \boldsymbol{u}_{j,t-1}^n), \quad (10)$$

where $t = 2, 3, \cdots, T$ and $\vee$ represents the logical "or" operation. The newly served users at time slot t will be

$$\mathcal{O}_t^n = \{j | w_{j,t}^n(\mathcal{L}_t^n, \boldsymbol{u}_{j,t}^n) = 1, w_{j,t-1}^n(\mathcal{L}_{t-1}^n, \boldsymbol{u}_{j,t-1}^n) = 0\}, \quad (11)$$

Then, the number of successful served users in each time period n can be given by

$$R^n(\mathcal{L}_{:T}^n, \boldsymbol{u}_{j,:T}^n) = \sum_{t=1}^{T} |\mathcal{O}_t^n|, \quad (12)$$

where $\mathcal{L}_{:T}^n = \{\mathcal{L}_1^n, \cdots, \mathcal{L}_T^n\}$ and $\boldsymbol{u}_{j,:T}^n = \{\boldsymbol{u}_{j,1}^n, \cdots, \boldsymbol{u}_{j,T}^n\}$.

E. Problem Formulation

Given the defined system model, our goal is to effectively select the subset of optimal VAPs to provide accurate positioning services and, then, determine the user-SBS association based on the obtained user positions so as to maximize the reliability of the studied VR network. Then, the reliability maximization problem is formulated as follows:

$$\max_{\mathcal{L}_t^n, \boldsymbol{u}_{j,t}^n} \quad \sum_{n=1}^{N} \frac{R^n(\mathcal{L}_{:T}^n, \boldsymbol{u}_{j,:T}^n)}{N}, \tag{13}$$

$$\text{s.t.} \quad |\mathcal{L}_t^n| = 3, \tag{13a}$$

$$0 \leqslant \sum_{i=1}^{B} u_{ij,t}^n \leqslant 1, \forall j \in \mathcal{U}, \tag{13b}$$

$$0 \leqslant \sum_{j=1}^{U} u_{ij,t}^n \leqslant 1, \forall i \in \mathcal{B}, \tag{13c}$$

$$u_{ij,t}^n \in \{0,1\}, \forall i \in \mathcal{B}, \forall j \in \mathcal{U}, \tag{13d}$$

where (13a) captures the fact that only three VAPs are selected at each time slot to provide positioning service. (13b), (13c), and (13d) indicate that each user can only associate with one SBS and each SBS can only serve one user at each time slot. From (13), we can see that the reliability depends on the selected VAPs and the user association with SBSs. Meanwhile, the VAP selection and the user association depend on the positions of VR users. However, the users' positions continuously change as time elapses. Therefore, real-time user positions are needed by the SBSs so as to generate corresponding VR images and build THz links without blockages. Moreover, due to the time-varying nature of VR applications, the user movement pattern varies over different time periods. Here, we define a position transition matrix $\boldsymbol{M}^n$ as the user movement pattern durning time period n, in which each element $\boldsymbol{M}^n_{\boldsymbol{v}_{j,t}^n, \boldsymbol{v}_{j,t+1}^n} = P(\boldsymbol{v}_{j,t+1}^n | \boldsymbol{v}_{j,t}^n)$ is the probability of the user moving from $\boldsymbol{v}_{j,t}^n$ to $\boldsymbol{v}_{j,t+1}^n$. Note that the studied THz/VLC-enabled VR network has no knowledge of user movement patterns and hence, the central controller must find and adapt to the time-varying movement pattern so as to control the VAPs and SBSs in advance. Hence, it is necessary to introduce a learning algorithm to sensitively adapt to new user movement patterns so as to proactively determine the VAP selection and the user association [13].

III. Meta-Learning for VAP Selection and User Association

Next, we introduce a policy gradient-based RL algorithm using meta-learning framework [14], called meta policy gradient (MPG), that can effectively solve problem (13). Traditional policy gradient algorithms can only determine the VAP selection and user association in a fixed environment (i.e., the fixed user movement patterns). Meta-learning is a novel learning approach that can integrate the prior reliability-enhancing experience with information collected from the new user movement patterns, thus training a fast-adaptive model. Therefore, the proposed MPG can obtain the VAP selection and user association policies that can be quickly updated to adapt to new user movement patterns using only a few further training steps. Next, we first introduce the components of the MPG algorithm. Then, we explain the entire procedure of using our MPG algorithm to select VAPs and determine the user association with SBSs.

A. Components of MPG Algorithm

An MPG algorithm consists of six components: a) agent, b) actions, c) states, d) policy, e) reward, and f) tasks, which are specified as follows

- *Agent*: Our agent is a central controller that can obtain the user positions and simultaneously control the VAPs and the SBSs.
- *Actions*: The action of the agent at each time slot t in period n is a vector $\boldsymbol{a}_t^n = [\mathcal{L}_t^n, \boldsymbol{u}_{1,t}^n, \boldsymbol{u}_{2,t}^n, \cdots, \boldsymbol{u}_{U,t}^n]$ that jointly considers the VAP selection and the user association. The action space $\mathcal{A}$ is the set of all optional actions.
- *States*: The state at time slot t in time period n is defined as $\boldsymbol{s}_t^n = [\boldsymbol{v}_t^n, \boldsymbol{w}_t^n, \mathcal{O}_t^n]$ that consists of: 1) the user position $\boldsymbol{v}_t^n = [\boldsymbol{v}_{1,t}^n, \cdots, \boldsymbol{v}_{U,t}^n]$, where $\boldsymbol{v}_{j,t}^n$ depends on $\boldsymbol{v}_{j,t-1}^n$ and the movement pattern $\boldsymbol{M}^n$ in time period n, which is unknown to the central controller; 2) the service state vector $\boldsymbol{w}_t^n = [w_{1,t}^n, \cdots, w_{U,t}^n]$ that implies each user whether has been successfully served until time slot t; 3) the set of newly served users $\mathcal{O}_t^n$ at time slot t. The state space $\mathcal{S}$ is the set of all optional states.
- *Policy*: The policy is the probability of the agent choosing each action at a given state. The MPG algorithm uses a deep neural network parameterized by $\boldsymbol{\theta}$ to map the input state to the output policy. Then, the policy can be expressed as $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{s}_{t-1}^n, \boldsymbol{a}_t^n) = P(\boldsymbol{a}_t^n | \boldsymbol{s}_{t-1}^n)$. Based on the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, an execution process in a time period n can be defined as a trajectory $\boldsymbol{\tau}^n = \{\boldsymbol{s}_0^n, \boldsymbol{a}_1^n, \cdots, \boldsymbol{s}_{T-1}^n, \boldsymbol{a}_T^n\}$.
- *Reward*: The benefit of choosing action $\boldsymbol{a}_t^n$ at state $\boldsymbol{s}_{t-1}^n$ is $r(\boldsymbol{s}_{t-1}^n, \boldsymbol{a}_t^n) = |\mathcal{O}_t^n(\boldsymbol{a}_t^n)|$. Therefore, the reward of a trajectory during a time period n will be $R(\boldsymbol{\tau}^n) = \sum_{t=1}^{T} r(\boldsymbol{s}_{t-1}^n, \boldsymbol{a}_t^n)$. The expected reward function of the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ durning a time period n can be given as
$$\bar{J}^n(\boldsymbol{\theta}) = \sum_{\boldsymbol{\tau}^n \in \mathcal{D}^n} R(\boldsymbol{\tau}^n) P_{\boldsymbol{\theta}}(\boldsymbol{\tau}^n), \tag{14}$$
where $\mathcal{D}^n$ is the set of sampled trajectories in time period n and $P_{\boldsymbol{\theta}}(\boldsymbol{\tau}^n) = P(\boldsymbol{s}_0^n) \prod_{t=1}^{T} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{s}_{t-1}^n, \boldsymbol{a}_t^n) P(\boldsymbol{s}_t^n | \boldsymbol{s}_{t-1}^n, \boldsymbol{a}_t^n)$ with $P(\boldsymbol{s}_t^n | \boldsymbol{s}_{t-1}^n, \boldsymbol{a}_t^n)$ depending on the movement patter $\boldsymbol{M}^n$. The total reward function of all time periods $\bar{R}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \bar{J}^n(\boldsymbol{\theta})$ is the objective function that the agent aims to optimize.
- *Tasks* : We use *task* $\mathcal{T}^n$ to refer to the reliability maximization problem $\max_{\mathcal{L}_t^n, \boldsymbol{u}_{j,t}^n} R^n(\mathcal{L}_{:T}^n, \boldsymbol{u}_{j,:T}^n)$ in each time period n. A task is thus defined as $\mathcal{T}^n = \{\boldsymbol{M}^n, \mathcal{D}^n, \bar{J}^n(\boldsymbol{\theta})\}$. Here, the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ is shared by all tasks. However, the expected reward $\bar{J}^n(\boldsymbol{\theta})$ for each task is affected by the user movement pattern $\boldsymbol{M}^n$ that is

Algorithm 1 MPG algorithm for VAP selection and user association.

Input: The set of VAPs $\mathcal{V}$, the set of SBSs $\mathcal{B}$, the user positions $\boldsymbol{v}_0^n$, and the transition matrix $\boldsymbol{M}^n$.
Initialize: Parameters $\boldsymbol{\theta}$ is initially generated randomly, $\Omega_0^n = 0$, $\boldsymbol{w}_0^n = [0, \cdots, 0]$, task learning rate α, meta-learning rate β, and the number of iterations E.
for $i = 1 \rightarrow E$ **do**
  for all each task $\mathcal{T}^n$ **do**
    Collect K trajectories $\mathcal{D}^n = \{\boldsymbol{\tau}_1^n, \cdots, \boldsymbol{\tau}_k^n, \cdots, \boldsymbol{\tau}_K^n\}$ using $\boldsymbol{\pi}_{\boldsymbol{\theta}}$.
    Compute $\nabla_{\boldsymbol{\theta}} \bar{J}^n(\boldsymbol{\theta})$ using $\mathcal{D}^n$ based on (15).
    Compute parameters $\tilde{\boldsymbol{\theta}}^n$ of the adapted policy based on (16).
    Collect K' trajectories $\mathcal{D}'^n = \{\boldsymbol{\tau}'^n_1, \cdots, \boldsymbol{\tau}'^n_{K'}\}$ using $\boldsymbol{\pi}_{\tilde{\boldsymbol{\theta}}^n}$.
  end for
  Compute total reward $R(\boldsymbol{\theta})$ using each $\mathcal{D}'^n$ based on (17).
  Update the parameters of the policy based on (18).
end for

unknown to the agent. Therefore, the agent must find the optimal policy that can quickly adapt to each task.

B. MPG for Optimization of Reliability

Next, we introduce the entire procedure of training the proposed MPG algorithm. Our purpose from training MPG is to find the optimal policy that maximizes the reliability of the THz/VLC-enabled wireless VR network over different time periods. The MPG algorithm enables the trained policy to quickly adapt to the time-varying user movements. The intuition behind the proposed MPG is that some parameters of MPG are task-sensitive while some parameters are broadly applicable to all tasks. Therefore, the training process of MPG has two steps: 1) task learning step and 2) meta-learning step. The task learning step enables the MPG to execute the policy gradient on task-sensitive parameters so as to make rapid progress on each new task. The meta-learning step aims to find the broadly applicable parameters that can improve the performance of all tasks. Specifically, the two steps can be given as follows:

1) *Task learning step*: For each task $\mathcal{T}^n$, the agent first collects K trajectories given a policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$. The set of collected trajectories is $\mathcal{D}^n = \{\boldsymbol{\tau}_1^n, \cdots, \boldsymbol{\tau}_k^n, \cdots, \boldsymbol{\tau}_K^n\}$ and the expected reward is $\bar{J}^n(\boldsymbol{\theta})$. The policy gradient for each task $\mathcal{T}^n$ based on (14) is

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \bar{J}^n(\boldsymbol{\theta}) =& \sum_{k=1}^{K} R(\boldsymbol{\tau}_k^n) P_{\boldsymbol{\theta}}(\boldsymbol{\tau}_k^n) \nabla \log P_{\boldsymbol{\theta}}(\boldsymbol{\tau}_k^n), \\ \approx& \frac{1}{K} \sum_{k=1}^{K} R(\boldsymbol{\tau}_k^n) \nabla \log P_{\boldsymbol{\theta}}(\boldsymbol{\tau}_k^n), \\ =& \frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{T} R(\boldsymbol{\tau}_k^n) \nabla \log \boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{s}_{t-1}^n, \boldsymbol{a}_t^n).\end{aligned} \tag{15}$$

To improve the expected reward of each task $\mathcal{T}^n$, the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ is updated using standard gradient ascent method

$$\tilde{\boldsymbol{\theta}}^n = \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} \bar{J}^n(\boldsymbol{\theta}), \tag{16}$$

where α is the learning rate which is assumed to be equal for all tasks. Finally, the agent collects K' trajectories for each task $\mathcal{T}^n$ using the corresponding updated policy $\boldsymbol{\pi}_{\tilde{\boldsymbol{\theta}}^n}$. Each trajectory set $\mathcal{D}'^n = \{\boldsymbol{\tau}'^n_1, \cdots, \boldsymbol{\tau}'^n_{K'}\}$ is used to optimize the broadly applicable parameters in the next meta-learning step to improve the all task performance.

2) *Meta-learning step*: For each trajectory set $\mathcal{D}'^n$, the agent computes the expected rewards $\bar{J}^n(\tilde{\boldsymbol{\theta}}^n)$. The total reward of all tasks is

$$R(\boldsymbol{\theta}) = \sum_{n=1}^{N} \bar{J}^n(\tilde{\boldsymbol{\theta}}^n) = \sum_{n=1}^{N} \bar{J}^n(\boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} \bar{J}^n(\boldsymbol{\theta})). \tag{17}$$

Then, to improve the total reward of all tasks, the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ will be updated by

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta \nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}), \tag{18}$$

where β is the learning rate for meta-learning. Here, note that the meta-learning step is performed over the parameters $\boldsymbol{\theta}$ of the policy instead of the parameters $\tilde{\boldsymbol{\theta}}^n$ updated in the previous task learning step.

By iteratively running the task learning and the meta-learning step, the optimal policy for determining the VAP selection and user association under different user movement patterns can be obtained [15]. The specific training process of the proposed MPG algorithm is summarized in **Algorithm 1**.

In this algorithm, the objective function is the total reward of all tasks that is consistent with the objective function of the optimization problem (13). Since the meta-learning step tends to optimize the broadly applicable parameters for all tasks, the proposed MPG algorithm enables the trained policy to quickly adapt to new tasks. This means that, for a new task with new user movement pattern, using the trained policy as initialization, the agent can quickly find the optimal policy by only executing the task learning step with a few trajectories.

TABLE I: System parameters

Parameters	Value	Parameters	Value
c	3×10^8m/s	f	1THz
P	1W	S	20Mbit
$K_B T_e$	-174dBm/Hz	T	15
K	50	K'	10
α	0.1	β	0.01

IV. Simulation Results and Analysis

For our simulations, a 5 m $\times$ 5 m square room is considered with $D = 5$ VAP and $B = 6$ SBSs evenly distributed at a fixed hight of $Z = 3$ m. A number of $U = 8$ wireless VR users are initially randomly distributed in the room and move according to a randomly generated user movement pattern in each time period. For comparison purposes, we consider the trust region policy optimization algorithm (TRPO) in [16] as the baseline scheme. All statistical results are averaged over a large number of independent runs. Other parameters are listed in Table I.

Fig. 2 shows the sum successful transmission probability of all users over training process of the proposed MPG algorithm. MPG model is trained for $N = 20$ tasks and $N = 50$ tasks to obtain fast-adaptive policy for VAP selection and user association, respectively. In Fig. 2, we can see that the proposed MPG algorithm can effectively converge to the optimal policy that maximizes the network reliability. This is due to the fact that the proposed MPG algorithm running gradient descent over the policy space toward the maximal reliability. Fig. 2 also shows that the training process of

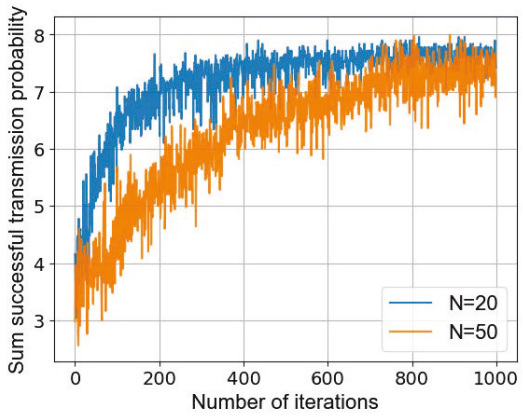

Fig. 2: The reliability of the THz/VLC VR network over training process of the proposed MPG algorithm.

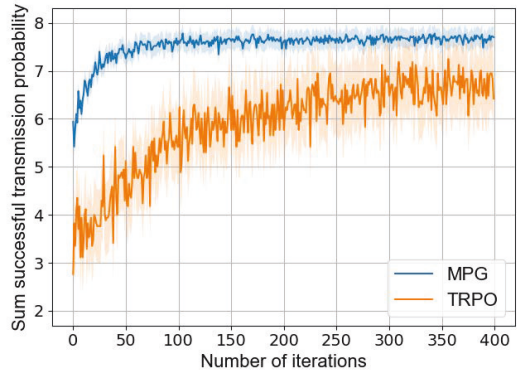

Fig. 3: Test adaptability on new tasks.

$N = 20$ tasks is more stable and converges faster than the training process of $N = 50$ tasks. This is because fewer tasks are more likely to find update gradients that work for most of the tasks in the meta-learning step.

Fig. 3 shows the adaptability of the proposed MPG algorithm testing for new tasks. In Fig. 3, the line and shadow are the mean and standard deviation computed over 5 random generated new tasks. From Fig. 3 we can see that, compared with random initialization for TRPO, our proposed algorithm that uses the policy trained on the old tasks as the initialization can achieve better performance at the beginning of the test process. Fig. 3 also shows that the proposed algorithm requires approximately 70 iterations to reach convergence for new tasks, which is 78% less than the traditional TRPO algorithm that requires about 320 iterations to reach convergence. This implies that the policy trained by the proposed MPG algorithm can record the knowledge that can be used in new tasks, thus quickly adapting to new tasks. In Fig. 3, we can also see that the proposed algorithm can yield up to 16.4% improvement in terms of the reliability compared with the TRPO algorithm. This is because the alternate iteration of the task learning step and the meta-learning step can find the broadly applicable parameters that can improve the performance of all tasks.

V. Conclusion

In this paper, we have developed a novel framework for maximizing reliability of THz/VLC-enabled wireless VR networks. To this end, we have formulated an optimization problem that jointly considers the user mobility, blockages of both THz and VLC links, VAP selection, and user association. To solve this problem, we have developed a novel MPG algorithm based on meta-learning framework, which can effectively find the policy of VAP selection and user association for maximizing reliability. The proposed MPG algorithm enables the trained policy to quickly adapt to new user movement patterns. Simulation results have shown that, compared with the traditional RL algorithm, the proposed algorithm can achieve better performance than traditional RL algorithms.

References

[1] F. Hu, Y. Deng, W. Saad, M. Bennis, and A. H. Aghvami, "Cellular-connected wireless virtual reality: Requirements, challenges, and solutions," *IEEE Communications Magazine*, vol. 58, no. 5, pp. 105–111, 2020.

[2] H. Zhang, H. Zhang, W. Liu, K. Long, J. Dong, and V. C. M. Leung, "Energy efficient user clustering, hybrid precoding and power optimization in terahertz MIMO-NOMA systems," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 9, pp. 2074–2085, Sep. 2020.

[3] V. Petrov, D. Moltchanov, and Y. Koucheryavy, "Interference and SINR in dense terahertz networks," in *Proc. IEEE Vehicular Technology Conference (VTC-Fall)*, Boston, MA, Sep. 2015, pp. 1–5.

[4] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low latency communications for wireless VR?," 2020, Available: https://arxiv.org/abs/2005.00536.

[5] T. Dang and M. Peng, "Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 7, pp. 1594–1607, July 2019.

[6] Y. Sun, Z. Chen, M. Tao, and H. Liu, "Communications, caching, and computing for mobile virtual reality: Modeling and tradeoff," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7573–7586, Nov. 2019.

[7] C. Perfecto, M. S. Elbamby, J. D. Ser, and M. Bennis, "Taming the latency in multi-user VR 360°: A QoE-aware deep learning-aided multicast framework," *IEEE Transactions on Communications*, vol. 68, no. 4, pp. 2491–2508, April 2020.

[8] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[9] Y. Zhuang, L. Hua, L. Qi, J. Yang, P. Cao, Y. Cao, Y. Wu, J. Thompson, and H. Haas, "A survey of positioning systems using visible LED lights," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1963–1988, Feb. 2018.

[10] S. Wang, M. Chen, X. Liu, C. Yin, S. Cui, and H. Vincent Poor, "A machine learning approach for task and resource allocation in mobile-edge computing-based networks," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1358–1372, Feb. 2021.

[11] Y. Wang, M. Chen, Z. Yang, T. Luo, and W. Saad, "Deep learning for optimal deployment of UAVs with visible light communications," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7049–7063, Nov. 2020.

[12] M. Yasir, S. Ho, and B. N. Vellambi, "Indoor positioning system using visible light and accelerometer," *Journal of Lightwave Technology*, vol. 32, no. 19, pp. 3306–3316, July 2014.

[13] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3039–3071, Fourthquarter 2019.

[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017, Available: https://arxiv.org/abs/1703.03400.

[15] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," in *submitted to IEEE Journal on Selected Areas in Communications*, Oct. 2020.

[16] J. Schulman, S. Levine, P.Moritz, M. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proc. International Conference on Machine Learning (ICML)*, Lille, UK, July 2015, pp. 1889–1897.