Placement Optimization of Aerial Base Stations with Deep Reinforcement Learning

Jin Qiu, Jiangbin Lyu, *Member, IEEE*, and Liqun Fu, *Senior Member, IEEE*

Abstract—Unmanned aerial vehicles (UAVs) can be utilized as aerial base stations (ABSs) to assist terrestrial infrastructure for keeping wireless connectivity in various emergency scenarios. To maximize the coverage rate of N ground users (GUs) by jointly placing multiple ABSs with limited coverage range is known to be a NP-hard problem with exponential complexity in N. The problem is further complicated when the coverage range becomes irregular due to site-specific blockage (e.g., buildings) on the air-ground channel in the 3-dimensional (3D) space. To tackle this challenging problem, this paper applies the Deep Reinforcement Learning (DRL) method by 1) representing the state by a *coverage bitmap* to capture the spatial correlation of GUs/ABSs, whose dimension and associated neural network complexity is invariant with arbitrarily large N; and 2) designing the action and reward for the DRL agent to effectively learn from the dynamic interactions with the complicated propagation environment represented by a 3D Terrain Map. Specifically, a novel two-level design approach is proposed, consisting of a preliminary design based on the dominant line-of-sight (LoS) channel model, and an advanced design to further refine the ABS positions based on site-specific LoS/non-LoS channel states. The double deep Q-network (DQN) with Prioritized Experience Replay (Prioritized Replay DDQN) algorithm is applied to train the policy of multi-ABS placement decision. Numerical results show that the proposed approach significantly improves the coverage rate in complex environment, compared to the benchmark DQN and K-means algorithms.

I. Introduction

With their high mobility and reducing cost, unmanned aerial vehicles (UAVs) have attracted increasing interests in military and civilian domains in recent years. In particular, integrating UAVs into cellular networks as aerial base stations (ABSs) to assist terrestrial communication infrastructure in various emergency scenarios such as battlefields, disaster scenes and hotspot events, has been regarded as an important and promising technology [1].

One of the key problems in UAV-aided communication is to find applicable placement of ABSs aiming to achieve maximum coverage of ground users (GUs). To maximize the coverage rate of N GUs by jointly placing multiple ABSs with limited coverage range is known to be a NP-hard problem with exponential complexity in N [2]. However, it has still spurred enthusiasm of many researchers in this theme [2]–[8]. The authors in [2] propose a spiral algorithm to place ABSs along a spiral path to cover all GUs with the minimum number of ABSs, which reduces the complexity to polynomial-time. A heuristic K-means clustering algorithm is applied in [3], which finds suitable ABS locations to serve the partitioned GUs. In

The authors are with School of Informatics, Xiamen University, China 361005 (email: qeauty@stu.xmu.edu.cn; {ljb, liqun}@xmu.edu.cn). *Corresponding Author: Jiangbin Lyu*.
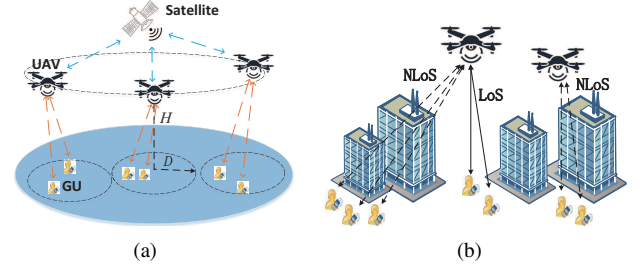
Fig. 1. Placement optimization of ABSs by (a) preliminary design based on dominant-LoS channel and (b) advanced design based on 3D Terrain Map.

terms of maximizing the coverage area, the authors in [4] optimize the altitude of a single ABS based on the probabilistic line-of-sight (LoS) channel model, while circle packing theory is used in [5] to maximize the total coverage area of multiple ABSs. On the other hand, controlling ABS movement to cover moving users is another challenging task [6] [7], for which [6] applies a majority rule to control the direction and distance of UAV displacement towards the cell with the highest user density, while [7] uses the K-means algorithm to partition GUs into clusters, and further applies the Q-learning algorithm for ABS movement. In addition, in terms of ABS coverage and energy consumption trade-off, a Deep Reinforcement Learning (DRL)-based approach is proposed in [8] to achieve energy-efficient and fair communication coverage.

In the aforementioned works, the ABS-GU communication range is determined by a certain signal-to-noise ratio (SNR) threshold, by assuming the air-ground channel to follow the dominant-LoS or probabilistic LoS model [4], thus resulting in uniform coverage range (or disk coverage area). However, the above channel models fail to capture the fine-grained structure of the LoS or non-LoS (NLoS) propagation at specific ABS and GU locations, which in turn critically affects the ABS-GU channel and hence the coverage performance of practical ABS deployment. For example, with slight change of its position, an ABS might transit from LoS to NLoS propagation to the GU due to building edges. Such site-specific LoS/NLoS propagation has been exploited in [9] to find the optimal UAV-relay position for a given pair of ground BS and user. In the paradigm of cellular-connected UAV [10] [11], the LoS/NLoS channel state can be estimated by the UAV on-site [12], or obtained from a given 3D Terrain Map [13], based on which the trajectory of the aerial (UAV) user can be optimized to avoid cellular coverage holes and/or minimize flying distance. However, these works [9] [12] [13] consider a single UAV as an aerial relay or user, with different setup and objective from those in our considered

multi-ABS/multi-GU coverage problem.

Due to the site-specific propagation in the 3-dimensional (3D) space, the LoS/NLoS channel states for all pairs of possible ABS-GU locations in a given environment ensemble an enormous and irregular state space, which cannot be readily handled by conventional optimization methods to achieve maximum coverage rate, especially when the number of ABSs/GUs is large. To tackle this challenging problem, we propose a novel two-level design approach, consisting of a preliminary design based on the dominant-LoS channel model, and an advanced design to further refine the ABS positions based on the 3D Terrain Map, as shown in Fig. 1. For each design, we apply the state-of-the-art double deep Q-network (DQN) with Prioritized Experience Replay (Prioritized Replay DDQN) method, with tailored incorporation of the domain knowledge, by 1) representing the state by a *coverage bitmap* to capture the spatial correlation of GU/ABS locations, which is well fit as the input of the underlying deep neural network (DNN), whose dimension and associated DNN complexity is invariant with arbitrarily large N; and 2) designing the action and reward for the DRL agent to effectively learn from the dynamic interactions with the complicated propagation environments. Numerical results show that the proposed approach significantly improves the coverage rate of GUs compared to the benchmark DQN and K-means algorithms. Moreover, the advanced design further improves the accuracy of GU coverage over the preliminary design, by exploiting the fine-grained structure of the complex propagation environment.

II. SYSTEM MODEL

Consider a UAV-aided communication system with M UAV-mounted ABSs to serve a group of N GUs with given locations denoted by $\mathbf{w}_n \in \mathbb{R}^2$, $n \in \mathcal{N} \triangleq \{1, \cdots, N\}$. Consider downlink communication from ABSs to GUs, while the proposed approach can be similarly applied to uplink communication. Assume that the UAVs fly at a fixed altitude H meters (m), with horizontal locations denoted by $\mathbf{u}_m \in \mathbb{R}^2$, $m \in \mathcal{M} \triangleq \{1, \cdots, M\}$. To focus on the coverage performance, we assume for simplicity that the available spectrum is equally divided into N orthogonal channels, each allocated to one GU, and thus there is no intra or inter-cell interference. Next, we introduce the channel models and coverage criteria for ABS-GU communications.

A. Dominant-LoS Channel Model

Due to the high altitude of the UAV, LoS channel exists with a high probability for practical ABS-GU links [14]. In the preliminary design without site-specific information, we assume the dominant-LoS channel model for the ABS-GU communication. As a result, the channel power gain between ABS m and GU n is given by

$$g_{m,n} \triangleq \frac{\beta_0}{d_{m,n}^2 + H^2}, \tag{1}$$

where $\beta_0 = (\frac{4\pi f_c}{c})^{-2}$ denotes the channel power gain at a reference distance of 1 m, with f_c denoting the carrier frequency and

c denoting the speed of light; and $d_{m,n} \triangleq \|\mathbf{u}_m - \mathbf{w}_n\|$ denotes the horizontal distance, with $\|\cdot\|$ denoting the Euclidean norm.

Assume that each ABS or GU is equipped with a single omni-directional antenna with unit gain. Assume that each ABS transmits with power P Watt (W) to its served GU, whose receiver noise power is denoted by σ^2 W. The SNR received by GU n from ABS m is then given by

$$\gamma_{m,n} \triangleq g_{m,n} P / \sigma^2. \tag{2}$$

A GU is said to be *covered* by an ABS, if the received SNR is not smaller than a certain threshold $\bar{\gamma}$, which corresponds to $g_{m,n} \geq \bar{\gamma}\sigma^2/P \triangleq \bar{g}$, with $\bar{g}$ denoting the corresponding threshold of channel power. For the LoS channel model in (1), $\bar{g}$ further corresponds to a distance threshold D (also known as *coverage range*) such that $d_{m,n} \leq D$, which is given by

$$D \triangleq \sqrt{\beta_0/\bar{g} - H^2}, \tag{3}$$

as illustrated in Fig. 1(a). Finally, denote C_n as the *coverage indicator* for GU n, which is given by

$$C_n \triangleq \begin{cases} 1, & \text{if } \min_{m \in \mathcal{M}} d_{m,n} \leq D, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

B. Site-Specific LoS/NLoS Channel Model

Despite the high LoS probability, the air-ground channel could be occasionally obstructed by obstacles, resulting in NLoS propagation. To investigate the large-scale coverage performance, we assume that the small-scale fading effect is averaged out, and thus focus on the dominant LoS and NLoS path-loss components, as in [4]. In the case when the 3D Terrain Map for a specific environment can be obtained, e.g., from geographic information system (GIS), we can extract the LoS/NLoS information for any pair of ABS and GU locations, as shown in Fig. 1(b). Therefore, the channel power gain between ABS m and GU n can be expressed as

$$g_{m,n} \triangleq \begin{cases} g_{\mathrm{L}}(\mathbf{u}_m, \mathbf{w}_n), & \text{without obstacles in between;} \\ g_{\mathrm{NL}}(\mathbf{u}_m, \mathbf{w}_n), & \text{otherwise,} \end{cases} \tag{5}$$

where g_{L} and g_{NL} denote the channel power gains of the LoS and NLoS channels, respectively, whose specific function forms can be referred to the empirical formula in [14]. In this case, the coverage indicator for GU n is given by

$$C_n \triangleq \begin{cases} 1, & \text{if } \max_{m \in \mathcal{M}} g_{m,n} \geq \bar{g}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

III. PROBLEM FORMULATION

Define the *coverage rate* of all GUs as the ratio of GUs covered by at least one of the ABSs, i.e., $\varphi \triangleq \frac{1}{N} \sum_{n \in \mathcal{N}} C_n$. We formulate the placement optimization problem to maximize the coverage rate of GUs with M ABSs, given by

$$\text{(P1):} \max_{\mathbf{u}_m, m \in \mathcal{M}} \varphi \triangleq \frac{1}{N} \sum_{n \in \mathcal{N}} C_n,$$

$$\text{s.t.} \quad C_n \text{ given by (4) or (6), for } n \in \mathcal{N}.$$

For the preliminary design with C_n given by (4), (P1) is a non-convex optimization problem due to the non-convex constraint of $\min_{m \in \mathcal{M}} d_{m,n} \leq D$. In fact, it is shown to be a NP-hard problem [2] in general, with exponential complexity in N. The problem is further complicated in the advanced design with C_n given by (6), where the LoS/NLoS channel states for all pairs of possible ABS-GU locations ensemble an enormous and irregular state space, which cannot be readily handled by conventional optimization methods, especially when the number of ABSs/GUs is large.

To tackle this challenging problem, we apply the state-of-the-art Prioritized Replay DDQN method, with tailored considerations of the domain knowledge, by 1) representing the state by a coverage bitmap to capture the spatial correlation of GU/ABS locations, which is well fit as the input of the underlying DNN, whose dimension and associated DNN complexity is invariant with arbitrarily large N; and 2) designing the action and reward for the DRL agent to effectively learn from the dynamic interactions with the complicated propagation environments. The DRL framework possesses general intelligence to solve complex problems, which is able to handle with the large and complicated state space involved in the problem (P1) and solve it effectively.

IV. PLACEMENT OPTIMIZATION WITH DRL

A. DRL Algorithm

This subsection gives a brief introduction on the DRL algorithm before presenting the proposed design. In general, DRL is the combination of DNN and RL. Specially, RL refers to the process in which an agent interacts with the environment and makes a series of decisions by using Markov Decision Process (MDP) [15]. At each time step t, the agent observes state s_t, executes action a_t, and then receives instant reward r_t, and transits to the next state s_{t+1}, which forms a sequence $\langle s_t, a_t, r_t, s_{t+1} \rangle$ of MDP. Define the return G_t as the sum of discounted rewards, given by

$$G_t \triangleq r_t + \beta r_{t+1} + \beta^2 r_{t+2} + \cdots = \sum_{k=0}^{\infty} \beta^k r_{t+k}, \quad (7)$$

where $0 < \beta < 1$ denotes the discount factor. Define *policy* π as the state-to-action mapping. The agent optimizes policy π in order to maximize the action-value function Q defined as

$$Q(s, a | \pi) \triangleq \mathbb{E}_\pi [G_t | s_t = s, a_t = a], \quad (8)$$

which is the expectation of the return G_t at the current state s and action a under policy π.

However, RL can only handle problems with small state space and action space, which is inappropriate for our problem. To this end, we use DNN as the approximator of the Q function in Q-learning [15], which constitutes a commonly-used DRL framework known as DQN. In particular, the algorithm applies *experience replay* to sample data offline, and *target network mechanism* that modifies action-value Q towards target values to improve algorithm convergence. The DQN is trained to minimize the loss function defined as

$$L(\theta) \triangleq \mathbb{E}\big[\big(y_t - Q(s_t, a_t | \theta)\big)^2\big], \quad (9)$$

where the vector θ represents the DQN weights that determine the policy π, and y_t is the target function given by

$$y_t \triangleq r_t + \beta \max_a Q(s_{t+1}, a | \theta_{\text{target}}), \quad (10)$$

where θ_{target} is copied from θ every fixed number of steps.

Despite the efficiency of DQN, it still has some critical limitations: 1) Overestimations have been attributed to the greedy algorithm used by the target function, which negatively affects the performance of policy; 2) Uniform samples have been applied in experience replay rather than weighted samples based on significance, which may lead to divergence in target with large state space. In order to overcome the above limitations, we apply the Prioritized Replay DDQN to address the problem, which mainly improves in two aspects. First, (10) is adapted as

$$y_t^{DQ} \triangleq r_t + \beta Q(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta) | \theta_{\text{target}}), \quad (11)$$

which untangles the selection and evaluation respectively in Q-learning to avoid overestimation [16]. Second, experiences are replayed with prioritized sampling using the sum-tree structure [17], which is updated efficiently.

B. Preliminary Design Based on LoS Channel Model

In this subsection, we aim to design the ABS placement to maximize the coverage rate of GUs under the dominant-LoS channel model using the DRL framework. To achieve fast convergence, we apply the DRL algorithm phase by phase. In each phase, we set a target coverage rate $\bar{\varphi}$ and train the underlying DNN towards achieving $\bar{\varphi}$. The target coverage rate $\bar{\varphi}$ is then gradually increased until it can no longer be achieved, by which a suboptimal solution to (P1) is obtained. Specifically, for each phase, we cast the placement problem into a MDP, and define the state-action-reward tuple $\langle s, a, r \rangle$ with our domain knowledge as follows.

1) State s: Normally, state represents the input of DNN. A straightforward choice of the state is the profile of all GU and ABS locations, whose dimension and associated complexity increases with N and M. Moreover, the neural network is not sensitive to the scalar-type location variables without proper quantification. Therefore, a more suitable form of state representation is desired, for which we propose the *coverage bitmap*. Specifically, we equally partition the considered (rectangular) ground area $\mathcal{G}$ into K-by-K grid regions $\mathcal{G}_{ij}$, $i, j \in \mathcal{K} \triangleq \{1, \cdots, K\}$. Denote the number of covered GUs in region $\mathcal{G}_{ij}$ as

$$f_{ij} \triangleq \sum_{\mathbf{w}_n \in \mathcal{G}_{ij}} C_n. \quad (12)$$

As a result, we choose the state $s = F \triangleq [f_{ij}]_{K \times K}$, where the matrix F is in the form of a 2D bitmap, which effectively captures the spatial correlation of the GU and ABS locations in terms of the number f_{ij} of covered GUs in each grid, and thus termed *coverage bitmap*. Moreover, the bitmap data structure is well fit as the input type of the state-of-the-art DNN (more specifically, the convolutional neural network (CNN)), whose

input dimension ($K \times K$) and associated DNN complexity is invariant with arbitrary large N, thus circumventing the curse of dimensionality in DNN. Note that the selection of K still needs to balance between the bitmap resolution and the complexity of DNN. However, a moderately large K would suffice since the detailed spatial correlation of GU and ABS locations are effectively represented (and weighted) by f_{ij} and nested into the state matrix F.

2) Action a: For simplicity, assume that the action space of each ABS in each time step consists of four moving operations {up, down, left, right} with a certain displacement size Δ m. The overall action a is then an M-dimensional vector.

3) Reward r: In our context of maximizing the coverage rate, the reward r_t in each time step t needs to encourage the state-action pair that brings the current coverage rate φ_t closer to the ideal value of 1. Therefore, we choose the negative error function as the reward for the intermediate steps with $\varphi_t < \varphi$. When the target coverage rate is achieved, i.e., $\varphi_t \geq \varphi$, we set a positive reward $r_t = 1$ and terminate the episode. In addition, when ABSs are out-of-border[1], we set a negative reward $r_t = -1$ to punish such behavior. Thus, the reward function is defined as

$$r_t \triangleq \begin{cases} -\alpha\left(\varphi_t - 1\right)^2, & \text{if } \varphi_t < \bar{\varphi}, \\ 1, & \text{if } \varphi_t \geq \bar{\varphi}, \\ -1, & \text{if ABSs are out-of-border}, \end{cases} \quad (13)$$

where α is a positive constant to scale the reward.

Based on the defined state-action-reward tuple $\langle s, a, r \rangle$, the proposed ABS placement optimization with Prioritized Replay DDQN is presented in Algorithm 1. The algorithm starts by initializing the parameters (Line 1), followed by E episodes. Each episode starts by resetting the state (Line 2), followed by T steps. Each step consists of the *exploration* part (Lines 4∼7) and the *training* part (Lines 8∼17). In the exploration part, the agent interacts with the environment by observing the current state s_t, choosing an action a_t based on policy π, and obtaining the next state s_{t+1} and instantaneous reward r_t. The transition $\{s_t, a_t, r_t, s_{t+1}\}$ is then stored in memory $\mathcal{H}$. After memory $\mathcal{H}$ is full, the training part starts by sampling a minibatch[2] of l transitions for the training process to update the weights θ in order to minimize the weighted loss function, which is given by

$$L(\theta) \triangleq \frac{1}{l} \sum_{j=1}^{l} \omega_j \delta_j^2 = \frac{1}{l} \sum_{j=1}^{l} \omega_j \left(y_j - Q(s_j, a_j | \theta)\right)^2. \quad (14)$$

In (14), $\delta_j \triangleq y_j - Q(s_j, a_j | \theta)$ denotes the Temporal-Difference (TD)-error, y_j is the target function given by (11), and ω_j denotes the importance-sampling weight [17] used to correct the bias, which is given by

$$\omega_j \triangleq \frac{\left(|\mathcal{H}| \cdot P(j)\right)^{-\nu}}{\max_i \omega_i}, \quad (15)$$

[1] We define *out-of-border* as the case with at least two ABSs flying beyond the border, in order not to receive too much negative rewards.

[2] Minibatch is used such that the model updates are fast (as opposed to processing the whole training data) and not too noisy (as opposed to processing every instance).

Algorithm 1: ABS placement optimization with Prioritized Replay DDQN

Input: GU locations, initial ABS locations (and 3D Terrain Map), and coverage range D (or channel power threshold $\bar{g}$).

Output: Final ABS locations and achieved coverage rate.

1 Initialize target coverage rate $\bar{\varphi}$, replay memory $\mathcal{H} = \emptyset$, $p_1 = 1$.
2 **for** $episode := 1, \cdots, E$ **do**
3   Initialize the environment and receive an initial state s_1.
4   **for** $step\ t := 1, \cdots, T$ **do**
5      $a_t = \pi(s_t)$;
6   Execute a_t, and obtain s_{t+1} and r_t in (13);
7   Store transition $\{s_t, a_t, r_t, s_{t+1}\}$ in $\mathcal{H}$ with maximal priority $p_t = \max_{i<t} p_i$.
8   **if** $\mathcal{H}$ *is full* **then**
9   **for** $j := 1, \cdots, l$ **do**
10   Sample transition j based on (16);
11   Compute importance-sampling weight in (15);
12   Compute TD-error δ_j;
13   Update transition priority $p_j \leftarrow |\delta_j|$.
14   **end**
15   Update weights θ of $Q(\cdot)$ by minimizing the loss function in (14);
16   Set $\theta_{\text{target}} = \theta$ every fixed number of steps.
17   **end**
18   Terminate the episode if $r_t = 1$ when $\mathcal{H}$ is full.
19   **end**
20 **end**
21 Increase the target coverage rate $\bar{\varphi}$ and repeat Lines 1∼20, until it can no longer be achieved.

---

where $|\mathcal{H}|$ is the memory size and ν is a positive constant. In the training process, transition j is sampled based on the probability given by

$$P(j) \triangleq \frac{p_j^{\mu}}{\sum_{i=1}^{l} p_i^{\mu}}, \quad (16)$$

where p_j denotes the priority of transition j, and $\mu > 0$ denotes the degree of priority. The priority p_j is initialized to be 1 for all samples before memory $\mathcal{H}$ is full, so that they all stand a chance to be sampled. After memory $\mathcal{H}$ is full, we set $p_j = |\delta_j|$ to attribute a higher priority to the transition with larger absolute TD-error (which suggests greater model mismatch). The episode is terminated if $\mathcal{H}$ is full, and the target coverage rate $\bar{\varphi}$ is achieved, i.e., $r_t = 1$ (Line 18).

Through the Prioritized Replay DDQN (Lines 1∼20), the DNN weights θ are trained to minimize the loss function, thus fitting the DNN towards achieving the target function in (11), which in turn approximates the maximum action-value Q (or expected sum of discounted rewards), and hence improving the achieved coverage rate. In particular, the proposed reward function in (13) encourages the state-action pair that brings the coverage rate closer to 1, and punishes ABS out-of-border behavior. Together with the proposed state representation by coverage bitmap, we have coherently incorporated the domain knowledge in our problem context into the DRL framework, and thus able to solve the complicated problem (P1) effectively.
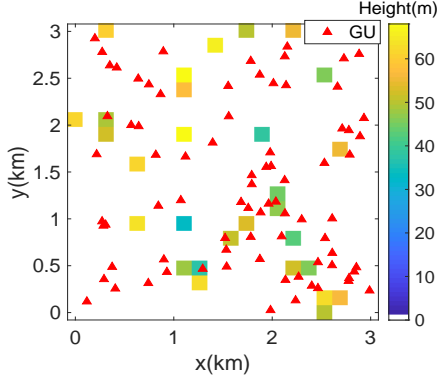
Fig. 2. An example of 3D Terrain Map with 80 GUs inside.

Algorithm 2: Obtaining the coverage bitmap based on 3D Terrain Map

Input: GU locations, ABS locations, 3D Terrain Map and channel power gain threshold $\bar{g}$.

Output: Matrix form of coverage bitmap $F = [f_{ij}]_{K \times K}$.

1   Initialize $F = 0$ and $C_n = 0, \forall n$.

2   **for** *GU* $n = 1, \cdots, N$ **do**

3   Sort the ABSs by ascending order of the GU-ABS distance (i.e., near to far), denoted by the ordered set $\mathcal{M}_{\text{sorted}}$.

4   **for** *ABS* $m \in \mathcal{M}_{\text{sorted}}$ **do**

5   Calculate $g_{m,n}$ based on (5);

6   **if** $g_{m,n} \geq \bar{g}$ **then**

7   Set $C_n = 1$; break.

8   **end**

9   **end**

10   **end**

11   Obtain f_{ij} by accumulating C_n for each grid region $\mathcal{G}_{ij}$ as in (12).

C. Advanced Design Based on 3D Terrain Map

The preliminary design optimizes the ABS placement under the dominant-LoS channel model. To achieve more accurate coverage in the site of interest, we assume that its 3D Terrain Map is available, based on which we propose the advanced design to further refine the ABS positions tailored to the specific environment. For illustration, we generate an example of 3D Terrain Map for a square region of side length 3 km, with 30 buildings and 80 GUs randomly located inside, as shown in Fig. 2. The buildings are modeled by cuboids with length and width of 150 m, and random height based on uniform distribution in $[30, 70]$ m.

In the advanced design, the definitions of $\langle s, a, r \rangle$ are similar to those in the preliminary design, despite the calculation of the coverage bitmap. Specifically, the coverage bitmap F now relies not only on the ABS-GU distance, but also the LoS/NLoS channel state for the specific ABS/GU locations. To this end, we propose Algorithm 2 to obtain the coverage bitmap based on the 3D Terrain Map. Algorithm 2 begins by resetting the coverage bitmap and coverage indicators (Line 1). For each GU n, we first sort the ABSs by ascending order of the GU-ABS distance (i.e., from near to far), and then check the GU coverage by the sorted ABS order (Lines 2∼ 10). This helps to reduce the computational complexity, since a nearer ABS is more likely to have LoS channel with the GU and hence cover it, thanks to the air-ground channel characteristics [14]. The obtained coverage indicator $C_n, n \in \mathcal{N}$ is then accumulated for each grid region $\mathcal{G}_{ij}$ to obtain f_{ij} as in (12) (Line 11), and hence the coverage bitmap $F = [f_{ij}]_{K \times K}$.

Based on the obtained coverage bitmap, we can then apply the proposed Algorithm 1 to further refine the ABS placement for the site of interest, by taking the placement result of the preliminary design as initial input. Note that the preliminary design based on the dominant-LoS channel model captures the main spatial correlation of ABSs/GUs by the *distance-based coverage rule*, while the advanced design based on the 3D Terrain Map exploits the *fine-grained structure* of the air-ground channel, thus able to achieve more accurate coverage in a specific environment.

V. NUMERICAL RESULTS

In this section, we present numerical results on the coverage performance of the proposed ABS placement design. We adopt the example of 3D Terrain Map and GU locations in Fig. 2. The following parameters are used if not mentioned otherwise: $M = 10$, $N = 80$, $H = 90$ m, $D = 0.5$ km (corresponding to $\bar{g} = -93$ dB), $f_c = 2$ GHz, $c = 3 \times 10^8$ m/s, $K = 20$, $\Delta = 10$ m, $l = 64$, $\mu = 0.6$, $\nu = 0.4$, $T = 100$ and $|\mathcal{H}| = 40000$.

In the preliminary design, the DRL model is trained for 900 episodes from an initial random state, where the obtained placement result is used to set the initial state in the advanced design, which is further trained for 1600 episodes. The final ABS placement and GU coverage result by the preliminary design with uniform coverage range is shown in Fig. 3(a), where 93.75% of GUs are covered. However, when the above result is applied in the considered 3D Terrain Map, the coverage rate drops sharply to 80%, as shown in Fig. 3(b), due to the NLoS path-loss caused by site-specific blockage in the 3D space. Fortunately, we further apply the advanced design based on the 3D Terrain Map, which raises the coverage rate back to 90%, as shown in Fig. 3(c). Note that in general, the achievable coverage rate based on the 3D Terrain Map is not greater than that based on the dominant-LoS channel model, since the additional signal blockage by obstacles results in NLoS path-loss and hence overall lower coverage rate. On the other hand, by comparing Fig. 3(b) and Fig. 3(c), it can be seen that some of the GUs (circled in the figures) originally not covered by the preliminary design result, are now covered in the advanced design with slight change of the ABS locations. For example, GUs 1∼4, originally not covered in the preliminary design, are now covered in the advanced design with slight movement of ABS 5. This thus demonstrates the benefit of considering site-specific LoS/NLoS channel for the site of interest, and the advantage of our proposed DRL-based ABS placement design in achieving higher coverage rate.

Finally, the coverage rates achieved by the proposed algorithm and the benchmark DQN and K-means algorithms are plotted in Fig. 4, respectively. It can be seen that the proposed algorithm achieves higher coverage rate in both the preliminary design and advanced design. On the other hand, in all three
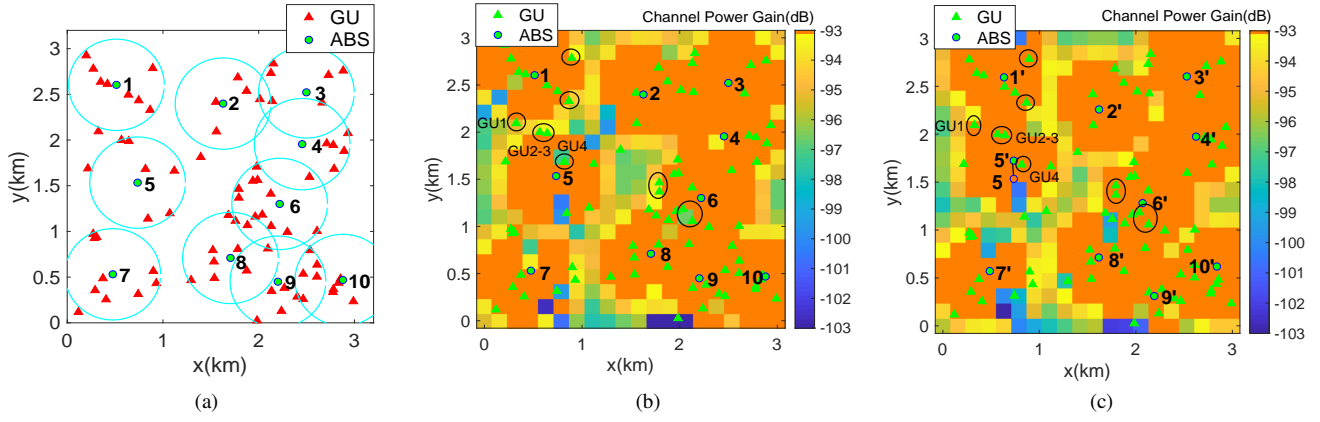
Fig. 3. ABS placement and GU coverage by (a) preliminary design with uniform coverage range; (b) preliminary design result applied in the 3D Terrain Map; and (c) advanced design based on the 3D Terrain Map.
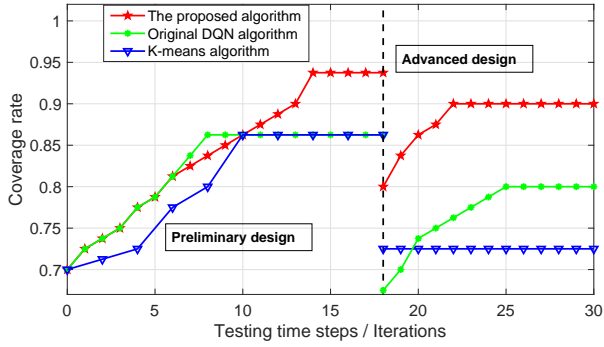

Fig. 4. Coverage rate achieved by the different algorithms.

algorithms, the achieved coverage rate under the LoS channel model drops when the placement results are applied in the 3D Terrain Map, due to the introduction of additional NLoS path-loss. Note that the K-means algorithm is distance-based and only applicable for the scenario with uniform coverage range. In comparison, our proposed algorithm with Prioritized Replay DDQN adapts well in the complex environment, and also outperforms the basic DQN algorithm.

VI. Conclusions

This paper investigates the placement optimization of multiple ABSs to maximize the coverage rate of GUs under the dominant-LoS channel model first, and further the site-specific LoS/NLoS model. The problem is NP-hard in general and further complicated by the complex propagation environment. We tackle this challenging problem using the DRL method by proposing the coverage bitmap as the state representation, which captures the spatial correlation of GUs/ABSs, and is well fit as the input of DNN with fixed dimension and associated complexity. Moreover, with our proposed action and reward, the DRL agent learns well from the dynamic interactions with the environment using the Prioritized Replay DDQN. Numerical results show that our proposed design significantly improves the coverage rate compared to benchmark DQN and K-means algorithms. Our next plan is to extend the current framework to the scenario with moving GUs.

References

[1] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, 2016.

[2] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, 2016.

[3] B. Galkin, J. Kibilda, and L. A. DaSilva, "Deployment of UAV-mounted access points according to spatial user locations in two-tier cellular networks," in *Proc. Wireless Days*, pp. 1–6, March 2016.

[4] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, pp. 569–572, Dec. 2014.

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

[6] Z. Wang, L. Duan, and R. Zhang, "Adaptive deployment for UAV-aided communication networks," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 4531–4543, Sep. 2019.

[7] X. Liu, Y. Liu, and Y. Chen, "Deployment and movement for multiple aerial base stations by reinforcement learning," in *Proc. IEEE GLOBE-COM*, pp. 1–6, Dec 2018.

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

[9] J. Chen and D. Gesbert, "Optimal positioning of flying relays for wireless networks: A LOS map approach," in *IEEE Int. Conf. Commun. (ICC)*, pp. 1–6, May 2017.

[10] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potentials, challenges and promising technologies," *IEEE Wireless Commun.*, vol. 26, pp. 120–127, Feb. 2019.

[11] J. Lyu and R. Zhang, "Network-connected UAV: 3-D system modeling and coverage performance analysis," *IEEE Internet Things J.*, vol. 6, pp. 7048–7060, Aug. 2019.

[12] Y. Zeng and X. Xu, "Path design for cellular-connected UAV with reinforcement learning," *[Online]. Available: http://arxiv.org/abs/1905.03440*, 2019.

[13] S. Zhang and R. Zhang, "Radio map based path planning for cellular-connected UAV," *[Online]. Available: http://arxiv.org/abs/1905.05046*, 2019.

[14] 3GPP-TR-36.777, "Enhanced LTE support for aerial vehicles," *3GPP Technical Report*, 2019.

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* MIT press, 2018.

[16] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artificial Intelligence*, 2016.

[17] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. Int. Conf. Learning Representations*, 2016.