Multi-Agent Reinforcement Learning Trajectory Design and Two-Stage Resource Management in CoMP UAV VLC Networks

Mohammad Reza Maleki, Mohammad Robat Mili, Mohammad Reza Javan, *Senior Member, IEEE*, Nader Mokari, *Senior Member, IEEE*, and Eduard A. Jorswieck, *Fellow, IEEE*

***Abstract*— In this paper, we consider unmanned aerial vehicles (UAVs) equipped with a visible light communication (VLC) access point and coordinated multipoint (CoMP) capability that allows users to connect to more than one UAV. UAVs can move in 3-dimensional (3D) at a constant acceleration, where a central server is responsible for synchronization and cooperation among UAVs. The effect of accelerated movement in UAV is necessary to be considered. Unlike most existing works, we examine the effects of variable speed on kinetics and radio resource allocations. For the proposed system model, we define two different time scales. In the frame, the acceleration of each UAV is specified, and in each slot, radio resources are allocated. Our goal is to formulate a multi-objective optimization problem where the total data rate is maximized, and the total communication power consumption is minimized simultaneously. To handle this multi-objective optimization, we first apply the scalarization method and then apply multi-agent deep deterministic policy gradient (MADDPG). We improve this solution method by adding two critic networks together with two-stage resource allocation.**

***Index Terms*— Visible light communication, UAV, CoMP, two-time scale, reinforcement learning, DDPG, MADDPG, resource allocation, 3D movements, constant acceleration.**

I. Introduction

A. State of the Art

As a cooperative communication system based on multiple transmission and reception points, coordinated multipoint (CoMP) is consolidated into the long-term evolution-advanced releases [1] as an adequate method for relieving inter-cell interference. It also enables symbol-level cooperation among unmanned aerial vehicles (UAVs) and base stations (BS) to enhance communication quality. CoMP technique significantly improves data-rate, and connection availability for cell center and edge users [2]. Despite the fact that CoMP can mitigate the effects of severe inter-cell interference (ICI), it is considered a key enabling technology for beyond fifth-generation (B5G) and sixth-generation (6G) networks. In order to improve network coverage for the next-generation mobile phone networks, CoMP is used to increase the received signal-to-interference-plus-noise ratio (SINR) [3]. UAVs are predicted to play an essential role B5G and 6G cellular networks [4]. On the one hand, for improving the communication and service range and enhancing the quality of service (QoS), UAVs with specific purposes such as aerial maneuvers can be linked directly with cellular BSs [5]. On the other hand, we can utilize UAVs as aerial wireless BSs in the sky to implement flexible and on-demand wireless services to mobile users, promoting communication performance and improved coverage [6]. Several technical opportunities and challenges are created with the advent of cellular-connected UAVs and wireless transmissions aided by UAVs. As a first consideration, UAVs usually have a strong line-of-sight (LoS) to users. As a result, channel gain and communication quality are improved, but inter-cell interference increases correspondingly. Second, UAVs offer high mobility in the 3-dimensional (3D) environment. Trajectory management becomes more complex due to 3D movement, but it provides more opportunity for UAV positioning and trajectory control, which can improve communication performance [7]. Visible light communication (VLC) is an evolving communication technology with low energy consumption and flexible coverage [8]. The VLC network can support a large number of services due to the available bandwidth in unlicensed spectrum, its ubiquitous presence, and low power consumption. Using light-emitting diodes (LEDs) in VLC, the technology offers illumination and communication in scenarios such as search and rescue. It will play an essential role in future generations [9], [10]. With the emergence of new technologies and applications, machine learning becomes more prevalent in B5G wireless applications in [11].

B. Related Works

A mobile UAV is considered, in which the UAV has a mission to fly from an initial location to a final location during the period of communications. As part of this scheme, the

Manuscript received 6 December 2021; revised 16 April 2022, 1 August 2022, and 11 September 2022; accepted 20 September 2022. Date of publication 4 October 2022; date of current version 18 November 2022. This work was supported by the Federal Ministry of Education and Research of Germany in the program of "Souverän. Digital. Vernetzt." Joint Project 6G-RIC, project identification number: 16KISK020K and 16KISK031. The associate editor coordinating the review of this article and approving it for publication was B. Makki. *(Corresponding author: Nader Mokari.)*

Mohammad Reza Maleki and Nader Mokari are with the Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran 14117-13116, Iran (e-mail: m.mohammadreza@modares.ac.ir; nader.mokari@modares.ac.ir).

Mohammad Robat Mili is with the Pasargad Institute for Advanced Innovative Solutions (PIAIS), Tehran 1991633361, Iran (e-mail: mohammad.robatmili@gmail.com).

Mohammad Reza Javan is with the Department of Electrical Engineering, Shahrood University of Technology, Shahrood 3619995161, Iran (e-mail: javan@shahroodut.ac.ir).

Eduard A. Jorswieck is with the Institute of Communications Technology, TU Braunschweig, 38106 Braunschweig, Germany (e-mail: jorswieck@ifn.ing.tu-bs.de).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2022.3211083.

Digital Object Identifier 10.1109/TCOMM.2022.3211083

authors optimize the UAV's 3D trajectory and power distribution to maximize the average secrecy rate during the whole communication period. In a MEC system with multiple UAVs, the authors address the sum power minimization problem by optimizing resource allocation, user association, and power control jointly. The authors suggest a centralized multi-agent reinforcement learning (MARL) algorithm for solving the non-convex problem. In contrast, the centralized method ignores a number of critical concerns, such as the distributed topology and privacy concerns. Federated reinforcement learning (FRL) algorithm is then utilized in a semi-distributed framework to handle mentioned issues [12]. Using a central ground controller to coordinate UAVs to serve multiple ground users, a UAV-assisted cellular network is employed in a downlink scenario [13]. The authors of [14] survey the top issues facing UAV-based wireless communication networks, and they investigate UAV networks with flying drones, the energy efficiency of UAVs, and seamless handover between UAVs and the ground BSs. The authors of [2] study performance optimization of UAV placement and movement in multi-UAV CoMP communication, where each UAV forwards its received signals to a central processor for decoding. Through using a trajectory design to exploit the high mobility of the UAV access point, it is possible to enhance the data rate significantly. However this may result in very long delays for users [15], making it problematic for delay-sensitive ultra-reliable low latency communications (URLLC) applications. It is demonstrated in [16] that maximizing sum rates for heterogeneous networks can lead to a remarkable enhancement of spectral efficiency for joint transmission CoMP-NOMA for a wide range of access distances. In [11], the problem of dynamical deployment of UAVs equipped with VLC capabilities to optimize the energy efficiency of UAV-VLC networks is studied. The optimization problem in [11] seeks to minimize the transmit power while guaranteeing the illumination and communication requirements of each user. The authors in [17] studied the problem of the trajectory design for a team of drone base stations (DBS) operating in dynamic wireless network environments where the DBSs cooperatively fly around the considered environment to provide on demand uplink communication service to ground users. In considered scenario, the DBSs navigate to maximize coverage of the dynamic requests of the ground users. In [18], the UAV coverage problem is addressed in order to either maximize coverage region or enhance the QoS. The authors of [19] examine the strategies for incorporating a UAV into a two-cell NOMA COMP system so that the BS can be sustained. A novel VLC/UAV framework is designed in [20] to both communicate and illuminate while also optimizing the locations of UAVs to minimize the total power consumption. In [21], the authors optimize the UAVs' trajectory and wireless resource allocation to maximize the uplink common (minimum) throughput of the two IoT devices in both interference coordination and CoMP scenarios. First, they consider the special case where the duration of the UAV missions is long enough to solve the common rate maximization problem. Following this, they discuss the general case with a finite mission duration. In order to enhance communication coverage, hybrid VLC/RF systems emerge, [22], so that mobile users can achieve higher rates of data transmission via integrated VLC/RF. A multi-agent reinforcement learning method is used to improve the QoS for the users in [23]. An RF/VLC aggregated system is discussed by the authors of [24] in order to maximize energy efficiency. However, they do not assume the user mobility, which is challenging in RF/VLC hybrid networks. By using NOMA-based hybrid VLC-RF with common backhaul, [25] addresses the problem of optimal resource allocation to maximize achievable data rate. An iterative algorithm is presented to train users on access networks of a hybrid RF/VLC system in [26]. To maximize the total achievable throughput, they formulate an optimization problem to assign power to RF APs and VLC APs. The symbol error rate of the code domain NOMA-based VLC system is investigated in [27], revealing that users exhibit identical error rate performance across locations, while recent works demonstrate that power domain-NOMA is an effective multiple access scheme for VLC systems.

DRL and its multi-agent extension, multi-agent DRL, are discussed in the literature cited above. Markov decision processes (MDPs) model the policy search as a Markov process [13], [28], [29], [30], [31], such that each agent updates its policy independently. Even though these algorithms can deal with many complex problems, they cannot be applied to multi-agent systems (MASs). Due to the fact that all agents act simultaneously, the MAS results in a non-stationary environment. As opposed to this, [32], [33], [34], and [12] are based on multi-agent discrete DRL. Discrete action spaces are not appropriate for power control scenarios, resulting in substandard results. Nevertheless, optimization theory is also used in aforementioned papers to obtain suboptimal and optimal solutions, e.g., trajectory design and resource allocation. As a consequence, solving such optimization problems usually involves a substantial amount of computational time and resources; also, they are not able to work properly in a dynamic environment. Recent efforts to address this problem have drawn much attention to DRL. Above papers, however, try to implement RF/VLC networks to assist each other in improving coverage and reliability. However, the implementation and synchronization of these various technologies still remain a challenging. The development of UAV-based networks also addresses poor channel conditions and enhances coverage in VLC networks, which is a current problem. Allowing users to assign to more than one AP improves coverage availability, particularly at the cell edge, where users suffer from a poor channel even in the LoS channel. In terms of resource allocation, considering the movement of UAVs is challenging. First, UAV movement with constant speed is practically impossible. In addition, it remarkably deteriorates maneuverability, however, flying at a constant acceleration with increasing maneuverability provides better opportunities for allocating resources and tracking the users. Likewise, 3D movement improves UAV performance by increasing maneuverability. We also provide a minimum data rate for each user to assist users with weaker channels usually located at the edge of the cell to enhance QoS. In [35] the authors proved that using NOMA in VLC systems results in better performance.

TABLE I

NOTATIONS AND SYMBOLS

Notation	Description	Notation	Description	Notation	Description
$m/M/\mathcal{M}$	Index/number/set of users	$\mathbf{v^m}$	Velocity of each user $[\mathrm{m/s} \times \mathrm{m/s} \times \mathrm{m/s}\ ]$	$t/T/\mathcal{T}$	Index/number/set of time frames
$n/N/\mathcal{N}$	Index/number/set time slots	$\mathbf{a}^f$	Acceleration of the f_{th} UAV $[\mathrm{m/s^2} \times \mathrm{m/s^2} \times \mathrm{m/s^2}]$	$\mathbf{q}^f$	Location of the f_{th} UAV $[m\times m\times m]$
$f/F/\mathcal{F}$	Index/number/set of UAVs	$\mathbf{v}^f$	Velocity of the f_{th} UAV $[\mathrm{m/s} \times \mathrm{m/s} \times \mathrm{m/s}\ ]$	$\mathbf{w}^m$	Location of each user $[\mathrm{m}\times\mathrm{m}\times\mathrm{m}\ ]$
$T_s(\cdot)$	Gain of the optical filter	$d^{m,f}$	Distance between the f_{th} UAV and the m_{th} user [m]	$g(\cdot)$	Optical concentrator gain at PD
ϕ	Angle of irradiance	ψ	Incidence angle between LED and device	ϖ	Order of Lambertian emission
A_r	Active area of PD i			ψ_c	Semi-angle field of view (FOV) of PD

C. Contribution

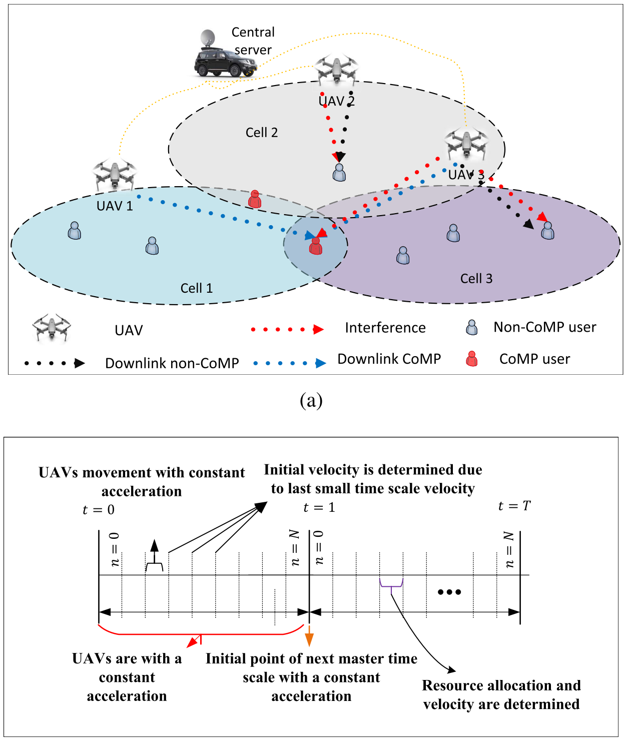
In this paper, we consider the downlink scenario for UAVs that are equipped with VLC APs. Studies show that 3D movement can improve UAV performance in terms of the allocation of radio resources [36]. However, we improve UAV movement by designing a two-time scale system applying constant acceleration movement. To enhance the resource allocation framework, we utilize constant acceleration movement to cover the weakness of low maneuverability. It also assists UAVs to hastily reach better locations in terms of allocating resources. Increasing in maneuverability manifests itself to have a higher system complexity. Obviously, solving complex problems requires a better and more robust learning method. We address this challenge using two-stage allocation method, which are entirely compatible with the proposed system model. In the frames, the constant acceleration is determined, whereas in the slot, radio resources are allocated, and the initial velocities are calculated. Our contributions can be summarized as follows:

- By implementing the constant acceleration movement, we enrich maneuverability, permit UAVs to move more flexibly, additionally, the optimization problem is combined with a movement model, consequently improving resource allocation performance. Although, implementing this model increases the complexity of the problem, by coupling with the proposed RL approach, we handle this complexity.
- We establish a two-time scale structure to accurately assess the trajectory of UAVs and deal with the resource allocation problem effectively. For each UAV, the constant acceleration is specified within a frame, while radio resources are determined within slots.
- In order to achieve maximum data rates and minimal power consumption, both objectives are optimized simultaneously. To calculate the movements of UAVs and users within our problem, we formulate a two-time scale model. In order to enhance coverage of VLC network, we employ CoMP. Consequently, with CoMP, cell edge users can be assured that their QoS is achieved.
- A two-time scale structure makes the problem more challenging when considering UAV trajectory and user movement. We present a novel solution method that adapts to UAV operations in resources allocation. The proposed RL-based solution is multi-agent-based and relies on a central server to manage and boost cooperation among agents as well as having a local critic that estimates the local expected reward. The global critic should evaluate the expected global reward and foster cooperation between several agents. Our approach is fundamentally different from other RL frameworks since we utilize two reward functions instead of one: a global reward function that takes its elements from the cumulative level of interference between UAVs, and a local reward function that takes inputs from the objective parts of the optimization problem we are solving. The TD3 algorithm [37] for the global critic is exploited in order to reduce the overestimation bias in Q-functions. Comparing our proposed method with existing methods, it has been shown that this method increases convergence time and effectively solves the problem.
- In the simulation section, we provide a comprehensive evaluation of the system model. The located baselines confirm the superior performance of our solution. The impacts of different terms in the objective function are inspected. The influence of various constraints on the objective function is also studied. Additionally, we examine the system with and without CoMP, which shows the positive impact of employing CoMP in our system model. Also, the constant acceleration and constant velocity are examined, showing that constant acceleration gives better performance than that of the constant velocity.

The rest of this paper is organized as follows: Section II describes the system model for our considered UAV-VLC-enabled CoMP. Section III formulates UAV resource allocation and movement optimization problems to maximize data-rate and the minimum total power consumption. We propose our reinforcement learning (RL) approach in Section IV. Section V includes simulation results and at the end, conclusion is in Section VI.

II. SYSTEM MODEL

In this paper, we consider some UAVs mounted with VLC AP, where the UAVs hover over ground with CoMP system while applying PD-NOMA to serve both communication and illumination simultaneously. Users are randomly distributed on ground. As shown in Fig. 1a, we consider the downlink transmission scenario where single-antenna UAVs are deployed as aerial BSs. For ease of exposition, the time horizon, T, is equally divided into N time slots with slot duration δ as shown in Fig. 1b where $\mathcal{T} = \{1, \ldots, T\}$, $\mathcal{N} = \{1, \ldots, N\}$ are the frame set and the slot set, respectively. We denote the set of UAVs by $\mathcal{F} = \{1, \ldots, F\}$, $F \in \mathbb{N}$, the set of users is indicated by $\mathcal{M} = \{1, \ldots, M\}$, $M \in \mathbb{N}$. We consider a 3D Cartesian coordinate system, with location, velocity, and acceleration are measured in m, m/s, and m/s, respectively, where the horizontal coordinate and velocity of user m are denoted by $\mathbf{w}^m[t,n] = (x^m[t,n], y^m[t,n], 0)^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$ and $\{\mathbf{v}^m[t,n]\}$, respectively. We assume that each UAV flies with the maximum speed constraint $v_{\max}$ and the maximum acceleration constraint $a_{\max}$. As such, the UAV trajectory,


Central server
UAV 2
Cell 2
UAV 1
UAV 3
Cell 1
Cell 3
UAV
Interference
Non-CoMP user
Downlink non-CoMP
Downlink CoMP
CoMP user
UAVs movement with constant acceleration
Initial velocity is determined due to last small time scale velocity
t = 0
t = 1
t = T
n = 0
n = N
n = 0
n = N
•••
UAVs are with a constant acceleration
Initial point of next master time scale with a constant acceleration
Resource allocation and velocity are determined

(a)

(b)

Fig. 1. (a) System Modal UAV VLC CoMP, (b) the two-time scale uses constant acceleration, allocates resources, and calculates the initial velocity to start a slot.

speed, acceleration, coordinates, velocity, and acceleration of UAV f over time T can be denoted by $\{\mathbf{q}[t,n]\}, \{\mathbf{v}[t,n]\}, \{\mathbf{a}[t]\}$, and $\mathbf{q}^f[t,n] = \left(x^f[t,n], y^f[t,n], z^f[t,n]\right)^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$, $\mathbf{v}^f[t,n] = \left(v_x^f[t,n], v_y^f[t,n], v_z^f[t,n]\right)^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$, and $\mathbf{a}^f[t,n] = \left(a_x^f[t,n], a_y^f[t,n], a_z^f[t,n]\right)^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$, respectively.

A. Two Time Scale

In order to handle and formulate the problem in a precise manner, we come up with defining the two-time scale structure. In spite of the fact that this structure increases the complexity of the problem, it allows us to accurately evaluate the trajectory of UAVs and appropriately deal with the resource allocation problem. We employ two time scales: slot and frame, with duration δ and $N \times \delta$, respectively. Therefore, each frame contains N slots. In the beginning time of each frame, an acceleration of UAV is ascertained in the frame. Besides that, communication resources are allocated, and the initial velocity of each UAV is calculated as well at each slot. So, we solve our optimization problem (22) for each $t \in \mathcal{T}$. Thereupon, we show Fig. 1b for better illustration.

B. Channel Gain

Assume that the communication channel from UAV to each user is dominated by a line-of-sight (LoS) link. In VLC and UAV networks without loss of generality, the LoS channel gain of the VLC link between UAV f and user m can be expressed as (1), shown at the bottom of the next page, where

$$d^{m,f}[t,n] = \left\|\mathbf{q}^f[t,n] - \mathbf{w}^m[t,n]\right\|, \tag{2}$$

where

$$\tilde{F}(\psi[t,n]) = T_s(\psi[t,n])\cos(\psi[t,n])g(\psi[t,n]), \tag{3}$$

where A_r is the active area of the photo detector (PD). $d^{m,f}[t,n]$ is the transmission distance from the UAV to the user and $\psi[t,n]$ denotes the angle of incidence between the UAV and the device, $\phi[t,n]$ is the angle of irradiance from the UAV to the device. $\tilde{m}$ is the order of Lambertian emission with $\tilde{m} = -\ln 2/\left(\ln\cos\phi_{1/2}\right)$ where $\phi_{1/2}$ is the LED's semi-angle at half power. $T_s(\psi[t,n])$ is the gain of the optical filter and $g(\psi[t,n])$ is the optical concentrator gain at the PD, $g(\psi[t,n])$ can be expressed as: $g(\psi[t,n]) = \eta/\sin^2\psi_c$ when $0 \geq \psi[t,n] \geq \psi_c$, and $g(\psi[t,n]) = 0$ if $\psi_c < \psi[t,n]$, where ψ_c and η are the semi-angle field of view (FoV) of the PD and the refractive index.

C. UAV Trajectory and User Mobility

We examine the UAV maneuvering in this sub-section. The UAV trajectory at a constant acceleration, which is introduced in the preceding section, can be formulated for each $t \in \mathcal{T}$ as follows:

$$\mathbf{q}^f[t,n] = \mathbf{q}^f[t,n-1] + \mathbf{v}^f[t,n-1]\delta + \frac{1}{2}\mathbf{a}^f[t]\left(\delta\right)^2, \quad \forall f \in \mathcal{F}, \quad n \in \mathcal{N}, \tag{4}$$

where

$$\mathbf{v}^f[t,n] = \mathbf{v}^f[t,n-1] + \mathbf{a}^f[t]\delta, \quad \forall f \in \mathcal{F}, \ n \in \mathcal{N}, \tag{5}$$

where $\mathbf{v}^f[t,n]$ and $\mathbf{a}^f[t]$ refer to the velocity at the previous slot and acceleration in the desired frame, respectively, and δ indicates the duration of the slot. Following that, the maximum speed and acceleration of UAVs are given below

$$\|\mathbf{v}^f[t,n]\| \leq v_{\max}, \|\mathbf{a}^f[t]\| \leq a_{\max}, \quad \forall f \in \mathcal{F}, \ n \in \mathcal{N}. \tag{6}$$

Additionally, a certain distance between UAVs should be established in order to avoid collisions stated as follows:

$$\|\mathbf{q}^f[t,n] - \mathbf{q}^{f'}[t,n]\| \leq d_{\min}, \quad \forall f, f' \in \mathcal{F}, \ n \in \mathcal{N}, \tag{7}$$

where $d_{\min}$ indicates the shortest distance between two UAVs. The second area of focus is the mobility of users. Recognizing that user movement is random, hence, we can consider it in the following way.

$$\mathbf{w}^m[t,n] = \mathbf{w}^m[t,n-1] + \mathbf{v}^m[t,n-1]\delta, \quad \forall m \in \mathcal{M}, \quad n \in \mathcal{N}, \tag{8}$$

and

$$\|\mathbf{v}^{\mathbf{m}}[t,n]\| \leq v'_{\max}, \quad \forall m \in \mathcal{M}, \ n \in \mathcal{N}, \tag{9}$$

where (9) indicates user step to be limited. Users and UAVs can move freely, so the limited simulation environment is defined as follows:

$$\|(x^f[t,n], y^f[t,n]) - (x_{\text{mid}}, y_{\text{mid}})\| \leq \hat{r}_{\text{Radius}}, \quad 0 \leq z^f[t,n] \ \leq z_{\max}, \quad \forall f \in \mathcal{F}, \ n \in \mathcal{N}, \tag{10}$$

and

$$\|\mathbf{w}^f[t,n] - (x_{\text{mid}}, y_{\text{mid}})\| \leq \hat{r}_{\text{Radius}}, \quad \forall f \in \mathcal{F}, \ n \in \mathcal{N}. \tag{11}$$

There are two types of users in this system, including CoMP and non-CoMP. A CoMP user is associated with more than one UAV, whereas, a non-CoMP is the same as a traditional network users. The SINR for the m_{th} user data, received at user m is expressed as

$$\gamma^{m,f}[t,n](m)=\frac{\mu^2 p^{m,f}[t,n]\rho^{m,f}[t,n]\left|h^{m,f}[t,n]\right|^2}{I_{\text{Intra}}^{m,f}[t,n]+B^{m,f}\left(\sigma^m\right)^2}, \quad (12)$$

where μ is the PD's responsibility, $p^{m,f}[t,n]$ is the allocated power between user m and UAV f, $\rho^{m,f}[t,n]$ is the user association variable where $\rho^{m,f}[t,n] \in \{0,1\}$, and $\left|\mathbf{h}^{m,f}\right|$ is channel coefficient between user m and UAV f. $I_{\text{Intra}}^{m,f}[n]$ is NOMA interference, $B^{m,f}$, and $(\sigma^m)^2$ are the channel bandwidth, and noise variance, respectively. Due to low interference in VLC system, we have only NOMA interference in the denominator of SINR. We can indicate $I_{\text{Intra}}^{m,f}$ due to the principle of NOMA as:

$$I_{\text{Intra}}^{m,f}[t,n]=\sum_{i\in\mathbf{U}_m}\rho^{i,f'}[t,n]p^{i,f'}[t,n]\left|h^{i,f'}[t,n]\right|^2, \quad (13)$$

where

$$\mathbf{U}^m=\left\{b\in M\left|\left|h^{b,f}[t,n]\right|^2>\left|h^{m,f'}[t,n]\right|^2\right.\right\}, \quad (14)$$

In (13), the user with the best channel ignores other signals and only senses noise as its interference as it is obvious in (14). We define the total date rate as summation of CoMP and nonCoMP users as:

$$\tilde{R}[t]=\sum_{m\in\mathcal{M}}\left(R_{\text{CoMP}}^m[t]+\sum_{f\in\mathcal{F}}R_{\text{nonCoMP}}^{m,f}[t]\right), \quad (15)$$

where

$$R_{\text{CoMP}}^m[t]=\frac{1}{N}\sum_{n=1}^{N}R'^{\,m}_{\text{CoMP}}[t,n], \quad (16)$$

$$R_{\text{nonCoMP}}^{m,f}[t]=\frac{1}{N}\sum_{n=1}^{N}R'^{\,m,f}_{\text{nonCoMP}}[t,n], \quad (17)$$

where

$$\begin{aligned}R'^{\,m}_{\text{CoMP}}[t,n]&=\nu^{m,f}[t,n]\sum_{f\in\mathcal{F}}\frac{B^{m,f}}{2}\\&\quad\times\log_2\left(1+\gamma^{m,f}[t,n]\right), \quad (18)\\R'^{\,m,f}_{\text{nonCoMP}}[t,n]&=\left(1-\nu^{m,f}[t,n]\right)\frac{B^{m,f}}{2}\\&\quad\times\log_2\left(1+\gamma^{m,f}[t,n]\right), \quad (19)\end{aligned}$$

where $\nu^{m,f}$ is a binary variable which indicates whether the user is CoMP or not. The scaling factor $1/2$ is due to the Hermitian symmetry [38]. As a first step in formulating the proposed problem, we must calculate the minimum transmitter power each UAV i uses to meet the data rate and illumination requirements. Based on the data rate constraint R_{min} of each user m located at the coordinates (x^m, y^m), the required power for the UAV f at time $[t,n]$ is [39]:

$$p^{m,f}[t,n]=\frac{\left(I_{\text{Intra}}^{m,f}[t,n]+B^{m,f}\left(\sigma^m\right)^2\right)\left(2^{\left(2/B^{\text{m,f}}\right)R'^f}-1\right)}{\mu^2\left|h^{m,f}[t,n]\right|^2}. \quad (20)$$

Once the maximum power requirement of the user has been met, a UAV can fulfill all the users' requirements. Thus, the minimum power of UAV f meeting the needs of its associated users corresponds to:

$$p_{\text{min}}^f[t,n]=\max\left\{p^{m,f}[t,n]\right\},\quad \forall f\in\mathcal{F},\ n\in\mathcal{N},t\in\mathcal{T}. \quad (21)$$

III. Problem Formulation

In this section, we introduce a multi-objective optimization problem (MOOP) where the data rates of CoMP users and non-CoMP users are maximized, and the total power of these users is minimized simultaneously for each $t\in\mathcal{T}$ as follows:

$$\max_{\{\mathbf{a}[t],\rho^{m,f}[t,n],p^{m,f}[t,n]\}}\ \tilde{R}[t], \quad (22a)$$

$$\min_{\{\mathbf{a}[t],\rho^{m,f}[t,n],p^{m,f}[t,n]\}}\ \frac{1}{N}\sum_{f\in\mathcal{F}}\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}p^{m,f}[t,n], \quad (22b)$$

$$\begin{aligned}s.t.\ &\sum_{f\in\mathcal{F}}\sum_{m\in\mathcal{M}}\rho^{m,f}[t,n]p^{m,f}[t,n]\le p_{\text{max}},\\&\forall n\in\mathcal{N}, \quad (22c)\\&p_{\text{min}}^f[t,n]\le\sum_{m\in M}\rho^{m,f}p^{m,f}[t,n]\le\tilde{p}_{\text{max}}^f,\\&\forall f\in\mathcal{F},\ n\in\mathcal{N}, \quad (22d)\\&R_{\text{CoMP}}^m[t,n]\ge R_{\text{min}}'^f,\quad\forall f\in\mathcal{F},\ m\in\mathcal{M},\\&n\in\mathcal{N}, \quad (22e)\\&R_{\text{nonCoMP}}^{m,f}[t,n]\ge R_{\text{min}}^f,\quad\forall f\in\mathcal{F},\ m\in\mathcal{M},\\&n\in\mathcal{N}, \quad (22f)\\&\sum_{m\in\mathcal{M}}\rho^{m,f}[t,n]\le J_K,\quad\forall f\in\mathcal{F},\ n\in\mathcal{N}, \quad (22g)\\&\gamma^{m,f}[t,n](i)-\gamma^{m,f}[t,n](m)\ge 0,\\&,|h^{i,f}[t,n]|^2>|h^{m,f}[t,n]|^2\\&\forall\ f\in\mathcal{F},\quad m,\ i\in\mathcal{M},\ n\in\mathcal{N}, \quad (22h)\\&\rho^{m,f}[t,n]\in\{0,1\},\quad\forall f\in\mathcal{F},\ m\in\mathcal{M},\\&n\in\mathcal{N}, \quad (22i)\\&\nu^{m,f}[t,n]\in\{0,1\},\quad\forall f\in\mathcal{F},\ m\in\mathcal{M},\\&n\in\mathcal{N}, \quad (22j)\\&(4),\ (5),(6),\ (7),(8),\ (9),\ (10),\ (11).\end{aligned}$$

(22c) demonstrates the maximum power of each UAV that can transmit, (22d) is UAVs power constrained between the

$$h^{m,f}[t,n]=\begin{cases}\frac{(\tilde{m}+1)A_r}{2\pi(d^{m,f}[t,n])^2}\cos^m(\phi[t,n])\tilde{F}(\psi[t,n]), & 0\le\psi[t,n]\le\psi_c,\\0, & \psi_c\le\psi[t,n],\end{cases} \quad (1)$$

maximum and zero. (22e) and (22f) are the minimum data rate constraint which all UAVs need to satisfy. (4)-(11) are movement constraints. The equations of movement of each UAV is shown in (4), (5) presents the velocity equation of UAVs, the location of users is obtained from equation (8). (10) and (11) illustrate the spatial constraint of the simulation for UAVs and users. (6), (7), and (9) indicate the maximum UAV speed, collusion avoidance, and the maximum speed of each user, respectively. (22g) demonstrates the number of users in the service of each UAV. (22h) are the NOMA constraints which shows the m_{th} user SINR at the i_{th} user must be bigger than at the m_{th} user. (22i) is the assignment index that specifies assignment of user to UAV and as a binary variable. (22j) is a binary variable which indicates whether the user is CoMP or not. Furthermore, both discrete and continuous variables are involved in the optimization problem (22), which makes this optimization MINLP. Noting that this MINLP is non-convex due to variable of transmission power. So, it is complicated to solve this MINLP optimization problems if we have as many UAVs as optimization problems. Fast decision-making is crucial for optimal resource allocation in a dynamic environment. To handle the complexity of the proposed optimization problem, we will be examining state-of-the-art RL methods.

IV. Multi-Agent Based Solution

Due to the complexity of the problem, we are not able to solve the problem by classical programming methods, so we move to use RL methods. Single-agent methods face problems in estimating and overloading information. Conventional multi-agent methods cannot obtain better results than single-agent methods in small environments due to utterly independent functionality. In this section, we form our environment, agents, and the relevant interaction among the agents, which is states, actions, and reward. This section ends up with formulating our multi-agent approach.

A. Environment

The purpose of each agent, in a multi-agent environment, is to maximize its policy function, which can be shown as:

$$\max_{\pi^f} \mathcal{J}^f\left(\pi^f\right), \quad f \in \mathcal{F},\ \pi^f \in \Pi^f, \tag{23}$$

where $\mathcal{J}^f\left(\pi^f\right) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \tilde{r}_{t+1}^f \mid s_0^f\right]$ is our conditional expectation, π_f is the policy of UAV f, and the set Π_j contains all feasible policies that are available for UAV f. Information from the aforementioned discussion about each UAV as an agent interacts with the UAV network environment and takes action relevant to its policy. The goal is to solve the optimization problem (22). The agents seek to improve their reward (23), which is related to the objective function. At each time n, the UAV immediately takes action, $a_{t,n}$ after observing environment state, $s_{t,n}$. The environment transfers to a new state $s_{t,n+1}$ in the transition step and UAV obtains a reward related to its action. Next we describe the state space $\mathcal{S}$, action space $\mathcal{A}$, and the reward function $r_{t,n}$, in our system model:

- State Space: The state of each agent includes all immediate channel gains at time n of users $h^{m,f}[t,n]$ for all $m \in \mathcal{M} \& f \in \mathcal{F}$, all UAV previous velocity and location $\mathbf{q}[t,n-1], \mathbf{v}[t,n-1]$ and users previous location $\mathbf{w}^{m,k}[t,n-1]$, all interferences involved in the $n-1$ slot, $I_{\text{Intra}}^{m,f'}[t,n-1]$ and $I_{\text{Inter}}^{m,f}[t,n-1]$. $\mathbf{s}_{t,n}^f = [h^{m,f}[t,n], \mathbf{q}[t,n-1], \mathbf{v}[t,n-1], \mathbf{w}^{m,k}[t,n-1],$ $I_{\text{Intra}}^{m,f'}[t,n-1]], \quad f \in \mathcal{F}$.
- Action Space: In each time step UAVs (agents) act as $a_{t,n}^f = \{\rho_{t,n}^f, \mathbf{p}_{t,n}^f, \nu_{t,n}^{m,f}, \mathbf{a}_t^f\}$, as we mentioned above, $\rho_{t,n}^f$ is our assignment variable, $\mathbf{p}_{t,n}^f$ is the power that is allocated to the users, $\nu_{t,n}^{m,f}$ is CoMP indicator that shows the user is either CoMP or not, and $\mathbf{a}_t^f$ is the acceleration matrix. Our variables contain both integer and continuous variables. It suggests to adopt policy gradients to solve the problem. Making use of this capability allows us to come up with better solutions.
- Reward function: The reward is one of the most important parts of RL because the major driver of RL is the reward. It is crucial to formulate a function accurately that can both represent the objective function and faster and more stable convergence. First we form our rewards then discuss it. In this system we employ two types of reward, one type per agent and another one is for global critic network. Agents' goals are to maximize their rates while minimizing power consumption. The global reward decreases the total interference among the users as mentioned in [40]. All agents have commitment only to their goal and it is not necessary to know about other agents policy, since they share global critic. The goal of the global critic is to connect all the agents together to aid faster convergence with the support of the reward. To maintain the stable convergence the central server connects the entire system and receives all the states and criticizes the actions of agents. Our global and per agent reward are defined as:

$$\begin{aligned}\tilde{r}_\ell^f = \alpha \sum_{m \in \mathcal{M}} \left(\frac{R_{\text{CoMP}}^{m,f}(\rho,p) + R_{\text{nonCoMP}}^{m,f}(\rho,p)}{B^{m,f}/2 \log_2 \left(1 + \tilde{p}_{\text{max}}^f / B^{m,f} \left(\sigma^m\right)^2\right)} \right) \\ - (1-\alpha) \frac{\sum_{m \in \mathcal{M}} \rho^{m,f} p^{m,f}}{\tilde{p}_{\text{max}}^f},\end{aligned} \tag{24}$$

and global reward

$$\tilde{r}_G = -\left(\sum_{f \in \mathcal{F}} \sum_{m \in \mathcal{M}} I_{\text{Intra}}^{m,f} \right). \tag{25}$$

We use linear scalarization to transform our multi-objective problem to single-objective using weight factor $\alpha \in (0,1)$ [41].

B. MADDPG

We assume $\boldsymbol{\pi}$ as a set for all UAVs policies, UAV f policy is π^f $\left(\pi^f \in \boldsymbol{\pi} = \{\pi^1, \ldots, \pi^F\}\right)$ with parameters $\tilde{\theta}^f$, $\tilde{Q}^f$ for UAV critic (Q-function) with parameter $\tilde{\phi}^f$ and $\tilde{Q}_G$ for global Q-function with parameter $\tilde{\psi}$. Now, we can form our neural network, after that, we can discuss gradient policy.

We consider L_{π}^{f}, L_{q}^{f}, and L_G as the number of layers in neural network for each agent action and critic and global critic, respectively. According to the above information, we can develop our neural network in this way $\tilde{\Theta}_i = \left(W_{\pi}^{(1)}, \ldots, W_{\pi}^{(L_{\pi})}\right)$, $\tilde{\Phi}_q = \left(W_{q}^{(1)}, \ldots, W_{q}^{(L_{q})}\right)$, and $\tilde{\Psi}_G = \left(W_{G}^{(1)}, \ldots, W_{G}^{(L_{G})}\right)$ as actor, UAV critic and global critic, respectively. According to the above, the gradient policy is

$$\nabla_{\theta^f} \mathcal{J}^f = \mathbb{E}\left[\nabla_{\tilde{\theta}^f} \pi^f\left(a^f \mid s^f\right) \nabla^{a^f} Q_{\pi}^{f}(\mathbf{s}, \mathbf{a})\Big|_{a^f=\pi^f(s^f)}\right], \tag{26}$$

where $\boldsymbol{a} = \left(a^1, \ldots, a^F\right)$ is all actions that are taken by each UAV with observation $\boldsymbol{s} = \left(s^1, \ldots, s^F\right)$. We integrate all actions and states in $Q_{\pi}^{f}(\mathbf{s}, \mathbf{a})$ as inputs to approximate Q-function for UAV f. Here we utilize two critic networks, and we reformulate the policy gradient. Although, this framework is able to reach quite good performance and converge in moderated steps, it still suffer from overwhelmed estimation and its loss in approximations deteriorates framework performance due to sub-optimal policy in Q-function. With results in [40], we swap global critic with twin delayed deterministic policy gradient in (27), then our new gradient policy is

$$\begin{aligned}\nabla_{\theta^f} \mathcal{J}^f = &\underbrace{\mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}\left[\nabla_{\theta^f} \pi^f\left(a^f \mid s^f\right) \nabla_{a^f} Q_{G_i}^{\psi_i}(\mathbf{s}, \mathbf{a})\right]}_{\text{TD3 Global Critic}} \\ &+ \underbrace{\mathbb{E}_{s^f, a^f\sim\mathcal{D}}\left[\nabla_{\theta^f} \pi^f\left(a^f \mid s^f\right) \nabla_{a^f} Q_{\phi^f}^{f}\left(s^f, a^f\right)\right]}_{\text{UAV critic}},\end{aligned} \tag{27}$$

as shown above $a^f = \pi^f(s^f)$ are actions which UAV f takes with observation s^f, respect to policy π^f. In (27), we have two terms, the first one shows global critic which receives actions and states of all UAVs and it estimates global Q-function using global reward $\tilde{r}_G$ and other term shows critic of UAV which receives only itself actions and states. For updating the loss function of global critic, we use

$$\mathcal{L}(\psi_i) = \mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{r},\mathbf{s}'}\left[\left(Q_{G_i}^{\psi_i}(\mathbf{s}, \mathbf{a}) - y_G\right)^2\right], \tag{28}$$

where y_G is a target value of estimation and

$$y^f = r^f + \gamma(1 - \tilde{d}) \min_{i=1,2} Q_{\phi_i'^{f_i}}^{f_i}\left(s'^f, a'^f\right)\Big|_{a'^f=\pi'^f(s'^f)}. \tag{29}$$

where our target policy is $\boldsymbol{\pi'} = \{\pi'^1, \ldots, \pi'^F\}$. We parameterize it with $\boldsymbol{\theta}' = \{\theta'^1, \ldots, \theta'^F\}$, and the UAV loss function and its target update as follows:

$$\mathcal{L}^f\left(\phi^f\right) = \mathbb{E}_{\mathbf{s}^f,\mathbf{a}^f,\mathbf{r}^f,\mathbf{s}'^f}\left[\left(Q_{\phi^f}^{f}\left(s^f, a^f\right) - y^f\right)^2\right], \tag{30}$$

and y^f:

$$y^f = r^f + \gamma Q_{\phi'^f}^{f}\left(s'^f, a'^f\right)\Big|_{a'^f=\pi'^f(s'^f)}. \tag{31}$$

By (30) and (31), the agents critic network are updated. Our proposed algorithm is summarized in Algorithm 1 and with intuitive illustrated in Fig. 2. Now, we discuss the reasons for choosing the rewards of system. Our first reward is directly related to the objective of the problem, but is calculated independently for each UAV. Each UAV strives to get as close to its maximum and best as possible by maximizing its reward. Due to the neural network's weakness in calculating interference, which plays a vital role in the management of radio resources, our global reward maximizes the symmetry of the total system interference. We apply TD3 because it trains global with with two extra networks, where $\tilde{d}$ is delay hyper-parameter.

Algorithm 1 Two-Time Scale Modified MADDPG

```
Initiate environment, generate UAVs and users
Inputs: Enter number of $a_t$, $s_t$ agents and users
Initialize all, global critic networks target global critic
networks and agents policy and critic networks.
for t=1 to T do
    for $n$ = 1 : $N$ do
        for each agnet $f$ do
            Observe state $s_t^f$ and take action $a_t^f$
        $\mathbf{s}_t = \left[s_t^1, \ldots, s_t^F\right], \quad \mathbf{a}_t = \left[a_t, \ldots, a_t^F\right].$
        Receive global and local rewards, $\tilde{r}_{G,t}$ and $\tilde{\mathbf{r}}_t^f$
        Store $\left(\mathrm{s}_t, \mathbf{a}_t, \tilde{\mathbf{r}}_t^f, \tilde{r}_{G,t}, \mathbf{s}_{t+1}\right)$ in replay buffer $\mathcal{D}$
    Sample minibatch of size S, $\left(\mathrm{s}^j, \mathbf{a}^j, \mathbf{r}_g^j, \mathbf{r}_\ell^j, \mathbf{s}'j\right)$, from
    replay buffer $\mathcal{D}$
    Set $y_g^j = r_g^j + \gamma \min_i Q_{\psi_i'}^{g_i}\left(\mathbf{s}', \mathbf{a}'j\right)$
    Update global critics by minimizing the loss:
    $\mathcal{L}\left(\psi_i\right) = \frac{1}{S} \sum_j \left\{\left(Q_{\psi_i}^{g_i}\left(\mathrm{s}^j, \mathbf{a}^j\right) - y_g^j\right)^2\right\}.$
    Update target parameters: $\psi_i' \leftarrow \tau\psi_i + (1-\tau)\psi_i'$
    if episode mod $d$ then
        Train actor and critic nerwork
        for for each agent $f$ do
            episode mod $d$
            $\mathcal{L}\left(\phi_i\right) = \frac{1}{S} \sum_j \left\{\left(Q_{\phi_i}^{i}\left(s_i^j, a_i^j\right) - y_i^j\right)^2\right\}.$
            Update local actors:
            $\nabla J_{\theta_i} \approx \frac{1}{S} \sum_j \{\nabla_{\theta_i}\pi_i\left(a_i \mid s_i^j\right) \nabla_{a_i} Q_{\psi_1}^{g_1}\left(\mathbf{s}^j, \mathbf{a}^j\right)$
                $\times \nabla_{\theta_i}\pi_i\left(a_i \mid s_i^j\right) \nabla_{a_i} Q_{\phi_i}^{i}\left(s_i^j, a_i^j\right)\}.$
            Update target networks parameters:
                $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i',$
                $\phi_i' \leftarrow \tau\phi_i + (1-\tau)\phi_i'.$
```

C. Computational Complexity

In this section, we investigate the computational complexity of the algorithm. We divide it into four sorts and address them in their dedicated section. Complexity is related to a) number of trainable variables, b) total neural network applied to network, c) the computational complexity, and d the

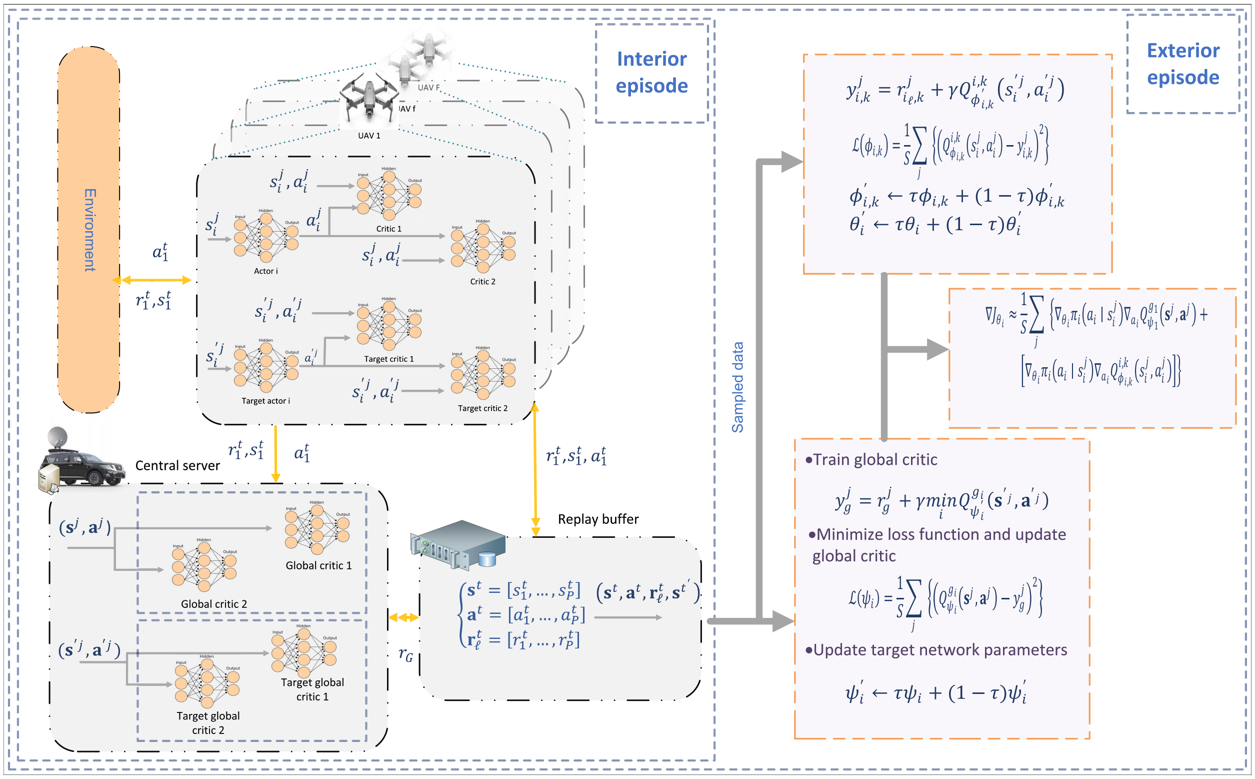

Interior episode
Exterior episode
Environment
UAV 1
UAV f
UAV F
a_1^t
r_1^t, s_1^t
s_i^j, a_i^j
s_i^j
a_i^j
Actor i
Critic 1
Critic 2
$s_i^{'j}, a_i^{'j}$
Target actor i
Target critic 1
Target critic 2
Input
Hidden
Output
Central server
r_1^t, s_1^t, a_1^t
Replay buffer
$(\mathbf{s}^j, \mathbf{a}^j)$
Global critic 1
Global critic 2
$(\mathbf{s}^{'j}, \mathbf{a}^{'j})$
Target global critic 1
Target global critic 2
$\mathbf{s}^t = [s_1^t, \ldots, s_P^t]$
$\mathbf{a}^t = [a_1^t, \ldots, a_P^t]$
$\mathbf{r}_\ell^t = [r_1^t, \ldots, r_P^t]$
$(\mathbf{s}^t, \mathbf{a}^t, \mathbf{r}_\ell^t, \mathbf{s}^{t'})$
r_G
Sampled data
$y_{i,k}^j = r_{i_\ell,k}^j + \gamma Q_{\phi_{i,k}'}^{i,k}(s_i^{'j}, a_i^{'j})$
$\mathcal{L}(\phi_{i,k}) = \frac{1}{S}\sum_j \left\{\left(Q_{\phi_{i,k}}^{i,k}(s_i^j, a_i^j) - y_{i,k}^j\right)^2\right\}$
$\phi_{i,k}' \leftarrow \tau\phi_{i,k} + (1-\tau)\phi_{i,k}'$
$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$
$\nabla J_{\theta_i} \approx \frac{1}{S}\sum_j \left\{\nabla_{\theta_i}\pi_i(a_i \mid s_i^j)\nabla_{a_i} Q_{\psi_1}^{g_1}(\mathbf{s}^j, \mathbf{a}^j) + \left[\nabla_{\theta_i}\pi_i(a_i \mid s_i^j)\nabla_{a_i} Q_{\phi_{i,k}}^{i,k}(s_i^j, a_i^j)\right]\right\}$
•Train global critic
$y_g^j = r_g^j + \gamma \min_i Q_{\psi_i'}^{g_i}(\mathbf{s}^{'j}, \mathbf{a}^{'j})$
•Minimize loss function and update global critic
$\mathcal{L}(\psi_i) = \frac{1}{S}\sum_j \left\{\left(Q_{\psi_i}^{g_i}(\mathbf{s}^j, \mathbf{a}^j) - y_g^j\right)^2\right\}$
•Update target network parameters
$\psi_i' \leftarrow \tau\psi_i + (1-\tau)\psi_i'$

Fig. 2. Proposed MADDPG two-stage resource allocation and trajectory planning.

TABLE II
SIMULATION PARAMETERS

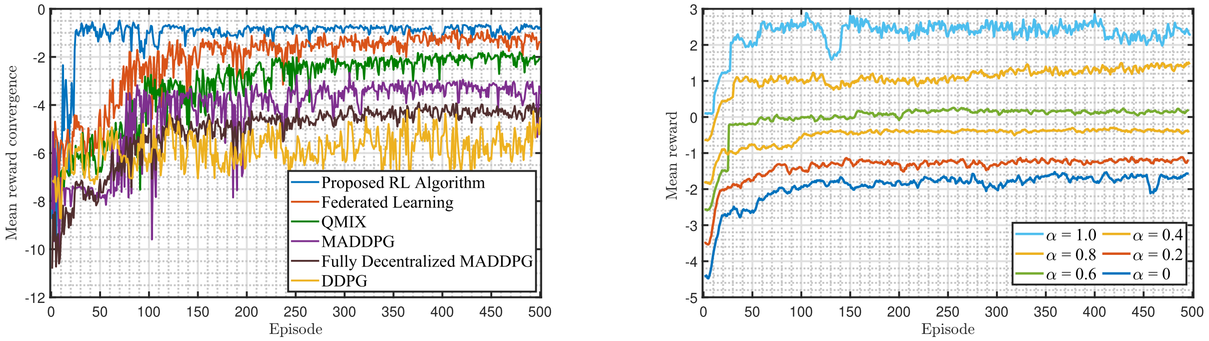
Environment parameters	Value	Environment parameters	Value	NN hyper-parameters	Value
FoV Ψ_c	60° [38]	Detector area of PD A_{PD}	1 cm^2 [38]	Experience replay buffer size	50000
$p_{\max}$	36 W	$\tilde{p}_{\max}^f$	12 W	Mini batch size	64
Number of user	9	Number of UAV	3	Number/size of local actor networks hidden layers	$2/1024, 512$
Half power angle,$\theta_{1/2}$	30° [38]	The order of the Lambertian emission ϖ	1	Number/size of local critic networks hidden layers	$2/512, 256$
PD responsivity	0.53 A/W [38]	Sigma-noise, $B^{m,f}$	10e − 12, 200 MHz	Number/size of global critic hidden layers	$3/1024, 512, 256$
$R_{\min}^{m,f}$	$0.1(kbps)$ [42]	$R_{\min}^{\prime m,f}$	$0.1(kbps)$ [42]	Critic/Actor networks learning rate	0.001/0.0001
J_K	3	x_{mid}	25 m	Discount factor	0.99
x_{mid}	25 m	$z_{\max}$	100 m	Target networks soft update parameter, τ	0.0005
$a_{\max}$	$2\sqrt{3}$ m/s^2	$v_{\max}$	$10\sqrt{3}$ m/s [34], [43]	Number of episodes	500
$v_{\max}'$	$5\sqrt{2}$ m/s [34]	T	500 s × 10e-1	Number of iterations per episode	100
N	100 ms	δ	1 ms		

communication overhead between the agents and the central server. By providing this comprehensive view, we can gain an adequate understanding of their applicability and feasibility. The following evaluations have also incorporated the MADDPG complexity analysis since the proposed algorithms are based on it.

1) Number of the Trainable Variables: All observations and actions of the agents are used as inputs to the centralized Q-functions in MADDPG. In the case where all agents have identical observation and action spaces, depicted by ω and α, then the number of MADDPG parameters to train is $\mathcal{O}\left(F^2(\omega+\alpha)\right)$, where F represents the number of agents. Even so, in the proposed RL methods, there are two types of critic networks: global and local critics. Neither algorithm is unique both share the same global centralized Q-function of linear increase in parametric space $(F(\omega+\alpha))$, expressed as $\mathcal{O}(F(\omega+\alpha))$. Comparatively, the local critics in the two RL methods consider only the agent's observations and actions. Due to the fact that local critics operate in the same parameter space, their manifestations can be depicted as $\mathcal{O}(\omega+\alpha)$. Since the agents solely regulate themselves based on their observations, the parameter space of their local critics is described by $\mathcal{O}(\omega+\alpha)$, as with FRL, QMIX, and MADDPG. Since DDPG has a single actor network, they must take all the states and actions of all the agents and recompute them using $\mathcal{O}(F(\omega+\alpha))$.

2) Number of the Nodes in the Neural Network: $2 \times \left(F\left(\underline{1}_Q + \underline{1}_A\right)\right)$ is employed by MADDPG, where the addition by 2 is the result of the target networks, $\underline{1}_Q$, and $\underline{1}_A$ represent the critic and actor networks that are respective to each agent, and F is the total number of agents. The modified MADDPG framework produces a total number of neural networks of $2 \times \left(F\left(\underline{1}_{Q_\ell} + \underline{1}_{A_\ell}\right) + \underline{1}_{Q_g}\right)$, the final amount indicating the global number of critics. Additionally, the TD3 algorithm will double the number of global critics. Thus, the number of neural networks in this case will be $2 \times \left(F\left(\underline{1}_{\mathrm{Q}_\ell} + \underline{1}_{\mathrm{A}}\right) + \underline{2}_{\mathrm{Q}}\right)$. It is expected that the same procedure will be followed for the decentralized MADDPG, QMIX, and FRL. Their neural networks will total $2 \times F\left(\underline{1}_{Q_\ell} + \underline{1}_{A_\ell}\right)$. Last but not least, for DDPG, this number will be $2 \times \left(\underline{1}_{Q_\ell} + \underline{1}_{A_\ell}\right)$.

3) Computational Complexity: Suppose we aim to express our analytics in mathematical terms, so we assume first that Γ_i^a and Γ_i^c are the number of neurons in the



(a)

(b)

Fig. 3. (a) Comparing our approach with existing baselines, (b) the rewards and converges: for $\alpha = 0.0$, $\alpha = 0.2$, $\alpha = 0.4$, $\alpha = 0.6$, $\alpha = 0.8$, and $\alpha = 1.0$.

i-th layer of the actor and critic networks, respectively. In the case of actor and critic networks that are fully interconnected, their computational complexity can be expressed as $\mathcal{O}\left(\sum_{i=2}^{i=L_a-1}\left(\Gamma_{i-1}^a\Gamma_i^a+\Gamma_i^a\Gamma_{i+1}^a\right)\right)$ and $\mathcal{O}\left(\sum_{i=2}^{i=L_c-1}\left(\Gamma_{i-1}^c\Gamma_i^c+\Gamma_i^c\Gamma_{i+1}^c\right)\right)$, respectively, where L_a and L_c represent the total number of layers in the respective actor and critic networks. Hence, the complexity of the mentioned frameworks can be as follows:

- MADDPG: $\mathcal{O}\left(F \cdot N_b \cdot E \cdot I \cdot \left(\mathfrak{C}_\ell^a + \mathfrak{C}_\ell^c\right)\right)$
- QMIX: $\mathcal{O}\left(F \cdot N_b \cdot E \cdot I \cdot \left(\mathfrak{C}_\ell^c\right)\right)$
- Modified MADDPG: $\mathcal{O}\left(N_b \cdot E \cdot I \cdot \left(\mathfrak{C}_g^a + \mathfrak{C}_g^c\right)\right) + \mathcal{O}\left(F \cdot N_b \cdot E \cdot I \cdot \left(\mathfrak{C}_\ell^a + \mathfrak{C}_\ell^c\right)\right)$
- FRL: $\mathcal{O}\left(F \cdot N_b \cdot E \cdot I \cdot \left(\mathfrak{C}_\ell^a + \mathfrak{C}_\ell^c\right)\right)$
- FDec. MADDPG: $\mathcal{O}\left(F \cdot N_b \cdot E \cdot I \cdot \left(\mathfrak{C}_\ell^a + \mathfrak{C}_\ell^c\right)\right)$
- DDPG: $\mathcal{O}\left(N_b \cdot E \cdot I \cdot \left(\mathfrak{C}_\ell^a + \mathfrak{C}_\ell^c\right)\right)$

where $\mathfrak{C}^a = \sum_{i=2}^{i=L_a-1}\left(\Gamma_{i-1}^a\Gamma_i^a+\Gamma_i^a\Gamma_{i+1}^a\right), \mathfrak{C}^c = \sum_{i=2}^{i=L_c-1}\left(\Gamma_{i-1}^c\Gamma_i^c+\Gamma_i^c\Gamma_{i+1}^c\right)$, F correspond to the number of UAVs (agents), N_b identifies the mini-batch sampling size, E represents the number of episodes, and I signifies the maximum training steps per episode.

4) Communication Overhead: Especially in communication systems based on MARL frameworks, communication overhead is one of those performance metrics commonly overlooked. This is because most MARL frameworks are built on agents' communication. Data exchange is necessary to stabilize the learning process and encourage cooperative behavior among agents. In spite of this, it is important to reduce the overhead. We have provided the total communication overhead for the proposed frameworks below. During the learning process, we have examined how often the agents need to interact with other agents and the central server. Our proposed algorithms, FRL, QMIX and DDPG, have an overhead of F, whereas MADDPG has an overhead of $F(F-1)$. Due to the fact that agents in this framework act independently of one another, Fully decentralized MADDPG does not require any overhead on the network side.

D. Convergence Analysis

While the Q-learning algorithm is forthright to prove the convergence [44], the policy-based RL algorithms, especially their multi-agent extensions, like the algorithms we propose, are complicated to prove because several agents simultaneously interact with their environment. Through simulations in the next section, we illustrated the convergence of the proposed algorithms.

V. Simulation

In this section, an evaluation of the performance of the proposed algorithm is presented via numerical results. In the absence of a direct method of determining a neural network's hyperparameters, we selected them by sophisticated trial and error. Rectified linear unit (ReLU) is high-performing, and we have incorporated it into RL as an activation function. The simulation settings are summarized in Table II. For the main simulation, the number of UAVs is 3 and 9 users are moving on the ground and the environment is limited to a cylindrical with radius and height of 50 m and 100 m, respectively. The time slot of the frame is considered to be 100 ms and the slots are considered to be 1 ms, in which the constant acceleration is specified in the frame, and in the slots the final speeds of each slot are considered as the initial speed of the next slot. The maximum speed for UAVs is 10 m/s in each direction, and the maximum acceleration is 2 m/s^2. All information about the neural network and the number of layers for each factor is given in Table II. In the following, we will discuss all the figures obtained from the simulation. First, we will explain the baselines, and then we will examine the reward and the overall data rate of the network and the allocated powers, the effect of the minimum value of the data rate, the trajectory of the UAVs, and finally, the impact of constant velocity and constant acceleration on the performers. In addition, the source code of the proposed modified MADDPG is available in [45]. In the presence of these hyperparameters settings, approximately 10 UAVs and 100 users would be able to operate, but at the cost of relatively low accuracy and a greater degree of instability in the reward distribution. Moreover, an agent would require more time than usual to develop a policy that achieves a high amount of reward or receives a lower amount of punishment. It is essential that we keep the state and action requirements as low as possible when evaluating our system model for different circumstances. By doing so, we will be able to maintain as incremental and flexible settings.

A. Solution Baselines

The baselines include three solution methods as show in Fig. 3, MADDPG, Federated Reinforcement Learning, QMIX, fully decentralized MADDPG, and DDPG, described below.

- **MADDPG:** A standard method that operates on a multi-agent basis and has DDPG neural networks. The various agents interact with each other through a central server.
- **Federated Reinforcement Learning:** In this algorithm, the UAVs are equipped with separate actor and critic networks and are trained using locally available information. In FRL, the central server is not responsible for training any neural network based on local states and actions of the UAVs motivate cooperation between the UAVs, as is the case in the proposed algorithms. Instead of local information, such as states and actions, an FRL agent transmits the weights of their actor and critic networks to a central server. The server collects these weights, and, running a pre-set algorithm, they are aggregated and then sent to the agents again. The central server's aggregation rules are formulated by [46]

$$\Theta^{t+1} = \Theta^{t} \cdot \Omega, \tag{32}$$

where $\mathbf{\Theta}^{t} = \left[\mathbf{\Theta}^{1,t}, \cdots, \mathbf{\Theta}^{F,t}\right]$ represents the vector of all the agents' parameters at the t-th learning epoch, and the $\mathbf{\Omega}$ is calculated by:

$$\mathbf{\Omega} = \begin{bmatrix} \omega & \dfrac{1-\omega}{F-1} & \cdots & \dfrac{1-\omega}{F-1} \\ \dfrac{1-\omega}{F-1} & \omega & \cdots & \dfrac{1-\omega}{F-1} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{1-\omega}{F-1} & \dfrac{1-\omega}{F-1} & \cdots & \omega \end{bmatrix}. \tag{33}$$

To aggregate parameters under the proposed scheme, each agent has its parameters preserved with weights ω, while other agents' parameters are mixed with weights $\left(\frac{1-\omega}{F-1}\right)$.
- **QMIX:** To produce its own Q value, each agent uses deep recurrent Q-learning (DRQN) [47] with the form of $Q_i\left(\tau_i, a_i; \theta_i\right)$. Recurrent neural networks (RNNs) are introduced by DQN to deal with the problem of partially observable data, which is $Q(o, a \mid \theta) \neq Q(s, a \mid \theta)$ in DQN. DQN derives the value for Q from the current observation $o_{\mathrm{i,\,t}}$ and the action $a_{i,t-1}$ of the previous time. There is only one prerequisite: a global argmax operation performed on Q_{tot} must yield the same result as a set of individual argmax operations done on each Q_a. Thus, each agent's local optimal action is a subset of the global optimal action and can be expressed as follows:

$$\underset{a}{\arg\max}\ Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \begin{pmatrix} \underset{a_1}{\arg\max}\ Q_1\left(\tau_1, a_1\right) \\ \vdots \\ \underset{a_n}{\arg\max}\ Q_n\left(\tau_n, a_n\right) \end{pmatrix}, \tag{34}$$

and

$$\frac{\partial Q_{\mathrm{tot}}}{\partial Q_i} \geqslant 0, \quad \forall i \in \{1, \ldots, n\}. \tag{35}$$

The constraint (35) enforces monotonicity for a given relationship between Q_{tot} and each Q_a. A given agent a can take distributed and greedy actions for Q_a. Therefore, w can calculate $\operatorname{argmax} Q_{\mathrm{tot}}$. Conversely, the strategy of each agent can be determined from Q_{tot}.
- **Fully decentralized MADDPG:** In this solution method, agents work independently of each other, have no contact, and only have their observations from the environment.
- **DDPG:** The single-agent interacts with the environment and has the whole environment as its neural network input.

B. System Model Baselines

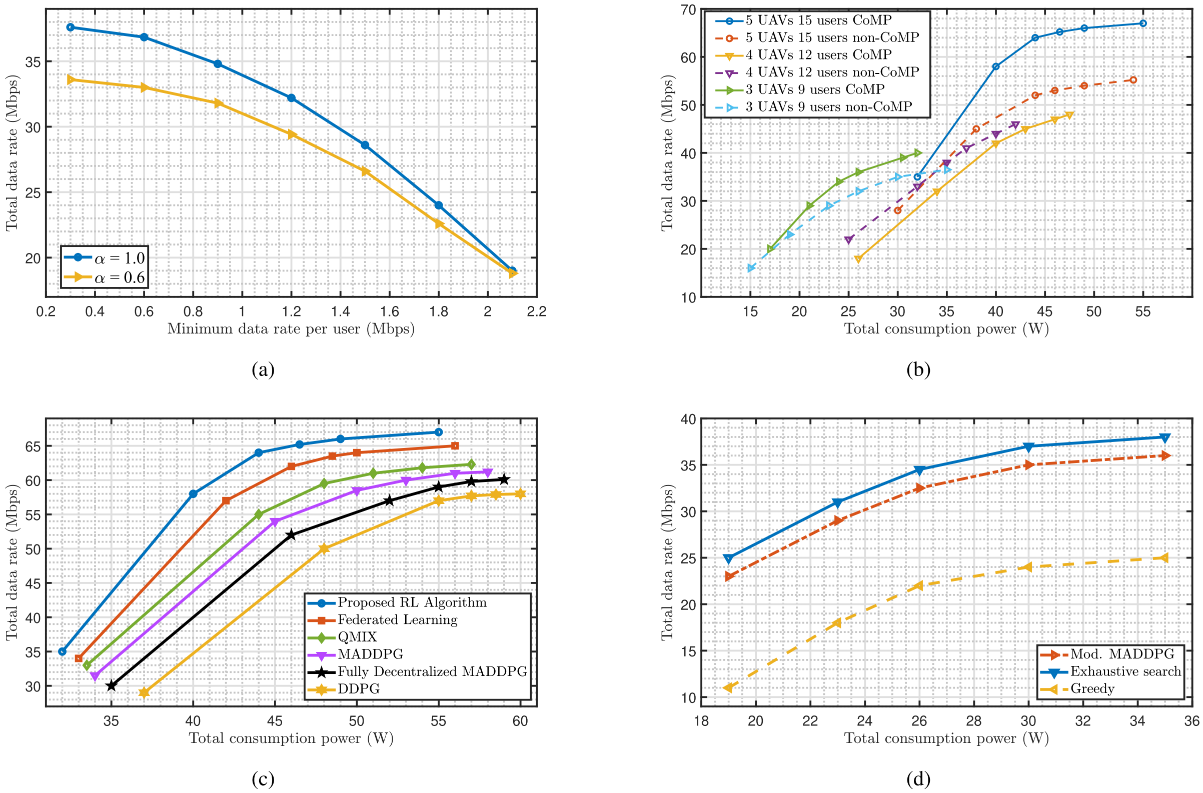
- **Non-CoMP:** A conventional system model without CoMP technology. Users with awful channels are deprived of receiving any assist from nearby UAVs.

C. Trade Off Between Data Rate and Power Consumption

Given the weight factor α, that changes the effect of each objective in the primary reward, we swipe this factor from 0 to 1 with step size of 0.2. Note that using this approach allows us to choose the priority between the power and the data rate flexibly. It can be seen in Fig. 3b that the goal is only to minimize the amount of power where only three constraints are considered. The required minimum data rate for each type of user and clipping the power between 0.1 and $p^{f}_{\max}$ helps stabilize the learning process. In this figure, the only goal of the agents is to meet the minimum rate for each type of user. In Fig. 3b, our reward is controlled by the power minimization instead of the data rate. In Fig. 3b, the sudden drop of the reward is due to the movement of users and handover, that the UAV is no longer able to track its users, and users switch among UAVs. Finally, in Fig. 3b, the effect of data rate exceeds the effect of power, and as we can observe in the figure, that rewards converge to a positive value. However, the effect of power minimization also remains strong, but the priority is to maximize the data rate. In Fig. 3b, the impact of the data rate is increasing, and almost all agents policies are affected by the data rate of the whole system, and the sudden drop can be related to user movement and change in users assignment. In Fig. 3b, the impact of data rate increases, and at the beginning of learning, it can be observed that the agents try to increase the reward. However, the commitment to reduce power consumption will reduce reward in episodes afterward. It converges to another point is due to the term of power minimization. Finally, in Fig. 3b, we see that the agents seek to maximize network data rates without considering power consumption. The fluctuation in Fig. 3b can be attributed to the complexity of the system model and corresponding constraints, and the agents that attempt to find the best case for the objective function.

D. Minimum Data Rate Constraint

The more we increase the minimum data rate constraint, the smaller the set of feasible answers gets. As illustrated in Fig. 4a, we continuously see a reduction in the data rate of the entire network by increasing the minimum data rate constraint, and this drop is more tangible in the higher numbers in the minimum data rate constraint.



(a)

(b)

(c)

(d)

Fig. 4. (a) Effect of the minimum data rate on the total data rate. (b) the data rate versus the total power consumption for CoMP and non-CoMP while varying the number of UAVs, (c) comparing the performance of the proposed approach with existing approaches as baselines, (d) performance of the proposed MARL method by comparing with two conventional approaches, exhaustive search, and Greedy.

E. CoMP Versus non-CoMP

Fig. 4b shows the variation of the number of UAVs and users, α value, and their impact on the data rate and the power. From this figure we can observe as the number of UAVs increases, the data rate increases. The power consumption of CoMP system is also significantly reduced in comparison with non-CoMP. The effect of CoMP is similar to the increasing of the number of users, i.e., using CoMP enhances the data rate. CoMP users also reduce the power consumption of each UAV. We compare the baselines to illustrate superiority of proposed algorithm respect to existing other RL method, in Fig. 4c.

F. Optimality

Comparing results based only on the different RL algorithms would seem to be an illegitimate way of comparing the results. As a baseline, we should have compared our proposed solution with other solutions, such as Successive Convex Approximation. To demonstrate the validity of our proposed methods, we have adopted two heuristic methods based on Successive Convex Approximation and Exhaustive Search. Consequently, we had to develop an idea for determining the values of p, ρ, and a. According to the interference levels of the users, we allocate the resources ρ using a graph-based resource allocation from [48]. Difference-of-Two-Concave-Functions (D.C.) Approximation: In this case, we first express the rate function (2) in a D.C. form as:

$$\sum_{n\in\mathcal{N}} r^{(n)}_{m,k(m,n)}\left(\mathbf{p}^{(n)}\right) := f_m(\mathbf{p}) - g_m(\mathbf{p}), \tag{36}$$

where $f_m(\mathbf{p})$ and $g_m(\mathbf{p})$ are the two concave functions that require approximation to transform into solvable problems. The authors adopt the heuristic successive hover-and-fly trajectory, in which the UAV successively hovers over these locations and then flies among them at the maximum speed. Next, the successive convex approximation (SCA) technique is further performed to refine the UAV trajectory by quantizing the continuous UAV trajectory into finite number of wayponits [49]. Two approaches are used for power allocation (p):

1) As a starting point, having mentioned that the non-CoMP setting was adopted to compare the proposed MARL with the conventional method. Divided problems into 3 separate ones are handled by the methods mentioned earlier, and for the power allocation, Greedy-based method is adopted.
2) In the present case, we are going to rely on the exhaustive search. Basically, the power allocation is done by a random method, in order to endeavor and find the best allocation order possible. For the exhaustive search criteria to function, we had to discretize the continuous power into different discrete power levels. In contrast, the dimensions of the problem grow exponentially as we increase the power.

As illustrated in Fig. 4d, our proposed algorithm outperforms greedy algorithm. Looking at the figure above, it is clear that with this method, the performance gap between the exhaustive search method and our proposed method is sufficiently narrow that we can claim the proposed method is optimal.

Additionally, an exhaustive search is insignificant in terms of complexity.

VI. Conclusion

In this paper, we proposed a new time scale structure to study the effect of constant acceleration on the CoMP UAV-enabled VLC networks. While giving two-time scale, and using constant acceleration, which assists to improve the system efficiency, we solved this complex problem using novel machine learning and multi-agent method. Also, we presented a solution method based on our system model. The results obtained from the simulation proved a better performance compared to other methods. Our proposed algorithms outperformed greedy and exhaustive search algorithms. As a future work, we will discuss energy consumption of UAV movement in the long run, and consider illumination for imaging applications predicting user movement. Despite the fact that massive MIMO influences the complexity of the system model, we will investigate the effect of cell-free to improve coverage of VLC network.

References

[1] D. Lee *et al.*, "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[2] L. Liu, S. Zhang, and R. Zhang, "CoMP in the sky: UAV placement and movement optimization for multi-user communications," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5645–5658, Aug. 2019.

[3] M. Elhattab, M.-A. Arfaoui, and C. Assi, "CoMP transmission in downlink NOMA-based heterogeneous cloud radio access networks," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7779–7794, Dec. 2020.

[4] Y. Zeng *et al.*, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Feb. 2019.

[5] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potential, challenges, and promising technologies," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 120–127, Feb. 2019.

[6] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[7] J. Yao and J. Xu, "Joint 3D maneuver and power adaptation for secure UAV communication with CoMP reception," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6992–7006, Oct. 2020.

[8] D. Bykhovsky and S. Arnon, "Multiple access resource allocation in visible light communication systems," *J. Lightw. Technol.*, vol. 32, no. 8, pp. 1594–1600, Mar. 15, 2014.

[9] C. Chen, W.-De Zhong, H. Yang, and P. Du, "On the performance of MIMO-NOMA-based visible light communication systems," *IEEE Photon. Technol. Lett.*, vol. 30, no. 4, pp. 307–310, Feb. 15, 2018.

[10] J. Chen, Z. Wang, and R. Jiang, "Downlink interference management in cell-free VLC network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9007–9017, Sep. 2019.

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

[12] Y. Nie, J. Zhao, F. Gao, and F. Yu, "Semi-distributed resource management in UAV-aided MEC systems: A multi-agent federated reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13162–13173, Dec. 2021.

[13] P. Luong, F. Gagnon, L.-N. Tran, and F. Labeau, "Deep reinforcement learning-based resource allocation in cooperative UAV-assisted wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7610–7625, Nov. 2021.

[14] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.

[15] Q. Wu and R. Zhang, "Common throughput maximization in UAV-enabled OFDMA systems with delay consideration," *IEEE Trans. Wireless Commun.*, vol. 66, no. 12, pp. 6614–6627, Dec. 2018.

[16] M. S. Ali, E. Hossain, A. Al-Dweik, and D. I. Kim, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3982–3998, Sep. 2018.

[17] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3177–3192, Oct. 2021.

[18] M. Hua, Y. Wang, M. Lin, C. Li, Y. Huang, and L. Yang, "Joint CoMP transmission for UAV-aided cognitive satellite terrestrial networks," *IEEE Access*, vol. 7, pp. 14959–14968, 2019.

[19] A. Kilzi, J. Farah, C. A. Nour, and C. Douillard, "Analysis of drone placement strategies for complete interference cancellation in two-cell NOMA CoMP systems," *IEEE Access*, vol. 8, pp. 179055–179069, 2020.

[20] Y. Yang, M. Chen, C. Guo, C. Feng, and W. Saad, "Power efficient visible light communication with unmanned aerial vehicles," *IEEE Commun. Lett.*, vol. 23, no. 7, pp. 1272–1275, Jul. 2019.

[21] L. Xie, J. Xu, and Y. Zeng, "Common throughput maximization for UAV-enabled interference channel with wireless powered communications," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3197–3212, May 2020.

[22] M. Kashef, M. Ismail, M. Abdallah, K. A. Qaraqe, and E. Serpedin, "Energy efficient resource allocation for mixed RF/VLC heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 883–893, Apr. 2016.

[23] J. Kong, Z.-Y. Wu, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Q-learning based two-timescale power allocation for multi-homing hybrid RF/VLC networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 443–447, Apr. 2020.

[24] S. Ma, F. Zhang, H. Li, F. Zhou, M.-S. Alouini, and S. Li, "Aggregated VLC-RF systems: Achievable rates, optimal power allocation, and energy efficiency," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7265–7278, Nov. 2020.

[25] V. K. Papanikolaou, P. D. Diamantoulakis, P. C. Sofotasios, S. Muhaidat, and G. K. Karagiannidis, "On optimal resource allocation for hybrid VLC/RF networks with common backhaul," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 1, pp. 352–365, Mar. 2020.

[26] M. Obeed, A. M. Salhab, S. A. Zummo, and M.-S. Alouini, "Joint optimization of power allocation and load balancing for hybrid VLC/RF networks," *J. Opt. Commun. Netw.*, vol. 10, no. 5, pp. 553–562, May 2018.

[27] J. Dai, K. Niu, and J. Lin, "Code-domain non-orthogonal multiple access for visible light communications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[28] J. Tang, J. Song, J. Ou, J. Luo, X. Zhang, and K.-K. Wong, "Minimum throughput maximization for multi-UAV enabled WPCN: A deep reinforcement learning method," *IEEE Access*, vol. 8, pp. 9124–9132, 2020.

[29] T. Zhang, Z. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Caching placement and resource allocation for cache-enabling UAV NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12897–12911, Nov. 2020.

[30] Q.-V. Pham, T. Huynh-The, M. Alazab, J. Zhao, and W.-J. Hwang, "Sum-rate maximization for UAV-assisted visible light communications using NOMA: Swarm intelligence meets machine learning," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10375–10387, Oct. 2020.

[31] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and A. Nallanathan, "Deep reinforcement learning based dynamic trajectory control for UAV-assisted mobile edge computing," *IEEE Trans. Mobile Comput.*, vol. 21, no. 10, pp. 3536–3550, Oct. 2022.

[32] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[33] H. Peng and X. Shen, "Multi-agent reinforcement learning based resource management in MEC-and UAV-assisted vehicular networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 131–141, Jan. 2021.

[34] F. Tang, Y. Zhou, and N. Kato, "Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G HetNet," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2773–2782, Dec. 2020.

[35] H. Ren *et al.*, "Performance improvement of M-QAM OFDM-NOMA visible light communication systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[36] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, "Optimal 3D-trajectory design and resource allocation for solar-powered UAV communication systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4281–4298, Jun. 2019.

[37] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn.*, Bali, Indonesia, 2018, pp. 1587–1596.

[38] H. Zhang, N. Liu, K. Long, J. Cheng, V. C. M. Leung, and L. Hanzo, "Energy efficient subchannel and power allocation for software-defined heterogeneous VLC and RF networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 658–670, Mar. 2018.

[39] Y. Wang, M. Chen, Z. Yang, X. Hao, T. Luo, and W. Saad, "Gated recurrent units learning for optimal deployment of visible light communications enabled UAVs," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[40] M. Parvini, M. Reza Javan, N. Mokari, B. Abbasi, and E. A. Jorswieck, "AoI-aware resource allocation for platoon-based C-V2X networks via multi-agent multi-task reinforcement learning," 2021, *arXiv:2105.04196*.

[41] O. Aydin, E. A. Jorswieck, D. Aziz, and A. Zappone, "Energy-spectral efficiency tradeoffs in 5G multi-operator networks with heterogeneous constraints," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5869–5881, Sep. 2017.

[42] Z. Xiang, F. Gabriel, E. Urbano, G. T. Nguyen, M. Reisslein, and F. H. P. Fitzek, "Reducing latency in virtual machines: Enabling tactile internet for human-machine co-working," *IEEE J. Sel. Areas Commun.*, vol. SAC-37, no. 5, pp. 1098–1116, May 019.

[43] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2042–2055, Jul. 2021.

[44] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[45] M. R. Maleki, M. Robat Mili, M. R. Javan, N. Mokari, and E. A. Jorswieck, "Multi agent reinforcement learning trajectory design and two-stage resource management in CoMP UAV VLC networks," *IEEE Dataport*, to be published, doi: 10.21227/4tg2-f112.

[46] Z. Zhu, S. Wan, P. Fan, and K. B. Letaief, "Federated multiagent actor–critic learning for age sensitive mobile-edge computing," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1053–1067, Jan. 2022.

[47] A. Subekti, H. F. Pardede, and R. Sustika, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. 35th Int. Conf. Mach. Learn.*, Bali, Indonesia, Dec. 2018, pp. 4295–4304.

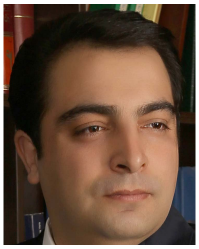
[48] L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, "Graph-based resource sharing in vehicular communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4579–4592, Jul. 2018.

[49] J. Xu, Y. Zeng, and R. Zhang, "UAV-enabled multiuser wireless power transfer: Trajectory design and energy optimization," in *Proc. 23rd Asia–Pacific Conf. Commun. (APCC)*, Dec. 2017, pp. 1–6.

Mohammad Reza Maleki received the B.Sc. degree in electrical engineering from Tabriz University and the M.Sc. degree in electrical engineering from Tarbiat Moders University, Tehran, Iran. In his current position, he is working as a Research Assistant with Tarbiat Modares University. Predominantly, his research interests are in wireless communications, radio resource allocation, UAV trajectory design, and VLC networks.

Mohammad Robat Mili received the Ph.D. degree in electrical and electronic engineering from the University of Manchester, U.K., in 2012. He held postdoctoral research positions at the Department of Telecommunications and Information Processing, Ghent University, Belgium and the Department of Electrical Engineering, Sharif University of Technology, Iran. His main research interests are in the area of design and analysis of wireless communication networks with particular focus on 5G and 6G cellular networks using mathematical methods such as optimization theory, game theory, and machine learning

Mohammad Reza Javan (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, in 2003, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2006, and the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, in 2013. At present, he is a Faculty Member of the Department of Electrical Engineering, Shahrood University, Shahrood, Iran. His research interests include design and analysis of wireless communication networks with emphasis on the application of optimization theory.



Nader Mokari (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 2014. He joined the Department of Electrical and Computer Engineering, Tarbiat Modares University, as an Assistant Professor, in October 2015. He has been elected as an IEEE Exemplary Reviewer in 2016 by IEEE Communications Society. He is currently an Associated Professor with the Department of Electrical and Computer Engineering, Tarbiat Modares University. His research interests cover many aspects of wireless technologies with a special emphasis on wireless networks. In recent years, his research has been funded by Iranian Mobile Telecommunication Companies, Iranian National Science Foundation (INSF). His thesis received the IEEE Outstanding Ph.D. Thesis Award. He received the Best Paper Award at ITU K-2020. He was also involved in a number of large scale network design and consulting projects in the telecom industry. He is on the Editorial Board of the IEEE TRANSACTIONS ON COMMUNICATIONS.

Eduard A. Jorswieck (Fellow, IEEE) is currently the Managing Director of the Institute of Communications Technology, the Head of the Chair for Communications Systems, and a Full Professor at Technische Universitaet Braunschweig, Brunswick, Germany. From 2008 to 2019, he was the Head of the Chair of Communications Theory and a Full Professor at TU Dresden, Germany. His main research interests are in the broad area of communications. He has published some 150 journal papers, 15 book chapters, three monographs, and 300 conference papers on these topics. In 2006, he received the IEEE Signal Processing Society Best Paper Award. Since 2017, he serves as an Editor-in-Chief of the EURASIP JOURNAL ON WIRELESS COMMUNICATIONS AND NETWORKING. He currently serves as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS. He has served on the Editorial Boards for IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE SIGNAL PROCESSING LETTERS, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.