Harnessing UAVs for Fair 5G Bandwidth Allocation in Vehicular Communication via Deep Reinforcement Learning

Tingting Yuan, Christian Esteve Rothenberg, Katia Obraczka, Chadi Barakat, Thierry Turletti

Harnessing UAVs for Fair 5G Bandwidth Allocation in Vehicular Communication via Deep Reinforcement Learning

Tingting Yuan, Christian Esteve Rothenberg, Katia Obraczka, Chadi Barakat, and Thierry Turletti

***Abstract*—Terrestrial infrastructure-based wireless networks do not always guarantee their resources will be shared uniformly by nodes in vehicular networks. This is due mainly to the uneven and dynamic geographical distribution of vehicles and path loss effects. In this paper, we leverage multiple fifth-generation (5G) unmanned aerial vehicles (UAVs) to enhance fairness in network resource allocation among vehicles by positioning UAVs on-demand as "flying communication infrastructure". We propose a deep reinforcement learning (DRL) approach to determine UAVs' position to improve network resource allocation fairness and efficiency while considering the UAVs' flying range, communication range, and energy constraints. We use a parametric fairness function to attain a number of resource allocation objectives ranging from maximizing the total throughput of vehicles, maximizing minimum throughput, and achieving proportional bandwidth allocation. Simulation results show that the proposed DRL approach to UAV positioning can improve network resource allocation according to the targeted fairness objective.**

***Index Terms*—Unmanned aerial vehicles (UAV), fifth-generation (5G), fairness, deep reinforcement learning (DRL).**

I. Introduction

Even though wireless communication is expected to become more ubiquitous and accessible in urban regions through base stations (BSs) and road-side units (RSUs), wireless users' quality of service varies significantly. This is primarily due to the uneven and highly dynamic physical location and spatial distribution of vehicles as well as communication channel impairments (e.g., path loss). Employing unmanned aerial vehicles (UAVs) to complement existing wireless communication infrastructure deployments, in particular, to assist in vehicular communication, has been the topic of several recent research efforts [1]–[3]. UAV's ability to provide line-of-sight (LoS) links, coupled with their agility and rapid deployability enable them to be deployed on-demand to serve as "flying access points". UAVs can assist the installed communication infrastructure in providing adequate communication services, especially in highly dynamic scenarios such as vehicular networks in urban, suburban, and rural regions. This holds particularly for fifth-generation (5G) and beyond networks, in which LoS communication is critical due to diffraction and material penetration effects, which cause greater signal attenuation [4].

Deploying UAVs to improve communication coverage to vehicles "anywhere, anytime" is quite challenging given that UAVs have limited communication range and are power-constrained [5], coupled with the fact that the number of available UAVs to cover a specific region is often limited. Additionally, UAV's location in three-dimensional (3D) space means that their communication channel characteristics vary not only with their location relative to the ground but also their altitude. Therefore, UAV placement needs to consider many aspects, including vehicle location and bandwidth requirements, the capacity of the local terrestrial communication infrastructure, and UAV limitations, e.g., their energy budget and communication capability. Some previous efforts on UAV placement try to maximize aggregate throughput [6]–[8], while some other proposals aim to maximize the number of users being serviced [9]. However, these approaches are either specific in their optimization target or may create unfair resource allocation as they tend to provide resources to well-connected vehicles while under-serving poorly connected ones.

Recently, computational intelligence, in particular machine learning (ML), has gained significant traction as a powerful tool to address challenges posed by new, more complex problems in a variety of disciplines and domains. The widespread use of ML techniques has been fueled in part by the ever-increasing availability of computing power. As networks and their applications become increasingly complex and heterogeneous, they can benefit from ML approaches that learn and adapt to network- and application dynamics automatically and autonomously. ML techniques do not require prior knowledge of the operating environment and acquire this knowledge as they operate and adjust accordingly without the need for complex mathematical models of the system. In particular, deep reinforcement learning (DRL) [10], [11] has been viewed as a promising solution for dynamic UAV placement [8], [12]. DRL is well-suited to tackle problems that require longer-term planning using high-dimensional observations. Additionally, DRL uses powerful deep neural networks to build a higher-level understanding of the target environment with limited or no prior knowledge.

In this work, we propose a DRL-based approach to perform optimal, real-time UAV placement to achieve fair and efficient bandwidth allocation in 5G vehicular networks. To this end, we model the problem of dynamic UAV placement as an optimization problem whose goal is to maximize a global

Tingting Yuan is now with Computer Networks Group, Georg-August-University of Göttingen, Germany, tingt.yuan@hotmail.com.

Christian Esteve Rothenberg is with University of Campinas. chesteve@dca.fee.unicamp.br.

Katia Obraczka is with UC Santa Cruz. katia@soe.ucsc.edu.

Chadi Barakat and Thierry Turletti are with Inria, Université Côte d'Azur, Sophia Antipolis, France. chadi.barakat@inria.fr, thierry.turletti@inria.fr.

objective fairness function calculated over the communication bandwidth allocated to vehicles. This problem accounts for (i) aggregate communication resources, i.e., bandwidth, provided by ground wireless communication infrastructure and deployed UAVs, (ii) time-varying vehicle location, and (iii) UAV energy budget. We employ parametric fairness as the objective function, which is inspired by previous work on resource allocation for end-to-end Internet flows (e.g., [13]). UAVs are dynamically placed in such a way to realize different allocation strategies, including maximizing vehicle aggregate throughput, maximizing minimum vehicle throughput, as well as achieving vehicle throughput proportional fairness. We evaluate the proposed DRL-based UAV placement framework's performance through simulations driven by real-world vehicle mobility traces.

Overall, the main contributions of this paper can be summarized as follows:

1) We present the first study on bandwidth allocation fairness in UAV-assisted vehicular networks. Unlike previous work, we consider a novel and important objective for dynamic UAV placement in vehicular networks, namely network resource allocation fairness, such that vehicles have access to a fair share of the network bandwidth.
2) We present a DRL approach to dynamically place a swarm of UAVs such that vehicles in a vehicular network are allocated a fair share of the bandwidth as the network operates while considering UAV battery and coverage limitations.
3) Through extensive simulations, we experimentally assess the performance of the proposed dynamic UAV placement approach using different DRL algorithms using real-world vehicle mobility traces. Our results demonstrate that the proposed approach delivers superior bandwidth allocation fairness when compared to existing techniques.

The remainder of the paper is organized as follows. Section II provides an overview of related work. Section III describes our system model, including how we model the underlying communication channel and coverage, UAV energy consumption, and vehicle spatial and temporal distribution. In Section IV, we formulate UAV placement for fair resource allocation as a non-linear optimization problem, and in Section V, we introduce the proposed DRL-based approach. Section VI describes our experimental methodology and presents results. Finally, Section VII concludes the paper with directions for future work.

II. Related Work

Optimizing UAVs' dynamic placement to improve communication networks' performance is complex and challenging and has been the focus of some research efforts in recent years. Early work focused on single UAV placement [7], [9], [14], [15]. Other efforts explored the use of UAV swarms. For example, [16], [17] proposes to minimize the number of UAVs to ensure coverage of all users, and [18] suggests maximizing the total coverage area. The placement problem is even more complex when multiple UAVs are considered, with new challenges raised by UAVs' cooperation and placement algorithms' efficiency.

In high-dimensional state-space and time-varying environments, ML-based technologies have been recently utilized for solving challenging problems in UAV placement [19], [20]. DRL provides a promising solution as it can efficiently handle the challenges related to the state space's high dimensionality and the environment's dynamicity. Recent works have investigated the use of DRL for the placement of multiple UAVs. In [6], a DRL-based approach is proposed for UAV placement that complements terrestrial communication to maximize total network throughput. In [8], DRL is also used to place UAVs to maximize aggregate network throughput while simultaneously balancing traffic load across UAVs. The work reported in [21] explores DRL to maintain an acceptable quality of service for each vehicle and minimize the number of UAVs. The authors of [12] study energy-efficient UAV placement with DRL considering the fairness in coverage time. This work tries to evenly distribute service time across cells using UAVs but does not address quality-of-service on an end-user basis.

Unlike previous work, in this paper, we consider a novel and important objective for dynamic UAV placement in vehicular networks, namely network resource allocation fairness, such that vehicles have access to a fair share of the network bandwidth. To achieve this goal, we use a parametric fairness function inspired by early work on Internet resource allocation [13] as an objective function and maximize it over the offered load from vehicles.

III. System Model

As illustrated in Fig. 1, our target scenario is a geographic region covered by a BS equipped with ground communication infrastructure, e.g., RSUs. A swarm of UAVs, $U = \{u_1, \ldots, u_N\}$, acts as airborne wireless access points serving vehicles in this region. As such, the available communication infrastructure includes a BS, a set of RSUs, and UAVs denoted as $I = \{S, R, U\}$. Since most currently deployed infrastructure uses 4G technology, we assume the BS and RSUs are 4G devices. However, the proposed approach can be directly applied to 5G communication infrastructure technology. The BS and RSUs are stationary, and their locations and coverage radius are known. UAVs are outfitted with 5G transceivers, and their coverage radius $\mathcal{R}_u$ depends on their current flying altitude h_u. A UAV controller is co-located with the BS to determine UAVs' placements based on the vehicles' locations and UAVs' states. We assume that UAVs have limited energy and need to recharge periodically at charging stations. We assume that a UAV charging station (CS) is located in the region, and its location l^{charge} is known. We aggregate vehicles into groups based on their location, defined at the cell level, where G refers to the set of cells in the region. Table I lists and describes the parameters used in our system model.

A. Channel Model

The signal-to-noise-ratio (SNR) between an infrastructure node i (BS, UAV or RSU) and vehicle v can be expressed

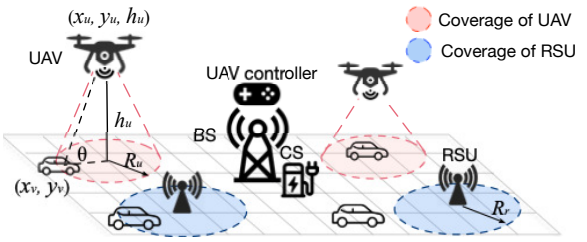


Fig. 1. Scenarios for UAV-assisted vehicular network.

TABLE I
LIST OF NOTATIONS

Notation	Description
I	Communication infrastructure set including BS, RSUs, and UAVs.
S, R	Set of BSs and RSUs.
U	Set of N UAVs.
V	Set of $\|V\|$ vehicles.
G	Set of $\|G\|$ cells in the region.
D	Total number of vehicles in all cells.
$\mathcal{R}$	Communication infrastructure coverage radius.
l	Device location defined by coordinates (l^x, l^y, h), where l^x and l^y are the horizontal coordinates, and h is the elevation.
$\mathbf{L}$	UAV locations $\mathbf{L} = \{l_{u,t} \| u \in U, t \in [1, T]\}$.
$d_{i,j}$	Distance between two nodes i and j.
φ	Channel path loss in dB.
X	Vehicles' access data rate $X = \{x_1, ..., x_{\|V\|}\}$.
x_v	Vehicle v's access data rate in bps.
$F^\alpha(X)$	Data rate fairness criteria for access rate vector X with parameter α.
$\pi_{v,i}$	Binary access indicator - denotes if infrastructure node i offers service to vehicle v.
$P(v)$	UAV energy consumption (J/s) in flying with speed v.
$\Delta\tau$	Time interval between two steps.
ΔE_u	UAV u's energy consumption.
$E_{u,t}$	UAV u's residual power at time t.
$T_{u,t}$	Cost of UAVs to fly during time step t.

based on the average path loss defined in [22] as:

$$\eta_{i,v} = \frac{\rho_{i,v}}{10^{\overline{\varphi}_{i,v}/10}\sigma^2} \tag{1}$$

where $\rho_{i,v}$ is the transmission power, $\overline{\varphi}_{i,v} \triangleq \mathbb{E}[\varphi_{i,v}]$ is the expected path loss, and σ^2 is the additive white Gaussian noise power at the receiver. We assume that the total bandwidth of an infrastructure node i is B_i and that this bandwidth is divided among the associated vehicles equally. The achievable rate of vehicle v (in bps) can be denoted as

$$x_v = \sum_{i \in I} \pi_{i,v} \frac{B_i}{N_i} log_2(1 + \eta_{i,v}) \tag{2}$$

where $N_i = \sum_{v \in V} \pi_{i,v}$ is the number of vehicles serviced by infrastructure node i, and $\pi_{i,v}$ is a binary access indicator, where $\pi_{i,v} = 1$ denotes that infrastructure node i offers access service to vehicle v, and $\pi_{i,v} = 0$ otherwise. We add the following two constraints: (1) $\pi_{i,v} \leq e_{v,i}$ to ensure that vehicles can only choose to associate with the infrastructure node whose communication range covers the vehicle, where $e_{v,i}$ is a Boolean indicator to describe whether vehicle v is in the coverage area of infrastructure node i or not; and (2) $\sum_{i \in I} \pi_{i,v} = 1$ to ensure that each vehicle can only be served by one infrastructure node. If a vehicle is in the coverage range of more than one infrastructure node, it will choose the one offering the highest rate.

Depending on the altitude of the communicating nodes, different channel models need to be used to account for different propagation conditions. For UAV-assisted communication, we consider two propagation scenarios, ground-to-ground (G2G) and air-to-ground (A2G) [3] channels.

1) Ground-to-Ground Channels: G2G channels are typically below 10 meters and 22.5 meters for suburban and urban environments, respectively, [3]. G2G channels include communication between BSs, RSUs, and vehicles. Their large-scale channel attenuation (in dB) include distance-dependent path loss, whose classical model is log-distance path loss [2], [23] defined as follows:

$$\varphi_{i,v} = 10\zeta \log_{10}(d_{i,v}) + X_0 + X_\sigma \tag{3}$$

In this formula, $d_{i,j} = \|l_i - l_j\|$ is the Euclidean distance between two nodes in the space where l_i is the location of node i. ζ is the path loss exponent that usually falls in the range between 2 and 6; X_0 is the path loss at a reference distance with definition as $X_0 = 20\log(\frac{4\pi f_c d_0}{c})$, d_0 being the free-space reference distance, f_c is the carrier frequency and c is the light speed. $X_\sigma \sim \mathcal{N}(0, \tilde{\sigma}^2)$ is the shadowing effect, which is modeled as a normal (Gaussian) random variable with zero mean and a certain variance $\tilde{\sigma}^2$. Thus, the average path loss is equal to $\overline{\varphi}_{i,v} = 10\zeta \log_{10}(d_{i,v}) + X_0$.

2) Air-to-ground Channels: Obstructed A2G channel is in the range of 22.5-100 meters for urban environments [3]. Such a channel usually experiences a higher LoS probability than ground channels; however, it is still not 100%. In our scenario, UAVs communicate with ground vehicles in their coverage simultaneously by employing orthogonal frequency-division multiple access (OFDMA). For UAV-Vehicle (U2V) communication, the mmWave channel of 5G is used. The mmWave propagation channel of the U2V link is modeled using the standard log-normal shadowing model with LoS and non-line-of-sight (NLoS) links by choosing specific channel parameters [22]. In this model, the general path loss (in dB) between UAV u and vehicle v is defined as:

$$\varphi_{u,v} = \begin{cases} 10\zeta_{LoS} \log(d_{u,v}) + X_0 + X_{\sigma_{LoS}} & \text{if LoS} \\ 10\zeta_{NLoS} \log(d_{u,v}) + X_0 + X_{\sigma_{NLoS}} & \text{if NLoS} \end{cases} \tag{4}$$

where ζ_{LoS} and ζ_{NLoS} are the path loss exponents of LoS and NLoS links; $X_{\sigma_{LoS}}$ and $X_{\sigma_{NLoS}}$ are the shadowing random variables which are, respectively, represented as the Gaussian random variables (in dB) with zero means and σ_{LoS} and σ_{NLoS} as standard deviations. $X_0 = 20\log(\frac{4\pi f_m d_0}{c})$ is the reference path loss in the band of carrier frequency f_m of mmWave and at reference distance d_0.

The probability of having LoS transmission between UAVs and vehicles is expressed in [24] as:

$$P_{u,v}^{LoS} = \frac{1}{1 + ae^{-b(-\theta_{u,v} - a)}} \; \forall i \in U, v \in V \tag{5}$$

where a and b are constants that depend on the environment (e.g., the environment density, such as rural, urban); and θ is the elevation angle of a point on the ground with respect to a UAV (measured in degrees), which can be expressed as

$\theta_{u,v} = \frac{180}{\pi} \sin^{-1}(\frac{h_u}{d_{u,v}})$. Thus, the average path loss of A2G channels can be expressed as:

$$\overline{\varphi}_{u,v} = P_{u,v}^{LoS} \varphi_{u,v}^{LoS} + P_{u,v}^{NLoS} \varphi_{u,v}^{NLoS} \tag{6}$$

where the probability of having NLoS is defined as $P_{u,v}^{NLoS} = 1 - P_{u,v}^{LoS}$. Above a certain altitude, which depends on the environment, is called high-altitude A2G. In high-altitude A2G, all channels are in LoS, so the propagation is close to the free-space with $\overline{\varphi}_{u,v} = \varphi_{u,v}^{LoS}$.

B. Communication Coverage Model

We assume that the BS and RSUs are unmovable, so their coverage radius is fixed and denoted as $\mathcal{R}_b$ for the BS, and $\mathcal{R}_r$ for RSU r, respectively. For each UAV, given its height h_u, the path loss increases with the distance. Thus, the coverage of UAV u is defined as the maximum radius in which the path loss is below a value ($\overline{\varphi}_0$). The radius is defined as $\mathcal{R}_u = \sqrt{d_u^2 - h_u^2}$, where d_u is the maximum distance between UAV u and the ground defined as $d_u = \arg\max_d(\overline{\varphi}_{u,v} \leq \overline{\varphi}_0)$.

As introduced earlier, we use Booleans $e_{v,i}$ to describe whether vehicles are in the coverage scope of infrastructure nodes or not, and they are defined as:

$$\forall\ i \in I,\ v \in V:\ e_{v,i} = \begin{cases} 1, & d_{i,v} \leqslant \mathcal{R}_i \\ 0, & d_{i,v} > \mathcal{R}_i \end{cases}$$

where $\mathcal{R}_i$ is the coverage radius of infrastructure node i.

C. UAV Energy Consumption Model

According to [25], energy consumption for rotary-wing UAVs with speed υ is modeled as the sum of three items, namely, blade profile power, induced power, and parasite power. Blade profile power is the energy used to overcome the blades' profile drag. Induced power and parasite power are used to overcome the blades' induced drag and the UAV's fuselage drag, respectively. The power sum physically represents the UAV energy consumption per second in Joule/second (J/s) with speed υ:

$$P(\upsilon) = P_0(1 + \frac{3\upsilon^2}{\upsilon_0^2}) + P_1 \left(\sqrt{1 + \frac{\upsilon^4}{4\upsilon_1^4}} - \frac{\upsilon^2}{2\upsilon_1^2} \right)^{1/2} + \frac{1}{2} P_2 \upsilon^3 \tag{7}$$

where the P_0, P_1, and P_2 are coefficients of blade profile power, induced power, and parasite power, respectively. These coefficients are related to the aircraft's weight, air density, fuselage drag ratio, rotor solidity, etc. υ_0 and υ_1 denote the rotor blade's tip speed and the mean rotor-induced velocity in hovering. We note that when the UAV is hovering with speed $\upsilon = 0$, the energy consumption is $P_h \triangleq P_0 + P_1$.

Combining both the propulsion energy and the communication-related energy, the energy consumption of each UAV at time interval T can be expressed as:

$$\Delta E_u = \int_0^T \left(P(\upsilon_t) + P_c \sum_{i \in V_t} \pi_{u,i,t} \right) dt,\ \forall u \in U \tag{8}$$

We assume that the communication-related power is a constant expressed in Watt (W) and is denoted as P_c. V_t is the set of vehicles at time t.

D. Spatial and Temporal Distribution of Vehicles

As previously described, we aggregate vehicles that are in close geographic proximity into groups called cells. Vehicle spatial and temporal distribution characteristics, which will be used in our simulation experiments, are described below.

1) Spatial Variation (SV): The spatial variation can be used to show the variability in the number of vehicles between the cells. It is defined as the coefficient of variation in the number of vehicles per cell. It is expressed as the ratio of the standard deviation $\hat{\sigma}$ to the mean μ:

$$SV(D) = \frac{\hat{\sigma}(D)}{\mu(D)}, \tag{9}$$

where D is the set of the number of vehicles in all cells denoted as $D = \{|V_1|, |V_2|, ...|V_{|G|}|\}$, and $|V_i|$ is the number of vehicles in the cell i.

2) Temporal Variation (TV): For each cell, we define two factors to describe the temporal variation in the number of vehicles per cell. First, to normalize the number of vehicles per cell, we define the proportion of the number of vehicles per cell at time step t as:

$$\forall g \in G, t \in [0, T]:\ \xi_{g,t} = \frac{|V_{g,t}|}{\sum_{i \in G} |V_{i,t}|}$$

where $|V_{g,t}|$ is the number of vehicles in the cell g at time step t, and G is the whole set of cells. The time coefficient of variation (TCV) is then proposed to describe the variation on vehicles' distribution over time from 0 to T. The TCV is the mean coefficient of variation of the time-varying number of vehicles overall cells in G:

$$TCV(\xi_g) = \frac{1}{|G|} \sum_{g \in G} \frac{\hat{\sigma}(\xi_g)}{\mu(\xi_g)} \tag{10}$$

where ξ_g is the vector of proportional number of vehicles per cell g over the time period, which is a defined as $\xi_g = \{\xi_{g,t} | t \in [0, T]\}$. Besides, to describe the variation on each cell over adjacent time steps, the step coefficient of variation (SCV) is proposed. It is the average change in ξ_g between two adjacent time steps overall cells g, which is defined as:

$$SCV(\xi_g) = \frac{\sum_{g \in G} \sum_{t=1}^{T} |\xi_{g,t} - \xi_{g,t-1}|}{|G|T} \tag{11}$$

IV. Problem Formulation

A. UAV Placement Optimization Model

The problem of UAVs' dynamic placement can be formulated as follows: given a limited region in which some terrestrial communication infrastructure has been deployed, we try to find the optimal placement of a limited number of UAVs, say N, with an objective function calculated over the bandwidth allocated to vehicles. Note that we also need to consider the UAV charging duration during which UAVs are out of service. Recall that our objective function needs to account for the fairness of allocating aggregate network resources, i.e., aerial and terrestrial. We consider that at the beginning of the control period ($t = 0$), all UAVs are located at charge stations. The optimization of UAV dynamic placement

where UAV location is given by $\mathbf{L} = \{l_{u,t} | u \in U, t \in [1, T]\}$ can then be modeled as follows:

$$\max_{\mathbf{L},\pi} \int_0^T \frac{F^\alpha(X_t)}{|V_t|} dt$$

$$\text{s.t.} \quad \forall u \in U, t \in T:$$

$$l^x_{min} \leq l^x_{u,t} \leq l^x_{max} \tag{12}$$

$$l^y_{min} \leq l^y_{u,t} \leq l^y_{max} \tag{13}$$

$$h_{min} \leq h_{u,t} \leq h_{max} \tag{14}$$

$$E_{u,t} > 0 \tag{15}$$

$$\pi_{i,v} d_{i,v} \leq \mathcal{R}_i \tag{16}$$

$$\sum_{i \in I} \pi_{i,v} = 1 \tag{17}$$

This problem has the following constraints. Firstly, UAV locations should be within the region under consideration defined by BS's coverage radius and between l_{min} and l_{max} as defined by Equations (12), (13) and (14). Secondly, the UAVs need to go back for charging before they run out of power as indicated by Equation (15), i.e., $l_{u,t} = l^{charge}$. Depending on whether a UAV's target position is a charging station or not, UAVs can be either "working" or "charging". Since while charging UAVs cannot provide service, vehicles previously connected to them need to find a new infrastructure node to connect. Thirdly, Equation (16) ensures that vehicles can only choose to associate with the infrastructure node whose communication range covers the vehicle, while Equation (17) stipulates that each vehicle can only be served by one infrastructure node at a time. If a vehicle is in the coverage range of more than one infrastructure node, it will choose the one offering the highest data rate.

The objective is an aggregated parametric function, which models the fairness and efficiency of network resource allocation for UAVs and terrestrial infrastructure. The objective function is normalized by the time-varying number of vehicles to not bias the optimization against periods with fewer vehicles. To account for as many bandwidth allocation scenarios as possible, we use a parametric fairness function widely known in the literature on resource allocation on the Internet and TCP congestion control. Given a positive constant α and access rate vectors of vehicles $X = \{x_1, ..., x_{|V|}\}$, this fairness function is given in [13] by:

$$F^\alpha(X) = \begin{cases} \frac{1}{1-\alpha} \sum_{i=1}^{|V|} x_i^{1-\alpha} & \text{if } \alpha \neq 1 \\ \sum_{i=1}^{|V|} \log x_i & \text{otherwise} \end{cases} \tag{18}$$

This parametric function covers different bandwidth allocation strategies according to the chosen value for parameter α. We can get the case of maximizing the sum of vehicles' access rates (namely total throughput) as $\alpha \to 0$; the proportional fairness case when $\alpha \to 1$; the harmonic mean fairness case when $\alpha \to 2$; the max-min fairness case when $\alpha \to \infty$, the latter tries to maximize the minimum access rate overall vehicles by giving priority to vehicles that have achieved lowest data rate. Proportional fair and harmonic mean fair allocations are trade-offs between maximizing total access rates and max-min access rate allocation. The first scenario of $\alpha = 0$ is the mere scenario where the total access rate is maximized independently of any fairness consideration.

B. Time Discretization

An intuitive way to maximize fairness is to select the optimal locations at each time interval of $\Delta\tau$. The time discretization variables are $\mathbf{L} = \{L_t | t \in [1, T]\}$, and $L_t = \{l_{u,t} | u \in U\}$. The UAVs are assumed to have two states in each time interval, namely, 1) flying to the target locations and hovering to serve the incoming flows, and 2) flying back to the charge station and charging. During the flying period, we assume that the UAVs have a constant velocity of v_m. The time cost by UAV u for flying from the current location $l_{u,t-1}$ to the target location $l_{u,t}$, which may be different from each other, is denoted as:

$$T_{u,t} = \frac{||l_{u,t} - l_{u,t-1}||}{v_m}, \ \ \forall u \in U$$

The UAVs' arriving order at step t is denoted as $\mathcal{I}_t$. ΔT refers to the time interval between two consecutive UAV arrivals as shown in Fig. 2. The time interval of the i-th arrival in time slot t can be expressed as:

$$\Delta T_{i,t} = \begin{cases} \Delta\tau - T_{u_i,t} & \text{if } u_i \text{ is the last to arrive} \\ T_{u_{i+1},t} - T_{u_i,t} & \text{otherwise} \end{cases} \tag{19}$$

where $T_{u_i,t}$ is the time of arrival of the i-th UAV u_i to its target location. During $\Delta T_{i,t}$, only the first i arriving UAVs can offer services. It follows that when the last UAV arrives at its target location, all working UAVs can offer service till the end of this time slot $\Delta\tau$. In addition, a UAV with $T_{u,t} \geq \Delta\tau$ will not have time to work as an access point in this time slot.

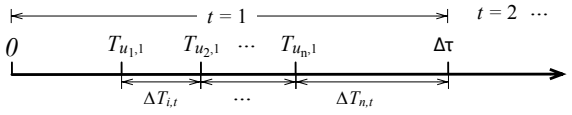


Fig. 2. UAV arrivals in time.

During the hovering period, UAVs hover over their target locations and serve the incoming flows. The mean fairness value of each step is defined as:

$$f_t^\alpha = \frac{\sum_{i \in \mathcal{I}_t} F^\alpha(X_{i,t}) \Delta T_{i,t}}{|V_t| \Delta\tau} \tag{20}$$

where $X_{i,t}$ is the vector of vehicles' access rates when the i-th UAV arrives to offer service in time slot t. Note that $X_{i,t}$ is calculated based on the mean distribution of vehicles in time slot t, and $|V_t|$ in Equation (20) is the mean number of vehicles in time slot t. The objective function can thus be approximated as $\sum_{t=0}^{T} f_t^\alpha$.

To ensure each UAV does not run out of energy, we assume that when a UAV's energy $E_{u,t}$ falls below level $\tilde{E}$ (e.g., 10%), the UAV will go back to the charging station. Boolean parameters ϕ_u indicate whether UAV u is in service (with value 1) or out of service to charge (with value 0).

$$\forall\ u \in U : \phi_{u,t} = \begin{cases} 0 & \text{if } E_{u,t} \leqslant \tilde{E} \\ 1 & \text{otherwise} \end{cases} \tag{21}$$

Equation (8) can be reformulated as:

$$\Delta E_{u,t} = P(\upsilon_m)T_{u,t} + \phi_{u,t}(P_h + P_c N_{u,t})(\Delta\tau - T_{u,t}) \quad (22)$$

where $N_{u,t}$ is the number of vehicles that are connected with UAV u with definition as $N_{u,t} = \sum_{i \in V_t} \pi_{u,i,t}$. The first item of Equation (22) is the flying power, and the second item is the hovering and service power. If the UAVs are selected to go back to charge with $\phi_{u,t} = 0$, there would be no power cost for hovering and offering services. The updating of the battery of UAVs is expressed as:

$$\forall\ u \in U : E_{u,t+1} = \begin{cases} E_{u,t} - \Delta E_{u,t} & \text{if } \phi_{u,t} = 1 \\ E^{full} & \text{if } \phi_{u,t} = 0 \end{cases} \quad (23)$$

where E^{full} denotes the capacity of UAVs' battery.

V. DRL for Dynamic UAV Placement

In this section, we introduce the proposed DRL-based method to solve the UAVs' dynamic placement problem.

A. DRL-Based Approach

We model trial-and-error learning as a Markov decision process (MDP). At each time t, the agent observes the current state s_t of the interactive environment and gives an action a_t according to its policy. Then, the environment returns reward r_t as feedback, and moves to the next state s_{t+1} according to the transition probability $P(s_{t+1}|s_t, a)$. The goal to find an optimal policy can thus be formulated as the mathematical problem of maximizing the expectation of cumulative discounted return $R_t = \sum_{k=t}^{T} \gamma^{k-t} r_k$, where $\gamma \in [0, 1]$ is a discount factor for future rewards to dampen the effect of future rewards on the action; r_k is the reward of each step, and T is the time horizon before game over.

In our scenario, we assume that the intelligent agent who can control UAVs is deployed in the BS. The vehicles and UAVs in this region send their states (e.g., geographic locations) to the agent through the network periodically, e.g., every 2 minutes in the bus tracing dataset in Rio provided by CRAWDAD [26]. Thus, the agent can obtain the vehicular traffic and the state of UAVs. Then, the agent makes decisions on UAVs' following locations, considering their current states and vehicles' information. Next, the agent sends the decisions to UAVs. Notice here that the UAVs' locations cannot be out of the agent's communication coverage. Thus, there are some constraints on the actions to be accounted for. The UAVs follow instructions from the agent and fly to their target locations. The key definitions of DRL for UAV placement are expressed as follows.

State Space: The state is defined as $s_t = \{D_t, L_t, E_t\}$ for time step t. Firstly, the state includes the distribution of vehicles D_t, a vector of the mean number of vehicles in all cells over the time interval of $\Delta\tau$. Secondly, the current locations of UAVs would affect the decisions on the next locations. Thus, the vector of current UAVs' locations $L_t = \{l_{u,t} | u \in U\}$ is also used as input. Then, the last item is the vector of UAVs' residual energy $E_t = \{E_{u,t} | u \in U\}$, which can also affect the decision on next locations. For example, UAVs may prefer to fly nearby to save energy and offer service for a longer time when the reward is inadequate to let them fly far away. Besides, residual energy is critical in making decisions on whether to charge or not.

Action Space: The action is defined as a vector of UAVs' normalized locations $a_t = \{w_{u_1,t}, ..., w_{u_N,t}\}$ of each time step t, in which $w_{u,t} = \{w^x_{u,t}, w^y_{u,t}, w^h_{u,t}\}$ is a vector denoting the normalized 3D coordinates of UAV u. The next location of UAV u can be calculated with $l^x_{u,t} = w^x_{u,t}(l^x_{max} - l^x_{min}) + l^x_{min}$ for the x coordinate, the other two coordinates y and z can be calculated in the same way. This allows the UAVs' locations to be limited to some certain range. Besides, the UAVs' next locations depend not only on the action a_t but also on the charging factor ϕ_t. The definition for the next location of UAV u is then updated as follows:

$$\forall\ u \in U : l_{u,t} = \begin{cases} l_{u,t} & \text{if } \phi_{u,t} = 1 \\ l^{charge} & \text{if } \phi_{u,t} = 0 \end{cases} \quad (24)$$

If a vehicle is in the coverage range of more than one infrastructure node, it will greedily choose the one offering the highest data rate. Therefore, π is not in the action space.

Reward: The value of reward depends both on the current state and on the action taken. In our problem, the reward is defined as the incremental ratio of the fair value of vehicles' access rates to the original fairness value without UAVs working for communication, which is:

$$r_t = \frac{(f^\alpha_t - f^\alpha_t(0))}{|f^\alpha_t(0)|} \quad (25)$$

where f^α_t is expressed in (20), and $f^\alpha_t(0)$ is the original fairness value without UAVs working for communication.

B. DRL-based Training Algorithm

DRL is known to be well-suited to tackle problems that require longer-term planning using high-dimensional observations which is the case of dynamic UAV placement to achieve fair bandwidth allocation in vehicular networks. There is a variety of DRL-based algorithms. Value-based algorithms, e.g., Deep Q-Networks (DQN) [27], use a deep neural network to learn the action-value function. However, they do not support continuous action space like the one in our problem. In the case of policy-based algorithms, e.g., Policy Gradient (PG) [28], they explicitly build a representation of a policy. However, evaluating a policy without action-value estimation is typically inefficient and causes high variance. Actor-critic algorithms learn the value function (critic) in addition to the policy (actor), since knowing the value function can assist policy updates, for example, by reducing variance in policy gradients. Many existing approaches are based on actor-critic, for example, Deep Deterministic Policy Gradient (DDPG) [29], Proximal Policy Optimization (PPO) [30], Asynchronous Advantage Actor-Critic (A3C) [31], Distributed Distributional Deterministic Policy Gradients (D4PG) [32], Twin Delayed Deep Deterministic Policy Gradients (TD3) [33], and Soft Actor-Critic (SAC) [34]. Both A3C and D4PG, which are asynchronous algorithms, can enable multiple worker agents to train in parallel, allowing faster training. However, in the

proposed problem, the model doesn't support multiple worker agents. SAC has been shown to exhibit good and stable performance in various benchmarks and robot control tasks and it enables stronger exploration capability. As such, we chose SAC as the training algorithm which is illustrated in Fig. 3.

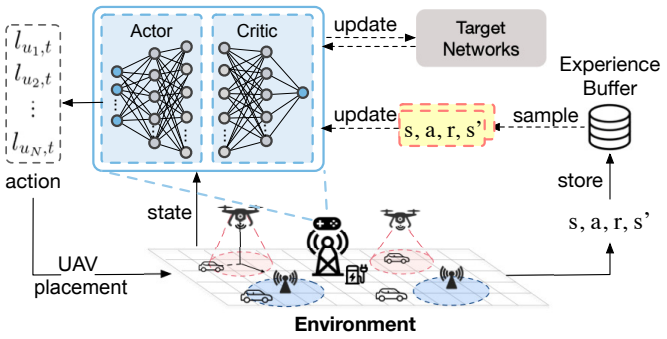


Fig. 3. DRL for UAV dynamic placement.

Experience Replay Buffer: For off-policy training, the experience replay buffer, denoted as $\mathcal{M}$, is used to store experiences and offer samplings for training. It is in the form of tuples, denoted as (s, a, r, s') with states, actions, rewards, and successor states of the next time.

Actor: The actor is responsible for taking actions a_t based on the observed states s_t. SAC employs a tractable policy $\pi(a_t|s_t;\theta^\pi)$ as the actor parameterized through neural networks with parameters θ^π, which targets planning of UAVs placement in continuous action spaces. SAC trains the actor with entropy regularization to maximize both the expected return and entropy. The optimal policy is $\pi^* = \arg\max_\pi \sum_t \mathbb{E}[r(s_t,a_t) + \beta H(\pi|s_t)]$, where $H(\pi|s_t)$ is the entropy of the policy at state s_t, and β controls the trade-off between exploration and exploitation. Then, the policy parameters can be learned by minimizing the expected Kullback-Leibler (KL) divergence [35] as

$$J_\pi(\theta^\pi) = \mathbb{E}_{s_t\sim\mathcal{M},a_t\sim\pi}[\beta log\pi(a_t|s_t) - Q(s_t,a_t)] \quad (26)$$

Critic: SAC employs value function V and Q-function Q in the critic. The value function $V(s_t;\theta^V)$ and the Q-function $Q(s_t,a_t;\theta^Q)$ are parameterized using neural networks with parameters θ^V and θ^Q, respectively. The soft value function networks are trained to minimize the squared residue error, according to:

$$\nabla J_V = \nabla_{\theta^V} V(s_t)(V(s_t) - Q(s_t,a_t) + \log\pi(a_t|s_t)) \quad (27)$$

The Q-function can be optimized with stochastic gradients:

$$\nabla J_Q = \nabla_{\theta^Q} Q(s_t,a_t)(Q(s_t,a_t) - r(s_t,a_t) - \gamma V'(s_{t+1})) \quad (28)$$

where V' is from the target value network with parameters $\theta^{V'}$.

Target Network: The target value network is a copy of the estimated value function, which can improve stability by simply using a separate set of periodically updated parameters. The parameters of the target value network are periodically updated as:

$$\theta^{V'} \leftarrow \varepsilon\theta^V + (1-\varepsilon)\theta^{V'} \quad (29)$$

where ε is a coefficient between 0 and 1.

C. Overhead Analysis

Although DRL is well-suited to tackle problems that need longer-term planning using high-dimensional observations, it requires more time to collect data and memory resources for storing experiences. In this section, we analyze both the computational- and storage complexity of the proposed DRL approach. For complexity analysis purposes, the deep neural network can be viewed as a matrix multiplication problem. Thus, the complexity for the actor is $\mathcal{O}(|G|HN)$, where $|G|$ is the number of cells in the region being considered, H is the number of hidden layers, and N stands for the number of UAVs. Similarly, the critic neural network's complexity is $\mathcal{O}(|G|HN)$. SAC training procedure's complexity is $\mathcal{O}(N_{ep}K_sNH|G|)$, where N_{ep} is the number of training episodes, and K_s is the batch size. Thus, the training is affected by the number of UAVs, which means using more UAVs will increase complexity. Real-time execution, namely calculating UAV locations based on the current state, the complexity is $\mathcal{O}(|G|HN)$. However, if the brute force method is used to find optimal locations, which chooses locations with the highest reward from K possible locations, the complexity is $\mathcal{O}(K^N)$, which has higher complexity. The proposed DRL storage footprint depends on the maximum number of transitions stored in the experience replay buffer, denoted as $|\mathcal{M}|$. All data is assumed to be a 32-bit float. Thus, each state is $4(|G|+4N)$ bytes, each action is $12N$ bytes, and each reward is 4 bytes. Then, one transition is $8|G|+44N+4$ bytes, and the total size of the memory for the experience reply buffer is $|\mathcal{M}|(8|G|+44N+4)$ bytes.

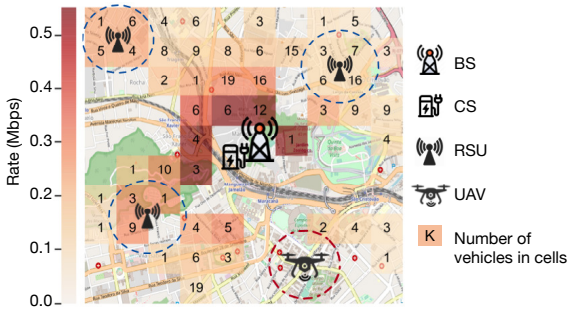


Fig. 4. UAV-assisted vehicle communication in Rio de Janeiro, Brazil.

VI. Experiments and Results

We use simulation experiments to evaluate the performance of the proposed DRL-based UAV dynamic placement algorithm. In this section, we first present our evaluation methodology and then discuss our experimental results. Our code and the datasets used in our simulations are publicly available.[1]

[1]https://github.com/TingtingYuan/UAV_fairness

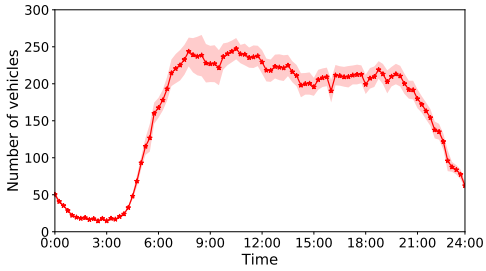


Fig. 5. Number of vehicles over time in Rio de Janeiro's 3×3 km^2 target region.

A. Evaluation Methodology

Fig. 4 shows the region used as the geographical footprint for the experiments. This region consists of a 3×3 km^2 area of Rio de Janeiro, Brazil. This region is centered around a BS, co-located with a charging station. 3 RSUs are located at the corners of the region since they are the farthest away areas from the BS with the highest vehicle traffic. The BS has a coverage radius of around 2 km, which is 4G's nominal coverage in urban areas, and each RSU is assumed to have a 300 m coverage radius. For UAVs, the coverage radius depends on their altitude, ranging from 50 m to 100 m. The region was divided as a square grid of 10×10 cells, thus, the total number of cells $|G|$ is 100, and the size of each cell is 300×300 m^2. Such division yields a suitable distribution matrix for the number of vehicles in cells. Note that increasing the number of cells also increases the algorithm's computational complexity.

TABLE II
LIST OF NOTATIONS

Notation	Description	Value
f_c, f_m	Carrier frequency of cellular and mmWave	2, 38 Ghz
B_i	Bandwidth of the BS, UAVs, and RSUs	60, 1, 1Mhz
υ_0	Tip speed of the rotor blade	120 m/s
υ_1	Mean rotor induced velocity in hover	4.03 m/s
υ_m	Fixed flying speed of UAVs	20 m/s
P_0	Coefficient of blade profile power	79.85 J/s
P_1	Coefficient of induced power	88.63 J/s
P_2	Coefficient of parasite power	0.018 kg/m
P_c	Communication power of UAVs	1 W
$\overline{\varphi}_0$	The path loss rate limitation of UAVs	-138 dBm
a, b	The constants of (5)	10, 0.6
ζ	The path loss exponent of LoS and NLoS	2, 2.4
σ^2	Additive white Gaussian noise in (1)	-95 dBm
E^{full}	UAV battery capacity	700 kwh

For vehicle mobility and to be as realistic as possible, we use an available dataset [26] with real-time position data reported by buses in Rio. The vehicle mobility dataset is in 24h time format. For our experiments, we select one week's worth of vehicle data, i.e., $\varpi = 7$, and consider every 15 minutes to be one step $\Delta\tau = 15$, which means that the agent takes actions every 15 minutes. For one day, there are 96 steps in total ($T = 96$). Fig. 5 shows a time series of number of vehicles $|V_t|$ over time t for the 7-day period considered, as well as the 68% confidence interval. Note that the intervals $\Delta\tau$ can be set smaller, e.g., 5 minutes and 3 minutes. However, we find 15 minutes a reasonable interval for this dataset because this interval can provide an apparent variation on vehicular distribution between steps. For this dataset, the mean spatial variation over 7 days is 195%. It is defined as $\frac{\sum_{d=1}^{\varpi}\sum_{t=1}^{T} SV(D_{d,t})}{T\varpi}$, where SV is in (9), and $D_{d,t}$ is the set of the number of vehicles in all cells at time t in day d. The mean of TCV defined in (10) and SCV in (11) over 7 days are 152.4% and 43.5%, respectively. These values show that the distribution of vehicles varies significantly over time (15 minutes as a step) and between cells (10×10). It is possible to accommodate this variability by placing UAVs differently and improving resource allocation according to the predetermined fairness objective.

In our experiments, we use the Adam optimizer [36] with the learning rate 0.0001 for the actor and 0.001 for the critic, and the hyper-parameters $\varepsilon = 0.01$, $\gamma = 0.95$. The size of the replay buffer $|\mathcal{M}|$ is set to be 50000, whereas the size of memory is around 50 MB, and the batch size is 800. The actor and critic are designed as 3-hidden-layer with 256-128-32 units neural networks. These values for the hyper-parameters are chosen empirically after extensive experiments. Other simulation parameters and their values are listed in TABLE II.

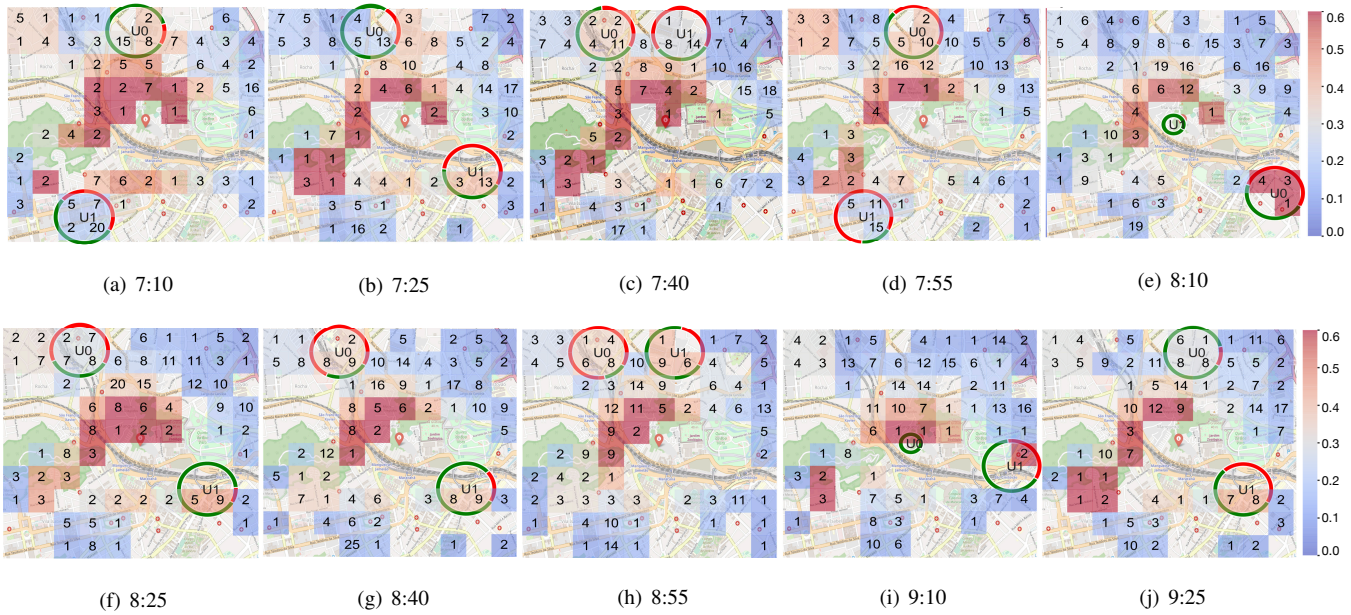
B. Results and Performance Evaluation

Fig. 6 illustrates the results of running the proposed UAV dynamic placement algorithm with 2 UAVs. The placement's objective is to maximize the sum of the vehicles' access data rates (i.e., total throughput) with $\alpha = 0$. The figure also shows $U0$'s and $U1$'s battery levels where the residual battery is represented in green and the used charge in red. Notice that $U1$'s remaining battery charge is lower than 10% in Fig. 6(e) and consequently $U1$ has to go back to the charging station.

Comparative Performance of DRL algorithms with Learning. Fig. 7 shows the comparative performance of DRL algorithms, i.e., PPO2, DDPG, SAC, and TD3, with $\alpha = 1$ and using 2 UAVs[2]. The horizontal axis is the learning episodes whereas the vertical axis presents the mean value of accumulated rewards $\frac{1}{\varpi}\sum_{d=1}^{\varpi}\sum_{t=1}^{T} r_{t,d}$, which reflect fairness' accumulated incremental ratio. The curves show the 68% confidence interval of the accumulated rewards for 7 days. From the comparison, we find that DDPG has a relatively good learning speed at the early stages and is the fastest to converge, but SAC can converge relatively fast and achieves better policy with higher rewards. Additionally, we evaluate SAC's performance with different numbers of UAVs as shown in Fig. 8 and observe that, with more UAVs, the algorithm needs more episodes to converge due to increased state and action dimensionality. We notice a diminishing returns type behavior where the difference in accumulated reward drops from 37% when the number of UAVs increases from 2 to 4 to 9% when the number of UAVs increases from 8 to 10.

Fairness Performance. We show simulation results for our proposed assignment methodology (called **FD** for short) compared against three other approaches: (1) Fairness-based place-

[2]We also obtained similar performance behavior with other α values. They are not shown here due to space limitations.



(a) 7:10 (b) 7:25 (c) 7:40 (d) 7:55 (e) 8:10

(f) 8:25 (g) 8:40 (h) 8:55 (i) 9:10 (j) 9:25

Fig. 6. UAV locations and remaining battery levels with $\alpha = 0$ and $N = 2$ from 7:00 to 9:30hs.

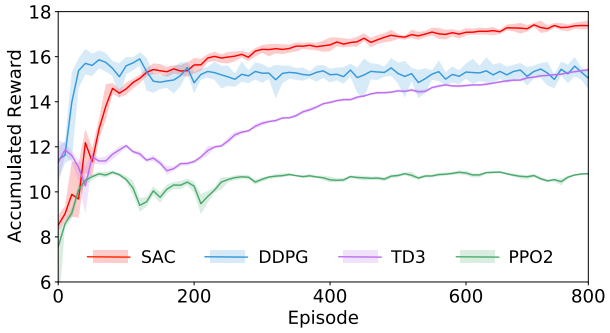


Fig. 7. Accumulated reward of DRL algorithms with $\alpha = 1$ and $N = 2$.

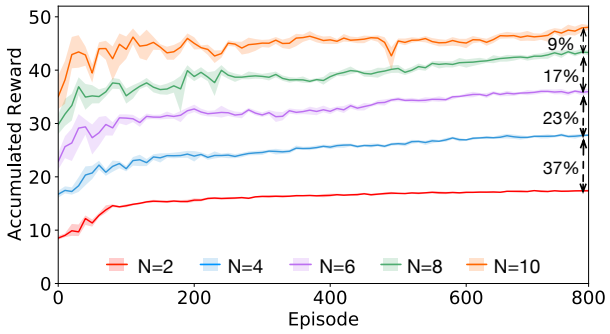


Fig. 8. Accumulated reward using SAC training with $\alpha = 1$ and different number of UAVs.

ment using enumerations (called **FE** for short), which chooses locations with the highest reward from $K = 100$ possible locations at each step. It is used to show DRL's superior performance when compared with the brute force method. (2) Maximum coverage placement [9] (called **MC** for short) using DRL. Although MC can cover as many vehicles as possible using UAVs, it does not consider that some vehicles already have adequate service. We used it as a comparison baseline to show the importance of accounting for cooperation with terrestrial infrastructure. (3) Coverage time fairness [12] based placement using DRL (called **TF** for short). It considers the fairness in coverage time and ignores the difference between covered areas due to uneven traffic and the location of terrestrial infrastructure. (4) No UAVs (called **NU**).

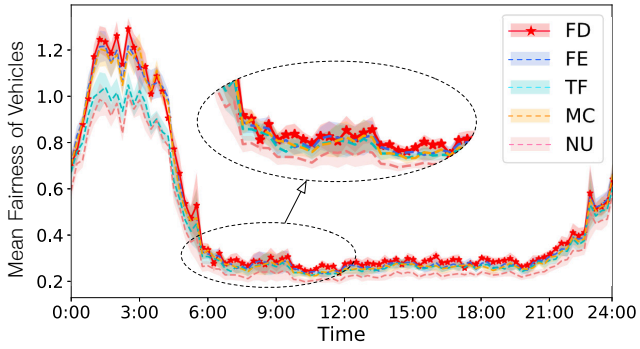


Fig. 9. Mean fairness value of vehicles with $\alpha = 1$ and $N = 2$.

Fig. 9 shows the advantages of our approach in terms of fairness compared to the four baselines. The Y-axis is the mean fairness value of vehicles $\frac{F^{\alpha}(X_t)}{|V_t|}$ with $\alpha = 1$ and $N = 2$. The curves show the 68% confidence interval for 7 days. Combined with Fig. 5, we observe that, in general, mean fairness decreases as the number of vehicles increases. For example, during rush hour 6:00-21:00, the mean fairness is around 0.3, and between 2:00-4:00 pm, it is much higher at around 1.2. However, when compared with FE, MC, TF, and NU, FD improves mean fairness. More specifically, FD's mean improvement is around 2.12%, 6.60%, 11.08%, and 23.88%, respectively. Furthermore, we show the mean fairness improvement of our approach (FD) compared to the baselines

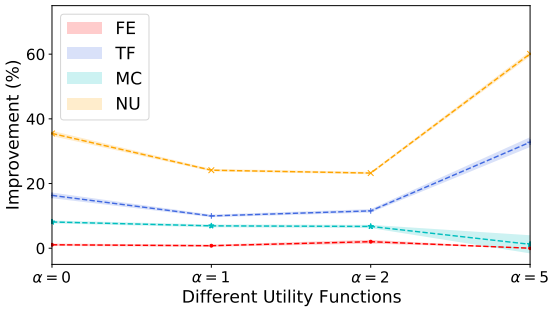
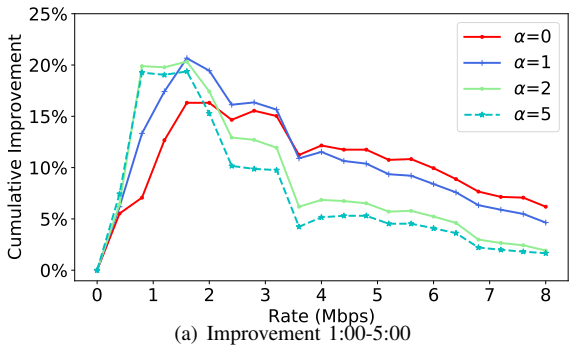
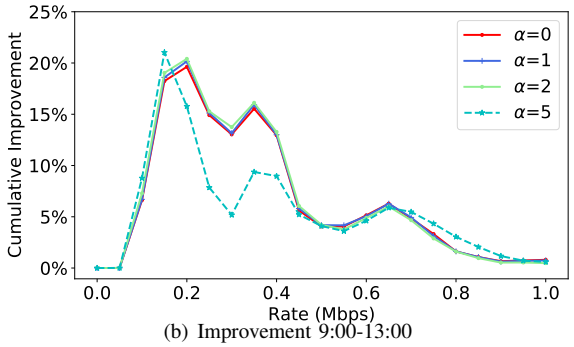


Fig. 10. Mean fairness improvement of FD compared with four approaches on baselines with $N = 2$ and different α values.

for different α in Fig. 10. The improvement is not significant compared to FE, but since FE is a brute force method, it cannot efficiently support real-time execution. Compared with the other approaches, there is a noticeable improvement in fairness for all α values, where MC yields relatively better performance among the 4 other algorithms. From these two figures, we observe that FD can improve fairness compared to the baselines because 1) neither Time Fairness (TF) nor Maximum Coverage (MC) can guarantee bandwidth allocation fairness, and because 2) DRL is also able to outperform the brute force method (FE).



(a) Improvement 1:00-5:00



(b) Improvement 9:00-13:00

Fig. 11. Cumulative distribution of vehicles' access rate with $N = 2$ showing improvement over NU.

Performance considering different fairness parameters. Fig. 11 shows the improvement in terms of the cumulative distribution of vehicles' access rates compared with NU. Fig. 11(a) is from 1:00 am to 5:00 am in 7 days, and Fig. 11(b) represents the rush hour from 9:00 am to 13:00 am in 7 days. These curves show performance with 2 UAVs and different α values. Seen from Fig. 11(a) and Fig. 11(b), compared with NU, all curves with different α values have a significant improvement. Comparing different α values, we can conclude that UAVs are more likely to serve vehicles at a lower rate with a higher α value. This is expected as when α increases, individual fairness becomes more important than the overall performance (e.g., the total sum of access rates).

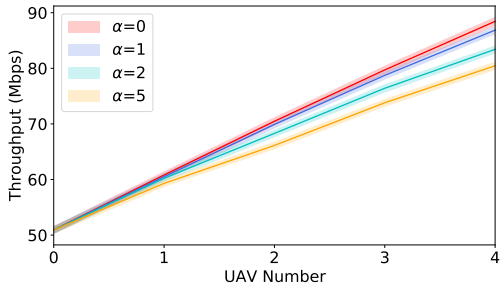


Fig. 12. Throughput with different numbers of UAVs and α values.

Fig. 12 shows the total throughput defined as the sum of vehicles' access rates with different values of α and different numbers of UAVs. The lines show the mean value with intervals for 7 days. We observe that when more UAVs are deployed in the same environment, the overall throughput increases. The total throughput grows with decreasing α, since, as said above, lower α more inclines to improve the overall performance. In particular, $\alpha = 0$ has the highest total throughput as expected since this allocation policy aims to maximize the sum of vehicles' access rates.

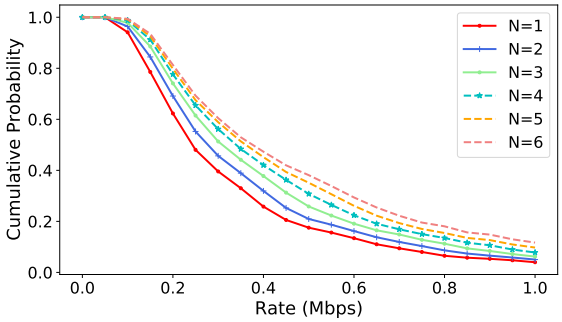
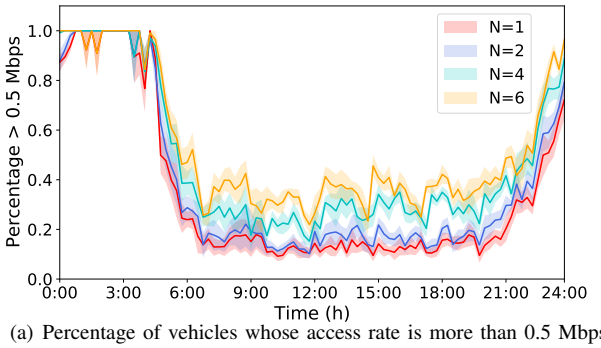


Fig. 13. Cumulative distribution of vehicles' access rates with $\alpha = 1$.

Performance with different UAV swarm sizes. To illustrate performance with an increasing number of UAVs, we show the cumulative distribution of vehicle access data rates in Fig. 13 and vehicle access data rates over time in Fig. 14 for different numbers of UAVs varying from 1 to 6. As previously discussed, the more UAVs, the higher the access rates, but a diminishing returns effect in access rate improvement is observed. For example, around 27%, 35%, 40%, 45%, 47%, and 49% of vehicles have more than 0.4 Mbps with one to six UAVs, respectively. With the other α values, similar results can be observed. According to Fig. 14 which shows the percentage of vehicles whose access rate is higher than 0.5 Mbps, the more UAVs are deployed, the higher the access rate, whether during rush hour or not. Note again the diminishing returns



(a) Percentage of vehicles whose access rate is more than 0.5 Mbps

Fig. 14. Performance with different numbers of UAVs.

behavior previously discussed where the mean access rate increase from 2 to 4 UAVs is more pronounced than that from 4 to 6 UAVs.

VII. Conclusions and Future Work

In this paper, we leverage multiple 5G UAVs to enhance network resource allocation among vehicles by dynamic placement on-demand. We propose a DRL approach to determine UAVs' position to improve the fairness and efficiency of network resource allocation while considering the UAVs' flying range, communication range, and limited energy resources. We use a parametric fairness function for resource allocation that can be tuned to reach different allocation objectives ranging from maximizing the total throughput of vehicles, maximizing minimum throughput, and achieving proportional bandwidth allocation. The results of our simulations show that the dynamic placement of UAVs can improve the fairness of communication. Besides, the UAVs' locations are affected by fairness criteria since they have different preferences to serve different vehicles. Furthermore, we demonstrate that by deploying more UAVs in the same environment, it is possible to improve fairness by serving more underserved vehicles, but with the cost of increasing the training time due to increasing computational complexity.

In future work, we plan to extend our solution to consider scenarios with multiple BSs as well as account for other criteria to drive UAV placement, e.g., adaptive bandwidth allocation and minimizing handover overheads. Furthermore, to reduce the delay and computation load of a centralized agent, we plan to extend our work to a multi-agent system by deploying agents in UAVs. We will also investigate predictive methods for vehicle mobility and resource demands in the UAV placement algorithm to further improve accuracy and efficiency.

Acknowledgment

This work was partly funded by Inria, supported by the French ANR "Investments for the Future" Program reference #ANR-11-LABX-0031-01, and UNICAMP, through the FAPESP Grant number #2017/50361-0, both in the context of the DrIVE #EQA-041801 Associated Team.

References

[1] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1123–1152, 2015.

[2] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," arXiv preprint arXiv:1903.05289, 2019.

[3] E. Vinogradov, H. Sallouha, S. De Bast, M. M. Azari, and S. Pollin, "Tutorial on UAV: A blue sky view on wireless communication," arXiv preprint arXiv:1901.02306, 2019.

[4] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks-with a focus on propagation models," IEEE Transactions on Antennas and Propagation, vol. 65, no. 12, pp. 6213–6230, 2017.

[5] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on uav cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3417–3442, 2019.

[6] R. Ghanavi, E. Kalantari, M. Sabbaghian, H. Yanikomeroglu, and A. Yongacoglu, "Efficient 3d aerial base station placement considering users mobility by reinforcement learning," in 2018 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2018, pp. 1–6.

[7] J. Chen and D. Gesbert, "Optimal positioning of flying relays for wireless networks: A los map approach," in 2017 IEEE international conference on communications (ICC). IEEE, 2017, pp. 1–6.

[8] V. Saxena, J. Jaldén, and H. Klessig, "Optimal UAV base station trajectories using flow-level models for reinforcement learning," IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 4, pp. 1101–1112, 2019.

[9] M. Alzenad, A. El-Keyi, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station for maximum coverage of users with different qos requirements," IEEE Wireless Communications Letters, vol. 7, no. 1, pp. 38–41, 2017.

[10] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3133–3174, 2019.

[11] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," IEEE SIGNAL PROCESSING MAGAZINE, 2017.

[12] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," IEEE Journal on Selected Areas in Communications, vol. 36, no. 9, pp. 2059–2070, 2018.

[13] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," IEEE/ACM Transactions on networking, vol. 8, no. 5, pp. 556–567, 2000.

[14] X. Li, H. Yao, J. Wang, X. Xu, C. Jiang, and L. Hanzo, "A near-optimal uav-aided radio coverage strategy for dense urban areas," IEEE Transactions on Vehicular Technology, vol. 68, no. 9, pp. 9098–9109, 2019.

[15] J. Wang, C. Jiang, Z. Wei, C. Pan, H. Zhang, and Y. Ren, "Joint uav hovering altitude and power control for space-air-ground iot networks," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1741–1753, 2018.

[16] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of uav-mounted mobile base stations," IEEE Communications Letters, vol. 21, no. 3, pp. 604–607, 2016.

[17] X. Li, H. Yao, J. Wang, S. Wu, C. Jiang, and Y. Qian, "Rechargeable multi-uav aided seamless coverage for qos-guaranteed iot networks," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 10 902–10 914, 2019.

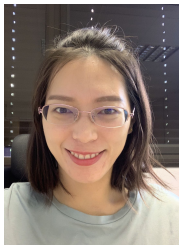
[18] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," IEEE Communications Letters, vol. 20, no. 8, pp. 1647–1650, 2016.

[19] P. S. Bithas, E. T. Michailidis, N. Nomikos, D. Vouyioukas, and A. G. Kanatas, "A survey on machine-learning techniques for UAV-based communications," Sensors, vol. 19, no. 23, p. 5170, 2019.

[20] T. Yuan, W. B. da Rocha Neto, C. Rothenberg, K. Obraczka, C. Barakat, and T. Turletti, "Harnessing machine learning for next-generation intelligent transportation systems: A survey," 2019.

[21] M. S. Shokry, D. Ebrahimi, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Leveraging UAVs for coverage in cell-free vehicular networks: A

deep reinforcement learning approach," IEEE Transactions on Mobile Computing, 2020.
[22] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," IEEE Journal on Selected Areas in Communications, vol. 35, no. 5, pp. 1046–1061, 2017.
[23] A. Goldsmith, Wireless communications. Cambridge university press, 2005.
[24] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," IEEE Wireless Communications Letters, vol. 3, no. 6, pp. 569–572, 2014.
[25] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," IEEE Transactions on Wireless Communications, vol. 18, no. 4, pp. 2329–2345, 2019.
[26] D. Dias and L. H. M. K. Costa, "CRAWDAD dataset coppe-ufrj/riobuses (v. 2018-03-19)," Downloaded from https://crawdad.org/coppe-ufrj/RioBuses/20180319, Mar. 2018.
[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., "Human-level control through deep reinforcement learning," nature, vol. 518, no. 7540, pp. 529–533, 2015.
[28] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour et al., "Policy gradient methods for reinforcement learning with function approximation." in Proceedings of Conference on Neural Information Processing Systems (NIPS), vol. 99. Citeseer, 1999, pp. 1057–1063.
[29] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.
[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.
[31] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in International conference on machine learning. PMLR, 2016, pp. 1928–1937.
[32] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. Tb, A. Muldal, N. Heess, and T. Lillicrap, "Distributed distributional deterministic policy gradients," arXiv preprint arXiv:1804.08617, 2018.
[33] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in International Conference on Machine Learning. PMLR, 2018, pp. 1587–1596.
[34] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in International conference on machine learning. PMLR, 2018, pp. 1861–1870.
[35] S. Kullback and R. A. Leibler, "On information and sufficiency," The annals of mathematical statistics, vol. 22, no. 1, pp. 79–86, 1951.
[36] Z. Zhang, "Improved adam optimizer for deep neural networks," in 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, 2018, pp. 1–2.



Tingting Yuan received her Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2018. During the year 2018-20, she was a postdoctor at INRIA, Sophia Antipolis, France. Since 2020, she joined the University of Göttingen, as a postdoctor. Her current interest interests are about next-generation network, including the software defined networking, deep reinforcement learning, vehicular ad-hoc networks and so on.

Christian Esteve Rothenberg is an Assistant Professor in the Faculty of Electrical & Computer Engineering (FEEC) at University of Campinas (UNICAMP), Brazil, where he received his Ph.D. and currently leads the Information & Networking Technologies Research & Innovation Group (INTRIG). His research spans all layers of distributed systems and network architectures and are often carried in collaboration with industry, resulting in multiple open source projects among other scientific results.

Katia Obraczka is Professor of Computer Engineering at UC Santa Cruz. Before joining UCSC, she held a research scientist position at USCs Information Sciences Institute and a joint appointment at USCs Computer Science Department. Her research interests span the areas of computer networks, distributed systems, and Internet information systems. She is the director of the Internetwork Research Group (i-NRG) at UCSC and has been a PI and a co-PI in a number of projects sponsored by government agencies (NSF, DARPA, NASA, ARO, DoE, AFOSR) as well as industry. Prof. Obraczka has edited one book, wrote a number of book chapters, and published over 200 technical papers in journals and conferences. She received the USC ISIs Meritorious Service Award in 1999. She is a fellow member of the IEEE.

Chadi Barakat is Senior Researcher at Inria - Sophia Antipolis since March 2002. He got his master, Ph.D. and HDR degrees in Computer Sciences from the University of Nice Sophia Antipolis in 1998, 2001 and 2009, respectively. He was general chair for ACM CoNEXT 2012, PAM 2004 and WiOpt 2005 workshops, guest editor for a JSAC special issue on sampling the Internet, area editor for the ACM CCR journal and is currently on the editorial board of Elsevier Computer Networks. His main research interests are in Internet measurements and network data analytics, user quality of experience, and mobile wireless networking. He is senior member of the IEEE and of the ACM.

Thierry Turletti received the M.S. (1990) and the Ph.D. (1995) degrees in computer science from the University of Nice - Sophia Antipolis, France. During the year 1995-96, he was a postdoctoral fellow at LCS, MIT and worked in the area of Software Defined Radio. He is currently a senior research scientist at the DIANA team at INRIA. His current research interests include Wireless Networking, Programmable Networks and Networking Experimental Platforms. He has been serving on the Editorial Boards of the following journals: Wireless Communications and Mobile Computing (2001-2010), Wireless Networks (since 2005) and Advance on Multimedia (since 2007). He is senior member of the IEEE and of the ACM.