Reinforcement Learning Optimized Throughput for 5G Enhanced Swarm UAS Networking

Jian Wang, Yongxin Liu, Shuteng Niu and Houbing Song
Embry-Riddle Aeronautical University, Daytona Beach, FL 32114 USA
Email: {WANGJ14, LIUY11, SHUTENGN}@my.erau.edu, h.song@ieee.org

***Abstract*—The ubiquitous of 5G New Radio (5G NR) accelerates the massive implementations in many fields including swarm Unmanned Aircraft System (UAS) networking. The ultra capacities of 5G NR can provide more sufficient networking services for the swarm UAS networking which can enable swarm UAS to deploy in more complex and challenging scenarios to achieve missions. However, the conventional swarm UAS networking are mainly centralized or hierarchical which is vulnerable to the dynamics and the deployment of swarm UAS networking on a large scale. In this paper, we formulate a cell wall communications for the heterogeneous swarm UAS networking with the inspiration of biological cell wall communication. Fueled by reinforcement learning, we resolve the edge-coloring problem of cell wall communication scheduling to achieve the maximum throughput between the heterogeneous swarm UAS networking globally. The evaluation shows our proposed reinforcement learning enabled algorithm can surpass the conventional scheduling algorithms over 90% when the time piece is less than $0.01s$ and achieve the optimal throughput for the heterogeneous swarm UAS networking.**

***Index Terms*—reinforcement learning, throughput optimization, swarm UAS networking, 5G new radio, edge-coloring**

I. Introduction

The compact and the affordable of next generation NodeB in 5G New Radio (5G NR) stimulus the deployment of 5G NR networking on a large scale. The lightweight and energy-saving of 5G NR devices lead the 5G NR networking spread from Mobile Ad-hoc Networking (MANET) to Vehicle Ad-hoc Networking (VANET) and Flying Ad-hoc Networking (FANET) [1]. As the mutual effects, FANET also extends the 5G NR networking to a large scale with more flexibility and mobility. The 5G NR enabled swarm Unmanned Aircraft System (UAS) networking can achieve more sufficient networking services from the inter and the intra networking which can provide the swarm UAS networking more capacities to finish the complex missions which have high requirements of collaborations and corporations simultaneously and sequentially [2], [3]. With ultra-high frequencies carriers, 5G NR can provide ultra wideband wireless communication for swarm UAS networking with the sacrifice of transmission range and energy divergence which attracts many researchers to make effort to amend the trade-off to extend the reliance of 5G NR enabled swarm UAS networking. With reliable wideband wireless communication, the feasible throughput of swarm UAS networking can assure the quality of services for swarm UAS which is critical to the perception and complement of mission from remote terminals in real time. The instantaneous feedback and intrusion loops between remote terminals and swarm UAS networking on the light is essential to the complex mission complement.

As the predominant stimulus for the evolution in many fields, machine learning fueling everything trends is prevailing and obtains remarkable achievement. As one main branch of machine learning, reinforcement learning is playing a pivotal role in enhancing the swarm UAS networking. The advantages of learning from the environment and evolving itself with interaction with external feedback make reinforcement learning can achieve more robust performance (delay, bandwidth, and throughput) for the swarm UAS networking [4]. Concurrently, the distributed deployment, the flexibility of interaction, and the convenience of transfer learning from sophisticated models of reinforcement learning can enable the swarm UAS networking to achieve sufficient packet delivery underneath the dynamics of swarm UAS networking. The experience exchanging between different UAS also extends the local optimal throughput of swarm UAS networking to global optimization.

The conventional reinforcement learning enabled throughput optimization focus on the central or hierarchical architecture of swarm UAS networking. The main optimal factors are trajectory, deployment, and coverage of communication which can enhance the swarm UAS networking temporarily in some specific scenarios and can not deploy the optimization on a large scale. Most optimizations will collapse as the amount of swarm UAS networking is over some specific scale. However, the amount of swarm UAS networking can decide the complexity of mission complement and quality, as well as the collaborations and cooperation between heterogeneous swarm UAS networking.

Different from the previous work on swarm UAS networking, in this paper, we focus on the communication between heterogeneous swarm UAS networking in bio-inspired behaviors, cell wall communication. The cell wall communication is not dependent on the specific UAS on the swarm UAS networking to afford gateway services for other peers which can be flexible and elastic to the dynamics of swarm UAS networking on a flight. To achieve the maximum throughput between the heterogeneous swarm UAS networking, we formulate the routing scheduling into an edge-coloring problem which is an NP hard problem. With reinforcement learning enhancement, we resolve the edge-coloring problem with the minimum colors to extend the utility of each link between the cell

wall. The evaluation shows that the reinforcement learning enabled approach can improve the throughput of over 90% when the time piece is less than 0.01s and elasticity of swarm UAS networking significantly. Simultaneously, the DQN can achieve more throughput as the number of installed beams in UAS rises.

The paper is organized as follows: Section II illustrates the related work of swarm UAS networking. Section III describes the methodology of our proposed approach. Section IV presents the evaluation of the proposed methodology. Section V concludes the paper.

II. Related Work

The ubiquitous implementations of swarm UAS require reliable, efficient, and high performance UAS networking to provide sufficient networking services. Fueled by reinforcement learning, swarm UAS networking is being capable of implementation on a large scale which accelerates the evolution of reinforcement learning enabled approaches.

Due to the advantages of learning from interaction with the environment, reinforcement learning can assure the robust optimization for swarm UAS networking. Comparable with the distributed characteristics of swarm UAS networking, Multi-Agent Reinforcement Learning (MARL) is adopted in many research to improve the throughput of swarm UAS networking. Each UAS as an agent is formulated to MARL to achieve the local optimization of throughput and coverage based on its observations [5], and maximization of long-term rewards. In [6], MARL is implemented to optimize the movement of UAS with off-line exploration and on-line learning. The joint optimization combining off-line and on-line movements will be generated to achieve global optimization of throughput for UAS networking and cellular networking. Similarly, MARL optimizes the UAS's path and time resource allocation to the ground IoT devices jointly to achieve the maximum throughput between UAS and IoT devices. The policy is rewarded as the minimum throughput to reinforce each agent to achieve local optimization [7]. A long-term resource allocation is formulated to achieve maximum throughput for UAS networking and optimized by MARL. The evaluation shows a good trade-off strike between the throughput gain and the information exchange overhead [8]. The trajectory and power allocation are critical to the performance of swarm UAS networking. In [9], MARL enhances the trajectory of UAS and power allocation of UAS networking in the mission to achieve the maximization of throughput between UAS and ground users. The evaluation shows the networking utility and system overhead can be optimized jointly. With the feedback from primary users and UAS fusion nodes, a distributed reinforcement learning approach is proposed in [10] which aims to improve the throughput allocation to improve the utility of the whole system and mitigate the security threat with congestion of spectrum and hopping of frequencies. However, the optimization fluctuates in the performance significantly and is lack of robust.

The conventional approaches of reinforcement learning are to deploy Q-learning or Deep Q-learning Network (DQN) into swarm UAS networking and achieve the local optimization and global optimization for the throughput of swarm UAS networking. In [11], DQN aims to reduce the time consumption of deployment of UAS on the networking relaying and to obtain seamless video offloading services. A liquid Q-learning is implemented to predict the users' requests and states, and deploy corresponding UAS with specific content to aid the throughput of users [12]. Based on the real-time interaction and historical data, Q-learning amends the policy of relay power to reduce the bit error rate and energy consumption underneath random jamming attacks [13]. With services providing seamless cellular networking for vehicles on highways, actor-critic algorithms learn the vehicle movement environment to obtain the knowledge of the signal coverage and the dynamics of vehicles to handle the continuous action space [14]. To extend the range of 5G NR, the swarm UAS with 5G network slice extension can enable the extension of 5G accessing scale to the other swarm UAS networking on the flight. The extension can provide computing aid for the swarm UAS networking which is under job offloading and minimizes the energy consumption and queuing delay [15]. The private base station can only provide services to the specific UAS which is a lack of coordination between different base stations. The throughput between the base stations and the UAS can not satisfy the requirement of the massive heterogeneous swarm UAS networking. A two-level architecture is proposed to optimize the base stations' behaviors and to achieve the long term payoffs. The self-interested and independent behaviors are deployed on the lower level of purchasing the noncooperation subgames, and the cooperative game is implemented to obtain the global optimizations [16]. To improve Multiple Input and Multiple Output (MIMO) throughput, DQN learns the environment policies through interacting with the external networking which can achieve the 20% improvement of throughput [17].

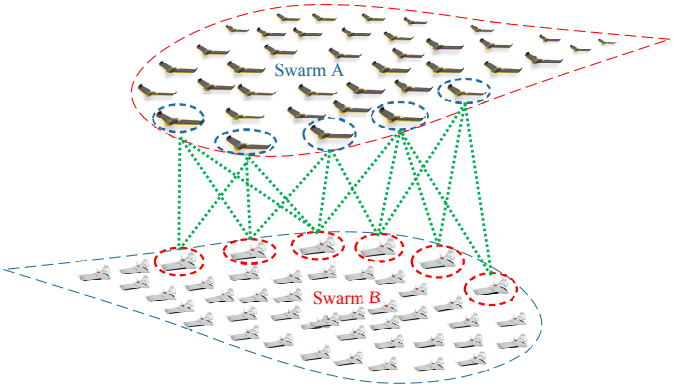

Fig. 1. Communication between heterogeneous swarm UAS networking

III. System Modeling

To achieve a complex mission complement, there needs several swarm UAS that play different roles in the processing of different mission execution stages in some scenarios. The conventional communication between heterogeneous swarm

UAS networking is hierarchical and central which needs specific UAS, as gateways, to provide packet delivery services. The hierarchical and central architectures of communication are lack flexibility, elasticity, and reliance on the dynamics of swarm UAS networking and can not be deployed on a large scale. The cell wall communication between heterogeneous swarm UAS networking is decentralized and dynamic. As Fig. 1 depicts, swarm A and swarm B are two heterogeneous swarm UAS networking which contains $\mathcal{N}$ and $\mathcal{M}$ UAS in each swarm respectively and need packet delivery for collaborations and cooperation. To achieve a feasible communication volume, swarm A and swarm B select qualified UAS (marked in dash line circles) in each peer to provide packet delivery services. The mechanism of selection for the qualified UASs are mainly dependent on the estimation from Automatic Dependent Surveillance-Broadcast (ADS-B). The UASs are in the range of communication for swarm UAS networking can be qualified. Due to the dynamics of each swarm UAS networking, the qualified UAS are variable to achieve stable and sufficient connections between heterogeneous swarm UAS networking.

In this scenario, each UAS is equipped with a compact and energy saving mmWave devices that can generate H mmWave radio frequency (RF) beams for connections. Along with Time Division Multiple Access (TDMA), each UAS in the cell wall can provide gateway services for the inter swarm UAS networking. We define that there are n UAS in swarm A and m UAS of swarm B to construct the cell wall for communication between swarm A and swarm B. Here, $A = \{A_n | n \leq \mathcal{N}\}$ and $B = \{B_m | m \leq \mathcal{M}\}$.

For a convenience, we simplify the cell wall from swarm UAS networking which is depicted as Fig. 2. The simplification of the cell wall can be mapped into a directed graph $G = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Here, $\mathcal{V} = A \cup B$, $\mathcal{V}$ is the union of swarm A and swarm B. $\mathcal{E}$ denotes the connections between swarm A and swarm B. $c_\mathcal{E}$ denotes the throughput capacity of $\mathcal{E}$. With the radio transmission, $c_\mathcal{E}$ is formulated as $c_\mathcal{E} = g_\mathcal{E} log(1 + \frac{p_\mathcal{E}}{10^{PL_\mathcal{E}/\sigma^2}})$ and $c_\mathcal{E} > 0$. Here, $g_\mathcal{E}$ denotes the direct gain between linked vertexes. $p_\mathcal{E}$ denotes the transmission power from vertex $\mathcal{V}$ and σ is Gaussian noise distributed with zero mean. $PL_\mathcal{E}$ is the path loss for beam transmission in line of sight which is formulated in: $PL_\mathcal{E} = 20log(d_\mathcal{E}) + 20log(f_\mathcal{E}) - 147.55$ where $d_\mathcal{E}$ and $f_\mathcal{E}$ denote the corresponding distance and frequencies of $\mathcal{E}$ respectively.

The packet delivery duration for the cell wall is definite which denotes $\mathcal{T}$ and obtains $\mathcal{F}$ frames for sequential connections. Simultaneously, each $\mathcal{F}$ contains multiple time unit time that can be divided into N time slots t, where $N \geq 1$. The i^{th} slot t_i is subjected to: $\sum_{i=1}^{N} t_i = 1$. Here, $0 \leq t_i \leq 1$. In t_i, a set of connections $\mathcal{E}_i \subseteq \mathcal{E}$ are active for packet delivery.

IV. Reinforcement Learning Enabled Throughput Optimization for Cell Wall Communication

With constructed cell between swarm A and swarm B, we assume the topology of the cell wall keeps stable in each frame

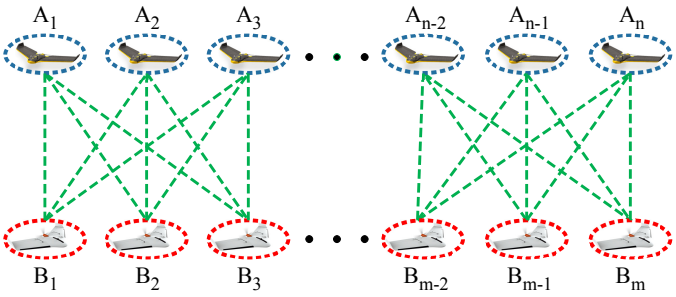

Fig. 2. Cell wall communication between heterogeneous swarm UAS networking

$\mathcal{F}$ which also means that the qualified UAS in the cell wall of swarm A and swarm B stay unchanged in each time slot t_i. During each time slot, the feasible throughput for the each edge is formulated into $c_{\mathcal{E}_f} = t_i c_\mathcal{E}$. With the avoidance of collision between different beams, the more time slot the edge $\mathcal{E}$ occupies, the more feasible throughput will the cell wall achieve. The whole optimization can be given:

$$\max_{\mathcal{V}} \sum_{i=1}^{N} \sum_{j=1}^{\mathcal{E}} t_i \cdot s_i \cdot c_j \tag{1a}$$

$$subject\ to \sum_{i=1}^{N} t_i \leq 1, \forall t_i \geq 0 \tag{1b}$$

Where, s_i is the selection of link $\mathcal{E}$ in i^{th} slot t_i with throughput capacity s_j, $s_i \in \{0, 1\}$.

To achieve the maximum throughput for the cell wall communication, we schedule different pieces in each time slot that activates different links in each piece. Here, we install a link time t^g for $\lceil \frac{t_i}{t^g} \rceil - 1$ links and the left time $mod(t_i, t^g)$ (mod is the modulo operation) is assigned for the last links. We perform edge-coloring on the cell wall networking to achieve the maximum throughput for the swarm A and swarm B. We assume there are just α colors that are sufficient for G in the time slot t_i. Each color is corresponding to one set of active links in time duration t^g expect from the last one with time assignment of $mod(t_i, t^g)$. In the time slot t_i, the maximum utility of link active time is given:

$$\alpha t^g = (\lceil \frac{t_i}{t^g} \rceil - 1) + mod(t_i, t^g) \tag{2a}$$

$$< (\lceil \frac{t_i}{t^g} \rceil - 1) + t^g \tag{2b}$$

$$\leq \lceil t_i \rceil \leq 1 \tag{2c}$$

The proof (2) shows the whole links can be scheduled into the t_i, and we do not need extra scaling for the αt^g to fit the time slot t_i. Combined with (1), we can rend the maximum throughput for edge-coloring scheduling which is given:

$$\max_{\mathcal{E}} \sum_{i=1}^{N} (\sum_{j=1}^{\alpha-1} t_j^g \cdot \sum_{i \in \mathcal{E}_j} s_i \cdot c_i + \sum_{i \in \mathcal{E}_\alpha} mod(t_i, t^g) \cdot s_i \cdot c_i) \tag{3a}$$

$$subject\ to\ (1b) \tag{3b}$$

Based on the optimization of the throughput between cell wall communication, the less color we use for the edge-coloring, the more throughput we will achieve. However, an edge-coloring problem for a graph is NP hard problem which has been proofed in [18]. Theoretically, the maximum colors of edge-coloring for G is $3\lceil\triangle(G)/2\rceil$. Here, $\triangle(G)$ is the maximum node degree of G. Based on the proof in [18], the $\triangle(G)$ is given:

$$\triangle(G) = \sum_{i=1}^{N}\lceil\frac{t_i}{t^g}\rceil < \sum_{i=1}^{N}(\frac{t_i}{t^g}+1) \tag{4a}$$

$$\leq \frac{t^g(m+n-1)+1}{t^g} \tag{4b}$$

Based on the rendering of $\triangle(G)$, we can have α for edge coloring of G.

$$\alpha = 3\cdot\lceil\frac{t^g(m+n-1)+1}{2\cdot t^g}\rceil \tag{5}$$

Theoretically, we have α options for the graph G to fill all the edges without collisions. The conventional resolution to the edge-coloring is a NP hard problem. Here, we leverage the Q learning approaches to solve the edge-coloring of G. Generally, the fewer colors the agents adopt, the more rewards they will get. The updating processing of agent is given:

$$\begin{aligned} Q^{t+1}(s,a) = Q^t(s,a) + \phi\{r^t - Q^t(s,a) \\ + \max_{a'\in a} Q^t(s',a')\} \end{aligned} \tag{6}$$

Where ϕ is the learning rate, $\phi \in (0,1)$, s is the current state observed by agent, s' is the next state predicted by the agent. a and a' are the current action adopted and the next action predicted by agents respectively. Q^t is what the value the agent can achieve in the current sequential time t and the r^t is the cumulative rewards the agent achieved. The whole processing is to maximize the Q-value at each step.

We incorporate Deep Neural Network (DNN) into the framework of Q-learning to achieve the edge-coloring problem resolving. The Q-value updating function from (6) can be modified with weight derived from DNN: $Q^t(s,a;\theta)$. The DNN can be trained to minimize the loss function $L^t(\theta)$ which is given:

$$L^t(\theta) = E_{(s,a;\theta)}[(r^t + \max_a Q^{t+1}(s',a;\theta^-) - Q^t(s,a;\theta))^2] \tag{7}$$

Here, θ is the parameter networking for the Q-network at iteration t and θ^- is the parameters at iteration $t+1$. θ^- is fixed when optimizing the $L^t(\theta)$ at t. By differentiating $L^t(\theta)$, we can have:

$$\begin{aligned} \nabla_{\theta^-}L^t\theta = E_{(s,a;\theta)}[(\max_a Q^{t+1}(s',a;\theta^-) - Q^t(s,a;\theta) \\ + r^t)^2\nabla_{\theta^-}Q(s,a,\theta^-)] \end{aligned} \tag{8}$$

To bridge the connection between the cell wall networking and DQN, we need to decompose the edge-coloring problem into modeling which can interact with DQN. With interaction with the modeling of the cell wall, the DQN can derive optimal operations for the edge-coloring resolving. The optimization can be given:

$$\max\frac{1}{\beta} \tag{9a}$$

$$subject\ to\ (1b) \tag{9b}$$

$$\beta \leq \alpha \tag{9c}$$

$$\mathcal{E}_\lambda^{t^g} \cap \mathcal{E}_\mu^{t^g} = \emptyset, \forall\lambda,\mu\in\mathcal{E} \tag{9d}$$

$$\sum_{\varphi=1}^{\beta} t_\varphi^g \leq t_i, t_{_\varphi}^g > 0 \tag{9e}$$

$$\sum_{i=1}^{N}\sum_{\varphi=1}^{\beta} t_\varphi^g \leq 1, t_{_\varphi}^g > 0 \tag{9f}$$

Here, β is the colors adopted by DQN, which is not bigger than the theoretical colors derived from [18], α. $\mathcal{E}_\lambda^{t^g} = \{\mathcal{V}_{s\lambda}^{t^g}, \mathcal{V}_{t\lambda}^{t^g}\}$. $\mathcal{V}_{s\lambda}^{t^g}$ and $\mathcal{V}_{t\lambda}^{t^g}$ denotes the endpoints of connection $\mathcal{E}_\lambda^{t^g}$ at time piece t^g. Simultaneously, $\mathcal{E}_\mu^{t^g} = \{\mathcal{V}_{s\mu}^{t^g}, \mathcal{V}_{t\mu}^{t^g}\}$. $\mathcal{V}_{s\mu}^{t^g}$ and $\mathcal{V}_{t\mu}^{t^g}$ denotes the endpoints of connection $\mathcal{E}_\mu^{t^g}$ at time piece t^g. $\mathcal{E}_\mu^{t^g}$ and $\mathcal{E}_\lambda^{t^g}$ are two connections active at time piece t^g. Correspondingly, all the active scheduling in the time slot t_i can not be over the allocation of t_i.

The action space for agent is the coloring $a_s \in \{1,2,...,\beta\}$, and the state s of the environment can be denoted as the edge $\mathcal{E}^{t^g}$ in time piece t^g. Thus, the reward function of action can be given:

$$r^t = \begin{cases} \frac{1}{a_k},\ (9d)\ (9e)\ (9f) \\ 0,\ otherwise \end{cases} \tag{10}$$

In (10), the reward of action is based on the counts of colors which input the colored G to detect the state of G and output the desired operations to maximize the reward. Each successful operation can be reward as $\frac{1}{a}$. The total reward for one iteration can be given:

$$\max_{\mathcal{E}}\sum_{s=1}^{\mathcal{E}}\frac{1}{a_k} \tag{11}$$

Here, $\frac{1}{a_k}$ denotes the average reward of edge k in the i^{th} time slot. To achieve the successful coloring for G, the state keeps unchanged if the subjection is broken.

We combined (6) and (11), and the updating processing agent can be given:

$$\begin{aligned} Q^{t+1}(s,a) = \frac{1}{t-t_0}\cdot\sum_{t_0}^{t}\frac{1}{a_k} + \ \phi\{\frac{1}{a_k} - \frac{1}{t-t_0}\cdot\sum_{t_0}^{t}\frac{1}{a_k} \\ + \max_{a'\in a}(\frac{1}{a_k'}|\ s')\} \end{aligned} \tag{12}$$

The detailed pseudocode is shown as Algorithm 1. We integrate DQN updating processing into the max throughput of σ for the cell wall.

The above optimization shows the single active link for UAS in time piece t^g in the cell wall. With different active frequencies, the UAS can generate multiple beams at the same time to deliver packages to the multiple UAS simultaneously.

Algorithm 1: Render the max throughput σ for cell wall

Initial $G = (\mathcal{V}, \mathcal{E})$ for Initial Schedule $\mathcal{S}$
setting t^g;
Calculate $\triangle(G)' \leftarrow \frac{t^g(m+n-1)+1}{t^g}$;
Calculate $a = 3\lceil \triangle(G)/2 \rceil$;
Initial random $a_k^{t_0}$; Initial s^{t_0};
$Q^{t_0}(a^{t_0}, s^{t_0}) \leftarrow \frac{1}{a_k^{t_0}}$;
$\phi = \frac{1}{t^\omega}$
while $\nabla_{\theta^-} L^t\theta \ > \ 0$ **do**
  Input G^t into DNN ;
  $s^t \leftarrow$ DNN;
  if $\mathcal{E}_\lambda^{t^g} \cap \mathcal{E}_\mu^{t^g} \neq \emptyset$ **then**
    Break;
  else
    $r^t \leftarrow \frac{1}{a_k}$;
    $\max_{a' \in a} Q^t(s', a') \leftarrow \max(\frac{1}{a_k'} | \ s^t)\}$;
    $Q^t(s, a) \leftarrow \frac{1}{t-t_0} \cdot \sum_{t_0}^{t} \frac{1}{a_k}$;
    $Q^{t+1}(s, a) \leftarrow Q^t(s, a) + \phi\{r^t - Q^t(s, a) + \max_{a' \in a} Q^t(s', a')\}$;
    $L^t(\theta) \leftarrow (r^t + \max_a Q^{t+1}(s', a; \theta^-) - Q^t(s, a; \theta))$;
    $\nabla_{\theta^-} L^t\theta \leftarrow L^t(\theta)$;
if $\nabla_{\theta^-} L^t\theta \ = \ 0$ **then**
  $\sigma \leftarrow \sum_{i=1}^{N} \sum_{j=1}^{\mathcal{E}} t^g \cdot s_i \cdot c_j$;

TABLE I
CELL WALL CONFIGURATION

Transmission Power, p	20 dBm
Distances between centers of S_1 and $S2$, d	100 m
Direct gain, g	30 dB
Carrier frequency, f	28 GHz
Noise power, N_0/B	-174 dBm/Hz
Bandwidth	1 GHz
Minimum SINR threshold	-5 dB

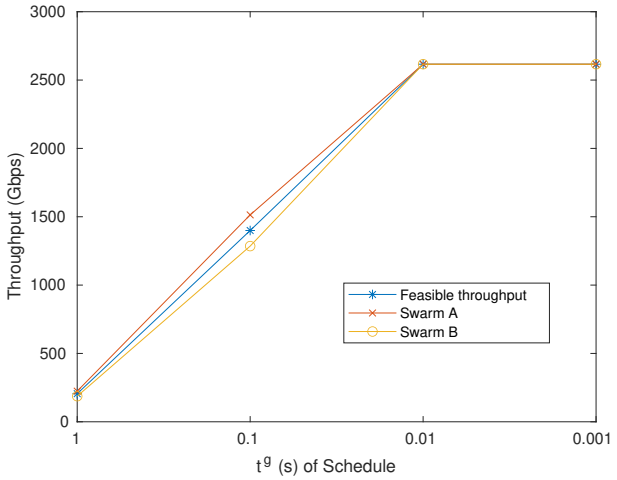

Fig. 3. Scheduled throughput between swarm A and swarm B

Here, we will loose the constraint of UAS in (1b) to Γ which means there are Γ beams active in different frequencies. The optimization of (1) can be modified to:

$$\max_{k \in \mathcal{V}} \sum_{k=1}^{\mathcal{N}} \sum_{j=1}^{\mathcal{E}} t_k \cdot s_k \cdot c_j \tag{13a}$$

$$subject\ to \sum_{k=1}^{\mathcal{N}} t_k \leq \Gamma, t_k \geq 0, \tag{13b}$$

Comparable with the modification of (13), (2) can be given:

$$\alpha t^g = (\lceil \frac{t_i}{t^g} \rceil - 1) + mod(t_i, t^g) \tag{14a}$$

$$< (\lceil \frac{t_i}{t^g} \rceil - 1) + t^g \tag{14b}$$

$$\leq \lceil t_i \rceil \leq \Gamma \tag{14c}$$

With the consideration of minimum interference between frequencies, we keep α unchanged to achieve the stability of the cell wall communication and maximum throughput between heterogeneous swarm UAS networking.

V. Evaluation

In this part, we will evaluate the throughput optimization for cell wall which is critical to the performance of swarm UAS networking. The configuration is shown as the TABLE I. With 5G NR, each UAS is installed with a mobile beamforming device which can generate mmWave beams over the carrier of 28 GHz. To approximate the practical dynamics of swarm UAS, we scatter the cell wall distances between heterogeneous swarm UAS range from 15 m to 100 m with the distribution of Poisson ($\lambda = 20$) randomly. There are 30000 UAS for each swarm (swarm A and swarm B) scattering randomly in the space: 50 $m \times 200\ m \times 50\ m$ with the spatial constraints: $\frac{x^2}{(a-b)x+ab} + \frac{y^2+z^2}{c^2} = 1$. Here, $a = 20\ m$, $b = 30\ m$, $c = 25\ m$.

Fig. 3 shows the edge-coloring based scheduling solved by the Karloff algorithm for one beam installed on the swarm UAS. We compared the throughput from swarm A and swarm B and calculated the feasible throughput between swarm A and swarm B with consideration of collisions between beams occurring. The result shows that the two heterogeneous swarm UAS networking can obtain more throughput with the reduction of t^g from $1s$ to $0.01s$. The beam utility reaches the maximum when t^g is less than 0.01 which keeps steady for the setting from $0.01s$ to $0.001s$.

Fig. 4 shows the throughput and normalized improvement of the cell wall for DQN. With trained by the best result, the DQN can achieve better performance over the Karloff algorithm than 90% when t^g is set in the range of $0.01s$ to $0.001s$. There are still some collisions for DQN based edge coloring resolving when $t^g = 1s$. Simultaneously, the DQN can achieve more throughput as the number of installed beams in UAS rises.

Fig 5 shows the training processing of DQN in the edge-coloring resolving. As the episode number increases the av-

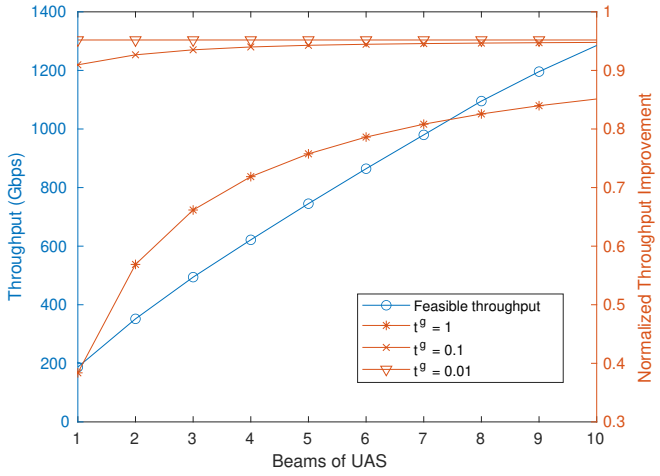

Fig. 4. Throughput and normalized improvement of cell wall

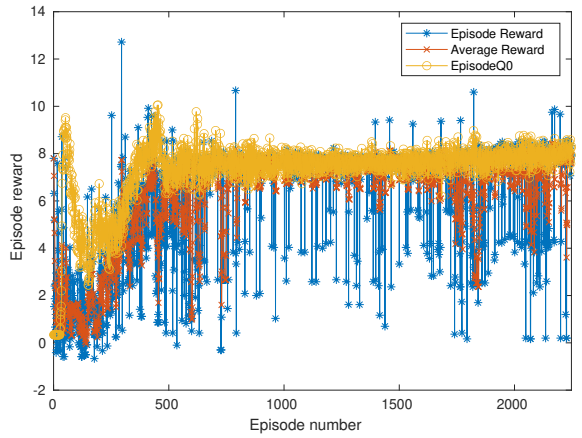

Fig. 5. DQN training processing

erage reward and episode reward are becoming convergence and the total reward can reach 1714.95. With colored edges for cell wall in multiple beams of UAS, the whole throughput can be improved significantly.

VI. Conclusion

In this paper, we resolve the edge-coloring problem of cell wall communication scheduling to achieve the maximum throughput between the heterogeneous swarm UAS networking with DQN. The evaluation shows our algorithm can surpass the conventional scheduling algorithms over 90% when the time piece is less than $0.01s$ and achieve the optimal throughput. As the beams installed on UAS increases, the whole throughput between cell wall of swarm A and swarm B can be enhanced remarkably. In the near future, we will explore the efficiency of the reinforcement learning with variable training frameworks to achieve more flexibility and elasticity of cell wall communication for heterogeneous swarm UAS networking.

Acknowledgement

This work was supported in part by the National Science Foundation under Grant No. 1956193.

References

[1] J. Hu, H. Zhang, and L. Song, "Reinforcement learning for decentralized trajectory design in cellular uav networks with sense-and-send protocol," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6177–6189, 2019.

[2] H. Cao, S. Wu, Y. Hu, G. S. Augla, and L. Yang, "Virtual resource allocation for tactile and flexible services in uavs-integrated 5g networks," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[3] X. Yue, Y. Liu, J. Wang, H. Song, and H. Cao, "Software defined radio and wireless acoustic networking for amateur drone surveillance," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 90–97, 2018.

[4] W. Guo, "Partially explainable big data driven deep reinforcement learning for green 5g uav," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.

[5] J. Cui, Y. Liu, and A. Nallanathan, "The application of multi-agent reinforcement learning in uav networks," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019, pp. 1–6.

[6] S. E. Hammami, H. Afifi, H. Moungla, and A. Kamel, "Drone-assisted cellular networks: A multi-agent reinforcement learning approach," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

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

[8] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for uav networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 729–743, 2020.

[9] N. Zhao, Z. Liu, and Y. Cheng, "Multi-agent deep reinforcement learning for trajectory design and power allocation in multi-uav networks," *IEEE Access*, vol. 8, pp. 139 670–139 679, 2020.

[10] A. Shamsoshoara, M. Khaledi, F. Afghah, A. Razi, and J. Ashdown, "Distributed cooperative spectrum sharing in uav networks using multi-agent reinforcement learning," in *2019 16th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2019, pp. 1–6.

[11] K. Zheng, Y. Sun, Z. Lin, and Y. Tang, "Uav-assisted online video downloading in vehicular networks: A reinforcement learning approach," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[12] M. Chen, W. Saad, and C. Yin, "Liquid state machine learning for resource and cache management in lte-u unmanned aerial vehicle (uav) networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1504–1517, 2019.

[13] W. Wang, X. Lu, S. Liu, L. Xiao, and B. Yang, "Energy efficient relay in uav networks against jamming: A reinforcement learning based approach," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[14] 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*, pp. 1–1, 2020.

[15] G. Faraci, C. Grasso, and G. Schembra, "Reinforcement-learning for management of a 5g network slice extension with uavs," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 732–737.

[16] A. Asheralieva and D. Niyato, "Hierarchical game-theoretic and reinforcement learning framework for computational offloading in uav-enabled mobile edge computing networks with multiple service providers," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8753–8769, 2019.

[17] N. Nurani Krishnan, E. Torkildson, N. B. Mandayam, D. Raychaudhuri, E. Rantala, and K. Doppler, "Optimizing throughput performance in distributed mimo wi-fi networks using deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 135–150, 2020.

[18] H. J. Karloff and D. B. Shmoys, "Efficient parallel algorithms for edge coloring problems," *Journal of Algorithms*, vol. 8, no. 1, pp. 39 – 52, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0196677487900265