Dispatch of UAVs for Urban Vehicular Networks: A Deep Reinforcement Learning Approach

Omar Sami Oubbati, Mohammed Atiquzzaman, Abdullah Baz, Hosam Alhakami, Jalel Ben-Othman

Dispatch of UAVs for Urban Vehicular Networks: A Deep Reinforcement Learning Approach

Omar Sami Oubbati, *Member, IEEE*, Mohammed Atiquzzaman, *Senior Member, IEEE*, Abdullah Baz, *Senior Member, IEEE*, Hosam Alhakami, *Senior Member, IEEE*, Jalel Ben-Othman, *Senior Member, IEEE*

***Abstract*—Due to the dynamic nature of connectivity in terrestrial vehicular networks, it is of great benefit to deploy unmanned aerial vehicles (UAVs) in these networks to act as relays. As a result, a remarkable number of studies have exploited UAVs to bridge the communication gaps between terrestrial vehicles, and sometimes despite their unoptimized mobility, their restricted communication coverage, and their limited energy resources. However, it was noted that for an intermittently connected vehicular network, UAVs could not cover all sparse areas all the time. Even worse, when deploying enough UAVs to cover all these areas, the probability of inter-UAV collisions increases, and it will be complex to control their movements efficiently. Consequently, it is required to dispatch an organized and intelligent group of UAVs to perform communication relays in the long term while keeping their connectivity, minimizing their average energy consumption, and providing an efficient coverage strategy. To meet these requirements, we propose a deep reinforcement learning (DRL) framework, called DISCOUNT (Dispatch of UAVs for Urban VANETs). Extensive simulations have been conducted to evaluate the performance of the proposed framework. It has been shown that the proposed framework significantly outperforms two commonly-used baseline techniques and some reinforcement learning methods in terms of energy consumption, coverage, and routing performances.**

***Index Terms*—UAV; Deep Reinforcement Learning; VANET; Energy Efficiency; Routing.**

I. Introduction

THE deployment of Unmanned Aerial Vehicles (UAVs) has become a practical solution in various applications in wireless systems [1]. Due to their many desirable advantages, including ease of deployment, swift mobility, and low-cost operation, UAVs turn into a mainstream tool of our everyday life. UAVs can be quickly deployed to complement existing terrestrial networks by providing various assistance services ranging from increasing network capacity and aerial monitoring to removing coverage holes and improving connectivity [2]. In the literature, UAVs have been mostly deployed to provide temporary coverage for overloaded terrestrial networks or as relays for terrestrial mobile networks suffering from frequent intermittent connectivity [3]. Recently, UAVs have been deployed as mobile relays to cooperate with existing vehicular ad hoc networks (VANETs) on the ground to get around disruptive ground obstacles, highly dynamic topology, frequent disconnections, and participate reliably in data transmission between vehicles [4], [5]. Also, the UAV assistance to VANETs is only done on the basis of several UAVs forming an aerial subnetwork, which can cover a broad region, detect any sparse connections appearing on the ground, and efficiently play the role of relays between disconnected vehicles.

To provide sufficient long-term communication coverage to a VANET deployed on a large urban area, a sufficient number of mobile UAVs would be required to operate as an autonomous connected group that intelligently covers every disconnected zone, as illustrated by a motivating scenario plotted in Fig. 1. However, the UAV deployment faces several constraints and challenges. For instance, since sparsely connected zones can appear anywhere and at any time, it is required that UAVs hover and continuously change their positions to cover these zones for a reasonable amount of time while linking the maximum number of disconnected vehicles. As a consequence, three main challenges could be encountered. First, UAVs could quickly exhaust their on-board energy due to their constant mobility, and thus restricting their communication capabilities significantly. Second, the probability of inter UAV collisions increases considerably when UAVs are flying close to each other. Finally, since UAVs are constantly moving mainly at the same and low altitude to communicate with vehicles, the possibility of colliding with obstacles (*e.g.,* skyscrapers or high trees) is extremely high. To overcome all these challenges, it is required to make intelligent decisions regarding each UAV trajectory so that the available energy among UAVs should be optimally and fairly exploited to provide coverage to as many disconnected zones as possible. For this purpose, it is supposed that UAVs are backhaul-connected through satellite links, which is a common assumption in different works [6], [7]. Then, the current state of the environment (*i.e.,* positions, energy usages, number of connected zones, etc.) is continually observed by the control agent located at the satellite, which allows controlling

O.S. Oubbati is with LIGM, University of Gustave Eiffel, Marne-la-Vallée, France. E-mail: omar-sami.oubbati@univ-eiffel.fr

M. Atiquzzaman is with the University of Oklahoma, Norman, OK USA. E-mail: atiq@ou.edu

A. Baz is with the Department of Computer Engineering, College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia. Email: aobaz01@uqu.edu.sa

H. Alhakami is with the Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia. Email: hhhakam@uqu.edu.sa

J. Ben-Othman is with is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, and Université Sorbonne Paris Nord, France. E-mail: jalel.benothman@centralesupelec.fr

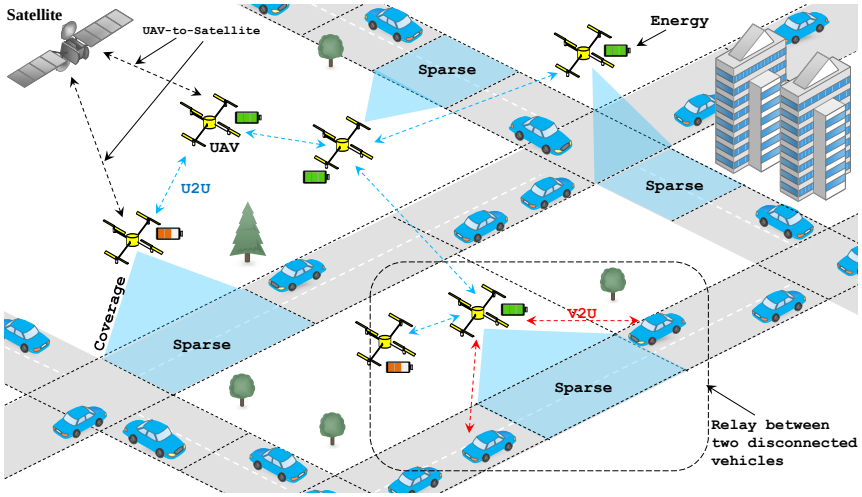


Fig. 1: Motivating scenario.

each UAV movement towards achieving an intelligent UAV coverage over VANETs. It is worthy to note that UAVs preserve their connectivity to participate in the data delivery between ground vehicles reliably, and especially when the VANET is highly fragmented (*i.e.,* there is no routing path between communicating vehicles).

The main focus of this piece of work is to address the above issues and challenges by jointly optimizing the trajectories of UAVs and providing as much intelligent coverage as possible for sparse areas within a terrestrial VANET. This task is a complicated problem, which makes it difficult to process using traditional optimization techniques. Hence, we propose to leverage Deep Reinforcement Learning (DRL) methods that have proven their efficiency by handling complicated state space and time-varying environments. Moreover, these methods can provide peak performance on a good number of learning tasks with little or no domain knowledge.

To ensure reliable connectivity of VANETs based on the assistance of UAVs, a set of contributions are carried out in this paper as follows:

- Maximizing the UAV coverage all over sparsely connected areas, while increasing the number of linked vehicles through relay UAVs.
- Reducing collisions between UAVs, while avoiding obstacles during their mobility.
- Minimizing the average energy consumption while ensuring the integrity of the UAV network by optimally using available energy on each UAV.
- Preserving the UAV network connectivity during the whole flight period.
- Conducting a series of simulations to evaluate the performance of the proposed work.

The remainder of this paper is outlined as follows. In Section II, the relevant works of the state-of-art are reviewed. Section III describes the proposed system model as well as the problem formulation. In Section IV, we present both the required preliminaries about DRL and the proposed DRL-based approach. The simulation results for performance evaluation are illustrated in Section V. Concluding remarks are outlined in Section VI.

II. Related Work

The optimization of control, deployment, and trajectory of UAVs have been widely studied in recent years. Various solutions have been proposed towards different goals, which take into account different requirements. Moreover, there has been a renewed interest in applying machine learning (ML) techniques to the mobility management of UAVs. Therefore, as part of our work, we focus our attention here on recent studies proposed in the context of several research areas, such as the UAV deployment and coverage optimization, UAV-assisted VANETs, and DRL-based for UAV trajectory, deployment, and coverage optimization.

A. Deployment of UAVs and Coverage

The optimization of UAV deployment and coverage is considered a complex task in emerging UAV-assisted systems. In [8], the authors proposed an ellipse clustering algorithm to allow UAV base stations (UAV-BSs) to increase the probability of covering the maximum of ground users while avoiding inter-cell interference. Moreover, an energy-efficient 3D deployment method is proposed to reduce the total amount of consumed energy by UAV-BSs, while ensuring a certain level of Quality of Service (QoS) for every covered ground user. The authors of [9] proposed a path planning optimization to enhance wireless coverage using UAVs during a disaster scenario by complementing affected infrastructures. An optimal separation distance between UAVs was proposed in [10] to both mitigate co-channel interference and increase the coverage of multiple UAVs in urban areas. In [11], the authors presented an iterative approach where a minimal number of UAVs are deployed to improve the communication coverage for user equipments (UEs) with unknown positions. Two fast UAV deployment problems were studied in [12] for maximizing wireless coverage. The first problem is to reduce the total deployment delay for efficiency, and the second is to minimize the deployment delay for fairness consideration.

TABLE I: Features comparison with the related works.

	Deployment of UAVs and Coverage		UAV-assisted VANETs		DRL-based UAV-assisted systems				
Features	Ref. [9]	Ref. [11]	Ref. [13]	Ref. [14]	Ref. [15]	Ref. [16]	Ref. [17]	Ref. [18]	Our framework
Basic ideology	Optimization of UAV coverage over disaster areas	Deploying a reduced number of UAVs to serve UEs while preserving their connectivity	Optimization of VANET routing process using UAVs	Improvement of VANET connectivity	Integration of cellular-connected UAVs and UAV-BSs through 3D cellular networks	Minimization of latency based on interference-aware path planning of cellular-connected UAVs	A DRL-based method to provide optimized energy-efficient UAV trajectory while minimizing AoI	A multi-UAV long-term coverage using DRL	DRL-based UAV-assisted VANET
Type of assisted networks	VANET	UEs	VANET	VANET	UAV Network	BSs/UEs	UEs	IoT	VANET
Optimized metrics	UAV trajectory, Throughput, Fairness	Coverage, Load balance, Throughput	Packet loss, Delivery delay	Routing path availability, Delivery delay	Latency, Throughput	Latency, Throughput	AoI, Energy consumption, Resource utilization	Energy consumption, Coverage, Fairness	Energy consumption, Coverage, Routing path, Delivery Delay
UAV density	Multiple UAVs	Multiple UAVs	Multiple UAVs	Single UAV	Multiple UAVs/UAV-BSs	Multiple UAVs	Multiple UAVs	Multiple UAVs	Multiple UAVs
Mobility of UAV(s)	Dynamic	Dynamic	Dynamic	Dynamic	Dynamic	Static	Dynamic	Dynamic/Static	Dynamic/Static
Mobility of assisted nodes	Dynamic	Static	Dynamic	Dynamic	Static	Static	Static	Static	Dynamic
Major advantage	Throughput is fairly maximized among vehicles	A high coverage ratio is ensured through a reduced number of UAVs	UAVs are used as alternatives (relays) in case of disconnections	Fill in communication gaps between vehicles using UAVs	Latency and spectral efficiency of cellular-connected UAVs are optimized	A better latency and rate are achieved per each UAV and UE, respectively	Enhancement of energy efficiency and collected data freshness	Maximizing the fair temporal coverage of IoT devices	Optimization of UAV trajectory while considering its energy consumption and connectivity
Major Limitation	UAV energy consumption is not considered	Unexpected disconnections and UAV energy consumption are not considered	Optimization of trajectory and energy consumption of UAVs are overlooked	High delivery delays due to UAV movements	Energy consumption and mobility of different UAVs are not considered	Energy consumption of UAVs is not considered and evaluated	Mobility of UEs is not considered	Collisions between UAVs are overlooked during the training process	Overlooking fairness in coverage of sparse areas

However, reducing collisions between UAVs and avoiding obstacles, especially in urban areas, can allow UAVs to perform coverage missions reliably. Moreover, the terrestrial environment is generally unknown for UAVs, which requires some interaction to explore it. These constraints have not been considered in the above-cited works.

B. UAV-assisted VANETs

Much research has been recently conducted to address various issues in UAV-assisted VANETs [13], [19]–[22]. For instance, in [13], the authors used cooperative UAVs, playing the role of relays over an urban VANET. This scheme estimates the link expiration time to make the data transmission between vehicles more reliable. In [23], a game-theory-based UAV-assisted VANET scheme is proposed. The main purpose of this scheme is to predict disconnected road segments and deploy UAVs accordingly. The authors of [24] proposed a novel UAV-assisted data dissemination strategy, where a recursive least square algorithm predicts the movement of vehicles, and thus maximizes throughput and minimizes the delay. Fawaz *et al.* [14] deployed UAVs as store-carry and forward nodes over a fully-collaborative VANET with the aim to enhance connectivity and minimize the delivery delay. The authors of [25] designed a UAV-assisted cooperative data dissemination strategy in VANETs to minimize network transmission delay and speed up the file downloading by deploying UAVs as relays with data caching features.

However, deploying UAVs without adequately controlling their movements, energy consumption, and continuous connectivity cannot provide better control of UAVs to achieve the assigned tasks efficiently.

C. DRL-based UAV-assisted systems

Actually, we are witnessing a resurgence of interest in ML techniques in the context of UAV-assisted systems. This is due to their ability to address complicated issues that are difficult, if not impossible, to solve using traditional optimization techniques. For example, Mozaffari *et al.* [15] proposed a novel machine learning approach to optimize a novel 3D UAV cell association scheme within a cellular network consisting of a set of UAV users and UAV-mounted BSs. To ensure fair coverage of each ground user and energy efficiency, the authors in [26] proposed a 3D UAV deployment algorithm based on Deep Deterministic Policy Gradient to schedule the mobility planning of multiple UAVs for energy replenishment. In [16], the authors exploited a DRL method based on an echo state network to support an interference-aware path planning method for cellular-connected UAVs. This method allows the UAV to optimize its direction, transmission power, and cell association, which can minimize the transmission latency and interference on the ground. In [27], the authors leveraged DRL to find the appropriate trajectories for a minimum number of UAVs to provide coverage and acceptable QoS to vehicles on a highway. In [17], the trajectory of UAVs is optimized to minimize both their energy consumption and the average age of information (AoI) of collected information under a predefined threshold. In [18], a decentralized DRL algorithm is developed to navigate a network of UAV-BSs for providing coverage to mobile ground users while considering their energy consumption and preserving their connectivity. Wand *et al.* [28] proposed a DRL approach for UAV navigation and control signals in complex urban areas without any knowledge of mapping information provided by sensed data. The authors of [29] proposed a DRL method to control multiple UAV-BSs for providing effective coverage over terrestrial users.

However, due to the extensive computational power of DRL methods and the restricted energy capacity of UAVs, it is required to preserve connectivity among UAVs while avoiding collisions between them or with obstacles. These requirements are essential to provide an updated coverage status and for the environment enlightenment.

In this context, when deploying a set of UAVs over a terrestrial VANET, this paper considers all the requirements related to each aforementioned research area, while assuming no prior knowledge of the terrestrial environment. To the best of our knowledge, filling communication gaps of terrestrial urban VANETs using an intelligent connected group of UAVs controlled by a DRL method has not been attempted so far. Table I highlights the novelties of our approach compared with the previously discussed works based on various features.

III. System Model and Problem Statement

As depicted in Fig. 1, we consider a terrestrial vehicular network with a medium density of vehicles. Traditionally, such a network suffers from frequent disconnections that disrupt communications between vehicles. In this work, we deploy a set $\mathcal{M} \triangleq \{i = 1, 2, \ldots, M\}$ of UAVs, which are intended to serve as relays by flying horizontally between disconnected vehicles moving within road segments, while maintaining a fixed optimal altitude h. A particular 2D urban square area of width W is considered where the road segments, as well as the other areas, are divided into a set $\mathcal{Z} \triangleq \{z = 1, 2, \ldots, Z\}$ of fixed size square zones. Then, the vehicles are designated as $\mathcal{V} = \{V_1, \ldots, V_Z\}$, where V_z is a subset of vehicles belonging to Zone$_z$, where $z \in \mathcal{Z}$. Each vehicle is supposed to belong to only one single zone, thus $V_z \cap V_{z'} = \emptyset$, where $z \neq z', \forall z, z' \in \mathcal{Z}$. It is noteworthy that the density of vehicles present within the square area is variable. However, to avoid the complexity of having variable densities of vehicles, we suppose that the maximum number of vehicles within the target area is $|\mathcal{V}| = N$. We consider that communicating vehicles are disconnected if, and only if there is at least one zone Zone$_z$ separating them, where $|\mathrm{V}_z| = 0$ (*i.e.,* V_z is empty). The initial distribution of vehicles and the zone partitioning will be discussed in Section III-B. Our system is analyzed over a given flight period of T time-slots of length $t \in \mathcal{T} = \{1, 2, \ldots, T\}$. Each UAV$_i$ is aware of its own coordinate at each time-slot t, which is given by: $p_i^t = [(x_i^t, y_i^t, h_i)]^T \in \mathbb{R}^{3\times 1}$. In addition, it is required that each UAV$_i$ with a communication range R_i maintains at least a single connectivity to the whole UAV network. This allows that a connected UAV network could be exploited as a relay bridge in the case when the terrestrial vehicular network is highly fragmented. It should be stressed that the deployed satellite has a role to operate as a central DRL agent to control the movements of the UAV network, which will also be discussed in Section IV.

To ensure reliable relay between disconnected vehicles, we assume that the center of each empty zone on the road, called Center-of-Interest (CoI), must be covered by at least a single UAV for a required amount of time. The covered zones cannot remain empty all the time due to the dynamic topology of vehicles. Thus, the UAVs need to fly around to cover other disconnections formed elsewhere. Once UAVs reach their target CoIs, they hover there and start their role of relays for a given number of time-slots until that covered CoIs become connected (*i.e.,* become occupied by vehicles). Due to the limited number of UAVs, we prioritize covering certain zones over others, according to the following priorities in the order listed: (i) Covering empty zones that directly connect the maximum occupied zones and (ii) Covering isolated empty zones. This task is very challenging due to the limited capacity of resources of UAVs in terms of energy resources and communication coverage. Indeed, at the initial stage of the task, UAVs take off from a random origin, and learn to fly horizontally with a direction $\Omega_i^t \in \{0, \ldots, 2\pi\}$ and distance $d_i^t \in \{0, \ldots, d_{max}\}$, or hovering at their current locations (*i.e.,* $d_i^t = 0$).

The energy consumption follows a linear model, which increases with the increase of flying distance. It should be stressed that the energy consumption follows the same energy consumption model proposed in [30] in which the propulsion energy can be expressed as follows:

$$\begin{aligned} Prop(V) = & \underbrace{P_b \left(1 + \frac{3V^2}{V_{tip}^2}\right)}_{blade\ profile\ power} \\ & + \underbrace{P_i \left(\sqrt{1 + \frac{V^4}{4u_0^2}} - \frac{V^2}{2u_0^2}\right)}_{induced\ power} \\ & + \underbrace{\frac{1}{2} f_0 a n R V^3}_{parasite\ power} \end{aligned} \tag{1}$$

where P_i and P_b represent the induced power in a hover state and the blade profile power, respectively. V denotes the flying speed of the UAV, V_{tip} represents the tip speed of the rotor blade, and u_0 and n indicate the mean induced velocity and the solidity of the rotor, respectively. f_0, a, and R are the fuselage drag ratio, the air density, and the rotor disc area, respectively. In each time-slot $t \in \mathcal{T}$ and after performing either hovering or flying horizontally, UAV$_i$ consume an energy of $E_i^t \in [f, E_{max}]$, where $f = P_i + P_b$ is the hovering energy and E_{max} is the maximum energy capacity that could be consumed by each UAV. The total energy consumption of UAV$_i$ during the whole period t can be calculated as follows:

$$E_i^t = \underbrace{\int_0^t Prop(\|v(t)\|)dt}_{propulsion\ energy} \tag{2}$$

For more clarity, Table II shows the major notations used in this paper.

TABLE II: List of notations.

Symbol	Explanation
$\mathcal{M}, M, i$	Set, number, and index of UAVs
$\mathcal{V}, N, n$	Set, number, and index of vehicles
$\mathcal{Z}, Z, z$	Set, number, and index of zones
$\mathcal{T}, T, t$	Set, number, and index of time-slots
R_i, E_i^t, p_i^t	Range, Consumed energy, and position of UAV$_i$
$d_i^t, \Theta_i^t, \Omega_i^t$	Distance, Activity, and Direction of UAV$_i$
R_n, p_n^t	Range and position of vehicle C_n
CoI$_z$, NBZ_{iz}	Center-of-Interest of zone z, Number of linked zones by UAV$_i$ from zone z.
$Con_i(t)$, $Con(t)$	Number of covered zones by UAV$_i$, Number of covered zones by all UAVs.
$Hov_{i,z}^t$, $Con_{i,z}^t$	Hovering and Coverage states of CoI$_z$ by UAV$_i$.
s_t, a_t, r_t	State, action, and reward of all UAVs.
$\mathcal{D}$, D, F	Replay buffer, Size of the buffer, Size of the mini-batch.
$Q(.)$, $r(.)$, tgt_{DDQN}	Q function, Reward function, Target value in DDQN.
γ, ε, ϱ	Discount factor, Exploration probability, Decrement factor.
$\pi(.)$, $\pi^*(.)$, $L(.)$	Policy function, Optimal policy function, Loss function.
η^Q, η^Q_{target}	Weights of $Q(.)$ network, Weights of target network.

A. Channel Model

Generally, the channel model of the UAV-assisted systems includes three different models, *i.e.,* Air-to-Ground (A2G) channel, Ground-to-Air (G2A) channel, and Air-to-Air (A2A) channel, which describe the links between UAVs and vehicles and the links between solely UAVs, respectively. To avoid interference, different carrier waves are used for each kind of link. Moreover, as in [31], the Doppler effect is assumed to be perfectly compensated based on the GPS, where the speeds and locations of UAVs can be accurately predicted. To illustrate channel modeling formulation, we consider the descriptive scenario depicted in Fig. 2 by zooming in the scenario named (*Relay between two disconnected vehicles*) in Fig. 1.

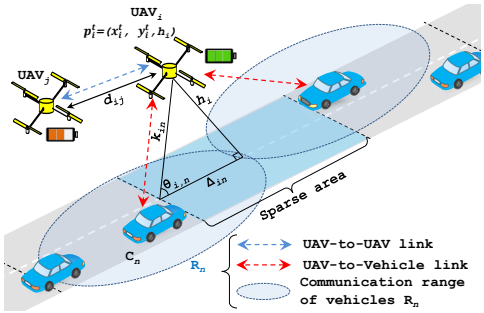


Fig. 2: A relay scenario with a single UAV linking two disconnected vehicles.

1) A2A Channel: The A2A channels are mainly dominated by line-of-sight (LoS) links due to the lack of obstacles in the sky. According to [32], the LoS pathloss (in dB) between a pair of communicating UAVs (*e.g.,* UAV$_i$ and UAV$_j$) can be modeled by the free space propagation loss (FSPL), which can be expressed as follows:

$$PL_{ij}^{A2A}[dB] = 10\zeta \log_{10}\left(\frac{4\pi\psi d_{ij}}{\lambda}\right) + L^{LoS} \tag{3}$$

where ζ is the free space path loss exponent, which is set to 2 according to FSPL. ψ is the DSRC carrier frequency, $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, $\lambda = 3 \times 10^8 m/s$ is the speed of the light, and L^{LoS} representing the attenuation that is added to the LoS environment. It should be stressed that the A2A channel model is less impacted by fading. To ensure that any pair of UAVs is connected, the signal-noise ratio (SNR) at the receiver should be greater than a certain threshold of τ_{A2A} as follows:

$$SNR_{ij}^{A2A} = Pow_{ij}^{t,A2A} - PL_{ij}^{A2A} - \sigma_{A2A} \geq \tau_{A2A} \tag{4}$$

where $Pow_{ij}^{t,A2A}$ is the sender transmit power, σ_{A2A} is the noise power at the receiver side. It is worthy to note that the maximum value of d_{ij}, which satisfies the equation (4) is considered to be the communication range $R_i = d_{ij}$, $\forall i, j \in \mathcal{M}$.

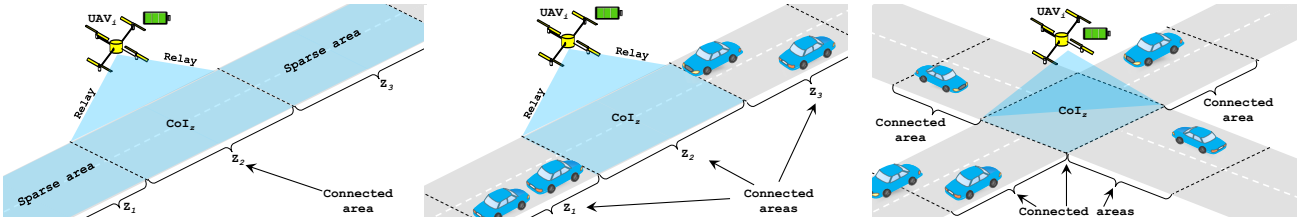
2) A2G/G2A Channels: A growing number of studies related to the wireless coverage of UAVs deployed over a terrestrial network have been studied. A vehicle may have LoS connection with the UAV with which it communicates. However, different obstructions (*e.g.,* buildings, high trees, or bridges) might exist between the vehicle and the UAV, causing a non-LoS (NLoS) connection, where the vehicle still receives a signal from the UAV due to reflections and diffractions. Consequently, the A2G/G2A channels can be modeled based on [33], [34]. As illustrated in Fig. 2, the received power from both UAV$_i$ (x_i, y_i, h_i) and vehicle C$_n$ (x_n, y_n) is given by the below equation:

$$\begin{aligned} Pow_{in}^{r,A2G} &= \begin{cases} Pow_{in}^{t,A2G} - L_{LoS}^{i,n}, & \text{LoS links} \\ Pow_{in}^{t,A2G} - L_{NLoS}^{i,n}, & \text{NLoS links} \end{cases} \\ Pow_{in}^{r,G2A} &= \begin{cases} Pow_{in}^{t,G2A} - L_{LoS}^{i,n}, & \text{LoS links} \\ Pow_{in}^{t,G2A} - L_{NLoS}^{i,n}, & \text{NLoS links} \end{cases} \end{aligned} \tag{5}$$

where $Pow_{in}^{t,A2G}$ and $Pow_{in}^{t,G2A}$ are the transmission powers of UAV$_i$ and vehicle C_n, respectively. $L_{LoS}^{i,n}$ and $L_{NLoS}^{i,n}$ are the average LoS and NLoS path-losses (in dB), and they can be expressed as:

$$\begin{aligned} L_{LoS}^{i,n} &= 20\log\left(\frac{4\pi\omega k_{in}}{\lambda}\right) + \delta(LoS), \\ L_{NLoS}^{i,n} &= 20\log\left(\frac{4\pi\omega k_{in}}{\lambda}\right) + \delta(NLoS), \end{aligned} \tag{6}$$

where ω is the carrier frequency of UAV-to-vehicle channel and λ is the speed of the light. k_{in} is defined as the distance between UAV$_i$ and C$_n$, which is given by $k_{in} = \sqrt{h_i^2 + \Delta_{in}^2}$, where $\Delta_{in} = \sqrt{(x_i - x_n)^2 + (y_i - y_n)^2}$ represents the horizontal euclidean distance between UAV$_i$ and C$_n$. $\delta(.)$ is the average additional path-loss, which is dynamic according to LoS and NLoS environments. Therefore, we deduct that A2G/G2A links are connected, if their respective SNR



(a) $NBZ_{iz} = 1$. (b) $NBZ_{iz} = 3$. (c) $NBZ_{iz} = 5$.

Fig. 3: UAV coverage based on hovering zones.

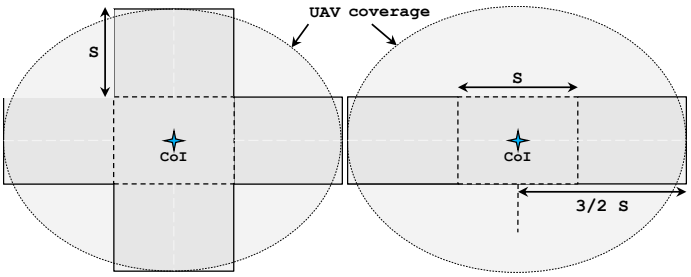
(SNR_{in}^{A2G} and SNR_{in}^{G2A}) at the receiver side are larger than the given thresholds (τ_{A2G} and τ_{G2A}). That is:

$$\begin{aligned} SNR_{in}^{A2G} &= Pow_{in}^{r,A2G} - \sigma_{A2G} \geq \tau_{A2G} \\ SNR_{in}^{G2A} &= Pow_{in}^{r,G2A} - \sigma_{G2A} \geq \tau_{G2A} \end{aligned} \tag{7}$$

where σ_{A2G} and σ_{G2A} are the noise powers at their respective receiver sides.

B. Problem Statement

Before describing the problem statement, it is essential to present some assumptions and concepts. Initially, we suppose that vehicles are non-uniformly distributed over the roads of the target area and the traffic of vehicles follows a normal distribution. Moreover, the speeds of vehicles are generated based on a truncated Gaussian distribution with a mean equal to 50 km/h. Then, the mobility traces of vehicles are generated during T time-slots based on a customized version of VanetMobiSim [35]. We consider that all road segments are divided into equal fixed square zones of width S, which approximately equals to the transmission range R_n of ground vehicles for vehicle-to-vehicle (V2V) communications that are established based on other embedded wireless interfaces in vehicles, such that $0 < (R_n - S) \leq \epsilon$, where $0 < \epsilon \ll 1$. This will cause a disconnection between two communicating vehicles if there is at least one empty zone between each other. By considering the constraint of the equation (7), we assume that each UAV$_i$ and vehicles C_n could establish a successful communication between each other at a maximum horizontal distance $\Delta_{in} = \frac{3S}{2}$ with a maximum altitude of h_i. Based on this assumption, we can say that each UAV$_i$ that will be positioned (hovering) over the CoI of a given zone, can both cover the entire current zone as well as its adjacent zones as shown in Fig. 4, and serve multiple vehicles simultaneously by employing a time division multiple access (TDMA).



(a) Intersection. (b) Road segment.

Fig. 4: Decomposition of zones.

1) Total coverage of the disconnected zones: The objective of this paper is to link the maximum number of disconnected vehicles while minimizing the energy consumption of UAVs. For the simplicity of analysis, maximizing the number of linked vehicles is considered the same as maximizing the number of covered zones. For this purpose, we have to find a control policy to define how UAVs should move in each time-slot over a set of dynamic CoIs (*i.e.,* disconnected zones). To mathematically formulate the problem, we introduce a binary variable $Hov_{i,z}^t \in \{0,1\}$, to indicate whether CoI$_z$ $\in$ Zone$_z$ is hovered (covered) by UAV$_i$ in a time-slot t, $Hov_{i,z}^t$ is expressed as follows:

$$Hov_{i,z}^t = \begin{cases} 1, & \text{if } \sqrt{(x_i^t - x_z)^2 + (y_i^t - y_z)^2} \leq \epsilon \\ 0, & \text{Otherwise} \end{cases} \tag{8}$$

To update the relay status of a given UAV$_i$ hovering over a given CoI$_z$, we calculate the SNR between UAV$_i$ and each vehicle C_n in range. In this manner, we get all possible connections between UAV$_i$ and all vehicles around satisfying a certain QoS level. Also, we calculate the distances of all vehicles from the hovered CoI$_z$ to know precisely how many zones are being linked and connected. The example of Fig. 3 represents the calculation process of connected zones in different scenarios. The detailed process is shown in Algorithm 1.

Algorithm 1: Zones connectivity

```
input : CoIz, UAV_i, V, B.
output: NBZ_iz      /* Number of connected zones. */
begin
    Initialization:
    NBZ_iz ← 1
    B ← ∅
    foreach C_n ∈ V do
        if (SNR_in^A2G ≥ τ_A2G) ∧ (SNR_in^G2A ≥ τ_G2A) then
            if Δ_zn ≤ 3S/2 then
                subset ← Search_subset(C_n, V)
                if subset ∩ V_z = ∅ then
                    if subset ∉ B then
                        B ∪ subset
                        NBZ_iz ← |B|+1
                else
                    NBZ_iz ← 0
                    Break
```

where NBZ_{iz} is the number of zones that are covered by UAV$_i$ from CoI$_z$. We also define a connectivity variable $Con_{i,z}^t$ indicating the coverage state by UAV$_i$ of CoI$_z$ at a time-slot t, $Con_{i,z}^t$ is defined as follows:

$$Con_{i,z}^t = \begin{cases} NBZ_{iz}, & \text{if } Hov_{i,z}^t = 1 \wedge NBZ_{iz} \geq 1 \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

The total coverage carried out by the i^{th} UAV at a time-slot t is given by (10):

$$Con_i(t) = \sum_{z=1}^{Z} Con_{i,z}^t \tag{10}$$

The total number of the entire covered zones at time interval t, $Con(t)$ is given by (11):

$$Con(t) = \sum_{i=1}^{M} \sum_{z=1}^{Z} \sum_{t=1}^{t} Con_{i,z}^t \tag{11}$$

2) UAV network connectivity: To have an accurate view of the UAV network connectivity, we need to define the status of the different links among UAVs at each time-slot t. To do so, a symmetric matrix $\mathbf{l}= [l_{i,j}]_{M\times M}$ is introduced to denote the existing links among UAVs, where $l_{i,j} = 1$ means that UAV$_i$ is connected with UAV$_j$ (*i.e.,* UAV$_i$ is within UAV$_j$'s communication range), otherwise, $l_{i,j} = 0$. UAVs periodically broadcast Hello packets to their one-hop neighbors, which contain their respective geographical positions. Based on this information, each UAV can calculate the distances with one-hop neighbors, which are periodically collected by the centralized entity to update the matrix $\mathbf{l}$ as shown in Algorithm 2.

Algorithm 2: UAV network connectivity

```
input : M
output: l
begin
    for ∀i ∈ M and ∀j ∈ M do
        if d_ij ≤ R_i then
            l_{i,j} ← 1
        else
            l_{i,j} ← 0
```

3) Optimal UAV mobility problem: For simplicity, after the mobility of each UAV$_i$ at each time-slot t, it remains $E - \sum_{t=1}^{t} E_i^t$ of energy, $\forall i \in \mathcal{M}$, where E is the total energy capacity within each UAV. Therefore, the average energy consumption of all UAVs during the whole flight period is given by $\left(\frac{\sum_{i=1}^{M}\sum_{t=1}^{T} E_i^t}{M}\right)$. For the convenient presentation, let $P = \{p_i(t), \forall i \in \mathcal{M}\}$ be the set of locations of all UAVs at a time-slot t. To find the optimal UAV mobility to provide reliable relaying services to disconnected zones, we propose solving the following optimization problem:

$$\max_{P} \quad \mathbb{E}\left[\frac{\sum_{i=1}^{M}\sum_{z=1}^{Z}\sum_{t=1}^{T} Con_{i,z}^{t}}{\left(\frac{\sum_{i=1}^{M}\sum_{t=1}^{T} E_i^t}{M}\right)}\right]$$

$$\begin{aligned}
\text{s.t.} \quad & \textbf{C1:} && (E - \sum_{t=1}^{t} E_i^t) > 0 && \forall i \in \mathcal{M}, \forall t \in \mathcal{T} \\
& \textbf{C2:} && Hov_{i,z}^{t} \leq 1 && \forall i \in \mathcal{M}, \forall t \in \mathcal{T}, \forall z \in \mathcal{Z} \\
& \textbf{C3:} && \|p_i[t] - p_j[t]\|^2 \geq (L)^2 && \forall i \neq j \in \mathcal{M}, \forall t \in \mathcal{T} \\
& \textbf{C4:} && 0 \leq x_i^t \leq W && \forall i \in \mathcal{M}, \forall t \in \mathcal{T} \\
& \textbf{C5:} && 0 \leq y_i^t \leq W && \forall i \in \mathcal{M}, \forall t \in \mathcal{T} \\
& \textbf{C6:} && \exists m \leq M, (\boldsymbol{l}^m)_{i,j} \neq 0 && \forall i \neq j \in \mathcal{M}
\end{aligned} \tag{12}$$

where $\mathbb{E}[.]$ is calculated by considering the random appearance of sparse areas on the roads. **C1** makes sure that the energy consumed of UAV$_i$ at each episode is within the limit of its available energy. **C2** imposes to each UAV to serve at most one CoI at each time-slot t. **C3** ensures that there is a safety distance L at each time-slot tt with obstacles or UAVs. **C4** and **C5** guarantee that each UAV$_i$ will not cross the target area boundaries during the whole flight period. **C6** denotes that there is at least one routing path between any pair of deployed UAVs. To summarize, our goal is to find a control policy to move a team of UAVs for providing coverage to disconnected zones that appear dynamically on the roads while simultaneously: (i) maximize the total/average CoI coverage score, (ii) minimize the energy consumption while making sure that the consumed energy by each UAV is within the limit of its available energy, and (iii) ensuring that the UAV network is connected in every time-slot t, while keeping UAVs neither crossing the area border nor colliding between each other or with obstacles. It is quite challenging to achieve all these objectives at once due to two reasons. On the one hand, to maximize the average CoI coverage, UAVs should continually move around to find out the maximum number of zones to be connected for a maximum of time-slots. On the other hand, to ensure that UAVs stay connected during the whole task and reduce energy consumption, it would be better to minimize the mobility of UAVs to save more energy and ensure durable connectivity between them. It is true that in theory, some optimization methods could be involved to relieve this problem. However, it is impractical to explore the dynamic movement of the vehicles, which denotes that it is challenging to adapt to all possible changes in the environment. Moreover, it is observed that (12) is a mixed-integer non-linear program (MINLP) due to the existence of the binary variable $Hov_{i,z}^{t}$ in **C2** and also includes a non-convex constraint **C1**, which is generally computationally complex to solve it efficiently, and especially for large-scale networks. Consequently, DRL methods are considered as efficient solutions to solve this multi-objective problem and learn the dynamics of the environment.

IV. DRL Background

Recently, one of the branches of Machine learning (ML) is Reinforcement Learning (RL) which deals with the multi-state decision process of an agent (a UAV in our case) while interacting (*i.e.,* taking a set of possible actions $\mathcal{A}$) with a system environment $\mathcal{S}$ at each of a sequence of discrete time-slots. At each time-slot t, the agent observes an environment state ($s_t \in \mathcal{S}$), executes an action ($a_t \in \mathcal{A}$) that updates the environment state ($s_{t+1} \in \mathcal{S}$). The agent receives a numerical reward r_t based on s_t, a_t, and s_{t+1}. The goal of RL is to learn a policy $\pi(s_t, a_t)$ mapping from states to actions, which maximizes the discounted future cumulative reward as,

$$r_t = \sum_{t'=t}^{T} \gamma^{t-t'} r(s_{t'}, a_{t'}) \tag{13}$$

where T is the number of time-slots, the discount factor $\gamma \in [0, 1]$, and $r(.)$ is the reward function at a time-slot t'. Classical RL algorithms are based on Q-learning that allows agents (UAVs) to optimally interact with the environment represented by a Markov Decision Process (MDP). Moreover, the Q-learning method uses a table to store the state-action value function $Q(s_t, a_t) = \mathbb{E}[r_t|s_t, a_t]$, which are iteratively enhanced by estimating the reward r_t when action a_t is

taken. However, Q-learning is dedicated to only a finite discrete action state space. In the case when the action state space is large, a deep neural network (DNN) can be used to approximate the $Q(.)$ function. This concept is called Deep Q Network (DQN), which is considered as a variant of Q-learning and it can be trained by minimizing the following loss function:

$$L(\eta^Q) = \mathbb{E}[(r_t(s_t, a_t) + \gamma \max_{a_{t+1} \in \mathcal{A}} Q(s_{t+1}, a_{t+1}|\eta^Q_{target}) - Q(s_t, a_t|\eta^Q))^2] \quad (14)$$

where η^Q are the weights or the function parameters of a DNN and $tgt_{DQN} = r_t(s_t, a_t) + \gamma \max_{a_{t+1} \in \mathcal{A}} Q(s_{t+1}, a_{t+1}|\eta^Q_{target})$ represents a target value. However, sometimes, a DNN can cause divergence, and that is certainly not wanted. To address this issue, DRL uses generally two major techniques: (i) experience replay and (ii) target network. In the experience replay mechanism, to effectively update the DNN, DRL uses a random mini-batch of samples from a big experience replay buffer to store $\mathcal{D}$ state transition experience tuples (s_t, a_t, r_t, s_{t+1}) which are generated during learning by π at a time-slot t. Experience replay helps break the correlations between sequentially generated samples, which can reduce the variance of updates, avoid divergence, and smooth out the learning. As for the target network mechanism, DRL estimates target values tgt_{DQN} for DNN training based on an additional target network.

tgt_{DQN} has the same structure as $Q(.)$, but its weight η^Q_{target} is slowly updated with the original $Q(.)$ weight η^Q. This may cause the problem of over-fitting in the DQN in which under certain conditions, the different action values get overestimated. Moreover, in the experience replay buffer, uniform samples have been selected rather than importance-weighted samples, which may cause divergence with large state spaces. To address all these issues, the double DQN (DDQN) with Prioritized Replay of experiences based on the sum-tree algorithm [36] has been proposed in [37]. Indeed, tgt_{DQN} in DQN is replaced by tgt_{DDQN} as follows:

$$tgt_{DDQN} = r_t(s_t, a_t) + \gamma Q(s_{t+1}, argmax_{a_{t+1} \in \mathcal{A}} Q(s_{t+1}, a_{t+1}|\eta^Q)|\eta^Q_{target}) \quad (15)$$

In our work, we leverage the advantages of DDQN to control the movements of UAVs under the same framework, called DISCOUNT. A DRL agent periodically inspects the status of the UAV network and the vehicular environment, chooses the appropriate actions based on DISCOUNT, and deploys them by transmitting commands to move UAVs at each time-slot t. Therefore, the aim of DISCOUNT is to maximize the long-term system rewards by attempting various actions, learning from the observation of the environment, and then reinforcing the actions until the best outcome is obtained.

A. State Space and Action Space

The state s_t, at each time-slot t consists of five parts:

- $Con_i(t), \forall i \in \mathcal{M}$: the total number of covered zones by each UAV$_i$.
- $E^t_i, \forall i \in \mathcal{M}$: the energy consumption of each UAV.
- $p^t_i = (x^t_i, y^t_i), \forall i \in \mathcal{M}$: the current position of each UAV, which corresponds to CoI of a given zone.
- $p^t_n = (x^t_n, y^t_n), \forall n \in \mathcal{V}$: the current position of each vehicle.
- $\Theta^t_i \in \{0, 1\}, \forall i \in \mathcal{M}$: takes the value of 0 if UAV$_i$ is deployed and does not exhaust its available energy $(E - \sum_{t=1}^{t} E^t_i) > 0$, and 1 otherwise (and therefore consider UAV$_i$ as inactive and update $\mathcal{M}$).

To reduce the training time for the network and the computational load of each UAV trajectory planning, the movements of UAVs are restricted at each state s_t. Indeed, each UAV$_i$ at a time-slot t, an action $a_t = \{a^t_i | i \in \mathcal{M}\}$ is taken, which consists of two parts:

- $\Omega^t_i \in \{0, \pi/4, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4, 2\pi\}$: the horizontal flying direction of UAV$_i$.
- $d^t_i \in \{0, \ldots, d_{max}\}$: the horizontal flying distance of each UAV$_i$. If $d^t_i = 0$, it means that UAV$_i$ is hovering at the same position, otherwise UAV$_i$ moves to a certain distance $d^t_i \in \{1, \ldots, d_{max}\}$, where $d_{max} = 3$.

In this way, as shown in Fig. 5, 25 basic actions are designed numbered from 0 to 24.

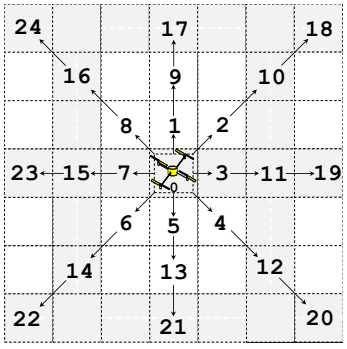


Fig. 5: Discret actions of each UAV.

For a given task, the action space $\mathcal{A} \triangleq \{a_t | t = 1, 2, \ldots, T\}$. The format of the action $a_t = [\Omega^t_1, \ldots, \Omega^t_M, d^t_1, \ldots, d^t_M]$ with a cardinality of $(2M)$. Moreover, Formally, the format of state s_t would be $s_t = [Con_1(t), \ldots, Con_M(t), E^t_1, \ldots, E^t_M, p^t_1, \ldots, p^t_M, p^t_1, \ldots, p^t_N, \Theta^t_1, \ldots, \Theta^t_M]$ with a cardinality of $(4M + N)$. The states and the actions are defined in this way to allow the DRL agent to both make decisions based on the current number of connected zones and energy consumption. For this purpose, the DRL agent uses the control policy to figure out the movement of each UAV at each time-slot $t \in \mathcal{T}$. To accelerate the learning process, all the input elements (*i.e.,* s_t, a_t, and r_t) are normalized to [0,1] by dividing them by their range.

B. Reward Function

The idea behind the reward function is to maximize the number of connected zones while maintaining the initial number of UAVs hovering without having exhausted their batteries during the whole episode. Therefore, the reward r_t is defined as:

$$r_t = \mathbb{E}\left[\frac{\sum_{i=1}^{M} Cont_i(t)}{\left(\frac{\sum_{i=1}^{M} E_i^t}{M}\right)}\right] - \sum_{i=1}^{M} \Theta_i^t \tag{16}$$

The reward has to satisfy two main objectives: (i) maximizing the coverage and (ii) minimizing the energy consumption of each UAV at a time-slot t, which have to be properly treated by the reward. $\mathbb{E}\left[\frac{\sum_{i=1}^{M} Cont_i(t)}{\left(\frac{\sum_{i=1}^{M} E_i^t}{M}\right)}\right]$ is considered as an energy-efficient incremental coverage, which aims to incite UAVs to move in a way the number of connected zones is increased at each step while consuming less energy. Moreover, to ensure that the initial number of UAVs remains the same at the end of the episode without exhausting their available energy, the number of inactive UAVs $\sum_{i=1}^{M} \Theta_i^t$ is subtracted from the reward.

C. Expected Penalties

Several conditions should be satisfied when UAVs take actions, such as flying in the area of interest, avoiding collisions with UAVs and obstacles, and maintaining the connectivity among UAVs. Therefore, three penalties are applied to UAVs:

1) **Penalty for crossing the area boundaries**: a penalty is incurred whenever an action a_t by UAV$_i$ would result in crossing the target area boundaries, which is given by:

$$PEN_i^1 = \begin{cases} 0, & \text{if } 0 \le x_i^t \le W \text{ and } 0 \le y_i^t \le W \\ \Upsilon_1^0, & \text{Otherwise} \end{cases} \tag{17}$$

2) **Penalty for collisions**: to ensure that UAV$_i$ avoid collisions with a given obstacle obs or another UAV$_j$, a safety distance $L \le d_{i,obs}$ or $L \le d_{ij}$ is defined, where d_{ij} and $d_{i,obs}$ are the distances between UAV$_i$ and UAV$_j$, and between UAV$_i$ and obstacle obs, respectively. For simplicity, the safety distance with UAVs and obstacles is set at the same length, where $L < R_i, \forall i \in \mathcal{M}$. Therefore, a penalty is incurred every time an action a_t would result in the UAV going over the safety distance, which is given by:

$$PEN_i^2 = \begin{cases} \Upsilon_2^0, & \text{if } \quad d_{i,obs} < L \\ \Upsilon_2^1, & \text{if } \quad d_{ij} < L \\ 0, & \text{Otherwise} \end{cases} \tag{18}$$

3) **Penalty for losing connectivity**: to satisfy the constraint **C6** of the optimization problem (12), the connectivity of the UAV network should be maintained at each time-slot t. To do so, we adopt the breadth-first search (BFS) [38] to check whether there is a path from a given UAV$_i$ to any other UAV in the network. Indeed, after updating the UAV network connectivity using Algorithm 2, a Boolean value $bool = 1$ is used if the constraint **C6** is satisfied. This process is described in Algorithm 3.

Algorithm 3: Connectivity checking using BFS

```
input : M, l
output: bool
begin
    Execute Algorithm 2
    bool ← 1
    for ∃i ∈ M and ∀j ∈ M and Θ_j^t ≠ 1 do
        path(i, j) ←BFS(i, j) /* Find a path from i to
          j based on l, where i ≠ j. */
        if path(i, j) then
            Continue
        else
            bool ← 0 and break
```

Therefore, a penalty is incurred every time an action a_t would cause a problem of disconnection within the UAV network, which is given by:

$$PEN_i^3 = \begin{cases} \Upsilon_3^0, & \text{if } \quad bool = 0 \\ 0, & \text{Otherwise} \end{cases} \tag{19}$$

DISCOUNT derives the optimal solution by obtaining the optimal policy π^* that defines the set of optimal actions a_t to execute at each state s_t. This process is repeated until the expected sum of discount rewards will be maximized over a finite flight period T. As an example, based on the policy π, the total expected reward of the system from an initial state s_1 is expressed as follows:

$$R_\pi = \sum_{t=1}^{T} \mathbb{E}_\pi[r_t | s_1] \tag{20}$$

Then, the optimal policy π^* is easily derived by maximizing the total expected reward, *i.e.,* $\pi^* = argmax_\pi R_\pi$. To be clear, the objective of DISCOUNT is to find the optimal policy π^* to maximize the long-term system reward, *i.e.,* obtaining $argmax_{\pi^*}\mathbb{E}\left[\frac{\sum_{i=1}^{M}\sum_{z=1}^{Z}\sum_{t=1}^{T} Con_{i,z}^t}{\left(\frac{\sum_{i=1}^{M}\sum_{t=1}^{T} E_i^t}{M}\right)}\right]$.

D. Markov Decision Model

Since we are dealing with a non-stationary environment, we consider the non-stationary MDP consisting of the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{C}_t, \mathcal{R}_t, \gamma)$ in the same way like in [27], where:

- $\mathcal{S}$ is the state space, at any time-slot t, the environment state $s_t \in \mathcal{S}$.
- $\mathcal{A}_t$ is the action space at time-slot t, which contains the possible actions of each UAV$_i$ at time-slot t, *i.e.,* $a_t \subset \mathcal{A}_t$.
- $\mathcal{C}_t \subset \mathcal{S} \times \mathcal{A}_t$ is a measurable subset of $\mathcal{S} \times \mathcal{A}_t$ and defines the set of possible state-action combinations at

the beginning of the tth time-slot. $\mathcal{C}_t$ contains the graph of a measurable mapping.

- $\mathcal{R}_t : \mathcal{C}_n \rightarrow \mathbb{R}$ is a measurable function where $\mathcal{R}_t(s_t, a_t)$ provides the instant reward at time-slot t if the current state s_t and a set of actions a_t is selected.

For a deep understanding, the current environment state, such as the positions of UAVs, their energy levels, their covered zones, their active status, and the positions of vehicles is quite a sufficient statistic of the history to make a decision. That is to say, the future is independent of the past given some current aggregate statistics about the present, which satisfy the Markov property. The objective of our DRL agent is to interact with the dynamic vehicular environment and select the appropriate actions for all UAVs, which maximize cumulative discounted future rewards in the given time T.

E. Algorithm of DISCOUNT

The DISCOUNT structure is depicted in Fig. 6. During their interaction with the vehicular environment ①, UAVs (controlled by a single agent on the satellite) select a set of optimal actions a_t generated by the policy $\pi^*(.)$ ①. Then, the tuple (s_t, a_t, r_t, s_{t+1}) is stored in the experience replay buffer ②. During the training process, we sample a prioritized mini-batch ③. In the step ④, based on (s_t, a_t, s_{t+1}) as an input in the predicted network $Q(.)$, we obtain the Q value as an output. As for the target network ⑤, DISCOUNT selects the action for the next state s_{t+1}, which gives $argmax_{a' \in \mathcal{A}} Q(s_t, a'|\eta^Q)$ in the predicted network $Q(.)$ and identifies the corresponding Q value of state-action pair in the target network $tgt_{DDQN}(.)$. Meanwhile, the predicted network is updated by the loss function provided in (14). Based on the updated Q-value, the agent selects a set of actions according to the current state s_t with the adopted ε-greedy policy ⑥.

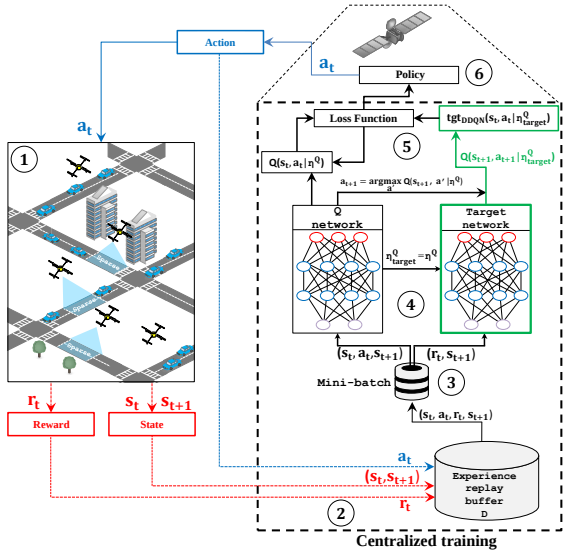


Fig. 6: DISCOUNT structure.

The pseudo-code to train DISCOUNT and obtain the appropriate trajectories of UAVs is formally presented in Algorithm 4.

Algorithm 4: Centralized training of DISCOUNT

1 **begin**
2 Initialize replay buffer $\mathcal{D}$ to capacity D, where ($\mathcal{D} = \emptyset$);
3 Randomly initialize η^Q, where $\eta^Q_{target} = \eta^Q$;
4 $Episode \leftarrow 0$; Number of episodes$\leftarrow G$;
5 Initialize ε and ϱ;
6 Number of epochs$\leftarrow T$;
8 **while** $Episode < G$ **do**
9 Collect the environment characteristics;
10 Initialize state $s_0 = [Con_i(0), E_i^0, p_i^0, p_n^0, \Theta_i^0]$, $\forall i \in \mathcal{M}, \forall n \in \mathcal{V}$;
12 **for** $t \leftarrow 1, \ldots, T$ **do**
13 $a_t = \begin{cases} Random\ actions, & \varepsilon\ probability \\ argmax_{a \in \mathcal{A}} Q(s_t, a|\eta^Q), & 1-\varepsilon\ probability \end{cases}$
14 **Execute**: actions $a_t^i = [\Omega_i^t, d_i^t]$, $\forall i \in \mathcal{M}$;
15 **Evaluate**: get reward r_t based on (16) and obtain a new state s_{t+1};
16 **Observe**: $s_{t+1} = [Con_i(t+1), E_i^{t+1}, p_i^{t+1}, p_n^{t+1}, \Theta_i^{t+1}]$, $\forall i \in \mathcal{M}, \forall n \in \mathcal{V}$;
18 **foreach** $UAV_i \in \mathcal{M}$ **do**
19 $PEN_i = \sum_{n=1}^{3} PEN_i^n$;
21 **if** $PEN_i > 0$ **then**
22 Cancel action a_t^i of UAV$_i$ and update s_{t+1};
23 $r_t \leftarrow r_t - \sum_{i=1}^{M} PEN_i$;
24 Store transition sample (s_t, a_t, r_t, s_{t+1}) into the replay memory $\mathcal{D}$ with maximal priority $priority_t = max_{i<t} priority_i$;
26 **if** $\mathcal{D}$ *is full* **then**
28 **for** $f \leftarrow 1, \ldots, F$ **do**
29 Sample transition $(s_f, a_f, r_f, s_{f+1}) \sim$ based on (23);
30 Calculate importance-sampling weight $\sim$ based on (22);
31 Calculate TD-error $\sim$ $(tgt_{DDQN} - Q(s_f, a_f|\eta^Q))$;
32 Update transition priority $\sim priority_f \leftarrow |(tgt_{DDQN} - Q(s_f, a_f|\eta^Q))|$;
33 Update weights η^Q of $Q(.)$ by minimizing the loss function $\sim$ based on (21);
34 **Every M steps:** $\eta^Q_{target} = \eta^Q$
36 **if** $\varepsilon > 0.01$ **then**
37 $\varepsilon \leftarrow \varepsilon - \varrho$;
38 $Episode \leftarrow Episode + 1$;
39 **Output:** The optimal policy π^*.

Since there is no feasible manner to get full knowledge about the dynamics of vehicles in advance before the deployment of UAVs, the algorithm is considered as an offline training phase to estimate the optimal policy π^*. Then, π^* is extracted to control the movement of UAVs to cover sparse areas with the maximal long-term system reward during the online testing phase. Indeed, at the initial phase of the algorithm, we initialize the reply buffer $\mathcal{D}$ of size D (Line 2). We randomly initialize the weights η^Q of $Q(.)$ and the target network with weights η^Q_{target} (Line 3). In Lines 4-6, we initialize the number of episodes, the number of epochs (time-slots), the exploration rate ε, and the decrement factor ϱ. In the second part of the algorithm (Lines 8-22), the training process of DISCOUNT is made with G episodes, and in each episode, there are T time-slots. In Lines 9-10, the environment is initialized, and UAVs get the state s_0. As for Lines 12-16, at

each time epoch (time-slot), DISCOUNT selects a trajectory action for each UAV based on the $\varepsilon - greedy$ mechanism according to output value $Q(.)$. When the action is executed, the system gets a reward r_t and go to the next state s_{t+1}. Then, an observation is carried out to see whether the UAVs have exceeded their restrictions (17), (18), and (19). If it is the case, a set of penalties is calculated in Lines 18-19. These penalties are deducted from the calculated reward, the involved movements are canceled, and the state s_{t+1} is updated accordingly (Lines 19-23). Then, the transition (s_t, a_t, r_t, s_{t+1}) is stored in memory $\mathcal{D}$ according to the adopted priority method. When memory $\mathcal{D}$ is full, the training process starts by sampling a prioritized mini-batch of F transitions to update the weights η^Q (Lines 26). This process aims to minimize the following weighted loss function:

$$L(\eta^Q) = \frac{1}{F}\sum_{f=1}^{F} \Psi_f (tgt_{DDQN} - Q(s_f, a_f|\eta^Q))^2 \qquad (21)$$

where $(tgt_{DDQN} - Q(s_f, a_f|\eta^Q))$ is the temporal different (TD) error and Ψ_f indicates the prioritized sampling weight, which is used to correct the bias. According to [36], Ψ_f is given by:

$$\Psi_f = \frac{(|\mathcal{D}| \times PR(f))^{-\varsigma}}{max_i \Psi_i} \qquad (22)$$

where $|\mathcal{D}|$ is the size of the memory and the constant $\varsigma > 0$. During the training process, transition f is sampled according to the following probability:

$$PR(f) = \frac{priority_f^\nu}{\sum_{f=1}^{F} priority_f^\nu} \qquad (23)$$

where $priority_f$ indicates the priority of transition (s_f, a_f, r_f, s_{f+1}) and $\nu > 0$ is the degree of priority. It should be stressed that before memory $\mathcal{D}$ is full, $priority_f$ is initialized at the same value for all samples so that they can have an equal chance of being sampled. When $\mathcal{D}$ is full, $priority_f$ is set to $|(tgt_{DDQN} - Q(s_f, a_f|\eta^Q))|$ to assign higher priority to the transition with higher absolute TD error (Lines 28-32). In line 33, the weights η^Q are updated by minimizing the loss function (21). In Line 37, a decrement ϱ is subtracted from ε at each epoch. It should be stressed that DISCOUNT ensures a low variance and avoids highly correlated data for successive updates. This can be achieved thanks to the adopted prioritized mini-batch in which multiple important experiences are sampled from $\mathcal{D}$. This kind of mini-batch can significantly optimize the weights of the Q network such that the loss function in (14) is significantly minimized during the training phase, and thus calculating optimized actions and cope with the non-stationarities induced by the dynamic speeds of vehicles. It is worthy to note that to ensure a rapid adaptation to the non-stationary environment, the agent should keep learning by adopting a higher exploration rate and quickly adapt its policy to maximize the long-term cumulative rewards. In addition, by decoupling the predicted Q value and the target Q value, DISCOUNT can also overcome the overestimation problem, and thus reduce the bias.

V. Performance Evaluation

A set of experiments is conducted to evaluate the performances of DISCOUNT. We first present the different simulation settings and then interpret the obtained results.

A. Simulation settings

In the conducted simulations, a target area of 1.4 km $\times$ 2.8 km is considered in Fig. 7. This area is divided into 98 fixed-size zones with two high-rise buildings (in red) that constitute two obstacles for moving UAVs. To obtain the most realistic results possible, the traffic of vehicles entering the target area follows a normal distribution.

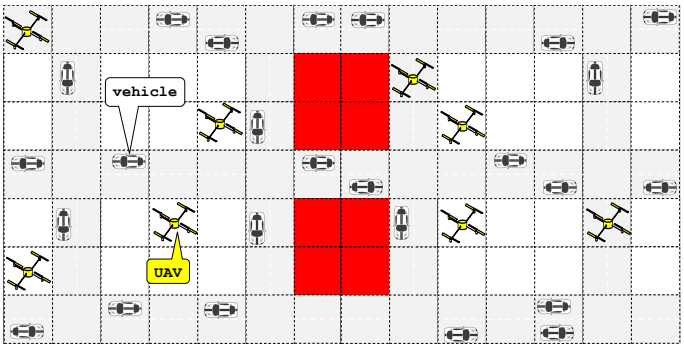
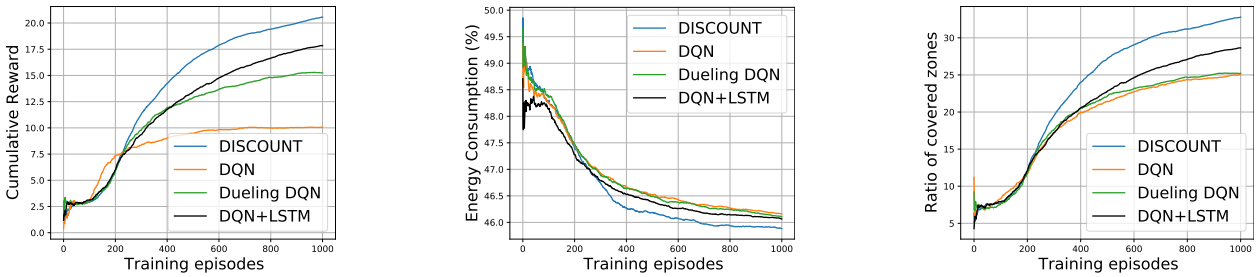


Fig. 7: Map of our simulation.

The simulation runs are carried out with Tensorflow 1.14 and Python 3.6.9, through which DISCOUNT is trained for 1000 episodes, each having 100 epochs. Then, a test is performed for 100 epochs and we calculate some crucial metrics. The simulated DDQN consists of an input layer of the neural network in the form of a state representative and two fully-connected hidden layers composed of 512 neurons for each one. The hidden layers are using the rectified linear unit (ReLU) function for activation. The output layer is a fully-connected linear layer, in which the number of neurons is equal to the size of the action space. The values of all input parameters used in the simulation are provided in Table III.

TABLE III: Simulation Parameters.

Parameter	Value
Replay memory	Prioritized
Memory size $\mathcal{D}$	10000
Mini-batch size F	64
Discount factor γ	0.94
Initial ε exploration	1.0
Optimizer method	RMSProp
RMSProp learning rate	0.00005
Υ_1^0, Υ_2^0, Υ_2^1, Υ_3^0	1.0
Zone size S	200 m
Prioritization sampling weight Ψ_f	0.3 $\rightarrow$ 1.0
Surface of the area	3.92 km^2
Number of UAVs	[5 ... 50]
Altitude of UAVs	100 m
Density of vehicles per km^2	[10 ... 100]
Activation function	ReLU



(a) Cumulative reward. (b) Average of energy consumption. (c) Ratio of covered zones.

Fig. 8: Performance comparisons over episodes (Number of UAVs=5 and Vehicles/km^2=50).

B. Numerical results and discussions

In this part, simulation results are presented to validate the performance of DISCOUNT based on three phases: (i) Training phase, (ii) Testing phase, and (iii) Routing phase. In the first phase, DISCOUNT is trained for 1000 episodes, and η^Q is updated accordingly. As for the second phase, the last updated η^Q value is used to study the behaviors and test the performance of DISCOUNT, where another set of mobility traces will be generated. While in the last phase, DISCOUNT is adopted within the routing protocol proposed in [13] to show its benefits in terms of several metrics.

1) Training phase: Firstly, we calculate and analyze the obtained reward per each episode (*see* Fig. 9). It is clearly observed that the obtained rewards converge to stable values after 400 training episodes for DISCOUNT. This is because UAVs in DISCOUNT can learn the vehicular network's dynamics efficiently by interacting with its entities while making the right decisions to increase the coverage of empty zones and reduce energy consumption.

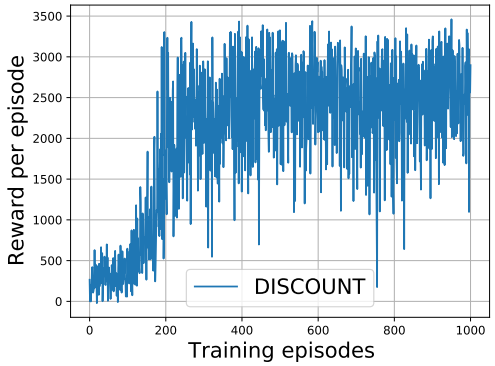
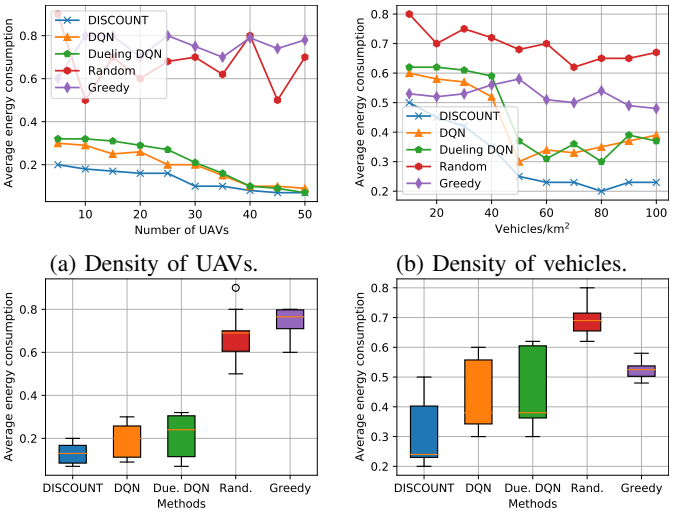


Fig. 9: Reward per episode in DISCOUNT (Number of UAVs=5 and Vehicles/km^2=50).

To study the effectiveness of DISCOUNT, a comparative study is performed, where the performances of DISCOUNT are compared with those obtained by the original DQN [39] and Dueling DQN [40] under the same parameters. It should be stressed that Dueling DQN is an enhanced version of DQN, which separately estimates state-value and the advantages of each action. Moreover, we have also considered another DRL variant consisting of a DQN based on an LSTM network, namely DQN+LSTM, which follows the same principle as in [41]. The objective of DQN+LSTM is to define the long-term correlated patterns over the input and output states. It is worthy to note that the sets of actions and environment states are recorded in the LSTM network and its output is used as an input in the DQN. At a first step, we start by analyzing the convergence performance in terms of average reward, the average energy consumption, and the ratio of covered zones for the three considered algorithms (*c.f.,* Fig. 8). Generally, as illustrated in Fig. 8(a), it is shown that the average reward per episode tends to increase over episodes until convergence. Indeed, at the beginning of the experiment, the average reward is approximately the same for the three algorithms, which is explained by the insufficient information provided by the interaction with the environment. However, with the increased number of episodes, UAVs accumulate more and more rewards by enlightening the environment. It can be distinguished that DISCOUNT significantly outperforms by alleviating the overestimation problem and enhancing the training stability, which is not the case for DQN and Dueling DQN. Additionally, compared with DISCOUNT, the cumulative rewards of DQN and Dueling DQN converge more slowly, which is caused by two different issues: (i) the over-fitting issue of DQN and (ii) the extra layers used by Dueling DQN which slow down the training speed. It can also be shown that DISCOUNT overcomes DQN+LSTM, which is due to the time taken by DQN+LSTM to preserve more important experiences and try to define the long-term correlated patterns between them to estimate the optimal policy π^*.

To top it off, the obtained results related to the energy consumption and the ratio of covered zones show that DISCOUNT achieves better performance compared with the other algorithms. Indeed, as shown in Fig. 8(b), it is clearly distinguished that DISCOUNT obtains the lowest energy consumption after convergence, which is 15% lower than that of the DQN+LSTM algorithm. This is because DISCOUNT learns faster and places UAVs at the appropriate locations so that they move less to reach frequent sparse areas. As for Fig. 8(c), DISCOUNT covers more CoIs after convergence, which is explained by the fact that UAVs in DISCOUNT can predict better the locations of CoIs that allow connecting the maximum number of zones. As a result, DISCOUNT covers more than 10% of zones compared with DQN, Dueling DQN, and DQN+LSTM.

2) Testing phase: To evaluate the performances of DISCOUNT, in addition to DQN and Dueling DQN, we consider two baseline methods, (i) Greedy and (ii) Random. The first method selects at each time-slot the action that can maximize the reward while considering the different constraints mentioned above. While in the second method, random action is selected for each UAV at each time-slot. The optimized actions of DISCOUNT, DQN, and Dueling DQN, are calculated by using the latest updated η^Q during the training phase. It is worthy to note that each point of the following obtained results represents the mean of 30 simulation runs with 95% confidence interval.



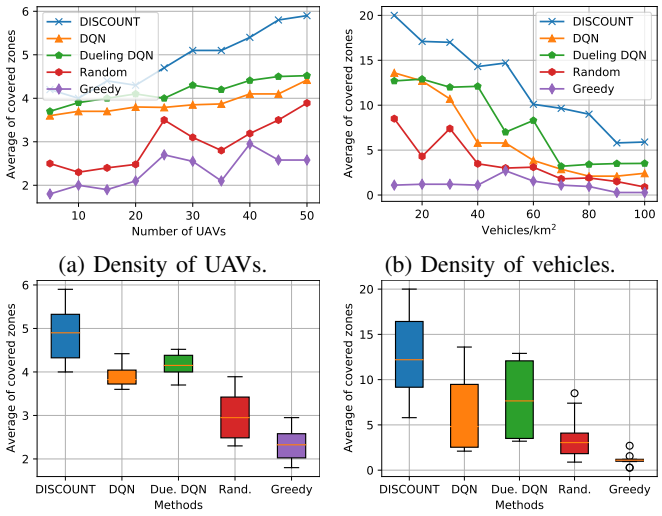
(a) Density of UAVs. (b) Density of vehicles.

(c) Average EC vs. UAVs. (d) Average EC vs. Vehicles.

Fig. 10: Average energy consumption results.

Initially, we study the impact of the density of UAVs and vehicles on average energy consumption (EC) (*see* Fig. 10). As depicted in Figs. 10(a) and 10(c), overall, DISCOUNT outperforms the other algorithms in terms of energy consumption for each density of UAVs, representing an improvement of 15% over RL algorithms and 60% over baseline methods. This is explained by the fact that UAVs in DISCOUNT can place themselves near the zones that are likely to be CoIs in the future, and thus UAVs move less than other algorithms and methods. In Figs. 10(b) and 10(d), we distinguish that the energy consumption of UAVs is high at low densities of vehicles due to the constant movement of UAVs that are looking for the increasing number of CoIs. However, at high densities of vehicles, the number of CoIs is getting weaker and weaker, which reduces the movements of UAVs and thus their energy consumption. As for the random and greedy methods, UAVs are permanently moving either randomly or looking for a more accumulating reward, which considerably consumes more energy.

Secondly, as shown in Fig. 11, it can be observed that DISCOUNT consistently outperforms the RL algorithms and the two baseline methods in terms of covered zones' (CZ) average. Overall, in Figs. 11(a) and 11(c), when the density of UAVs is low (*i.e.,* $5 \leq Density \leq 20$), the number of covered zones is low, and it increases as UAVs become denser and



(a) Density of UAVs. (b) Density of vehicles.

(c) Average CZ vs. UAVs. (d) Average CZ vs. Vehicles.

Fig. 11: Average of covered zones results.

denser, which is obviously explained by the number of CoIs that could be covered. Generally, DISCOUNT achieves better performance compared with other schemes. This is explained by the fact that taking appropriate decisions certainly leads to better coverage, which is not the case of baseline methods or RL algorithms that require more learning steps. From Figs. 11(b) and 11(d), it can be noticed that the number of covered zones decreases as the density of vehicles increases. This is because the number of CoIs that should be covered decreases. In the high UAV densities, DISCOUNT tries to find the maximum number of CoIs to cover based on its robust learning algorithm.

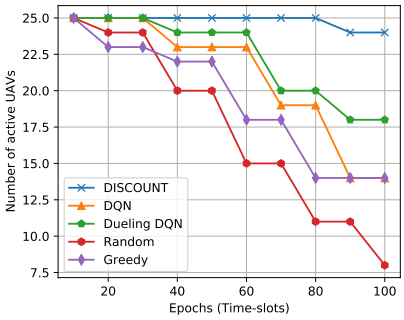
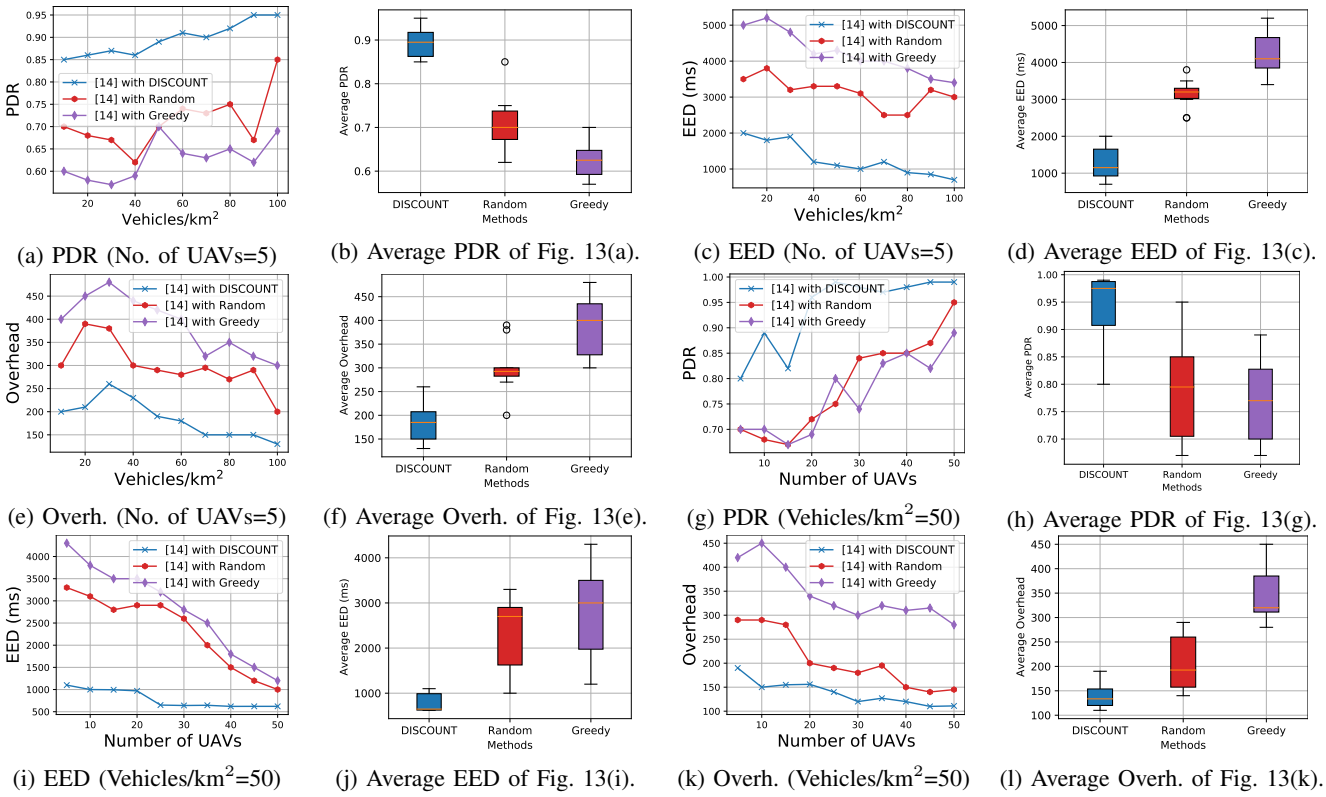


Fig. 12: Number of active UAVs during an episode.

In Fig. 12, we analyze the number of active UAVs during a whole episode. On average, DISCOUNT outperforms all other techniques by making the number of active drones approximately stable during the testing phase. This is because DISCOUNT is based on an energy-efficient strategy that allows UAVs to only move when necessary to cover CoIs. As for the baseline methods, specific UAVs are moving permanently according to their own rules, causing high energy consumption. Thus, some UAVs exhaust their batteries before the end of the episode.



(a) PDR (No. of UAVs=5) (b) Average PDR of Fig. 13(a). (c) EED (No. of UAVs=5) (d) Average EED of Fig. 13(c).

(e) Overh. (No. of UAVs=5) (f) Average Overh. of Fig. 13(e). (g) PDR (Vehicles/km^2=50) (h) Average PDR of Fig. 13(g).

(i) EED (Vehicles/km^2=50) (j) Average EED of Fig. 13(i). (k) Overh. (Vehicles/km^2=50) (l) Average Overh. of Fig. 13(k).

Fig. 13: Routing performances of Ref. [13] using DISCOUNT and baseline methods.

3) Routing phase: To demonstrate the different benefits of DISCOUNT when applied in conjunction with the routing protocol that is proposed in [13], we calculate three metrics under different scenarios (*see* Fig. 13). In the first scenario, we set the number of UAVs to 5 and vary the density of vehicles. While in the second scenario, we set the density of vehicles per km^2 to 50 and vary the number of UAVs. Three different metrics are calculated for each scenario, which are Packet Delivery Ratio (PDR), End-to-End Delay (EED), and Overhead. It is worthy to note that we consider the same routing parameters in [13].

In Fig. 13(a), we clearly notice that the DISCOUNT strategy generates a high PDR compared with the baseline strategies. This can be justified by the optimized UAV actions at each density of vehicles, where UAVs are trying to fill the communication gaps wherever it is possible. Fig. 13(c) shows that the DISCOUNT strategy achieves the lowest EED for all vehicle densities. This is explained by the fact that routing paths stay connected for long periods, and thus avoiding the re-initialization of the discovery process and minimizing the delivery delay. In Fig. 13(e), we clearly distinguish that the DISCOUNT strategy generates less overhead compared to baseline strategies. This is mainly due to the fact that routing paths, once established, transit the maximum of data packets without generating control packets. As shown in Fig. 13(g), PDR with DISCOUNT strategy tends to be maximized to its higher levels as the number of UAVs increases. This is due to the increased number of covered empty zones, which can definitely help avoiding packet losses. In Fig. 13(i), we notice that the DISCOUNT strategy generates the lowest delays for all UAV densities. This is caused by the use of the greedy forwarding technique that can minimize the number of hops, and thus the delay of delivery. Finally, Fig. 13(k) shows that the DISCOUNT strategy generates the lowest control overhead for all UAV densities. This is because the increased number of UAVs tends to connect all possible terrestrial gaps, and thus minimizing the number of disconnections and avoiding the re-initialization of the discovery process at each disconnection.

VI. Conclusion

UAVs are considered as a flexible solution to provide wireless connectivity to disconnected vehicles. However, several constraints should be considered, such as collisions, energy restrictions, connectivity between UAVs, and movement optimization. Due to the complexity of these constraints, in this paper, we study an energy-efficient DRL-based framework called DISCOUNT, which intelligently controls the movements of multiple UAVs over a dynamic environment, and efficiently places them as relays between disconnected vehicles whenever possible. It has been proved that DISCOUNT was able to both learn quickly the terrestrial vehicular environment dynamicity and guide UAVs through the most appropriate trajectories for providing effective relay between disconnected vehicles. The obtained results show that DISCOUNT outperforms the considered RL algorithms and the baseline methods in terms of energy consumption and coverage. Further, DISCOUNT reduces inter-UAV collisions, maintains the connectivity among UAVs, and prevents UAVs

from exhausting their available energy. As a deduction, we believe that by deploying UAVs for a significant pretraining period of time (e.g., $\mathcal{T} > year$), UAVs can almost perfectly learn the traffic flow of vehicles for a given region and can accurately select the appropriate placements for coverage. As future work, we aim to extend this framework to control the unlimited action space of UAVs and achieve achieving fair and near-optimal coverage of an urban vehicular network.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by grant code 19-COM-1-01-0015.

REFERENCES

[1] B. Alzahrani, O. S. Oubbati, A. Barnawi, M. Atiquzzaman, and D. Alghazzawi, "UAV Assistance Paradigm: State-of-the-art in Applications and Challenges," *Journal of Network and Computer Applications*, vol. 166, p. 102706, 2020.

[2] X. Zhong, Y. Guo, N. Li, and Y. Chen, "Joint Optimization of Relay Deployment, Channel Allocation, and Relay Assignment for UAVs-Aided D2D Networks," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 804–817, 2020.

[3] S.-Y. Park, C. S. Shin, D. Jeong, and H. Lee, "DroneNetX: Network reconstruction through connectivity probing and relay deployment by multiple UAVs in ad hoc networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11 192–11 207, 2018.

[4] L. Deng, G. Wu, J. Fu, Y. Zhang, and Y. Yang, "Joint Resource Allocation and Trajectory Control for UAV-Enabled Vehicular Communications," *IEEE Access*, vol. 7, pp. 132 806–132 815, 2019.

[5] O. S. Oubbati, M. Atiquzzaman, P. Lorenz, M. H. Tareque, and M. S. Hossain, "Routing in flying Ad Hoc networks: Survey, constraints, and future challenge perspectives," *IEEE Access*, vol. 7, pp. 81 057–81 105, 2019.

[6] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8036–8049, 2019.

[7] M. Maimaitijiang, V. Sagan, P. Sidike, A. M. Daloye, H. Erkbol, and F. B. Fritschi, "Crop Monitoring Using Satellite/UAV Data Fusion and Machine Learning," *Remote Sensing*, vol. 12, no. 9, p. 1357, 2020.

[8] S.-C. Noh, H.-B. Jeon, and C.-B. Chae, "Energy-efficient Deployment of Multiple UAVs Using Ellipse Clustering to Establish Base Stations," *IEEE Wireless Communications Letters*, 2020.

[9] F. Malandrino, C.-F. Chiasserini, C. Casetti, L. Chiaraviglio, and A. Senacheribbe, "Planning UAV activities for efficient user coverage in disaster areas," *Ad Hoc Networks*, vol. 89, pp. 177–185, 2019.

[10] A. A. Khuwaja, G. Zheng, Y. Chen, and W. Feng, "Optimum deployment of multiple UAVs for coverage area maximization in the presence of co-channel interference," *IEEE Access*, vol. 7, pp. 85 203–85 212, 2019.

[11] H. Wang, H. Zhao, W. Wu, J. Xiong, D. Ma, and J. Wei, "Deployment algorithms of flying base stations: 5G and beyond with UAVs," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 009–10 027, 2019.

[12] X. Zhang and L. Duan, "Fast deployment of UAV networks for optimal wireless coverage," *IEEE Transactions on Mobile Computing*, vol. 18, no. 3, pp. 588–601, 2018.

[13] O. S. Oubbati, N. Chaib, A. Lakas, P. Lorenz, and A. Rachedi, "UAV-assisted supporting services connectivity in urban VANETs," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3944–3951, 2019.

[14] W. Fawaz, R. Atallah, C. Assi, and M. Khabbaz, "Unmanned aerial vehicles as store-carry-forward nodes for vehicular networks," *IEEE Access*, vol. 5, pp. 23 710–23 718, 2017.

[15] M. Mozaffari, A. T. Z. Kasgari, W. Saad, M. Bennis, and M. Debbah, "Beyond 5G with UAVs: Foundations of a 3D wireless cellular network," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 357–372, 2018.

[16] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, 2019.

[17] S. F. Abedin, M. S. Munir, N. H. Tran, Z. Han, and C. S. Hong, "Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[18] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1274–1285, 2019.

[19] O. S. Oubbati, M. Atiquzzaman, P. Lorenz, A. Baz, and H. Alhakami, "SEARCH: An SDN-enabled Approach for Vehicle Path-Planning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 523–14 536, 2020.

[20] O. S. Oubbati, A. Lakas, P. Lorenz, M. Atiquzzaman, and A. Jamalipour, "Leveraging communicating UAVs for emergency vehicle guidance in urban areas," *IEEE Transactions on Emerging Topics in Computing*, 2019.

[21] O. S. Oubbati, N. Chaib, A. Lakas, S. Bitam, and P. Lorenz, "U2RV: UAV-assisted reactive routing protocol for VANETs," *International Journal of Communication Systems*, vol. 33, no. 10, p. e4104, 2020.

[22] O. S. Oubbati, N. Chaib, A. Lakas, and S. Bitam, "On-demand routing for urban VANETs using cooperating UAVs," in *Proceedings of the International Conference on Smart Communications in Network Technologies (SaCoNeT)*. IEEE, 2018, pp. 108–113.

[23] H. Sedjelmaci, M. A. Messous, S. M. Senouci, and I. H. Brahmi, "Toward a lightweight and efficient UAV-aided VANET," *Transactions on Emerging Telecommunications Technologies*, vol. 30, no. 8, p. e3520, 2019.

[24] F. Zeng, R. Zhang, X. Cheng, and L. Yang, "UAV-assisted data dissemination scheduling in VANETs," in *Proceedings of the IEEE International Conference on Communications (ICC-2019)*. IEEE, 2018, pp. 1–6.

[25] R. Lu, R. Zhang, X. Cheng, and L. Yang, "Relay in the Sky: A UAV-Aided Cooperative Data Dissemination Scheduling Strategy in VANETs," in *Proceedings of the IEEE International Conference on Communications (ICC-2019)*. IEEE, 2019, pp. 1–6.

[26] H. Qi, Z. Hu, H. Huang, X. Wen, and Z. Lu, "Energy Efficient 3-D UAV Control for Persistent Communication Service and Fairness: A Deep Reinforcement Learning Approach," *IEEE Access*, vol. 8, pp. 53 172–53 184, 2020.

[27] 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.

[28] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2124–2136, 2019.

[29] P. Yang, X. Cao, X. Xi, W. Du, Z. Xiao, and D. Wu, "Three-Dimensional Continuous Movement Control of Drone Cells for Energy-Efficient Communication Coverage," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6535–6546, 2019.

[30] 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.

[31] L. Bai, R. Han, J. Liu, Q. Yu, J. Choi, and W. Zhang, "Air-to-ground wireless links for high-speed UAVs," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 12, pp. 2918–2930, 2020.

[32] Y. Chen, N. Zhao, Z. Ding, and M.-S. Alouini, "Multiple UAVs as relays: Multi-hop single link versus multiple dual-hop links," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6348–6359, 2018.

[33] 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.

[34] S. Jia and L. Zhang, "Modelling unmanned aerial vehicles base station in ground-to-air cooperative networks," *IET Communications*, vol. 11, no. 8, pp. 1187–1194, 2017.

[35] J. Härri, M. Fiore, F. Filali, and C. Bonnet, "Vehicular mobility simulation with VanetMobiSim," *Simulation*, vol. 87, no. 4, pp. 275–300, 2011.

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

[37] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the Thirtieth AAAI conference on artificial intelligence*, 2016.

[38] M. Weik, *Computer science and communications dictionary*. Springer Science & Business Media, 2000.

[39] 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.
[40] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of the International conference on machine learning*, 2016, pp. 1995–2003.
[41] Y. Lin, M. Wang, X. Zhou, G. Ding, and S. Mao, "Dynamic spectrum interaction of UAV flight formation communication with priority: A deep reinforcement learning approach," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 892–903, 2020.

Omar Sami Oubbati is an Associate Professor at University of Gustave Eiffel in the region of Paris, France. He received his degree of Engineer (2010), M.Sc. in Computer Engineering (2011), M.Sc. degree (2014), and a PhD in Computer Science (2018), all from University of Laghouat, Algeria. From Oct. 2016 to Oct. 2017, he was a Visiting PhD Student with the Laboratory of Computer Science, University of Avignon, France. He spent 6 years as an Assistant Professor at the Electronics department, University of Laghouat, Algeria and a Research Assistant in the Computer Science and Mathematics Lab (LIM) at the same university. His main research interests are in Flying and Vehicular ad hoc networks, Energy harvesting and Mobile Edge Computing, Energy efficiency and Internet of Things (IoT). He is the recipient of the 2019 Best Survey Paper for Vehicular Communications (Elsevier). He has actively served as a reviewer for flagship IEEE Transactions journals and conferences, and participated as a Technical Program Committee Member for a variety of international conferences, such as IEEE ICC, IEEE CCNC, IEEE ICCCN, IEEE WCNC, IEEE ICAEE, and IEEE ICAIT. He serves on the editorial board of Vehicular Communications Journal of Elsevier and Communications Networks Journal of Frontiersin. He has also served as guest editor for a number of international journals. He is a member of the IEEE and IEEE Communications Society.

Mohammed Atiquzzaman received the M.S. and Ph.D. degrees in electrical engineering and electronics from the University of Manchester, U.K., in 1984 and 1987, respectively. He currently holds the Edith J. Kinney Gaylord Presidential Professorship with the School of Computer Science, University of Oklahoma, USA. His research has been funded by the National Science Foundation, National Aeronautics and Space Administration, U.S. Air Force, Cisco, and Honeywell. He co-authored Performance of TCP/IP Over ATM Networks and has authored over 300 refereed publications. His current research interests are in areas of transport protocols, wireless and mobile networks, ad hoc networks, satellite networks, power-aware networking, and optical communications. He was the Chair of the IEEE Communication Society Technical Committee on Communications Switching and Routing. He received the IEEE Communication Society's Fred W. Ellersick Prize and the NASA Group Achievement Award for outstanding work to further NASA Glenn Research Center's efforts in the area of the Advanced Communications/Air Traffic Management's Fiber Optic Signal Distribution for Aeronautical Communications project. He also received the 2017 Distinguished Technical Achievement Award from IEEE Communications Society in recognition of outstanding technical contributions and services in the area of communications switching and routing. He is the Editor in Chief of Journal of Networks and Computer Applications, the founding Editor in Chief of Vehicular Communications, and serves served on the editorial boards of many journals, including IEEE Communications Magazine, IEEE Journal on Selected Areas in Communications, and IEEE Transactions on Mobile Computing.

Abdullah Baz received the B.Sc. degree in electrical and computer engineering from UQU, in 2002, the M.Sc. degree in electrical and computer engineering from KAU, in 2007, and the M.Sc. degree in communication and signal processing and the Ph.D. degree in computer system design from Newcastle University, in 2009 and 2014, respectively. He was a Vice-Dean, and then the Dean of the Deanship of Scientific Research with UQU, from 2014 to 2020. He is currently an Associate Professor with the Computer Engineering Department, a Vice-Dean of DFMEA, the General Director of the Decision Support Center, and the Consultant of the University Vice Chancellor with UQU. His research interests include data science, ML, AI, VLSI design, EDA/CAD tools, intelligent transportation, computer system and architecture, smart systems, smart health. Since 2015, he has been served as a Review Committee Member of the IEEE International Symposium on Circuits and Systems (ISCAS) and a member of the Technical Committee of the IEEE VLSI Systems and Applications. In 2017, IEEE has elevated him to the grade of IEEE Senior Member. He served as a Reviewer in a number of journals, including the IEEE Internet of Things, the IET Computer Vision, the Artificial Intelligence Review, IEEE Access, and the IET Circuits, Devices and Systems.

Hosam Alhakami received the B.Sc. degree in computer science from King Abdulaziz University, Saudi Arabia, in 2004, the M.Sc. degree in internet software systems from Birmingham University, Birmingham, U.K., in 2009, and the Ph.D. degree in software engineering from De Montfort University, in 2015. From 2004 to 2007, he worked with Software Development Industry, where he implemented several systems and solutions for a national academic institution. Dr. Alhakami was the Vice-Dean of the Deanship of Admission and Registration for Academic affairs with UQU, from 2015 to 2020. Currently, he is an associate professor of the computer science department with UQU. His research interests include algorithms, semantic web, and optimization techniques. He focuses on enhancing real-world matching systems using machine learning and data analytics in a context of supporting decision-making.

Jalel Ben-Othman received his B.Sc. and M.Sc. degrees both in Computer Science from the University of Pierre et Marie Curie, (Paris 6) France in 1992, and 1994 respectively. He received his PhD degree from the University of Versailles, France, in 1998. He is currently full professor at the University of Paris 13 since 2011 and member of L2S lab at CentraleSupélec. Dr. Ben-Othman's research interests are in the area of wireless ad hoc and sensor networks, VANETs, IoT, performance evaluation and security in wireless networks in general. He was the recipient of the IEEE COMSOC Communication Software technical committee Recognition Award in 2016, the IEEE computer society Meritorious Service Award in 2016, and he is a Golden Core Member of IEEE Computer Society, AHSN Exceptional Service and Contribution Award in 2018 and the VEHCOM Fabio Neri award in 2018. He has served as steering committee member of IEEE Transaction on Mobile computing (IEEE TMC), he is currently a senior Editor of IEEE communication letters (IEEE COMML) an editorial board member of several journals (IEEE Networks, IEEE IoT journal, JCN, IJCS, SPY, Sensors, etc.). He has also served as TPC Co-Chair for IEEE Globecom and ICC conferences and other conferences as (WCNC, IWCMC, VTC, ComComAp, ICNC, WCSP, Q2SWinet, P2MNET, WLN, etc.). He was the chair of the IEEE Ad Hoc and sensor networks technical committee January 2016-2018, he was previously the vice chair and secretary for this committee. He has been appointed as IEEE COMSOC distinguished lecturer from 2015 to 2018 and he is currently IEEE VTS distinguished lecturer where he did several tours all around the world. He is member of IEEE technical services board since 2016.