

Northumbria Research Link

Citation: Wang, Liang, Wang, Kezhi, Pan, Cunhua, Xu, Wei, Aslam, Nauman and Hanzo, Lajos (2021) Multi-Agent Deep Reinforcement Learning Based Trajectory Planning for Multi-UAV Assisted Mobile Edge Computing. IEEE Transactions on Cognitive Communications and Networking, 7 (1). pp. 73-84. ISSN 2372-2045

Published by: IEEE

URL: <https://doi.org/10.1109/TCCN.2020.3027695>
<<https://doi.org/10.1109/TCCN.2020.3027695>>

This version was downloaded from Northumbria Research Link:
<http://nrl.northumbria.ac.uk/id/eprint/44878/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

Multi-Agent Deep Reinforcement Learning Based Trajectory Planning for Multi-UAV Assisted Mobile Edge Computing

Liang Wang, Kezhi Wang, Cunhua Pan, Wei Xu, Nauman Aslam and Lajos Hanzo, *Fellow, IEEE*

Abstract—An unmanned aerial vehicle (UAV)-aided mobile edge computing (MEC) framework is proposed, where several UAVs having different trajectories fly over the target area and support the user equipments (UEs) on the ground. We aim to jointly optimize the geographical fairness among all the UEs, the fairness of each UAV’ UE-load and the overall energy consumption of UEs. The above optimization problem includes both integer and continuous variables and it is challenging to solve. To address the above problem, a multi-agent deep reinforcement learning based trajectory control algorithm is proposed for managing the trajectory of each UAV independently, where the popular Multi-Agent Deep Deterministic Policy Gradient (MADDPG) method is applied. Given the UAVs’ trajectories, a low-complexity approach is introduced for optimizing the offloading decisions of UEs. We show that our proposed solution has considerable performance over other traditional algorithms, both in terms of the fairness for serving UEs, fairness of UE-load at each UAV and energy consumption for all the UEs.

Index Terms—Multi-Agent Deep Reinforcement Learning, MADDPG, Mobile Edge Computing, UAV, Trajectory Control.

I. INTRODUCTION

AS a benefit of their compelling features, unmanned aerial vehicles (UAVs) are expected to play a vital role in wireless communication systems. To elaborate a little further, UAVs are capable of providing wireless connectivity even without network infrastructure, or complement the conventional base stations (BSs), whose coverage may suffer from severe blockage due to tall buildings or by the damage caused by natural disasters [1]. In order to support reliable communication links, UAVs can promptly adjust their locations according to the dynamic communication environment.

Corresponding Author: Kezhi Wang

The work of W. Xu was supported in part by the NSFC under grants 62022026 and 61871109. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/N004558/1, EP/P034284/1, EP/P034284/1, EP/P003990/1 (COALESCE), of the Royal Society’s Global Challenges Research Fund Grant as well as of the European Research Council’s Advanced Fellow Grant QuantCom.

Liang Wang, Kezhi Wang and Nauman Aslam are with the Department of Computer and Information Science, Northumbria University, Newcastle upon Tyne, NE1 8ST, U.K., emails: {liang.wang, kezhi.wang, nauman.aslam}@northumbria.ac.uk.

Cunhua Pan is with School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, U.K., email: c.pan@qmul.ac.uk.

Wei Xu is with the National Mobile Communications Research Lab, Southeast University, Nanjing 210096, China, and also with Purple Mountain Laboratories, Nanjing 211111, China, email: wxu@seu.edu.cn.

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K., email: lh@ecs.soton.ac.uk.

Furthermore, since UAVs can be deployed freely and flexibly in three-dimensional (3D) space, direct line-of-sight (LoS) communication with ground-UEs can be readily established, which can potentially boost the throughput in practical scenarios [2]. As a benefit of the above appealing features, in [3] and [4], both fixed-wing UAVs and rotary-wing UAVs were considered as the relaying nodes, for providing seamless connectivity. In [5], Wang *et al.* investigated a fixed-wing UAV-to-UAV communication system, and they proposed a path planning algorithm for minimizing the latency of information transmission, under the constraints of accelerations, location uncertainties and throughput. In [6], Cui *et al.* studied the problem of maximizing the average data rate among UEs in mobile-UAV-enabled networks both in orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) modes. Furthermore, in agricultural applications, as well as in weather monitoring and wildfire management, UAV can be utilized as a mobile data collector [7]. As a future development, in [8], the authors deployed the UAV as the mobile energy transmitter (ET) in a wireless power transfer (WPT) system.

In order to fully exploit the potential of UAVs in wireless communication systems, it is important to investigate their path planning, hovering altitude and trajectory control [9], [10], [11]. In [11], Wang *et al.* creatively proposed a joint UAV altitude and power allocation optimization method, which beneficially alleviated the inter-cell interference of each UAV network. In [12], Al-Hourani *et al.* optimized the latitude of UAVs in order to provide the maximum radio coverage area on the ground. In [13], both static and mobile UAVs were considered in device-to-device (D2D) networks. Additionally, the UAV’s altitude was optimized for maximizing the system’s sum-rate and coverage probability. To tackle the throughput maximization problem of UAV-aided mobile relaying systems, Zeng *et al.* [14] proposed an iterative algorithm to optimize the UAV’s trajectory and power allocation. In the content of multi-UAV enabled multiuser systems, Wu *et al.* [15] maximized the minimum throughput over all ground users by jointly optimizing the user scheduling, power control and UAV trajectories. In order to meet the different quality-of-service (QoS) requirement of users, Alzenad *et al.* [16] investigated coverage-placement problem of UAV-BSs and proposed an optimal placement algorithm for maximizing the number of users supported.

In recent years, mobile edge computing (MEC) has been shown to dramatically improve the user experience [17], [18]. By providing both computing and storage hardware at the

network edges, namely at the BSs or access points (APs), the resource-limited UEs have the option of offloading their computation-intensive and latency-critical applications to the MEC servers [19]. Due to the mobility of UAVs, recent years have seen research progress on the integration of UAVs with MEC [20], [21]. In [22], Motlagh *et al.*, were amongst the first who proposed UAV-enabled MEC, in which UEs can significantly reduce the energy consumption via offloading. In order to minimize the overall energy dissipation of UEs while meeting their QoS requirement, Jeong *et al.* [23] proposed an efficient successive convex approximation-based algorithm for jointly optimizing the bit allocation and UAV's trajectory. Considering a multi-UAV system, Hua *et al.* [24] investigated the multi-UAV scenario, and they optimized the UAVs' trajectories, transmit power and user scheduling.

Given the recent advances in machine learning [25], the combination of deep neural networks (DNNs) [26] and reinforcement learning (RL) [27], i.e., deep reinforcement learning (DRL) has become a hot research topic. In DRL, an agent is assumed to interact with the environment for learning the optimal policy with the aid of exploration. Compared to traditional RL, DRL facilitates more accurate convergence and approximation by exploiting the power of DNNs for estimating the associated functions in RL [28]. The great potential of DRL in solving complex control problems has also been demonstrated in [29], [30], [31], [32], [33]. In [29], Mnih *et al.* introduced the deep Q network (DQN) philosophy, which ignited the field of DRL. For instance, Wang *et al.* [30] systematically investigated the problem of distributed Q-learning aided heterogeneous network association in the content of energy-efficient Internet of things (IoT). In order to improve the training procedure, DQN relies on a pair of techniques namely, experience replay and target networks. For the sake of tackling the typical over-estimation problem of RL, a double DQN (D-DQN) was proposed by Van Hasselt *et al.* [31]. However, DQN may suffer from the curse of high-dimensional action spaces and cannot be readily applied to continuous domains. Thus, motivated by this, Lillicrap *et al.* [32] proposed a deep deterministic policy gradient (DDPG) technique based on the so-called actor-critic architecture, which can be readily applied for a range of challenging problems. A comprehensive survey of multi-agent RL, have also been provided by Bu *et al.* [33].

Against the above background, we conceive a multi-UAV assisted MEC framework, where each UAV is controlled by a dedicated agent. We aim for jointly maximizing the geographical fairness¹ among the UEs covered, the fairness of UE-load of each UAV², while minimizing the overall energy consumption of UEs by optimizing each UAV's trajectory and offloading decisions. This is a complex problem which includes both integer and continuous variables. Hence it is challenging to address it by traditional algorithms, such as convex optimization and dynamic programming. Therefore, we conceive a multi-agent deep reinforcement learning based solution, with the help of the popular Multi-Agent Deep

Deterministic Policy Gradient (MADDPG) [34] for solving it. Given the UAVs' trajectories, a low-complexity approach is introduced for optimizing the offloading decisions of UEs. Our simulation results will show that the proposed DRL based algorithm outperforms the benchmark algorithms. We summarize the difference between our work and the existing literature in Table I.

The rest of the paper is organized as follows. In Section II, we introduce the system model and the optimization problem. In Section III, our multi-agent based DRL algorithm is proposed. Our experimental results are shown in Section IV. Finally, our conclusions are drawn in Section V.

The main notations used in this paper are summarized in Table II.

II. SYSTEM MODEL

In this section, we describe the system model. As shown in Fig. 1, we assume that there are N UEs randomly distributed in a square-shaped area with side length l^{max} , and the set of UEs is denoted as $\mathcal{N} \triangleq \{n = 1, 2, \dots, N\}$. There are M UAVs flying at a fixed altitude H over the target area to serve the ground UEs, and the set of UAVs is denoted as $\mathcal{M} \triangleq \{m = 1, 2, \dots, M\}$. We also assume that UAVs can be deployed and easily charged on the building roof when UAVs run out of their energy. Assume that each UE has a computational task to be executed at each time slot (TS) over T consecutive TSs, $\mathcal{T} \triangleq \{t = 1, 2, \dots, T\}$. Each of the tasks can be executed either by the UE or offloaded to one of the UAVs. We define a new set $m \in \mathcal{M}' \triangleq \{0, 1, \dots, M\}$ to denote the possible places where the tasks can be executed, with $m = 0$ representing local execution. Then, we define the offloading decision variable $z_{n,m,t}$ as

$$z_{n,m,t} = \{0, 1\}, \forall n \in \mathcal{N}, m \in \mathcal{M}', t \in \mathcal{T}, \quad (1)$$

where $z_{n,m,t} = 1, m \neq 0$ means that UE n decides to offload the task to UAV m in TS t , while $z_{n,m,t} = 1, m = 0$ represents that UE n carries out the task itself in TS t , and otherwise $z_{n,m,t} = 0$. Furthermore, we assume that each task can only be executed at a single place. Thus, we have

$$\sum_{m=0}^M z_{n,m,t} = 1, \forall n \in \mathcal{N}, t \in \mathcal{T}. \quad (2)$$

Similarly to [35], in the TS t , we assume that UE n has a computationally intensive task $S_{n,t}$ to be executed, which is defined as

$$S_{n,t} = \{D_{n,t}, F_{n,t}\}, \forall n \in \mathcal{N}, t \in \mathcal{T}, \quad (3)$$

where $D_{n,t}$ denotes the data volume to be processed, while $F_{n,t}$ describes the total number of the CPU cycles required for executing this task. Both $D_{n,t}$ and $F_{n,t}$ can be characterized as in [36].

Furthermore, in TS t , each of UAV flies in a direction determined by the angle of $\alpha_{m,t} \in [0, 2\pi)$, distance of $d_{m,t} \in [0, d^{max}]$, and cannot go beyond the border of the target area. We assume that the initial coordinates of UAV m are set as $[X_{m,0}, Y_{m,0}, H]$. Then, the coordinates of UAV m in TS t can be calculated as $[X_{m,t}, Y_{m,t}, H]$, where $X_{m,t} = X_{m,0} +$

¹The geographical fairness reflects the QoS level of UEs served by UAVs from the initial time slot to the current time slot.

²The UE-load of UAV is defined in (18).

TABLE I: Comparison between our work and the existing literature.

Reference	Single UAV	Multi UAV	Mobile edge computing (incl.)	Path planning	Offloading decision	Reinforcement learning (e.g., Q-learning)	Multi agent learning	DNN
[2]	✓			✓				
[4]	✓			✓				
[5]		✓		✓				
[6]	✓			✓				
[7]	✓			✓				
[8]	✓			✓				
[9]		✓		✓				✓
[10]		✓		✓		✓	✓	
[11]		✓		✓				
[13]	✓			✓				
[14]		✓		✓				
[15]		✓		✓				
[16]	✓			✓				
[20]		✓	✓	✓	✓			
[21]	✓			✓	✓			
[23]	✓		✓	✓				
[24]		✓		✓				
[30]						✓	✓	
[34]						✓	✓	✓
Our work		✓	✓	✓	✓	✓	✓	✓

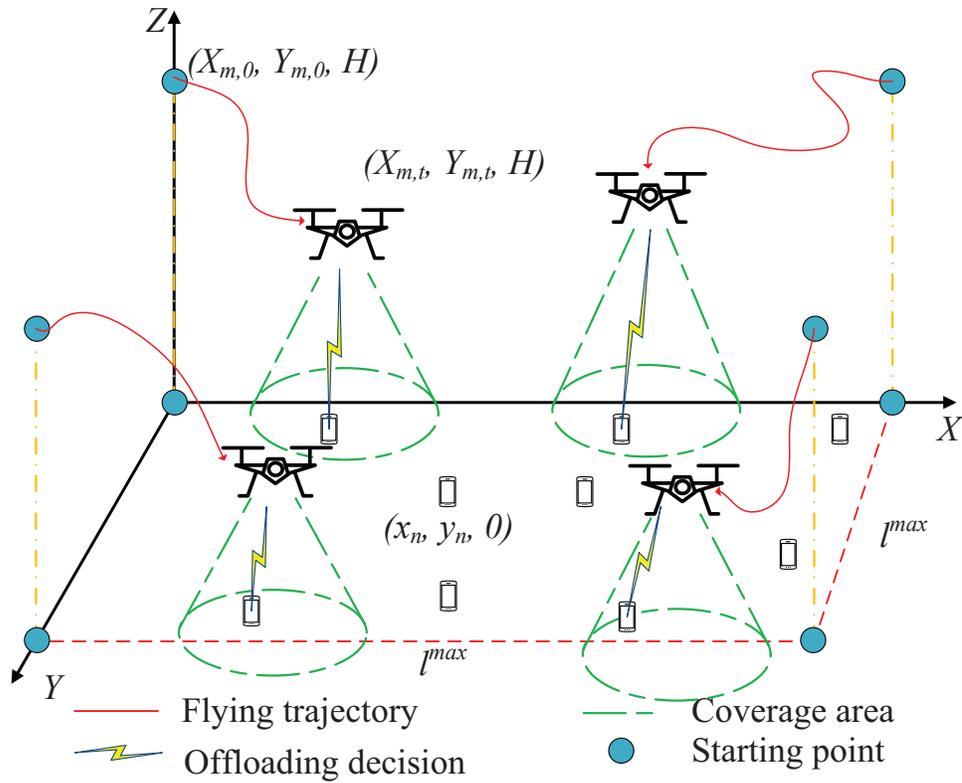


Fig. 1: Overall System Architecture

TABLE II: List of main notations

Notation	Description
n, N, \mathcal{N}	The index, number and the set of UEs
m, M, \mathcal{M}	The index, number and the set of UAVs
t, T, \mathcal{T}	The index, number and the set of TSs
$z_{n,m,t}$	Offloading decision of UE n
$S_{n,t}$	Computation task of UE n in TS t
$D_{n,t}$	Data volume of task $S_{n,t}$
$F_{n,t}$	Overall CPU cycles required for task $S_{n,t}$
$f_{n,m,t}$	Computation capacity of UAV m allocated to UE n
$T_{n,m,t}^C$	Execution time of UAV m to UE n in TS t
$T_{n,m,t}^{Tr}$	Transmission time of UE n to UAV m in TS t
T^{max}	Maximal time duration of each TS
$\alpha_{m,t}, d_{m,t}$	Flying angle and distance of UAV m in TS t
d^{max}	Maximal flying distance of UAV in each TS
$[X_{m,t}, Y_{m,t}, H]$	Coordinates of UAV m in TS t
$[x_n, y_n]$	Coordinates of UE n
$R_{n,m,t}$	Horizontal distance between UAV m and UE n
$R_{m,m',t}$	Horizontal distance between UAV m and UAV m'
R^{max}	Maximal horizontal coverage radius of UAV
$r_{n,m,t}$	Transmitting data rate of UE n to UAV m
$E_{n,m,t}^C$	Energy consumption for task execution
$E_{n,m,t}^{Tr}$	Energy consumption for offloading
$c_{m,t}$	Relative UE-load of UAV m in TS t
f_t^u	Fairness index of UE-load of each UAV in TS t
f_t^e	Fairness index of UEs in TS t

$\sum_{t'=1}^t d_{m,t'} \cos(\alpha_{m,t'})$ and $Y_{m,t} = Y_{m,0} + \sum_{t'=1}^t d_{m,t'} \sin(\alpha_{m,t'})$. Thus, we have

$$0 \leq X_{m,t} \leq l^{max}, \forall m \in \mathcal{M}, t \in \mathcal{T}, \quad (4)$$

and

$$0 \leq Y_{m,t} \leq l^{max}, \forall m \in \mathcal{M}, t \in \mathcal{T}. \quad (5)$$

Additionally, we denote the distance between UAV m and UAV m' in TS t as $R_{m,m',t}$, which can be expressed as

$$R_{m,m',t} = \sqrt{(X_{m,t} - X_{m',t})^2 + (Y_{m,t} - Y_{m',t})^2}. \quad (6)$$

We assume that the UAVs should keep a minimal distance of R^u for avoiding their collision in each TS. Then, we have

$$R_{m,m',t} \geq R^u, \forall m, m' \in \mathcal{M}, m \neq m'. \quad (7)$$

The horizontal distance between UE n and UAV m in TS t is calculated as

$$R_{n,m,t} = \sqrt{(X_{m,t} - x_n)^2 + (Y_{m,t} - y_n)^2}, \quad (8)$$

$$\forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T},$$

where $[x_n, y_n]$ is assumed to be the coordinate of UE n . Note that if UE n decides to offload a task to UAV m in TS t , it must be in the coverage of UAV m . Then, we have

$$z_{n,m,t} R_{n,m,t} \leq R^{max}, \forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T}, \quad (9)$$

where R^{max} is the maximal horizontal coverage radius of the UAVs.

Then, the offloading data rate can be expressed by

$$r_{n,m,t} = B \log_2 \left(1 + \frac{\rho P_n}{H^2 + R_{n,m,t}^2} \right), \quad (10)$$

$$\forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T},$$

where B is the channel's bandwidth, P_n is the transmission power of UE n , $\rho = g_0 G_0 / \sigma^2$, $G_0 \approx 2.2846$, g_0 is the channel's power gain at the reference distance of 1 m and σ^2 is the noise power [37]. Here we do not consider any particular modulation and coding scheme.

Thus, if UE n decides for offloading its task to UAV m in TS t , the time required for offloading the data is given by

$$T_{n,m,t}^{Tr} = \frac{D_{n,t}}{r_{n,m,t}}, \forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T}, \quad (11)$$

and the execution time of the task can be expressed as

$$T_{n,m,t}^C = \frac{F_{n,t}}{f_{n,m,t}}, \forall n \in \mathcal{N}, m \in \mathcal{M}', t \in \mathcal{T}, \quad (12)$$

where $f_{n,m,t}$ represents the computational capability of UAV m that can be allocated to UE n , and $m = 0$ indicates local execution. Thus, the overall time required for executing the task can be described as

$$T_{n,m,t} = \begin{cases} T_{n,m,t}^C & \text{if local execution,} \\ T_{n,m,t}^{Tr} + T_{n,m,t}^C & \text{if offloading.} \end{cases} \quad (13)$$

We also assume that all tasks should be executed within the maximal time duration T^{max} of TS. Then, we have

$$z_{n,m,t} T_{n,m,t} \leq T^{max}, \forall n \in \mathcal{N}, m \in \mathcal{M}', t \in \mathcal{T}. \quad (14)$$

According to [38], if the UE n decides to execute a task locally, the energy consumption is given by

$$E_{n,m,t}^C = k_n (f_{n,m,t})^{v_n} T_{n,m,t}^C, \forall n \in \mathcal{N}, t \in \mathcal{T}, \quad (15)$$

where $k_n \geq 0$, $v_n \geq 1$ are positive coefficients.

If UE n decides to offload a task, the energy consumption of offloading is

$$E_{n,m,t}^{Tr} = P_n T_{n,m,t}^{Tr}, \forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T}. \quad (16)$$

Thus, the energy consumption at UE n can be expressed as

$$E_{n,m,t} = \begin{cases} E_{n,m,t}^C & \text{if local execution,} \\ E_{n,m,t}^{Tr} & \text{if offloading.} \end{cases} \quad (17)$$

Then, we define $c_{m,t} \in [0, 1]$ as the relative UE-load of UAV m in TS t , as:

$$c_{m,t} = \frac{\sum_{n=1}^N z_{n,m,t}}{N}, \forall m \in \mathcal{M}, t \in \mathcal{T}. \quad (18)$$

In this paper, our first objective is to minimize the total energy consumption of UEs via optimizing both the offloading decisions and the UAVs' trajectories. However, this may lead to an unfair process since some UAVs may serve more UEs than others. To address this issue, we define a fairness index f_t^u as

$$f_t^u = \frac{(\sum_{m=1}^M \sum_{t'=1}^t c_{m,t'})^2}{M \sum_{m=1}^M (\sum_{t'=1}^t c_{m,t'})^2}, \quad (19)$$

where f_t^u reflects the level of fairness among the UAVs physically, if all the UAVs have a similar UE-load commencing from the initial TS up to TS t , the value of f_t^u is closer to 1.

Then, to avoid the situation that some UEs are served during many TSs, while others are never served at all, we define another geographical fairness f_t^e as follows

$$f_t^e = \frac{(\sum_{n=1}^N \sum_{t'=1}^t z_{n,m,t'})^2}{N \sum_{n=1}^N (\sum_{t'=1}^t z_{n,m,t'})^2}, \quad (20)$$

where f_t^e reflects the level of fairness among the UEs, explicitly, if all UEs are served for a similar number of TSs commencing from the initial TS to the TS t , the value of f_t^e is closer to 1.

Then, we formulate our optimization problem as follows

$$\mathcal{P}1 : \max_{\mathbf{P}, \mathbf{Z}} \sum_{t=1}^T \frac{f_t^u \cdot f_t^e}{\sum_{n=1}^N \sum_{m=0}^M z_{n,m,t} E_{n,m,t}} \quad (21a)$$

subject to:

$$z_{n,m,t} = \{0, 1\}, \forall n \in \mathcal{N}, m \in \mathcal{M}', t \in \mathcal{T}, \quad (21b)$$

$$\sum_{m=0}^M z_{n,m,t} = 1, \forall n \in \mathcal{N}, t \in \mathcal{T}, \quad (21c)$$

$$0 \leq X_{m,t} \leq l^{max}, \forall m \in \mathcal{M}, t \in \mathcal{T}, \quad (21d)$$

$$0 \leq Y_{m,t} \leq l^{max}, \forall m \in \mathcal{M}, t \in \mathcal{T}, \quad (21e)$$

$$0 \leq \alpha_{m,t} < 2\pi, \forall m \in \mathcal{M}, t \in \mathcal{T}, \quad (21f)$$

$$0 \leq d_{m,t} \leq d^{max}, \forall m \in \mathcal{M}, t \in \mathcal{T}, \quad (21g)$$

$$R_{m,m',t} \geq R^u, \forall m, m' \in \mathcal{M}, m \neq m', \quad (21h)$$

$$z_{n,m,t} R_{n,m,t} \leq R^{max}, \forall n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T}, \quad (21i)$$

$$z_{n,m,t} T_{n,m,t} \leq T^{max}, \forall n \in \mathcal{N}, m \in \mathcal{M}', t \in \mathcal{T}. \quad (21j)$$

where $\mathbf{P} = \{\alpha_{m,t}, d_{m,t}, \forall m \in \mathcal{M}, t \in \mathcal{T}\}$ and $\mathbf{Z} = \{z_{n,m,t}, \forall n \in \mathcal{N}, m \in \mathcal{M}', t \in \mathcal{T}\}$. Our objectives are to maximize the fairness of UE-load of each UAV and the fairness of the number of times that each UE is served by UAVs over all the TSs, while minimizing the overall energy consumption of UEs. It is readily observed that the optimization problem cannot be solved by traditional approaches, since it involves both the continuous variables \mathbf{P} and the discrete variables \mathbf{Z} . Thus, in this paper, a Multi-Agent deep reinforcement learning based Trajectory control algorithm (MAT) is proposed.

III. THE PROPOSED ALGORITHM

In this section, we present our proposed algorithm. First, some background knowledge on deep reinforcement learning is provided, followed by our MAT conceived for solving Problem (21).

A. Background Knowledge

In the traditional reinforcement learning setup, a Markov decision process (MDP) [39] is employed with the state space of $\mathcal{S} = \{s_t = s_1, s_2, \dots, s_T\}$ and the action space of $\mathcal{A} = \{a_t = a_1, a_2, \dots, a_T\}$. In the MDP, a decision agent interacts with the environment in discrete TSs. More specifically, the decision agent observes the current state s_t of the environment and takes the action a_t that is allowed in that state. As a benefit, the agent will obtain a reward r_t and traverses to a new state s_{t+1} . In [32], a policy $a_t = \pi(s_t)$ is introduced that maps the state to a legitimate action. During the process of

interacting with the environment, the agent aims for selecting the beneficial policy that maximizes the accumulated reward $R_t = \sum_{i=t}^T \gamma^{i-t} r_i$, where $\gamma \in (0, 1)$ is the discount factor and T is the number of TSs. Additionally, as a beneficial combination of DNN and RL, the philosophy of DQN [29] was proposed, which uses an action-value function $Q(\cdot)$ for approximately evaluating R_t by applying a_t, s_t and following $\pi(\cdot)$:

$$Q(s_t, a_t) = \mathbb{E}_\pi [R_t | s_t, a_t], \quad (22)$$

which is known as the Q-function, and can be obtained by a DNN. Then, the DNN can be trained with the aid of the loss function $L(\cdot)$ defined as:

$$L(\theta^Q) = \mathbb{E} [y_t - Q(s_t, a_t | \theta^Q)]^2, \quad (23)$$

where θ^Q denotes the network parameter of the DNN and y_t is formulated by

$$y_t = r_t + \gamma Q(s_{t+1}, a_{t+1} | \theta^Q). \quad (24)$$

where a_{t+1} denotes the action of the agent in the next TS, which is generated by the DNN, given the state s_{t+1} .

Furthermore, in order to make the network more stable, a pair of techniques, namely the experience replay and the target network are utilized. The experience replay employs a buffer for storing transitions for the sake of mitigating the correlations between consecutive transitions and hence increasing their independence. Compared to the original RL training procedure, the DQN relies on a mini-batch for randomly sampling the transitions from the experience replay buffer, rather than only selecting a single transition. Furthermore, the target network that has the same network structure as $Q(s_t, a_t)$ is employed for reducing the correlations. Note that the target network is only updated at certain intervals.

However, it is proved in [32] that DQN cannot be directly used for solving continuous-valued control problems. Thus, the popular actor-critic method at DDPG of [32] is resorted to. Specifically, DDPG consists of a DNN turned as actor and a DQN referred to as the critic network. The actor carries out the mapping function $\pi(s_t | \theta^\pi)$ while the critic performs the function $Q(s_t, a_t | \theta^Q)$. The actor can generate the optimal action a_t based on the state s_t and it can be trained by applying the policy gradient method of [40] defined as

$$\begin{aligned} \nabla_{\theta^\pi} J &\approx \mathbb{E} [\nabla_{\theta^\pi} Q(s, a | \theta^Q) |_{s=s_t, a=\pi(s_t | \theta^\pi)}] \\ &= \mathbb{E} [\nabla_a Q(s, a | \theta^Q) |_{s=s_t, a=\pi(s_t)} \nabla_{\theta^\pi} \pi(s | \theta^\pi) |_{s=s_t}], \end{aligned} \quad (25)$$

while the critic network can be updated by using the loss function of (23).

B. MAT

In this section, by applying the popular MADDPG [34], we conceive a multi-agent MDP, namely an observable Markov game [41]. It is assumed that there are M agents interacting with the environment characterized by a set of states $\mathcal{S} \triangleq \{s_t, t \in \mathcal{T}\}$ and a set of actions $\mathcal{A} \triangleq \{a_t, t \in \mathcal{T}\}$. The state s_t consists of the private observation $o_{m,t}$ and some other extra information known by each agent. Additionally, each UAV is controlled by its dedicated agent. In each TS, each agent obtains its private observation $o_{m,t}$ and takes its

own action $a_{m,t}$ as well as receives a reward $r_{m,t}$. Then, the environment updates the state and traverses to a new state. Note that each agent is equipped with an actor network $a_{m,t} = \pi^m(o_{m,t})$, a critic network $Q^m(s_t, a_t)$, their target networks $a_{m,t+1} = \pi^{m'}(o_{m,t+1})$ and $Q^{m'}(s_{t+1}, a_{t+1})$, as well as an experience replay buffer B_m .

The proposed algorithm is based on the framework of centralized training combined with decentralized execution. During the training process, each agent sends its own private observation $o_{m,t}$ and action $a_{m,t}$ to the environment, and then the states s_t which consist of the observations of all the agents and actions are sent back to each agent. Here, all the agents can exchange their private information simultaneously with each other, including coordinates. Furthermore, the critic network of each agent is trained with the states and actions that includes all the agents' observations and actions. Then, during the testing process, each agent can execute its action by only receiving its own private observations $o_{m,t}$, which can potentially maximize the accumulated rewards.

Thus, we define the observation, action and reward function for each agent in TS t as follows:

- 1) Observation $o_{m,t}$: we first add the coordinates $[X_{m,t}, Y_{m,t}]$ of UAV m in TS t into the observation of agent m . For avoiding collisions between each pair of UAVs, we define the set of relative UAV distances $\{R_{m,m',t}, m' \in \mathcal{M}, m' \neq m\}$ as part of the observation. Additionally, for better exploration, we also add the set of accumulated times of UEs served by UAVs and UE-load of UAVs commencing from the initial TS up to TS t , i.e., $\{\sum_{t'=1}^t z_{n,m,t'}, \forall n \in \mathcal{N}\}, \{\sum_{t'=1}^t c_{m,t'}, \forall m \in \mathcal{M}\}$, respectively into the observation set.
- 2) Action $a_{m,t}$: we define the UAV's flying direction and distance as the action $a_{m,t} = \{\alpha_{m,t}, d_{m,t}\}$ of the m -th UAV in the t -th TS.
- 3) Reward Function $r_{m,t}$: we define the reward function as:

$$r_{m,t} = \frac{f_t^u \cdot f_t^e}{\frac{1}{N} \sum_{n=1}^N \sum_{m=0}^M z_{n,m,t} \cdot E_{n,m,t}} - p_m, \quad (26)$$

where p_m is the penalty incurred if UAV m flies out of the target area or UAV m is collided with another UAV (i.e., the relative distance is under the defined limit).

Then, we define the entire state s_t , and action a_t as follows

- 1) State s_t : the state consists of the observations of all the agents, which is expressed as $s_t = \{o_{m,t}, \forall m \in \mathcal{M}\}$.
- 2) Action a_t : the action consists of the actions of all the agents, which is $a_t = \{a_{m,t}, \forall m \in \mathcal{M}\}$.

We show the structure of agent m in Fig. 2. During its interaction with the environment ①, each UAV (controlled by agent) ② selects the optimal action associated with its actor network $\pi^m(\cdot)$ ⑥, and then obtains the Q value from the critic network $Q^m(\cdot)$ ⑧ as well as its target action and target Q value from $\pi^{m'}(\cdot)$ ⑦ and $Q^{m'}(\cdot)$ ⑨ respectively. The profile of observation, action and reward, which determine the transition are defined as $e_{m,t} \triangleq \{s_t, a_t, r_{m,t}, s_{t+1}\}$ that are stored in the experience replay buffer ④. However, during the training procedure, randomly sampling the mini-batch ⑤ may have unpredictable effects, since some transitions associated

with poor attempts may lead to the termination of the training procedure or may not converge. As a result, Schaul *et al.* [42] pointed out that transitions having high Temporal Difference (TD)-error often indicate successful attempts. The TD-error δ_m of agent m can be defined as follows

$$\delta_m = r_{m,t} + \gamma Q^{m'}(s_{t+1}, a_{t+1} | \theta^{Q^{m'}}) - Q^m(s_t, a_t | \theta^{Q^m}), \quad \forall m \in \mathcal{M}, t \in \mathcal{T}. \quad (27)$$

Additionally, motivated by [42], we utilize a prioritized experience replay scheme, in which the absolute TD-error $|\delta_{m,k}|$ was used for evaluating the probability of the k -th sampled transition in the mini-batch. Then, the probability of sampling the k -th transition is expressed as

$$P_{m,k} = \frac{(|\delta_{m,k}| + \varepsilon)^\beta}{\sum_{k'=1}^K (|\delta_{m,k'}| + \varepsilon)^\beta}, \quad \forall m \in \mathcal{M}, \quad (28)$$

where K is the size of mini-batch, ε is a positive constant value, and β is 0.6. Thus, the loss function ⑩ of the agent m is defined as

$$L(\theta^{Q^m}) = \mathbb{E} \left[\frac{1}{(K \cdot P_{m,k})^\mu} (\delta_m)^2 \right], \quad (29)$$

where μ is given as 0.4.

Then, the critic network ⑧ of agent m can be updated by the loss function ⑩ provided in (29). Furthermore, the actor network ⑥ of agent m can be trained by the policy gradient ⑪ defined as

$$\begin{aligned} \nabla_{\theta^{\pi^m}} J &= \mathbb{E} \left[\nabla_{\theta^{\pi^m}} \pi^m(o_{m,t} | \theta^{\pi^m}) \right. \\ &\quad \left. \nabla_{a_{m,t}} Q^m(s_t, a_t) | \theta^{Q^m} \right], \quad \forall m \in \mathcal{M}, t \in \mathcal{T}. \end{aligned} \quad (30)$$

Given the UAVs' trajectories, we introduce a low-complexity approach for optimizing the offloading decisions of UEs. Here, we do not consider the constraint of the maximal available computing resource in each UAV. This can be readily extended to more practical scenarios, where each UAV can only have a certain amount of the computing resources, with the introduction of the matching algorithm. We will leave this idea for our future work. For each UE in TS t , we select the offloading decision based on the following expression

$$z_{n,m,t} = \begin{cases} 1, & m = \underset{m' \in \mathcal{M}'}{\operatorname{argmin}} \{E_{n,m',t}\}, \\ 0, & \text{otherwise.} \end{cases} \quad (31)$$

Specifically, after the movement of UAVs, each UE can select the most suitable UAV for offloading, which consumes the least energy. Otherwise, the UE may execute the task itself. If UE n decides to offload a task to UAV m , the computational capacity allocated to UE from the UAV is expressed as

$$f_{n,m,t} = \frac{F_{n,t}}{T_{n,m,t}^{\max} - T_{n,m,t}^{\text{Tr}}}. \quad (32)$$

We provide the pseudo code of proposed procedure in Algorithm 1. Specifically, we carry out the initialization between Line 1 and 5 at the beginning, where each UAV initializes its actor, critic and two target networks. Then, the training procedure starts from Line 6, where each UAV first obtains its observation from the environment ①. Note that each UAV is controlled by its dedicated agent ②. Then, based on the

Algorithm 1 MAT

```

1: for UAV  $m$  in  $\mathcal{M}$  do
2:   Initialize actor network  $\pi^m(\cdot)$ , critic network  $Q^m(\cdot)$  with parameters  $\theta^{\pi^m}$  and  $\theta^{Q^m}$ ;
3:   Initialize target networks  $\pi^{m'}(\cdot)$  and  $Q^{m'}(\cdot)$  with parameters  $\theta^{\pi^{m'}} = \theta^{\pi^m}$  and  $\theta^{Q^{m'}} = \theta^{Q^m}$ ;
4:   Initialize experience replay buffer  $B_m$ ;
5: end for
6: for Episode = 1, 2, ...,  $e^{max}$  do
7:   for UAV  $m$  in  $\mathcal{M}$  do
8:     Initialize observation  $o_{m,t}$ ;
9:   end for
10:  for TS  $t$  in  $\mathcal{T}$  do
11:    Obtain  $s_t$ ;
12:    for UAV  $m$  in  $\mathcal{M}$  do
13:      Obtain action  $a_{m,t} = \pi^m(o_{m,t} | \theta^{\pi^m}) + \epsilon$ ;
14:      Execute  $a_{m,t}$ . Note that the UAV will stay at the current location if it flies out of the target area or it is collided with another UAV;
15:    end for
16:    Obtain  $a_t$ ;
17:    for UE  $n$  in  $\mathcal{N}$  do
18:      Obtain the available offloading decision  $z_{n,m,t}$  that consumes the least energy according to (31);
19:      Calculate  $E_{n,m,t}$ ;
20:    end for
21:    for UAV  $m$  in  $\mathcal{M}$  do
22:      Obtain  $r_{m,t}$  according to (26);
23:      Obtain  $o_{m,t+1}$ ;
24:    end for
25:    Obtain  $s_{t+1}$ ;
26:    for UAV  $m$  in  $\mathcal{M}$  do
27:      Store transition  $\{s_t, a_t, r_{m,t}, s_{t+1}\}$  into experience replay buffer  $B_m$  with priority  $|\delta_m| + \epsilon$ ;
28:      if learning process starts then
29:        Sample a mini-batch of  $K$  transitions from  $B_m$  with probability  $P_{m,k}$ ;
30:        Update critic network according to (29);
31:        Update actor network according to (30);
32:        Update target networks with updating rate  $\tau$ :
33:           $\theta^{\pi^{m'}} \leftarrow \tau \theta^{\pi^m} + (1 - \tau) \theta^{\pi^{m'}}$ ;
34:           $\theta^{Q^{m'}} \leftarrow \tau \theta^{Q^m} + (1 - \tau) \theta^{Q^{m'}}$ ;
35:        Update priorities of  $K$  transitions;
36:      end if
37:    end for
38:  end for
39: end for

```

TABLE III: Simulation parameters

Notation	Description
N	50
T	20
l^{max}	100 m
$D_{n,t}$	[10, 14] Kb
$F_{n,t}$	[1800, 2000] cycles/bit
T^{max}	1 s
d^{max}	20 m
R^{max}	20 m
R^u	1 m
H	50 m
B	10 MHz
P_n	0.1 Watt
σ^2	-90 dBm
k_n	10^{-28}
v_n	3
g_0	1.42×10^{-4}
γ	0.95
K	256
τ	0.01
ϵ	0.001
B_m	10^5
e^{max}	3000
p_m	10

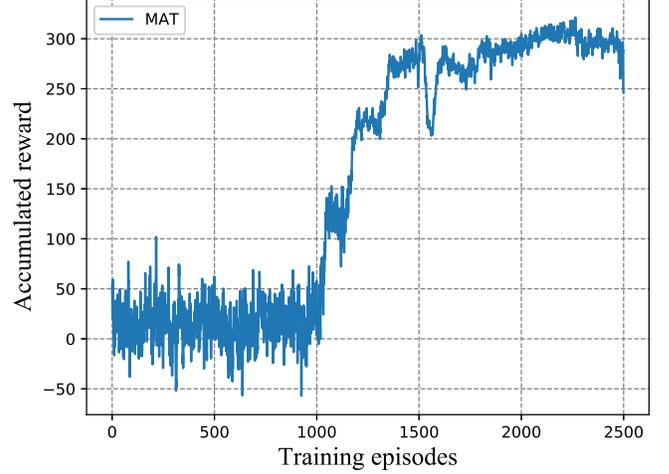


Fig. 3: Accumulated reward versus training episodes (with 3 UAVs).

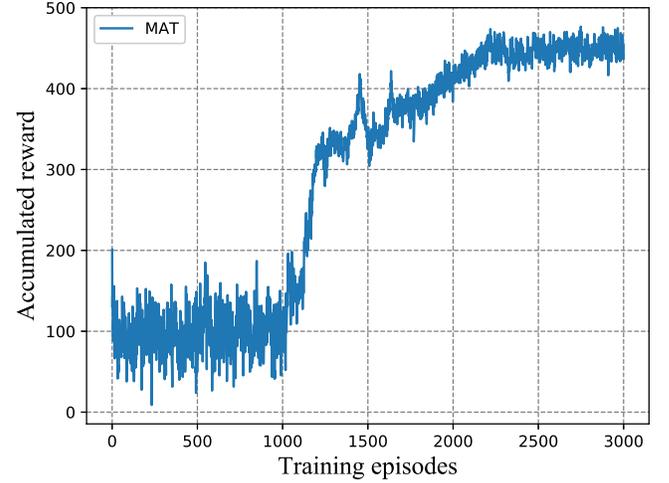


Fig. 4: Accumulated reward versus training episodes (with 4 UAVs).

solutions:

- **RANDOM:** In this setup, each UAV randomly selects a flying direction within $\alpha_{m,t} \in [0, 2\pi)$, and a flying distance $d_{m,t} \in [0, d^{max}]$. Note that the UAVs are restricted to the target area.
- **CIRCLE:** We group all the UEs into a single cluster according to the UEs' coordinates and then all the UAVs fly in a circle twice around the center of the cluster having a radius of R^{max} .

Note that the MAT, RANDOM, and CIRCLE benchmarks have the same starting points for the UAVs and their offloading decisions are described in Eq. (31).

We first depict the UAV trajectories in Fig. 5, where 3 UAVs are deployed. In this figure, dots represent the location of UEs. We apply a heat map to show the number of times that each UE

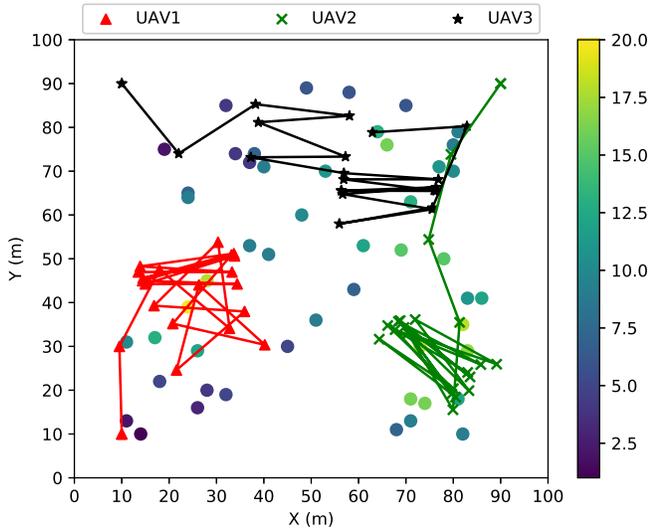


Fig. 5: UAVs' trajectories (with 3 UAVs and the locations of UEs are represented by dots.)

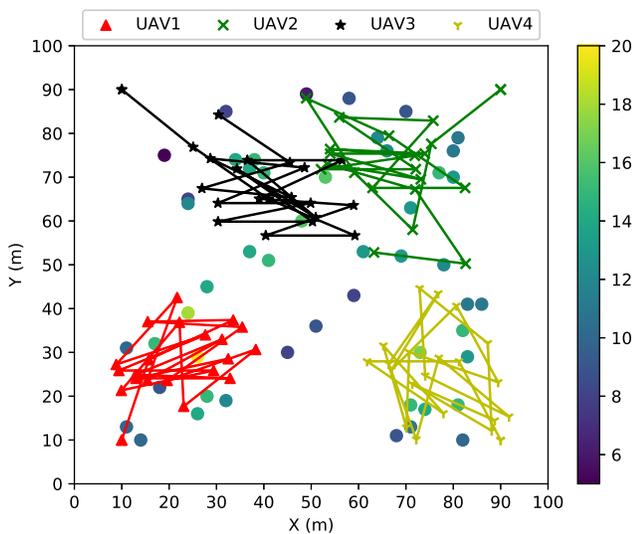
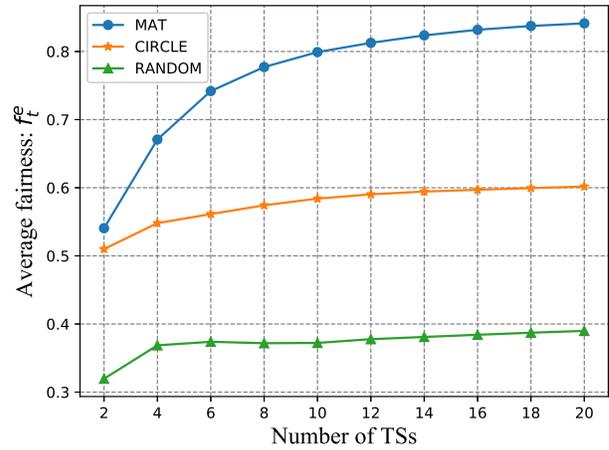


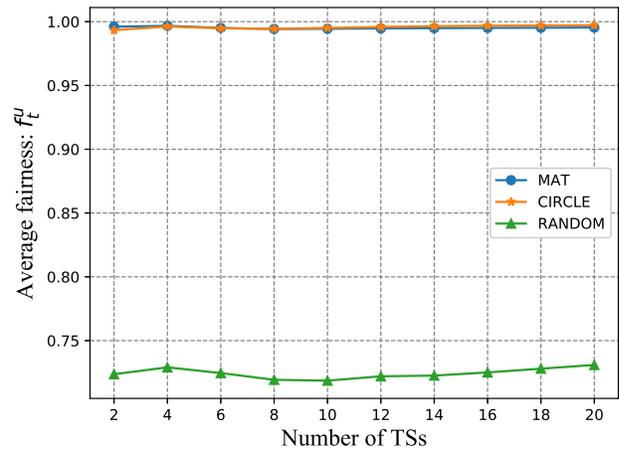
Fig. 6: UAVs' trajectories (with 4 UAVs and the locations of UEs are represented by dots.)

is served by the UAV commencing from the initial TS to the final TS. The darker the dots, the less amount of time that the UE is spent by the UAV serving. Observe from this figure that all the UAVs move around certain areas, since their coverage range is limited and they have to move for the sake of serving more UEs to increase the fairness index. Additionally, we can see that each UAV covers the particular area in a cooperative manner, so as to maximize the reward defined. For instance, 'UAV2' moves to the lower right corner from its initial location for serving more UEs, while 'UAV3' moves to the upper right corner to help users in this region.

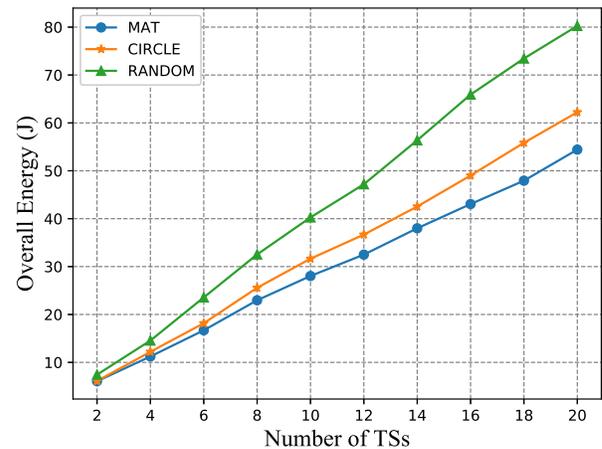
Then, we increase the number of UAVs to 4 and depict the trajectories in Fig. 6. Observe that more UAVs result in better coverage. Again, the UAVs cooperate for serving more UEs



(a)

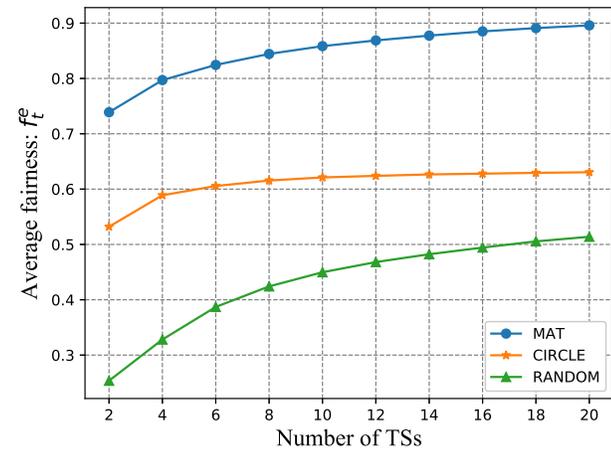


(b)

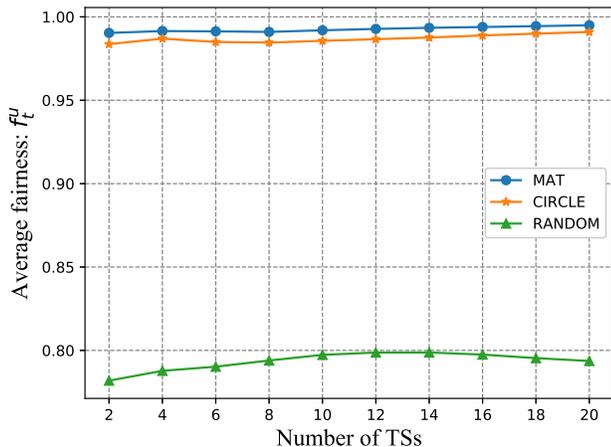


(c)

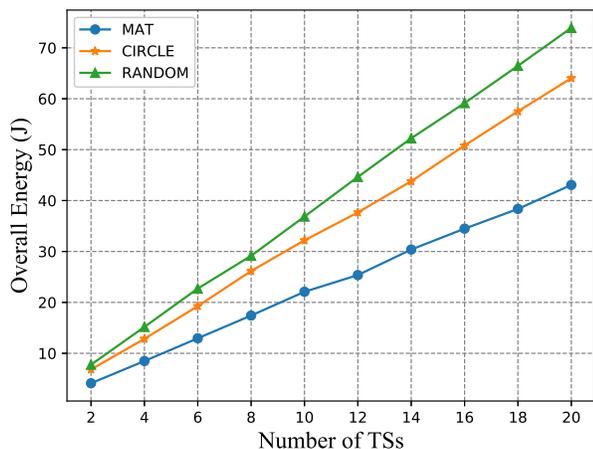
Fig. 7: The performance of MAT, CIRCLE and RANDOM versus different number of TSs, in terms of (a) fairness index f_t^e , (b) fairness index f_t^u and (c) overall energy consumption of all the UEs (with 3 UAVs).



(a)



(b)



(c)

Fig. 8: The performance of MAT, CIRCLE and RANDOM versus different number of TSs, in terms of (a) fairness index f_t^e , (b) fairness index f_t^u and (c) overall energy consumption of all the UEs (with 4 UAVs).

within the required number of TSs. Furthermore, compared to the heat map shown in Fig. 5, 4 UAVs can serve each UE more times than 3. More specially, 4 UAVs can increase the minimum number of serving occurrences from about 2.5 TSs in Fig. 5 to about 6 TSs in Fig. 6.

In Fig. 7, we show the fairness attained by 3 UAVs while serving all UEs, the fairness of each UAV's UE-load and the overall energy consumption of all the UEs. Observe from Fig. 7a that the average fairness f_t^e among all the served UEs achieved by the MAT, CIRCLE and RANDOM regimes increases with the increase of the number of TSs, as expected. Specifically, MAT increases from 0.53 to 0.85, while CIRCLE increases from about 0.5 to 0.6. Finally, RANDOM remains under 0.4.

Then, we show the fairness f_t^u of each UAV's UE-load achieved by the MAT, CIRCLE and RANDOM regimes in Fig. 7b. Observe that both MAT and CIRCLE approach the fairness of 1, because both solutions can control the UAVs to serve a similar number of UEs. However, RANDOM can only achieve a fairness of 0.75.

Next, in Fig. 7c, we analyse the energy consumed by UEs. We can see that our proposed MAT achieves the best performance, followed by CIRCLE and RANDOM. This is because after training, MAT assists the UAVs in a cooperative way serving the UEs. Hence, more UEs can offload their tasks to UAVs, which results in reduced energy consumption for all the UEs.

Next, in Fig. 8, we increase the number of UAVs to 4 and evaluate the performance of three compared solutions. One can see from Fig. 8a that the average fairness f_t^e increases with the increase of TSs, as expected. Our proposed MAT can achieve the best performance, reaching at 0.9, whereas the RANDOM performs the worst, which can only achieve about 0.5.

Then, in Fig. 8b, we draw the fairness of each UAV's UE-load f_t^u achieved by MAT, CIRCLE and RANDOM. One sees that MAT outperforms CIRCLE and RANDOM, as expected. CIRCLE performs worse than MAT but has much better performance than RANDOM.

Additionally, we show the performance of energy consumed by UEs in Fig. 8c. Similar with before, one can observe that MAT can always achieve the best performance and help UEs to save the energy consumption, while CIRCLE performs the second, followed by RANDOM. This further proves that with proper training, MAT can control the UAVs to provide better service to UEs.

V. CONCLUSIONS

In this paper, we have proposed a multi-agent deep reinforcement learning based trajectory control algorithm for jointly maximizing the fairness among all the UEs and the fairness of UE-load of each UAV, as well as minimizing the energy consumption of all the UEs by optimizing each UAV's trajectory and offloading decision from all the UEs. Simulation results show that the proposed MAT has the considerable performance gain over the compared benchmark algorithms.

REFERENCES

- [1] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, 2016.
- [2] Q. Wu and R. Zhang, "Common throughput maximization in UAV-enabled OFDMA systems with delay consideration," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6614–6627, Dec 2018.
- [3] L. Kong, L. Ye, F. Wu, M. Tao, G. Chen, and A. V. Vasilakos, "Autonomous relay for millimeter-wave wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2127–2136, Sep. 2017.
- [4] R. Fan, J. Cui, S. Jin, K. Yang, and J. An, "Optimal node placement and resource allocation for UAV relaying network," *IEEE Communications Letters*, vol. 22, no. 4, pp. 808–811, 2018.
- [5] H. Wang, J. Wang, G. Ding, J. Chen, F. Gao, and Z. Han, "Completion time minimization with path planning for fixed-wing UAV communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 7, pp. 3485–3499, 2019.
- [6] F. Cui, Y. Cai, Z. Qin, M. Zhao, and G. Y. Li, "Multiple access for mobile-UAV enabled networks: Joint trajectory design and resource allocation," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 4980–4994, 2019.
- [7] J. Lyu, Y. Zeng, and R. Zhang, "UAV-aided offloading for cellular hotspot," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3988–4001, 2018.
- [8] J. Xu, Y. Zeng, and R. Zhang, "UAV-enabled wireless power transfer: Trajectory design and energy optimization," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5092–5106, Aug 2018.
- [9] T. Q. Duong, L. D. Nguyen, H. D. Tuan, and L. Hanzo, "Learning-aided realtime performance optimisation of cognitive UAV-assisted disaster communication," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.
- [10] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7957–7969, 2019.
- [11] 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, April 2019.
- [12] 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, Dec 2014.
- [13] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3949–3963, June 2016.
- [14] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Transactions on Communications*, vol. 64, no. 12, pp. 4983–4996, 2016.
- [15] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, March 2018.
- [16] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 434–437, Aug 2017.
- [17] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.
- [18] K. Wang, P. Huang, K. Yang, C. Pan, and J. Wang, "Unified offloading decision making and resource allocation in ME-RAN," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8159–8172, Aug 2019.
- [19] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.
- [20] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, "Energy efficient resource allocation in UAV-enabled mobile edge computing networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 9, pp. 4576–4589, 2019.
- [21] Y. Zhou, C. Pan, P. L. Yeoh, K. Wang, M. ElKashlan, B. Vucetic, and Y. Li, "Secure communications for UAV-enabled mobile edge computing systems," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 376–388, 2020.
- [22] N. H. Motlagh, M. Baggaa, and T. Taleb, "UAV-based IoT platform: A crowd surveillance use case," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 128–134, 2017.
- [23] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2049–2063, March 2018.
- [24] M. Hua, Y. Wang, Q. Wu, H. Dai, Y. Huang, and L. Yang, "Energy-efficient cooperative secure transmission in multi-UAV-enabled wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7761–7775, 2019.
- [25] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Communications Surveys Tutorials*, pp. 1–46, 2020.
- [26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [27] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 2, no. 4.
- [28] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.
- [29] 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, p. 529, 2015.
- [30] J. Wang, C. Jiang, K. Zhang, X. Hou, Y. Ren, and Y. Qian, "Distributed Q-learning aided heterogeneous network association for energy-efficient IIoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2756–2764, April 2020.
- [31] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Thirtieth AAAI conference on artificial intelligence*, 2016.
- [32] 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.
- [33] L. Bu, R. Babu, B. De Schutter *et al.*, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.
- [34] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6382a–6393.
- [35] K. Wang, K. Yang, and C. S. Magurawalage, "Joint energy minimization and resource allocation in C-RAN with mobile cloud," *IEEE Transactions on Cloud Computing*, vol. 6, no. 3, pp. 760–770, July 2018.
- [36] L. Yang, J. Cao, Y. Yuan, T. Li, A. Han, and A. Chan, "A framework for partitioning and execution of data stream applications in mobile cloud computing," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 4, pp. 23–32, 2013.
- [37] H. He, S. Zhang, Y. Zeng, and R. Zhang, "Joint altitude and beamwidth optimization for UAV-enabled multiuser communications," *IEEE Communications Letters*, vol. 22, no. 2, pp. 344–347, 2017.
- [38] X. Lyu, H. Tian, W. Ni, Y. Zhang, P. Zhang, and R. P. Liu, "Efficient admission of delay-sensitive tasks for mobile edge computing," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2603–2616, June 2018.
- [39] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.
- [40] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14, 2014.
- [41] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.
- [42] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.
- [43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.