Multiplexing eMBB and mMTC Services over Aerial Visible Light Communications

Hosein Zarini[†], Mohammad Reza Maleki[⋆], Narges Gholipoor[⋆], Mohammad Robat Mili[§], Mehdi Rasti[††], Ali Movaghar[†], Derrik Wing Kwan Ng[⋆⋆], and Ekram Hossain[§§]

[†]Dept. of Computer Engineering, Sharif University of Technology, Tehran, Iran
[⋆]Dept. of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
[§]Pasargad Institute for Advanced Innovative Solutions (PIAIS), Tehran, Iran
[††]Centre for Wireless Communications, University of Oulu, Oulu, Finland
[⋆⋆]School of Electrical Engineering and Telecommunications, University of New South Wales, NSW, Australia
[§§]Dept. of Electrical and Computer Engineering, University of Manitoba, Canada

***Abstract*—Downlink transmission of non-orthogonal multiple access visible light communication systems empowered by an unmanned aerial vehicle (UAV) is considered for multiplexing enhanced mobile broadband (eMBB) and massive machine type communication (mMTC) services. Accordingly, a resource allocation problem of joint transmit power control and motion trajectory design of the UAVs is formulated, whose goal is to characterize a multi-objective trade-off as a weighted sum of the UAVs' power consumption and the perceived quality-of-experience (QoE) of eMBB users, while ensuring the eMBB and mMTC service-specific requirements. We leverage an alternative decomposition and tools from convex optimization and actor-critic multi-agent deep reinforcement learning to address this problem in an iterative fashion. We analytically derive the upper- and lower-bounds on the reward of the UAVs as the learning agents and demonstrate that the proposed resource allocation method outperforms the similar scheme in literature, by up to 17% average reduced power consumption, as well as 12% average perceived QoE gain.**

***Index Terms*—Visible light communication (VLC), unmanned aerial vehicle (UAV), enhanced mobile broadband (eMBB), massive machine type communication (mMTC).**

I. Introduction

Wireless cellular systems are gradually evolving to satisfy diverse heterogeneous service requirements. Massive machine type communication (mMTC) is one of the key services defined in the fifth-generation (5G) of wireless networks, aiming at connecting a large number of internet of things (IoT) devices simultaneously. The current generation of wireless networks though is incapable of supporting this service regarding the ever-growing number of connected terminals [1]. On the other hand, with respect to the noticeable popularization in multimedia services, a majority of universal mobile data traffic belongs to real-time video streaming [2]. Despite significant advancements in wireless communication technologies, satisfying the capacity requirement for multimedia services (categorized mainly in enhanced mobile broadband (eMBB) service) is still an undeniable challenge. Furthermore, the evaluation metrics for multimedia services are generally different from those of conventional services. More specifically, different from the conventional quality-of-service (QoS) metrics that analyze the overall or average network efficiency, the notion of quality-of-experience (QoE) is a user-centric metric, used for evaluating the performance of multimedia services. By definition, QoE is a service-based subjective performance measure for capturing the quality of perceived multimedia services from the end users' perspective [3].

In recent years, coexisting the eMBB and the mMTC services have been attended in research on 5G and beyond. For instance, the authors in [4] leveraged a radio access network with non-orthogonal multiple access (NOMA) to realize this coexistence. In [5], an unmanned aerial vehicle (UAV)-assisted cognitive radio network is proposed with mobile edge computing to support this coexistence, whereas the authors in [6] investigated the potentials of 5G dense small cells to this end. Further investigations reveal that despite the pivoting role of high-frequency bands in the next generation of wireless networks, the coexistence of mMTC and eMBB services in literature has never gone beyond the conventional sub-6GHz band thus far.

In this regard, visible light communication (VLC) is a promising technology with abundant spectral resources [7] for provisioning of eMBB service. Additionally, due to serving both illumination and communication purposes at the same time, VLC networks have been exploited by UAVs. Thanks to the great maneuverability and wide aerial coverage of UAVs as servicing flying base stations (BSs), a UAV-enabled VLC network based on NOMA is capable of covering a large number of ground users as a massive access communication system [8] or mMTC service. Therefore, an aerial (i.e., UAV-assisted) VLC system turns out to be of great promise for supporting the coexistence of mMTC and eMBB services.

In this paper, we propose a wireless resource allocation framework for coexisting the eMBB and mMTC services. Unlike the previous art [4]–[6], which considers the sub-6GHz band for coexisting eMBB and mMTC services, we consider a UAV-aided VLC system with NOMA for this purpose. In particular, the main contributions of this paper are as follows.

- We propose a resource allocation framework for multiplexing the eMBB and mMTC services in a UAV-assisted

VLC system with NOMA. To this goal, an optimization problem is formulated that jointly optimizes the transmit power and the motion trajectory of the UAVs. Regarding the limited energy storage of the UAVs, as well as the stringent data rate requirement of the eMBB service, a multi-objective function is proposed to characterize the trade-off for each UAV as a weighted sum of its power consumption and the perceived QoE of its eMBB users.

- An iterative solution methodology is invoked to address the resource allocation problem, wherein a convex optimization method is leveraged to control the UAVs' transmit power and an actor-critic multi-agent deep reinforcement learning (MA-DRL) model with prioritized experience replay optimizes the UAVs' motion trajectory.
- Simulations demonstrate an efficient trade-off that balances the UAVs' power consumption and the perceived QoE of its supported eMBB users. The upper- and lower-bounds for the reward of each UAV as a learning agent are analytically derived. It is as well shown that the proposed method achieves up to 17% average reduced average power consumption, as well as 12% average perceived QoE gain over the similar scheme in literature.

The remainder of this paper is organized as follows: Section II describes the system setup, assumptions, and problem statement. In Section III, the power control policy is elaborated, whereas the trajectory design policy is discussed in Section IV. Eventually, simulation results, complexity analysis, and conclusions are presented in Sections V, VI, and VII, respectively.

II. System Setup, Assumptions, and Problem Statement

A. Network Structure

An aerial-assisted VLC system in downlink transmission is considered, where randomly distributed single-antenna photodiode-equipped ground users are served by a set $\mathcal{G}$ of single-antenna UAVs. The UAVs are assumed to have the channel state information (CSI) on line-of-sight (LoS) VLC links [7]. We consider a set $\mathcal{U} = \{\mathcal{U}^E, \mathcal{U}^M\}$ of ground users, where $\mathcal{U}^E$ and $\mathcal{U}^M$ denote the set of eMBB and mMTC users, respectively. All the UAVs are supposed to fly at a fixed hovering altitude, with a maximum speed of $V^{\max}$ to serve ground users with diverse service requirements. Unlike [7], we assume that the illuminated area covered by UAVs are overlapping, and therefore, users are expected to experience inter-UAV interference. The LoS VLC channel gain between the UAV g and the ith user is denoted by $h_{g,i}$ and given by [9]: $h_{g,i} = \frac{(m+1)A^{\text{VLC}}}{2\pi d_{g,i}^2} G^{\text{VLC}}(\psi_{g,i})\cos^m(\phi_{g,i})\cos(\psi_{g,i})$ when $0 \le \psi_{g,i} \le \Psi_c$ and zero, otherwise, with A^{VLC} being the detector area of the user photodiode, $d_{g,i}$ denoting the distance between the ith user and the gth UAV, as well as the Lambert order $m = -\frac{\ln 2}{\ln(\cos\Phi_{1/2})}$ on a semi-angle of $\Phi_{1/2}$. In addition, $\psi_{g,i}$ stands for the angle of incidence, $\phi_{g,i}$ is the angle of irradiance, Ψ_c denotes the receiver field of vision (FOV) semi-angle and $G^{\text{VLC}}(\psi_{g,i})$ indicates the gain of the optical concentrator that is given by: $G^{\text{VLC}}(\psi_{g,i})$ when $0 \le \psi_{g,i} \le \Psi_c$, and zero, otherwise, where $n_R \ge 0$ is the refractive index. Given $(\dot{x}_i, \dot{y}_i)$ as the x- and y-coordinates of the ith user, the distance between this user and the gth UAV is modeled as $d_{g,i} = \sqrt{(x_g - \dot{x}_i)^2 + (y_g - \dot{y}_i)^2 + H_g^A}$ with x_g, y_g, and H_g^A, respectively, denoting the x- and y-coordinates, as well as the hovering altitude for the gth UAV. Hence, the channel gain between the UAV g and the users can be denoted by $\mathbf{h}_g = \{h_{g,1}, h_{g,2}, ..., h_{g,|\mathcal{U}|}\}$.

B. eMBB Service

Let Υ_g be the set of descending-ordered channel gains related to all the eMBB and mMTC users, multiplexed upon the UAV g, with $\Upsilon_g(i) = h_{g,i}$. Then the received signal at the ith eMBB user from the gth UAV will be expressed as: $\mathbf{y}_{g,i}^E = \iota^e \iota^r \Big(h_{g,i}\sqrt{p_{g,i}} s_i + \sum_{\substack{i' \in \mathcal{U}^E, i \neq i', \\ \Upsilon_g(i') > \Upsilon_g(i)}} h_{g,i}\sqrt{p_{g,i'}} s_{i'} + \sum_{\substack{i'' \in \mathcal{U}^M, i \neq i'', \\ \Upsilon_g(i'') > \Upsilon_g(i)}} h_{g,i}\sqrt{p_{g,i''}} s_{i''} + \mathcal{I}_{\text{I-UAV}} \Big) + \mathbf{n}_i^E, \forall g \in \mathcal{G}, i \in \mathcal{U}^E$, where $p_{g,i}$ indicates the transmit power the of the UAV g to the ith user, s_i represents the unit normalized transmit symbol of this user, $\mathbf{n}_i$ stands for the complex additive white Gaussian noise (AWGN) with unit variance $(\sigma_i^E)^2$ at this user and $\mathcal{I}_{\text{I-UAV}} = \iota^e \iota^r \Big(\sum_{\substack{g' \neq g, \\ g' \in \mathcal{G}}} \sum_{\substack{i_1 \neq i, \\ i_1 \in \mathcal{U}^E}} h_{g',i_1}\sqrt{p_{g',i_1}} s_{i_1} + \sum_{\substack{g' \neq g, \\ g' \in \mathcal{G}}} \sum_{\substack{i_2 \neq i, \\ i_2 \in \mathcal{U}^M}} h_{g',i_2}\sqrt{p_{g',i_2}} s_{i_2} \Big)$, represents the inter-UAV interference. Besides, ι^r and ι are the responsivity of the photodiode and the electrical-to-optical conversion coefficient, respectively. In such a superpositioning model, a typical eMBB user decodes its desired signal by canceling the prior eMBB/mMTC decoded signals (related to precedent eMBB/mMTC users in Υ_g) from its observation and treating the remaining signals in Υ_g, as the intra-service interference (corresponded to posterior eMBB users in Υ_g) and the inter-service interference (corresponded to posterior mMTC users in Υ_g). Therefore, the received SINR for the ith eMBB user can be characterized as $\gamma_{g,i}^E = \frac{2(\iota^e\iota^r)^2 (|h_{g,i}|^2 p_{g,i})}{\pi e \left[\frac{(\iota^e\iota^r)^2}{3} (\mathcal{I}_{\text{Intra}}^E + \mathcal{I}_{\text{Inter}}^E + \mathcal{I}_{\text{I-UAV}}) + (\sigma_i^E)^2 \right]}$, $\forall g \in \mathcal{G}, i \in \mathcal{U}^E$, where $\mathcal{I}_{\text{Intra}}^E = \sum_{\substack{i' \in \mathcal{U}^E, i \neq i', \\ \Upsilon_g(i') > \Upsilon_g(i)}} |h_{g,i}|^2 p_{g,i'}, \forall i \in \mathcal{U}^E$, and $\mathcal{I}_{\text{Inter}}^E = \sum_{\substack{i'' \in \mathcal{U}^M, i \neq i'', \\ \Upsilon_g(i'') > \Upsilon_g(i)}} |h_{g,i}|^2 p_{g,i''}, \forall i \in \mathcal{U}^M$. Relying on Shannon-Hartley theorem, the achievable rate of this user can be calculated as [7]: $R_i^E = \frac{1}{2} \sum_{g=1}^{|\mathcal{G}|} \log_2(1 + \gamma_{g,i}^E)$.

The eMBB users are supposed to be of video-service requirement, whose QoE is modeled according to the mean opinion score (MOS) [2]. To be more specific, the achievable MOS for a typical eMBB user i, requesting a H.264/AVC online video streaming is given by [10]:

$$\text{MOS}_i = \begin{cases} 4.5, & \text{PSNR}_i > C_{4.5}^{\text{MoS}}, \\ C_1^{\text{QoE}} \log_2(\text{PSNR}_i) + C_2^{\text{QoE}}, & C_1^{\text{MoS}} \le \text{PSNR}_i \le C_{4.5}^{\text{MoS}}, \\ 1, & \text{PSNR}_i < C_1^{\text{MoS}}, \end{cases}$$

where C_1^{QoE} and C_2^{QoE} are constants, respectively, expressed as $C_1^{\text{QoE}} = \frac{3.5}{\log_2\left(C_{4.5}^{\text{MoS}}\right) - \log_2\left(C_1^{\text{MoS}}\right)}$, and $C_2^{\text{QoE}} =$

$\frac{\log_2\left(C_{4.5}^{\text{MoS}}\right)-4.5\log_2\left(C_1^{\text{MoS}}\right)}{\log_2\left(C_{4.5}^{\text{MoS}}\right)-\log_2\left(C_1^{\text{MoS}}\right)}$. Besides, PSNR_i denotes the peak-signal-to-noise-ratio (PSNR) for this user and represented as:

$$\text{PSNR}_i = C_3^{\text{QoE}} + C_4^{\text{QoE}}\sqrt{R_i^E/C_5^{\text{QoE}}}\left(1-C_5^{\text{QoE}}/R_i^E\right), \forall i \in \mathcal{U}^E, \quad (1)$$

where the constants C_3^{QoE}-C_5^{QoE} are mainly determined by the adopted video coding scheme. Further, C_1^{MoS} and $C_{4.5}^{\text{MoS}}$ correspond to the minimum acceptable, as well as the maximum achievable MOS [11], respectively. Finally, each eMBB user i is required to satisfy a minimum MOS threshold $\text{MOS}_i^{\min}$ as

$$\text{MOS}_i = C_1^{\text{QoE}}\log_2\left(C_3^{\text{QoE}} + C_4^{\text{QoE}}\sqrt{\frac{BR_i^E}{C_5^{\text{QoE}}}}\left(1-\frac{C_5^{\text{QoE}}}{BR_i^E}\right)\right) + C_2^{\text{QoE}} \geq \text{MOS}_i^{\min}, \forall i \in \mathcal{U}^E. \quad (2)$$

C. mMTC Service

Under the described superpositioning model, the received signal at the ith mMTC user from the gth UAV is given by: $\mathbf{y}_{g,i}^M = \iota^e\iota^r\left(h_{g,i}\sqrt{p_{g,i}}s_i + \sum_{\substack{i'\in\mathcal{U}^M, i\neq i',\\ \Upsilon_g(i')>\Upsilon_g(i)}} h_{g,i}\sqrt{p_{g,i'}}s_{i'} + \sum_{\substack{i''\in\mathcal{U}^E, i\neq i'',\\ \Upsilon_g(i'')>\Upsilon_g(i)}} h_{g,i}\sqrt{p_{g,i''}}s_{i''} + \mathcal{I}_{\text{I-UAV}}\right) + \mathbf{n}_i^M, \forall g \in \mathcal{G}, i \in \mathcal{U}^M$, where $\mathbf{n}_i^M$ indicates the AWGN with unit variance $(\sigma_i^M)^2$. Accordingly, the received SINR for this user from the gth UAV can be posed as: $\gamma_{g,i}^M = \frac{2(\iota^e\iota^r)^2\left(|h_{g,i}|^2 p_{g,i}\right)}{\pi e\left[\frac{(\iota^e\iota^r)^2}{3}\left(\mathcal{I}_{\text{Intra}}^M + \mathcal{I}_{\text{Inter}}^M + \mathcal{I}_{\text{I-UAV}}\right) + \left(\sigma_i^M\right)^2\right]}, \forall g \in \mathcal{G}, i \in \mathcal{U}^M$, where $\mathcal{I}_{\text{Intra}}^M = \sum_{\substack{i'\in\mathcal{U}^M, i\neq i',\\ \Upsilon_g(i')>\Upsilon_g(i)}} |h_{g,i}|^2 p_{g,i'}, \forall g \in \mathcal{G}, i \in \mathcal{U}^M$, and $\mathcal{I}_{\text{Inter}}^M = \sum_{\substack{i''\in\mathcal{U}^E, i\neq i'',\\ \Upsilon_g(i'')>\Upsilon_g(i)}} |h_{g,i}|^2 p_{g,i''}, \forall g \in \mathcal{G}, i \in \mathcal{U}^M$, denote the intra-service and inter-service interference imposed to this user, respectively. Accordingly, the minimum QoS guarantee constraint for this user can be stated as: $R_i^M = \frac{1}{2}\sum_{g=1}^{|\mathcal{G}|}\log_2(1+\gamma_{g,i}^M) \geq R_i^{\min}, \forall i \in \mathcal{U}^M$.

D. Problem Formulation

We define a multi-objective trade-off as a weighted sum of the consumed power by the UAV g and the perceived QoE of its served eMBB users. This trade-off is formally expressed as: $\Lambda_g = \nu_g \frac{\sum_{i=1}^{\mathcal{U}^E}\varrho_{g,i}^E \text{MOS}_i}{\sum_{i=1}^{\mathcal{U}^E} 4.5\varrho_{g,i}^E} - (1-\nu_g)\mathbf{p}_g/P^{\max}, \forall g \in \mathcal{G}$, where $\nu_g \in (0,1)$ is a weighting factor for linear scalarization, $\varrho_{g,i}^E$ ($\varrho_{g,i}^M$) is the predetermined UAV-user association profile, indicating that the ith eMBB (mMTC) user is associated with the g UAV, and finally, $\mathbf{p}_g = \sum_{i\in\mathcal{U}^E}\sum_{i\in\mathcal{U}^M}\varrho_{g,i}^E\varrho_{g,i}^M p_{g,i}$. Accordingly, the optimization problem of transmit power control and motion trajectory design of the UAVs, aimed at maximizing Λ_g with respect to eMBB and mMTC service-specific constraints can be posed as follows:

$$\begin{aligned}
\mathcal{P}_{\text{Joint}}: \max_{\{\boldsymbol{x},\boldsymbol{y},\boldsymbol{P}\}} & \sum_{g=1}^{|\mathcal{G}|}\Lambda_g \\
\text{s.t. } & \text{MOS}_i \geq \text{MOS}_i^{\min}, \forall i \in \mathcal{U}^E, && (3a)\\
& R_i \geq R_i^{\min}, \forall i \in \mathcal{U}^M, && (3b)\\
& (x_g^{t+1}-x_g^t)^2 + (y_g^{t+1}-y_g^t)^2 \leq (V^{\max})^2, && (3c)\\
& p_{g,i} \geq 0 \ \forall i \in \mathcal{U}^E, \mathcal{U}^M, && (3d)\\
& \sum_{i\in\mathcal{U}^E}\sum_{i\in\mathcal{U}^M} p_{g,i} \leq P^{\max}. && (3e)
\end{aligned}$$

While (3c) is in quadratic form and thus convex [7], the optimization problem (3), nonetheless, is non-convex due to the intractable form of its objective function and the constraints (3a)-(3b). Hence, it is extremely complicated to obtain its globally optimal solution. In what follows, we propose a sub-optimal solution to address this problem in an iterative fashion.

III. Power Control Policy

Given the motion trajectory of the UAVs, the power control sub-problem can be stated as follows:

$$\begin{aligned}
\mathcal{P}_{\text{Power}}: \max_{\{\boldsymbol{P}\}} & \sum_{g=1}^{|\mathcal{G}|}\Lambda_g \\
\text{s.t. } & \text{(3a)-(3b), (3d)-(3e)}, && (4a)
\end{aligned}$$

which is still non-convex similar to (3). In following, we propose convex transformations to address this sub-problem in a computationally efficient manner.

First of all, each user i (regardless of its service requirement) is supposed to be associated with only one UAV, whose index from now on, is shown by g_i^*. After simple manipulations, (3b) can be restated as:

$$\begin{aligned}
&\frac{2(\iota^e\iota^r)^2\left(|h_{g,i}|^2 p_{g_i^*,i}\right)}{2^{R_i^{\min}/B}-1} - \pi e\frac{(\iota^e\iota^r)^2}{3}\Big(\sum_{\substack{i'\in\mathcal{U}^M, i\neq i',\\ \Upsilon_{g_i^*}(i')>\Upsilon_{g_i^*}(i)}} |h_{g_i^*,i}|^2 p_{g_i^*,i'} \\
&+\mathcal{I}_{\text{I-UAV}} + \sum_{\substack{i''\in\mathcal{U}^E, i\neq i'',\\ \Upsilon_{g_i^*}(i'')>\Upsilon_{g_i^*}(i)}} |h_{g_i^*,i}|^2 p_{g_i^*,i''}\Big) \geq \pi e\left(\sigma_i^M\right)^2, \forall i \in \mathcal{U}^M, g_i^* \in \mathcal{G},
\end{aligned} \quad (5)$$

which is affine and linear, so convex in $\boldsymbol{P}$. Also, let us restate (3a) as: $C_4^{\text{QoE}}\sqrt{BR_i^E/C_5^{\text{QoE}}}\left(1-C_5^{\text{QoE}}/BR_i^E\right) \geq X_i, \ \forall i \in \mathcal{U}^E$, with $X_i = 10^{\left\lceil\frac{(\text{MOS}_i^{\min}-C_2^{\text{QoE}})}{C_1^{\text{QoE}}}\right\rceil} - C_3^{\text{QoE}}, \ \forall i \in \mathcal{U}^E$. By further manipulations, it can be concluded that [11]:

$$BR_i^E - \frac{\sqrt{BR_i^E C_5^{\text{QoE}}}}{C_4^{\text{QoE}}}X_i - C_5^{\text{QoE}} \geq 0, \ \forall i \in \mathcal{U}^E, \quad (6)$$

where expanding R_i^E, results in

$$2(\iota^e\iota^r)^2\big(|h_{g_i^*,i}|^2 p_{g_i^*,i}\big)\Gamma^{-1}_{g_i^*,i} - \pi e\frac{(\iota^e\iota^r)^2}{3}\Bigg(\sum_{\substack{i'\in\mathcal{U}^M, i\neq i',\\ \Upsilon_{g_i^*}(i')>\Upsilon_{g_i^*}(i)}} |h_{g_i^*,i}|^2 p_{g_i^*,i'} + \mathcal{I}_{\text{I-UAV}} + \sum_{\substack{i''\in\mathcal{U}^E, i\neq i'',\\ \Upsilon_{g_i^*}(i'')>\Upsilon_{g_i^*}(i)}} |h_{g_i^*,i}|^2 p_{g_i^*,i''}\Bigg) \geq \pi e\big(\sigma_i^E\big)^2, \forall i\in\mathcal{U}^E, g_i^*\in\mathcal{G}, \quad (7)$$

with $\Gamma_{g_i^*,i} = 2\Bigg[\dfrac{\frac{\sqrt{BC_5^{\text{QoE}}}}{C_4^{\text{QoE}}}X_i + \sqrt{\frac{BC_5^{\text{QoE}}X_i^2}{(C_4^{\text{QoE}})^2} + 4BC_5^{\text{QoE}}}}{2B}\Bigg]^2 - 1, \forall i\in\mathcal{U}^E,\ \forall g_i^*\in\mathcal{G}$. Now we can redefine $\mathcal{P}_{\text{Power}}$ as follows:

$$\mathcal{P}'_{\text{Power}}: \max_{\{\boldsymbol{P}\}} \sum_{g=1}^{|\mathcal{G}|}\Lambda_g$$
$$\text{s.t. } \gamma^E_{g_i^*,i} \geq \Gamma_{g_i^*,i}, \forall i\in\mathcal{U}^E, g_i^*\in\mathcal{G}, \quad (8a)$$
$$(5),\ (7),\ (3d)\text{-}(3e), \quad (8b)$$

wherein (8a) is affine, yet linear-fractional w.r.t. $\boldsymbol{P}$. Let us declare auxiliary variables $\boldsymbol{\zeta}^{\{1\}}\in\mathbb{R}^{\mathcal{U}^E}$, $\boldsymbol{\zeta}^{\{2\}}\in\mathbb{R}^{\mathcal{U}^E}$ and $\boldsymbol{\zeta}^{\{3\}}\in\mathbb{R}^{\mathcal{U}^E}$ as:

$$C_3^{\text{QoE}} + C_4^{\text{QoE}}\sqrt{BR_i^E/C_5^{\text{QoE}}}\left(1 - C_5^{\text{QoE}}/BR_i^E\right) \geq \zeta_i^{\{1\}}, \forall i\in\mathcal{U}^E, \quad (9)$$

$BR_i^E \geq \zeta_i^{\{2\}}, \forall i\in\mathcal{U}^E$, and $\zeta_i^{\{3\}} = 2^{\zeta_i^{\{2\}}/B},\ \forall i\in\mathcal{U}^E$, respectively. By applying a linear transformation on (8a), $\mathcal{P}'_{\text{Power}}$ can be equivalently restated as following:

$$\mathcal{P}''_{\text{Power}}: \max_{\{\boldsymbol{P},\boldsymbol{\zeta}^{\{1\}},\boldsymbol{\zeta}^{\{2\}},\boldsymbol{\zeta}^{\{3\}}\}} \sum_{g=1}^{|\mathcal{G}|}\Xi_g$$
$$\text{s.t. } C_3^{\text{QoE}} + C_4^{\text{QoE}}\sqrt{\frac{\zeta_i^{\{2\}}}{C_5^{\text{QoE}}}}\left(1 - \frac{C_5^{\text{QoE}}}{\zeta_i^{\{2\}}}\right) \geq \zeta_i^{\{1\}}, \forall i\in\mathcal{U}^E, \quad (10a)$$
$$2(\iota^e\iota^r)^2\big(|h_{g_i^*,i}|^2 p_{g_i^*,i}\big)\big(\zeta_i^{\{3\}} - 1\big)^{-1} - \pi e\frac{(\iota^e\iota^r)^2}{3}\Bigg(\sum_{\substack{i'\in\mathcal{U}^M, i\neq i',\\ \Upsilon_{g_i^*}(i')>\Upsilon_{g_i^*}(i)}} |h_{g_i^*,i}|^2 p_{g_i^*,i'} + \mathcal{I}_{\text{I-UAV}} + \sum_{\substack{i''\in\mathcal{U}^E, i\neq i'',\\ \Upsilon_{g_i^*}(i'')>\Upsilon_{g_i^*}(i)}} |h_{g_i^*,i}|^2 p_{g_i^*,i''}\Bigg) \geq \pi e\big(\sigma_i^E\big)^2, \forall i\in\mathcal{U}^E, g_i^*\in\mathcal{G}, \quad (10b)$$
$$(5),\ (7),\ (3d)\text{-}(3e), \quad (10c)$$

where $\Xi_g = \nu_g\dfrac{\sum_{i=1}^{\mathcal{U}^E}\varrho^E_{g,i}\left[C_1^{\text{QoE}}\log_2\left(\zeta_i^{\{1\}}\right) + C_2^{\text{QoE}}\right]}{\sum_{i=1}^{\mathcal{U}^E}4.5\varrho^E_{g,i}} - (1-\nu_g)\mathbf{p}_g/P^{\max}\forall g\in\mathcal{G}$. By decoupling $\boldsymbol{P}$ and $\{\boldsymbol{\zeta}^{\{1\}},\boldsymbol{\zeta}^{\{2\}},\boldsymbol{\zeta}^{\{3\}}\}$ in an iterative manner, one can conclude that $\mathcal{P}''_{\text{Power}}$ is convex due to convexity of its objective function, as well as all the constraints within. This sub-problem can be efficiently solved through existing off-the-shelf optimization software packages like CVX [12].

IV. Trajectory Design Policy

Given the transmit power of the UAVs, the trajectory design sub-problem can be formulated as follows:

$$\mathcal{P}_{\text{Trajectory}}: \max_{\{\boldsymbol{x},\boldsymbol{y}\}} \sum_{g=1}^{|\mathcal{G}|}\Lambda_g$$
$$\text{s.t. } (3a)\text{-}(3b), \quad (11a)$$

which is non-convex yet. So, it is quite challenging to obtain its optimal solution using the conventional optimization methods at hand. Additionally, due to the frequent mobility of the UAVs, the trajectory design needs to be optimized in an online manner. For this reason, we propose an actor-critic MA-DRL model with prioritized experience replay, aiming at an online trajectory design for the UAVs. Regarding this model, each UAV g as an agent in specific takes the action $a_g(t)$ towards its policy π_g by observing the wireless system state $s_g(t)$, at instant t from its perspective. The adopted action is then evaluated through the agent's local reward function $\mathcal{R}_g$ a feedback returned by the wireless system. Each agent's policy is periodically updated toward a locally optimal action, based on which, the agent's reward is maximized. The goal in this context is to address the trajectory design sub-problem (11), whereby the state, action and reward of the UAVs are defined as follows:

- **State Space:** At instant t, the state of the agent g is defined as the channel gain of the UAV g at instant $t-1$, i.e., $\mathbf{h}_g(t-1)$.
- **Action Space:** The action adopted by the UAV g at instant t indicates its motion trajectory, i.e., $a(t) = \big[x(t), y(t)\big]$.
- **Reward:** The value returned by the reward function of the UAV g at instant t is categorized in manifold cases as follows:
  1) (11a) holds and $\Lambda_g(t) \geq \Lambda_g^*$; therefore, $\mathcal{R}_g(t) = \Lambda_g(t)$ and $\Lambda_g^* = \Lambda_g(t)$.
  2) (11a) holds, $\Lambda_g(t) \leq \Lambda_g^*$ and $\Lambda_g(t) \geq \Lambda_g(t-1)$; therefore, $\mathcal{R}_g(t) = \Lambda_g(t) - |\Lambda_g(t)|$.
  3) (11a) holds, $\Lambda_g(t) \leq \Lambda_g^*$, $\Lambda_g(t) \leq \Lambda_g(t-1)$; therefore, $\mathcal{R}_g(t) = \Lambda_g(t) - 2|\Lambda_g(t)|$.
  4) (11a) does not hold; therefore. $\mathcal{R}_g(t) = \Lambda_g(t) - 3|\Lambda_g(t)|$.

In all the cases, Λ_g^* stands for the highest Λ_g achieved by the UAV g thus far. By doing so, this UAV is localized such that the most gain on Λ_g is achieved, till the constraints in (11a) are all respected.

In brief, the actor-critic MA-DRL with prioritized experienced replay is composed of the following elements.

1) Actor Network: As an estimator, the actor function f^{Act} performs a mapping from the state space into the action space as $f^{\text{Act}}(\cdot|\theta^{\text{Act}}): \mathcal{S}\rightarrow\mathcal{A}$, with θ^{Act}, $\mathcal{S}$ and $\mathcal{A}$, denoting the parameter set of the actor-network, the state space and the action space of the model, respectively. According to this definition, the output of the actor-network (called $proto-action$) for the agent g with the state s_g is represented by $f^{\text{Act}}(s_g|\theta^{\text{Act}}) = a_g$.

2) Critic Network: Parameterized by θ^{Crt}, the critic network for a typical agent g uses a function for mapping its corresponding pair of state and paroto-action (s_g,a_g) made by the actor-network, into a scalar value in $\mathbb{R}$, expressed as: $f^{\text{Crt}}(\cdot,\cdot|\theta^{\text{Crt}}) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. In fact, the critic function is the same as the Q function in the Q-learning method that evaluates a state-action pair through its Q value.

3) Parameter Optimizer: For better training an MA-DRL model, it is necessary to update its parameters, i.e., the weights and biases of the embedded neural network in agents, towards maximizing the critic network output. For a typical agent g, this output can be declared as $\text{Crt}(\theta^{\text{Act}}) = \mathbb{E}\big[f^{\text{Crt}}\big(s_g, a_g|a_g = f^{\text{Act}}(s_g|\theta^{\text{Act}})\big)\big]$, given a specific state s_g and action a_t. The parameter updating rule for the actor-network of a typical agent g can be declared as

$$-\nabla_{\theta^{\text{Act}}}\text{Crt}(\theta^{\text{Act}}) \approx -\mathbb{E}\big[\nabla_a f^{\text{Crt}}(s_g, a_g)\nabla_{\theta^{\text{Act}}} f^{\text{Act}}(s_g|\theta^{\text{Act}})\big] \approx -\frac{1}{N_B} \sum_{b=1}^{N^b} \nabla_{a_g} f^{\text{Crt}}(s_g, a_g)|_{s_g=s_g^b, a_g=f^{\text{Act}}(s^b|\theta^{\text{Act}})} \nabla_{\theta^{\text{Act}}} f^{\text{Act}}(s_g|\theta^{\text{Act}})|_{s_g=s_g^b},$$

where N_B indicates the number of training batches. Note that like Q-learning method, for a typical agent g with the new state $s(t+1)$, the objective of the critic network for evaluating the agent's action $a_g(t+1)$ adopted by the actor network, can be specified as:$\text{Crt}(\theta^{\text{Act}}) = \mathbb{E}\big[r^{\text{Crt}} + \gamma^{\text{Crt}}\text{max}_{a_g(t+1)} f^{\text{Crt}}\big(s_g(t+1), a_g(t+1)|\theta^{\text{Act}}\big)\big]$. On this basis, the critic network parameters will be updated using the mean square error (MSE) per batch $b \in N^b$, given by $\text{MSE}^b = 1/N_B \sum_{b=1}^{N_B} \Big(\text{Crt}^b(\theta^{\text{Act}}) - f^{\text{Crt}}\big(s_g^b, a_g^b|\theta^{\text{Act}}\big)\Big)^2$. In the DRL model, duplicate networks (called the target critic and target actor networks) are conventionally introduced to original actor and critic networks, whereby the computational stability is assured [13]. The target networks are not trained like their original counterparts, but their parameters need to be updated according to the actor and critic network parameters. Denoted by $\theta^{\text{Act}}_{\text{Dup}}$ and $\theta^{\text{Crt}}_{\text{Dup}}$, the parameters of the target actor and target critic networks, respectively, they can be updated as $\theta^{\text{Act}}_{\text{Dup}} \longleftarrow \varpi_{\text{Dup}}\theta^{\text{Act}} + (1 - \varpi_{\text{Dup}})\theta^{\text{Act}}_{\text{Dup}}$, and $\theta^{\text{Crt}}_{\text{Dup}} \longleftarrow \varpi_{\text{Dup}}\theta^{\text{Crt}} + (1 - \varpi_{\text{Dup}})\theta^{\text{Crt}}_{\text{Dup}}$, with the hyperparameter ϖ_{Dup}, where $0 \leq \varpi_{\text{Dup}} \ll 1$.

4) Prioritized Experience Replay: This feature introduces a queue, whose length is proportional to the number of agents. A typical agent g commonly keeps a tuple of its experiences, as $(s_g(t),\ a_g(t),\ \mathcal{R}_g(t),\ s_g(t+1))$, from which, random mini-batches of experiences are uniformly sampled in replay memory. Conventionally, experiences are uniformly sampled in the replay memory of agents, whose distribution is linear. For promoting the training performance of the embedded neural network at the agents, it is better to adopt the key experiences for training purposes. To this end, each experience needs to be assigned a priority [14], based on which, key experiences are frequently sampled during the training process. The prioritization of rich experiences is characterized by a probability function, whereby the probability of the typical ith experience to be adopted from the replay memory can be expressed as: $\mathcal{P}(i) = \frac{Z_i^\eta}{\sum_i Z_i^\eta}$, where Z_i^η is the priority of the i experience with the prioritization level η. By doing so, key experiences play a pivoting role during the training process and faster training convergence is therefore anticipated [14].

V. Complexity Analysis and Solution Outline

A. Computational Complexity

We use the CVX to solve the convex sub-problem $\mathcal{P}''_{\text{Power}}$ with the computational complexity of: $O_{\text{PC}}\left(\log_2(N_{\text{Cons}})/q_0\psi_{\text{acc}}/log_2(\psi_{\text{step}})\right)$, where N_{Cons} represents the total number of constraints, $0 \leq \psi_{\text{acc}} \leq 1$ denotes the desired accuracy level of the interior point method for convergence with the initial point q_0 and ψ_{step} is the gradient step size. On the other hand, training a MA-DRL network with N_{lays} number of hidden layers to address the trajectory design sub-problem, has a computational complexity of [13]: $O_{\text{MA-DRL}}\left(\left(N_{\text{epi}} \times N_{\text{iter}}\big(|\mathcal{S}| \times N_{\text{lay}} + N_{\text{lay}} \times |\mathcal{A}|\big)\right)\right)$, where N_{epi} and N_{iter} indicate the number of training episodes, as well as the number of iterations per training episode, respectively. More so, $|\mathcal{S}|$ and $|\mathcal{A}|$ indicate the size of state and action space in the DRL network, respectively. Hence, the overall computational complexity of our solution approach can be calculated as: $O_{\text{Total}}\left(L^{\max}\left[\max\left\{O_{\text{MA-DRL}}, O_{\text{PC}}\right\}\right]\right)$. with $L^{\max}$, indicating the number of iterations to converge.

B. Solution Procedure

The proposed iterative procedure converges to a sub-optimal solution for the resource allocation optimization problem (3). There are two stopping criterion for the iterative procedure: the predefined maximum iteration threshold $L^{\max}$, as well as a predetermined convergence error threshold ε. The overall iterative procedure is as follows: $[\boldsymbol{P}^{(0)},\ \boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)}] \rightarrow [\boldsymbol{P}^{(1)}, \boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}] \rightarrow \ ... \ \rightarrow [\boldsymbol{P}(t), \boldsymbol{x}(t), \boldsymbol{y}(t)] \rightarrow \ ... \ \rightarrow [\boldsymbol{P}^{(\text{opt})}, \boldsymbol{x}^{(\text{opt})}, \boldsymbol{y}^{(\text{opt})}]$.

VI. Experimental Results

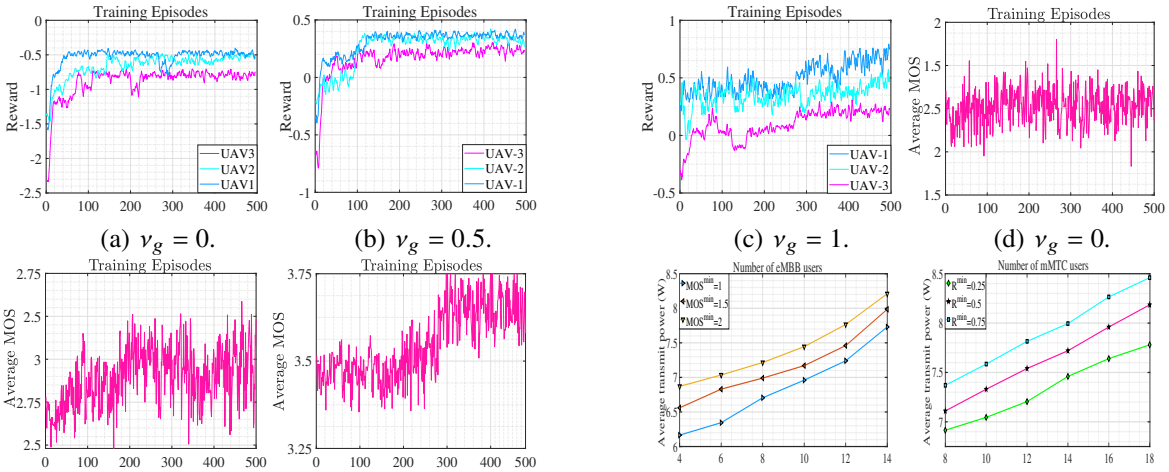
A. Simulation Parameters

The simulation parameters are given in Table I. In particular, we consider 10 UAVs with cylindrical illumination environments, each having a height of 20 m and covering a radius of 50 m. All the UAVs are of the maximum speed of $V^{\max}$ = 10 m/s and have zero acceleration. The QoE parameters are initialized from corresponding parameters in [10]. We assume that there are 8 eMBB users and 15 mMTC users randomly distributed within the coverage areas of the UAVs unless otherwise stated. The service-specific parameters and the information about the multi-agent DRL model are also given in Table I. Note that we have developed our framework using Python and TensorFlow library, as well as Matlab for optimization.

TABLE I: Simulation parameters

Environment parameters	Value	Environment parameters	Value
Field of vision	60°	Area of photo detector	1 cm^2
$P^{\max}$	12 W	Number of UAVs	10
Number of eMBB users	5	Number of mMTC users	15
Half power angle, $\theta_{1/2}$	30°	The order of the Lambertian emission	1
PD responsivity	0.53 A/W	AWGN	10^{-12}
Maximum hovering altitude	20 m	Convergence threshold	10^{-2}
UAV acceleration	0	$R_i^{\min}$	0.5
$V^{\max}$	10	$\text{MOS}_i^{\min}$	1.5
ι	0.58 [9] A/W	ι^r	0.56 A/W [9]

Neural network hyperparameters	Value	Neural network hyperparameters	Value
Experience replay buffer size	100000	Mini batch size	64
Number/size of local actor networks hidden layers	2/2048, 1024	Number/size of local critic networks hidden layers	2/512, 256
Number/size of global critic hidden layers	3/1024, 512, 256	Critic/Actor networks learning rate	0.001/0.0001
Discount factor	0.99	Target networks soft update parameter, τ	0.0005
Number of episodes	500	Number of iterations per episode	100



(a) $\nu_g = 0$. (b) $\nu_g = 0.5$. (c) $\nu_g = 1$. (d) $\nu_g = 0$.

(e) $\nu_g = 0.5$. (f) $\nu_g = 1$. (g) eMBB service. (h) mMTC service.

Fig. 1: Training convergence and system evaluation.

(a) $\nu_g = 0$. (b) $\nu_g = 1$.

Fig. 2: Overall convergence behavior.

B. Simulation Analysis

For better demonstrating the trade-off at Λ_g, we upgrade the value of the weighting factor ν_g with a step size of 0.5. By doing so, we can prioritize the objectives over one another. Fig. (1a)-Fig. (1c) represent the reward function during the learning process of 3 UAVs and they can be upper- and lower-bounded according to the following cases:

- In Fig. (1a), $\nu_g = 0$, each agent only cares about its energy efficiency, and therefore, $\Lambda_g \leq 0$. The upper-bound implies $\mathbf{p}_g \longrightarrow 0$, $\Lambda_g \longrightarrow 0$, and hence, $\mathcal{R}_g \longrightarrow 0$. The lower-bound, on the other hand, implies $\mathbf{p}_g \approx P^{\max}$ and one of the constraints in (11) at least is violated. So, $\Lambda_g \approx -1$ and $\mathcal{R}_g = \Lambda_g - 3|\Lambda_g| \longrightarrow -4$. All in all, $-4 \leq \mathcal{R}_g \leq 0$.
- In Fig. (1b), $\nu_g = 0.5$, both objectives at Λ_g have the same priority, and therefore, $\Lambda_g \leq 0.5$. The upper-bound implies $\text{MOS}_g \approx 4.5$, $\mathbf{p}_g \longrightarrow 0$, $\Lambda_g \longrightarrow 0.5$ and hence, $\mathcal{R}_g \longrightarrow 0.5$. The lower-bound, on the other hand, implies $\text{MOS}_g \approx 0$, $\mathbf{p}_g \approx P^{\max}$ and one of the constraints in (11) at least is violated. So, $\Lambda_g \approx -0.5$, $\mathcal{R}_g = \Lambda_g - 3|\Lambda_g| \longrightarrow -1.5$. All in all, $-1.5 \leq \mathcal{R}_g \leq 0.5$.
- In Fig. (1c), $\nu_g = 1$, each agent only cares about the achieved MOS of its eMBB users, and therefore, $\Lambda_g \geq 0$. The upper-bound on the other hand, implies $\text{MOS}_g \approx 4.5$, $\Lambda_g \longrightarrow 1$ and hence, $\mathcal{R}_g \longrightarrow 1$. The lower-bound implies $\text{MOS}_g \approx 4.5$ yet one of the constraints in (11) at least is violated. So, $\Lambda_g \approx 1$, $\mathcal{R}_g = \Lambda_g - 3|\Lambda_g| \longrightarrow -2$. All in all, $-2 \leq \mathcal{R}_g \leq 1$.

Next, we study the impact of upscaling the weighting factor ν_g with a step size of 0.5 on average achievable MOS of eMBB users. Fig. 1d demonstrates the learning convergence through variations of average achievable MOS of eMBB users as the number of training episodes increases. Due to $\nu_g = 0$, the objective is solely to minimize the transmit power. For each eMBB user i, its minimum MOS threshold, i.e., $\text{MOS}_i^{\min}$ is guaranteed and the average achievable MOS of eMBB users converges instantly. By scaling up the weighting factor to 0.5 in Fig. 1e, a balance is achieved between the objectives at Λ_g, whereby both are equally concerned. Another observation from the figure is that the average perceived MOS of eMBB users is higher at the convergence point, compared to the counterpart in Fig. 1d. This observation stems from increasing the weighting factor ν_g from 0 in Fig. 1d to 0.5 in Fig. 1e, for each UAV g. From Fig. 1f, one can evidently understand that UAVs as learning agents aim solely at maximizing the perceived MOS for their associated eMBB users, without any concern on energy efficiency. Therefore, the highest perceived MOS is observed in this figure in comparison with those in Fig. 1d and Fig. 1e.

According to Fig. (1g) and Fig. 1h, lower average MOS is achieved by the eMBB users when a more number of eMBB and mMTC users coexist. This is due to the intra-service interference $\mathcal{I}^E_{\text{Intra}}$, as well as the inter-service interference $\mathcal{I}^M_{\text{Inter}}$ in these figures, respectively. Also, the minimum MOS requirement of the eMBB users, as well as the minimum data rate requirement of the mMTC users are limiting factors for the average achievable MOS of the eMBB users. This observation can be argued from the perspective of constrained optimization problems and their feasible solution space w.r.t. constraints. Generally, imposing larger constraints on an optimization problem makes its feasible solution space more limited. As a result, guaranteeing a higher service requirement makes the feasible solution space more limited.

Eventually in Fig. (2a)-Fig. 2b, we investigate the convergence behavior of the overall proposed iterative solution approach at the presence of a comparative baseline scheme, namely alternating direction method of multipliers (ADMM) [10] for both transmit power control sub-problem $\mathcal{P}_{\text{Power}}$ and the trajectory design sub-problem $\mathcal{P}_{\text{Trajectory}}$. In Fig. (2a), we set $\nu_g = 0$ and the objective is to only minimize the transmit power of the UAVs. As observed, both baselines converge within limited number of iterations and our proposed scheme provides up to 17% average reduced transmit power for the UAVs over the ADMM counterpart. In Fig. (2b), the objective is to only maximize the perceived MOS of the eMBB users, by setting $\nu_g = 1$. It is evidently seen that at the convergence point, our proposed method achieves up to 12% average perceived MOS gain over the ADMM baseline. Note that the achieved performance gain in Fig. (2a) and Fig. (2b) is mainly due to this fact that only locally optimal solutions for the non-convex deterministic optimization problems is guaranteed by the ADMM.

VII. Conclusion

We have investigated the problem of multiplexing eMBB and mMTC services in UAV-aided downlink VLC NOMA system and formulated a resource allocation problem of joint transmit power control and motion trajectory design of the UAVs, aiming at characterizing the trade-off between the UAV's power consumption and the perceived QoE of eMBB users, while ensuring the eMBB and mMTC service-specific requirements. An iterative solution based on tools from convex optimization and learning have been proposed to solve this problem. Simulations have revealed remarkable reduced power consumption, as well as increased perceived MOS gain.

References

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive Access for 5G and Beyond," *IEEE Journal Sel. Areas Commun.*, vol. 39, no. 3, pp. 615-637, March 2021.

[2] A. A. Barakabitze, N. Barman, A. Ahmad, S. Zadtootaghaj, L. Sun, M. G. Martini, and L. Atzori, "QoE management of multimedia streaming services in future networks: A tutorial and survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 526–565, 2019.

[3] K. Kilkki, "Quality of experience in communications ecosystem," *J.UCS*, vol. 14, no. 5, pp. 615–624, 2008.

[4] E. N. Tominaga, H. Alves, O. L. A. López, R. D. Souza, J. L. Rebelatto, and M. Latva-aho, "Network Slicing for eMBB and mMTC with NOMA and Space Diversity Reception," *IEEE VTC2021-Spring*, Helsinki, Finland, pp. 1-6, 2021.

[5] S. R. Sabuj, D. K. P. Asiedu, K. -J. Lee, and H. -S. Jo, "Delay Optimization in Mobile Edge Computing: Cognitive UAV-Assisted eMBB and mMTC Services," *IEEE Trans. Cognitive Commun. Netw.*, vol. 8, no. 2, pp. 1019-1033, June 2022.

[6] N. H. Mahmood, M. Lauridsen, G. Berardinelli, D. Catania, and P. Mogensen, "Radio Resource Management Techniques for eMBB and mMTC Services in 5G Dense Small Cell Scenarios," *IEEE VTC-Fall*, Montreal, QC, Canada, pp. 1-5, 2016.

[7] 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.

[8] Q. Wu et al., "A Comprehensive Overview on 5G-and-Beyond Networks With UAVs: From Communications to Sensing and Intelligence," *IEEE Journal Sel. Areas Commun.*, vol. 39, no. 10, pp. 2912-2945, Oct. 2021.

[9] Y. Yang et al., "Joint LED selection and precoding optimization for multiple-user multiple-cell VLC systems," *IEEE Int. Things J.*, vol. 9, no. 8, pp. 6003-6017, 15 Apr., 2022.

[10] H. Zarini, A. Khalili, H. Tabassum, M. Rasti, and W. Saad, "AlexNet classifier and support vector regressor for scheduling and power control in multimedia heterogeneous networks," *IEEE Trans. Mobile Comput.*.

[11] H. Abarghouyi, S. M. Razavizadeh, and E. Björnson, "QoE-aware beamforming design for massive MIMO heterogeneous networks," *IEEE Trans. Veh. Technl.*, vol. 67, no. 9, pp. 8315-8323, Sept. 2018.

[12] S. P. Boyd and M. C. Grant. *The CVX users' guide release 2.2.* Accessed: Jan. 2020. [Online]. Available: http://cvxr.com/cvx/doc/CVX.pdf/.

[13] M. S. Allahham, A. A. Abdellatif, N. Mhaisen, A. Mohamed, A. Erbad, and M. Guizani, "Multi-Agent reinforcement learning for network selection and resource allocation in heterogeneous multi-RAT networks," *IEEE Trans. Cognitive Commun. Net.*, vol. 8, no. 2, pp. 1287-1300, 2022.

[14] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint, arXiv:1511.05952*, 2015.