Learning-Based Cooperative Aerial and Ground Vehicle Routing for Emergency Communications

Yuchao Zhu and Shaowei Wang
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
E-mail: dz20230030@smail.nju.edu.cn, wangsw@nju.edu.cn

***Abstract*—Unmanned aerial vehicle (UAV)-assisted communications has emerged as a promising technology in many domains, such as disaster relief and emergency scenarios. However, the limited battery capacity restricts UAVs from performing such persistent missions. In this paper, we consider a general hybrid trajectory planning problem to efficiently provide emergency communications in time-constrained disaster-affected regions, where a ground vehicle carrying backup batteries moves along with the UAV as a "mobile charging platform" to handle the energy issue of the UAV. Our optimization task is to minimize the total cost of the mission. We show that the optimization task is an extension of the traveling salesman problem with soft time window constraints. Due to the NP-hardness of the task, we propose a novel deep reinforcement learning with a sequential model strategy to learn the policy for the UAV's visiting order, based on which the collaborative routes of the UAV and ground vehicle are designed. Numerical results show that our proposed learning-based route planning scheme is effective and efficient.**

***Index Terms*—Emergency communications, reinforcement learning, trajectory planning, unmanned aerial vehicles.**

I. Introduction

Large-scale natural or man-made disasters (e.g., floods, fires) always inflict heavy loss of property and life. Swift response to the disaster is crucial to minimize further loss since disasters are generally unforeseeable and hard to be prevented. Maintaining real-time communications helps the relief personnel to have immediate access to emergency information, which improves the efficiency of the response mission. Unfortunately, the existing ground communication infrastructures will be unable to function normally due to the disaster. Therefore, emergency communication with rapid response is crucially important for search and rescue in the event of disasters.

Currently, acting as air base stations (BSs), relay nodes and mobile anchors, unmanned aerial vehicles (UAVs) are widely used in many domains [1]–[3]. Due to the inherent advantages of mobility and flexibility, the UAVs mounting BSs are a promising option for disaster relief [4]. Considering a post-disaster scenario with unknown user distribution, a UAV is used to scan the region and localize trapped users by received signal strength indicators in [5], where the task is decomposed into two subproblems: scanning points selection and trajectory planning. Note that the battery capacity of UAV is limited, two energy-efficient UAV path planning algorithms based on multi-armed bandit algorithm are proposed in [6] to maximize the system throughput. Taking limited user equipment energy into account, a trajectory optimization problem to balance uplink throughput and energy efficiency is studied in [7], where the task is transformed into a constrained Markov decision-making process and tackled by a deep Q network. In [8], an integrated and dynamic deployment of aerial and ground BSs is proposed to provide swift and stable area coverage.

Although the UAV emergency communication networks play a powerful role in disaster scenarios, the UAV still suffers severe flight time limitations. Recent works have studied the route planning task while taking UAV recharging into consideration [9], [10]. In [11], the path planning of UAVs and mobile charging stations is regarded as a vehicle routing problem with synchronized networks and finite candidate locations of charging stations, solving by a genetic algorithm-based heuristic method. It is worth noting that the urgency of communications varies (e.g., the "Golden 72 hours" for life savings) since the damage severity of stricken regions differs [12], indicating that time constraints should be considered.

Motivated by the points mentioned above, we consider a hybrid trajectory planning task for the emergency communication network with one UAV and one ground vehicle (GV), where the different parts of the disaster-affected region are assumed to have service deadlines reflecting their urgency. We show that this task can be formulated as an extended traveling salesman problem with soft time windows (TSPSTW), which is one canonical example of combinatorial optimization and is NP-hard in general. We propose a practical deep reinforcement learning-based approach to address the intractable time-constrained response mission. Numerical results verify that our proposal is effective and efficient.

The remainder of the paper is organized as follows. Section II introduces system model and formulates the optimization task. In Section III, our proposed learning-based cooperative path planning approach is given in detail. Section IV presents numerical results and performance evaluation. Finally, Section V concludes the paper.

II. System Model And Problem Formulation

A. Network Description

As depicted in Fig. 1, we consider a square disaster-affected region $\mathcal{A} \in \mathbb{R}^2$ with side length L, where the existing

This work was supported in part by the National Natural Science Foundation of China under Grants 61931023 and U1936202.

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001484

network infrastructure can not work normally. In this case, the UAV is used as a mobile aerial BS to provide temporary communication connection for the trapped people. Since the UAV operates in hovering to establish communication links with ground user equipments (UEs) with limited coverage radius, the entire region is divided into N regions of interests (ROIs) so that the UAV can provide service for each ROI through one taking-off and landing. We assume that the service

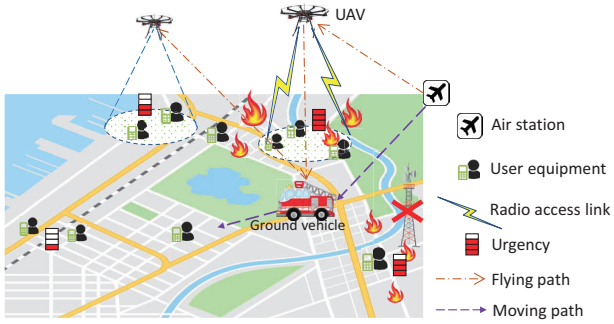


Fig. 1. UAV and GV assisted emergency communication networks.

urgency differs by the damage severity of ROI, which can be obtained from the remote sensing images. Considering the UAV's battery endurance, an emergency GV is dispatched to aid the UAV so that the UAV can fly back to the moving GV to replace its battery timely. The set of ROIs and UAV's 3D hovering locations are represented by $\mathcal{R} = \{R_1, R_2, ..., R_N\}$ and $\mathcal{G} = \{G_1, G_2, ..., G_N\}$, respectively. Each ROI R_i is assumed to have a service deadline τ_i based on its urgency. The UAV takes off from the air station $\boldsymbol{a}_0$ to traverse the ROIs along its trajectory, when the UAV hovers at $G_i = (g_i^x, g_i^y, g_i^z)$ with a duration T_h, it establishes communication links with potential UEs in ROI R_i. The UAV needs to return to the GV for battery replacement before it visits the next ROI, the positions at which they rendezvous with each other are denoted by "launch sites" $\mathcal{X} = \{\boldsymbol{x}_i = (x_i^x, x_i^y, 0), 1 \leq i \leq N\}$. Let $\mathcal{U}_N \ni \sigma(\cdot)$ denote the set of permutations of $\{1, 2, ..., N\}$ representing all the possible sequences of visiting ROIs, where we set $\sigma(N+1) = \sigma(1)$ to simplify the notation. Thus, the trajectories of the ground and aerial vehicle can be represented as $[\boldsymbol{x}_{\sigma(1)}, \boldsymbol{x}_{\sigma(2)}, ..., \boldsymbol{x}_{\sigma(N+1)}]$ and $[\boldsymbol{x}_{\sigma(1)}, G_{\sigma(1)}, \boldsymbol{x}_{\sigma(2)}, ..., G_{\sigma(N)}, \boldsymbol{x}_{\sigma(N+1)}]$, respectively, where the air station is the start/end point (i.e., $\boldsymbol{a}_0 = \boldsymbol{x}_{\sigma(1)}$).

B. Channel Model

We consider the air-to-ground channel model as in [13], where both line-of-sight (LoS) links and non-line-of-sight (NLoS) links are taken into consideration. The probability of LoS is of the following form:

$$P_{LoS}(\psi) = 1/\left(1 + a\exp\left(-b\left(\psi - a\right)\right)\right), \tag{1}$$

where a, b are environment constants. $\psi = \frac{180^\circ}{\pi}\arctan(H/r)$ is the elevation angle between the UEs and UAV, where H is the altitude of the UAV and r is the horizontal distance between the UEs and UAV. The probability of NLoS links is $P_{NLoS}(\psi) = 1 - P_{LoS}(\psi)$, and the average path loss is:

$$\begin{aligned} P_{loss}(H, r) = 20\log d + 20\log f_c + 20\log\left(4\pi/c\right) + \\ P_{LoS}(\psi)\mu_{LoS} + \left(1 - P_{LoS}(\psi)\right)\mu_{NLoS}, \end{aligned} \tag{2}$$

where μ_{LoS} and μ_{NLoS} are the mean values of the additional loss in LoS and NLoS links, respectively. $d = \sqrt{H^2 + r^2}$ is the distance between the UAV and UEs. c is the speed of light, f_c represents the carrier frequency. The average communication data rate between the UEs and the UAV is defined as:

$$\xi_{data} = B_w \log_2\left(1 + P_t/(P_{loss}N_0)\right), \tag{3}$$

where B_w is the communication bandwidth, P_t and N_0 are the transmission power and noise power, respectively.

C. Problem Formulation

For the purpose of scanning the entire region rapidly as well as trying not to exceed the deadline, we define the total cost of the mission as a weighted sum of operation time and late penalty, where the former is the completion time of serving all the ROIs reflecting the swiftness of recovery, and the latter is the penalty of service latency aiming at avoiding out of time. Let $V_0, V_1 > 0$ denote the speed of the GV and the UAV, respectively, with $V_0 < V_1$. The total cost can be mathematically written as follows:

$$\begin{aligned} C_{total} &= T_o + \alpha T_p \\ &= \sum_{i=1}^{N} \max\left\{t_{GV}^{\sigma(i)}, t_{UAV}^{\sigma(i)}\right\} + \alpha \sum_{k=1}^{N} \max\left\{0, t_{arr}^{\sigma(k)} - \tau_{\sigma(k)}\right\}, \end{aligned} \tag{4}$$

where $t_{GV}^{\sigma(i)} = \frac{1}{V_0}\|\boldsymbol{x}_{\sigma(i)} - \boldsymbol{x}_{\sigma(i+1)}\|$ is the time for the GV to move from one launch site to the next, $t_{UAV}^{\sigma(i)} = \frac{1}{V_1}\left(\|\boldsymbol{x}_{\sigma(i)} - G_{\sigma(i)}\| + \|G_{\sigma(i)} - \boldsymbol{x}_{\sigma(i+1)}\|\right) + T_h$ represents the amount of time for the UAV to leave one launch site, reach the i-th hovering location and provide communications, then return to rendezvous with the GV at the next launch site for battery replenishment, $t_{arr}^{\sigma(k)} = \sum_{j=1}^{k-1} \max\left\{t_{GV}^{\sigma(j)}, t_{UAV}^{\sigma(j)}\right\} + \frac{1}{V_1}\|\boldsymbol{x}_{\sigma(k)} - G_{\sigma(k)}\|$ corresponds to the arrival time at the k-th hovering location, and α is the penalty coefficient that adjusts the sensitivity of latency. Thus, the considered problem can be formulated as follows:

$$\begin{aligned} \underset{\sigma \in \mathcal{U}_N,\ \boldsymbol{x}_{\sigma(\cdot)}}{\text{minimize}} \quad & C_{total} \\ \text{s.t.} \quad & E_f t_{UAV}^{\sigma(i)} + E_h T_h \leq E_{max},\ 1 \leq i \leq N, \end{aligned} \tag{5}$$

where E_f, E_h are the power consumption of flying and hovering, respectively, and the UAV's battery capacity is denoted as E_{max}. Problem (5) is NP-hard since it is an extension of the TSPSTW [14] that requires considering the positions of launch sites and battery life of the UAV.

III. Our Proposed Approach

The hardest part of the task lies in determining the visiting order since problem (5) over variables $\boldsymbol{x}_i$ will be convex and easy to solve if the permutation σ is fixed. We propose a novel deep reinforcement learning-based scheme to solve the task.

A. Selection of Hovering Locations

First of all, we clarify the selection of hovering locations for the purpose of guaranteeing the data transmission rate between the UAV and the UEs. Assume that the rate requirement is ξ_{data}^{min}, we can correspondingly obtain a threshold of path loss P_{loss}^{max} according to (3). The UEs are supposed to connect to the UAV if $P_{loss}(H,r) \leq P_{loss}^{max}$, i.e., the maximum coverage radius can be written as $r_{max} = \{r|P_{loss}(H,r) = P_{loss}^{max}\}$. Note that the coverage radius rises first and then descends as the UAV altitude increases, we can search the value H satisfying $\partial r_{max}/\partial H = 0$ to get the optimal altitude H_{opt} that yields the widest coverage. To minimize the number of taking off and landing, the entire region is divided into multiple ROIs based on the maximum coverage radius, i.e., the ROIs are small squares with side length $\frac{L}{\sqrt{2}r_{max}}$, and the center of the ROIs are the horizontal part of hovering locations, the flying altitude of the UAV is set as H_{opt}.

B. Attention-Based Framework for UAV's Visiting Order

Given the hovering locations, the optimization task can be reconsidered as a TSPSTW with a given start point as follows when setting the battery constraint aside:

$$\begin{aligned}\underset{\sigma\in\mathcal{U}_N}{\text{minimize}}\ C(\sigma) = &\sum_{i=0}^{N} t_{UAV}^{\sigma(i+1)} + \\ &\alpha\left(\sum_{k=0}^{N-1}\max\left\{0, t_{arr}^{\sigma(k+1)} - \tau_{\sigma(k+1)}\right\}\right),\end{aligned} \tag{6}$$

where we set $\boldsymbol{x}_{\sigma(i)} = \boldsymbol{x}_{\sigma(i+1)} = G_{\sigma(i-1)}$, and $G_{\sigma(0)} = G_{\sigma(N+1)} = \boldsymbol{a}_0$ is the start point.

Problem (6) is still NP-hard, which is difficult to solve with exact methods. Note that the task can be viewed as a sequence decision problem by a policy, we propose a sequential model-based deep neural network to tackle the combinatorial optimization problem in an unsupervised manner. As depicted in Fig. 2, one network encodes the input start node and all hovering nodes, and then another network converts the encoded information to a visiting order as its output.

Our attention-based encoder-decoder model defines a stochastic policy $p(\sigma|\boldsymbol{s})$ for selecting a solution σ given a problem instance $\boldsymbol{s}$, which can be parameterized by $\boldsymbol{\theta}$ [15]:

$$p_{\boldsymbol{\theta}}(\sigma|\boldsymbol{s}) = \prod_{t=1}^{N} p_{\boldsymbol{\theta}}\left(\sigma(t)|\boldsymbol{s}, \sigma(1), ..., \sigma\left(t-1\right)\right), \tag{7}$$

where t is time step, $p_{\boldsymbol{\theta}}(\sigma(t)|\cdot)$ is the probability of the ROI being visited at the t-th time step based on $\boldsymbol{s}$ and the ROIs that have been visited at previous time steps.

1) Encoder: Following the Transformer architecture [16] but without positional encoding, the encoder reads and maps the low-dimensional input features into several high D_h-dimensional vectors. In order to allow the model to distinguish the start node from the regular nodes, we use separate parameters W_0^x and b_0^x to compute the initial embedding of the start point. Additionally, we provide the deadline τ_i and required service time T_h of each ROI as input features. The initial embeddings are computed through a learned linear projection:

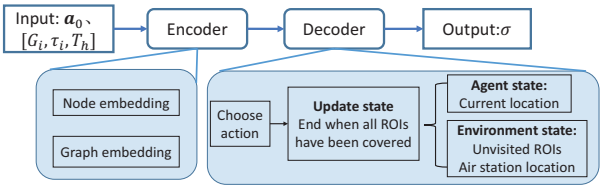


Fig. 2. The encoder-decoder framework.

$$\boldsymbol{h}_i^{(0)} = \begin{cases} W_0^x \boldsymbol{a}_0 + \boldsymbol{b}_0^x, & i = 0 \\ W^x\left[G_i, \tau_i, T_h\right] + \boldsymbol{b}^x, & i = 1, ..., N, \end{cases} \tag{8}$$

where W_0^x, W^x, $\boldsymbol{b}_0^x$ and $\boldsymbol{b}^x$ are learnable parameters.

The embeddings are updated by n attention layers, each consisting of a multi-head attention layer and a feed-forward layer. The attention mechanism passes weighted messages between the nodes in a graph, which can be expressed as:

$$u_{ij} = \begin{cases} \frac{\boldsymbol{q}_i^T \boldsymbol{k}_j}{\sqrt{D_k}}, & \text{if } i \text{ adjacent to } j \\ -\infty, & \text{otherwise,} \end{cases} \tag{9}$$

$$\boldsymbol{h}_i^{'} = \sum_j \text{softmax}(u_{ij})\boldsymbol{v}_j = \sum_j \frac{e^{u_{ij}}}{\sum_{j'} e^{u_{ij'}}}\boldsymbol{v}_j, \tag{10}$$

where $\boldsymbol{k}_i = W^K \boldsymbol{h}_i^{(\cdot)}$, $\boldsymbol{q}_i = W^Q \boldsymbol{h}_i^{(\cdot)}$ and $\boldsymbol{v}_i = W^V \boldsymbol{h}_i^{(\cdot)}$ are the key, query and value for each node, respectively. u_{ij} calculates the compatibility of the query $\boldsymbol{q}_i$ of node i with the key $\boldsymbol{k}_j$ of node j as the scaled dot-product, and we compute the attention weights using a softmax. Instead of using a single head, a multi-head attention with head size M is used to allow nodes to receive different messages from neighbors. The final multi-head attention value for node i is as follows:

$$\text{MHA}_i(\boldsymbol{h}_1, ..., \boldsymbol{h}_N) = \sum_{m=1}^{M} W_m^O \boldsymbol{h}_{im}^{'}. \tag{11}$$

The output of the MHA sublayer along with skip connection is passed through batch normalization and then a fully connected feed-forward layer with ReLU activation function, the operations are expressed as follows:

$$\begin{aligned} \hat{\boldsymbol{h}}_i &= \text{BN}^l\left(\boldsymbol{h}_i^{(l-1)} + \text{MHA}_i^l\left(\boldsymbol{h}_1^{(l-1)}, ..., \boldsymbol{h}_n^{(l-1)}\right)\right), \\ \boldsymbol{h}_i^{(l)} &= \text{BN}^l\left(\hat{\boldsymbol{h}}_i + \text{FF}(\hat{\boldsymbol{h}}_i)\right), \end{aligned} \tag{12}$$

where l is the number of the layer, $\text{FF}(\hat{\boldsymbol{h}}_i) = W^{ff,1} \cdot \text{ReLU}\left(W^{ff,0}\hat{\boldsymbol{h}}_i + \boldsymbol{b}^{ff,0} + \boldsymbol{b}^{ff,1}\right)$.

Similar to [17], an aggregated embedding $\bar{\boldsymbol{h}}^{(n)}$ of the input graph as the mean of final node embeddings $\boldsymbol{h}_i^{(n)}$ is computed by the encoder: $\bar{\boldsymbol{h}}^{(n)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}_i^{(n)}$. Both the node embeddings and the graph embedding are used as the input to the decoder.

2) Decoder: Decoding happens sequentially, the decoder outputs the selected node $\sigma(t)$ at time step $t \in \{1, ..., N\}$ based on the current state, which consists of the environment state (including the node embeddings and the already visited nodes) and agent state (i.e., the current location). We design a special context node representing the decoding context to utilize the information of the states [17]:

$$\boldsymbol{h}_{(c)}^{(n)} = \begin{cases} \left[\bar{\boldsymbol{h}}^{(n)}, \boldsymbol{h}_{\sigma(t-1)}^{(n)}\right], & t > 1, \\ \left[\bar{\boldsymbol{h}}^{(n)}, \boldsymbol{h}_{0}^{(n)}\right], & t = 1. \end{cases} \tag{13}$$

We compute a new context node embedding $\boldsymbol{h}_{(c)}^{(n+1)}$ using attention mechanism again to augment the exchange and fusion of information. Note that the start point can not be visited if not yet all nodes have been visited, and the ROIs can not be visited twice, the compatibility of the query with all nodes is given as:

$$u_{(c)j} = \begin{cases} -\infty, & j = 0, \text{ and } t < N, \\ -\infty, & j \neq 0, \text{ and } \exists t^{'} < t : \sigma(t^{'}) = j, \\ \frac{\boldsymbol{q}_{(c)}^{T} \boldsymbol{k}_j}{\sqrt{D_k}}, & \text{otherwise}, \end{cases} \tag{14}$$

where the keys $\boldsymbol{k}_i = W^K \boldsymbol{h}_i^{(n)}$ and values $\boldsymbol{v}_i = W^V \boldsymbol{h}_i^{(n)}$ come from the node embeddings, and the query $\boldsymbol{q}_{(c)} = W^Q \boldsymbol{h}_{(c)}^{(n)}$ is from the context node. Then, by applying the same multi-head self attention mechanism as described in (10)-(11), we get the result $\boldsymbol{h}_{(c)}^{(n+1)}$. With query from $\boldsymbol{h}_{(c)}^{(n+1)}$, we compute the compatibilities by (14) using a single attention head, and clip the result within $[-C, C]$: $u_{(c)j}^{'} = C \cdot \tanh\left(u_{(c)j}\right)$ [18]. These compatibilities are interpreted as unnormalized log-probabilities, and we compute the final output probability vector p using a softmax:

$$p_i = p_{\boldsymbol{\theta}}\left(\sigma(t) = i | \boldsymbol{s}, \sigma(1), ..., \sigma(t-1)\right) = \frac{e^{u_{(c)i}^{'}}}{\sum_j e^{u_{(c)j}^{'}}}. \tag{15}$$

The decoder outputs the selected node based on p_i and the process will end when all of the ROIs have been visited.

3) Training Method: The presented attention-based network must be trained through exploring actions and receiving feedback in a form of rewards. We define the training objective function as follows:

$$\mathcal{C}(\boldsymbol{\theta}|\boldsymbol{s}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\sigma|\boldsymbol{s})}\left[C(\sigma)\right], \tag{16}$$

which is the expectation of the total cost $C(\sigma)$ shown in (6). We optimize $\mathcal{C}$ by gradient descent, using REINFORCE [19] gradient estimator with baseline $\boldsymbol{B}(s)$:

$$\mathcal{C}(\boldsymbol{\theta}|\boldsymbol{s}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\sigma|\boldsymbol{s})}\left[\left(C(\sigma) - \boldsymbol{B}(s)\right) \nabla \log p_{\boldsymbol{\theta}}(\sigma|\boldsymbol{s})\right]. \tag{17}$$

The training procedure is shown in Algorithm 1. Here, greedy decoding and sampling decoding are employed for baseline policy and current policy, respectively. We compare the current model with the baseline model at the end of each epoch, and update the baseline parameters $\boldsymbol{\theta}^{bl}$ only if the improvement is significant in terms of a paired t-test. The optimizer used to train the parameters is Adam.

C. Cooperative Route Planning

Given one permutation σ to visit all the ROIs, the origin problem (5) can be rewritten as follows ($\forall i \in \{2, ..., N+1\}$):

$$\begin{aligned} & \underset{\boldsymbol{x}_{(\cdot)}, t_{(\cdot)}, c_{(\cdot)}}{\text{minimize}} \quad c_{N+1} \\ \text{s.t. } & C_1: \ t_i \geq t_{i-1} + \frac{\|\boldsymbol{x}_{i-1} - \boldsymbol{x}_i\|}{V_0}, \\ & C_2: \ t_i \geq t_{i-1} + \frac{\|\boldsymbol{x}_{i-1} - G_{i-1}\| + \|G_{i-1} - \boldsymbol{x}_i\|}{V_1} + T_h, \\ & C_3: \ c_i \geq c_{i-1} + \frac{\|\boldsymbol{x}_{i-1} - \boldsymbol{x}_i\|}{V_0}, \\ & C_4: \ c_i \geq c_{i-1} + \frac{\|\boldsymbol{x}_{i-1} - G_{i-1}\| + \|G_{i-1} - \boldsymbol{x}_i\|}{V_1} + T_h, \\ & C_5: \ c_i \geq c_{i-1} + \frac{\|\boldsymbol{x}_{i-1} - \boldsymbol{x}_i\|}{V_0} + \alpha\Big(t_{i-1} + \\ & \qquad \frac{\|\boldsymbol{x}_{i-1} - G_{i-1}\|}{V_1} - \tau_{i-1}\Big), \\ & C_6: \ c_i \geq c_{i-1} + \frac{\|\boldsymbol{x}_{i-1} - G_{i-1}\| + \|G_{i-1} - \boldsymbol{x}_i\|}{V_1} \\ & \qquad + T_h + \alpha\left(t_{i-1} + \frac{\|\boldsymbol{x}_{i-1} - G_{i-1}\|}{V_1} - \tau_{i-1}\right), \\ & C_7: \ \|\boldsymbol{x}_{i-1} - G_{i-1}\| + \|G_{i-1} - \boldsymbol{x}_i\| \leq V_1 \frac{E_{max} - E_h T_h}{E_f}, \\ & C_8: \ c_1 = 0, \ t_1 = 0, \ \boldsymbol{x}_1 = \boldsymbol{x}_{N+1}, \end{aligned} \tag{18}$$

where $t_{(\cdot)}$ and $c_{(\cdot)}$ represent the accumulative operation time and cost at each ordered node, respectively. $\boldsymbol{x}_1 = \boldsymbol{x}_{N+1} = \boldsymbol{a}_0$ is the fixed air station. After obtain the visiting sequence of the TSPSTW tour of $\mathcal{G}$, we can solve this convex optimization problem by using standard techniques, then the coordinated routes under ordered visiting assignment can be found.

Algorithm 1 REINFORCE with baseline algorithm

```
Input: number of epochs E, steps per epoch S, batch size
  B, training dataset S = {s_1, ...s_{S×B}}, significance δ.
Initialization: θ,θ^bl ← θ
for epoch = 1, ..., E do
    for step = 1, ..., S do
        Randomly choose training data s_k(∀k ∈ {1, ..., B})
        from S;
        Find routes σ_k(∀k ∈ {1, ..., B}) by sampling;
        Find routes σ_k^bl(∀k ∈ {1, ..., B}) by greedy decoding;
        ∇C ← Σ_{k=1}^{B} (C(σ_k) − C(σ_k^bl)) ∇_θ log p_θ(σ_k);
        θ ← Adam(θ, ∇C);
    end for
    if OneSidedPairedTTest(p_θ, p_{θ^bl}) < δ then
        θ^bl ← θ
    end if
end for
```

IV. Numerical Results

Consider a disaster-affected square region in urban with a size of $L \times L$ km^2, where the air station is located with

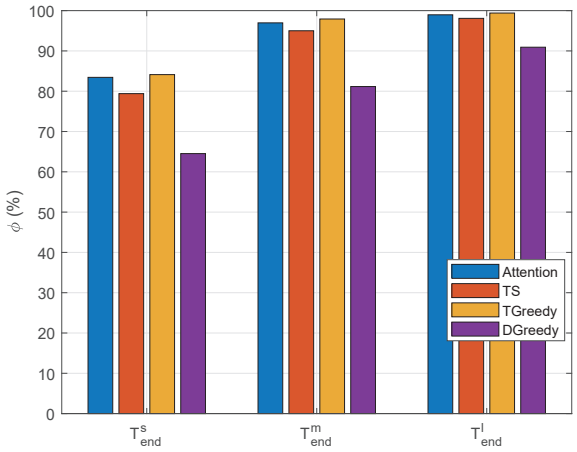
coordinate $\boldsymbol{a}_0 = [0\text{m}, 0\text{m}]$. The urban environment parameters for $f_c = 2$ GHz are $a = 9.61$, $b = 0.16$, $\mu_{LoS} = 1$, $\mu_{NLoS} = 20$, respectively [13]. The communication bandwidth is $B_w = 1$ MHz, the transmission power and the noise power are $P_t = 20$ dBm and $N_0 = -110$ dBm, respectively [20]. Assume that the required transmission rate is 5 Mbps, thus the flying altitude is chosen as $H_{opt} = 600$ m to achieve the maximum coverage radius $r_{max} = 600$ m. The number of divided ROIs is set as $\left(\lceil \frac{L}{\sqrt{2} r_{max}} \rceil\right)^2$. The battery parameters of the UAV are $E_f = 155$ W, $E_h = 206$ W and $E_{max} = 40$ Wh, respectively. The hovering time above each ROI is set as $T_h = 5$ minutes with the consideration of battery capacity. The speeds of GV and UAV are set as $V_0 = 20$ km/h and $V_1 = 80$ km/h, respectively. The time required for replacing battery can be ignored. Considering that T_o and T_p are of equal importance, the coefficient for late penalty is set as the mean value of the estimated operation time, which differs by the region size.

We train the model for 100 epochs with randomly generated data under the learning rate of 10^{-4}. In every epoch, 2500 batches of 512 instances are processed. Each element in any problem instance is embedded into a vector of size 128 by the encoder network with 3 layers and 8 attention heads. We evaluate the performance of our proposed attention-based cooperative route planning method (Attention) and compare with the following common baseline methods, which adopt different strategies for the selection of visiting sequence:

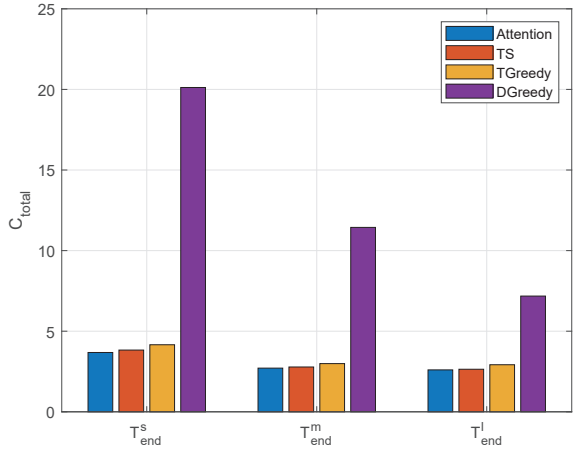
1) TGreedy: The UAV provides service according to the ROIs' urgency, i.e., the visiting sequence is arranged in ascending order of deadlines.
2) DGreedy: The UAV selects the one closest to its current location among all ROIs to be served.
3) TS: The UAV's route is designed by tabu search, a meta-heuristic for vehicle routing problems [21]. Here we initialize the route by TGreedy. The tabu length is set as 30, and the maximum number of iterations is 500.

In addition to the total cost C_{total}, the percentage of on-time served ROIs denoted as ϕ is also employed to measure the performance, which reflects the timeliness of the service. All results are averaged over 100 Monte Carlo simulations.

In Fig. 3, we show the performance of our proposal when given different ranges of deadline T_{end} (in hours) under a fixed region size. τ_i for each ROI is sampled from the uniform distribution $[0, T_{end}]$. Here we consider three representative deadlines for a region with 16 km^2, i.e., $T^s_{end} = 3.35$, $T^m_{end} = 6.7$, $T^l_{end} = 13.4$, reflecting the degree of urgency. Fig. 3(a) shows that the percentage of on-time served ROIs increases when broadening the range of deadline, which is expected since the probability of out-of-time decreases when extending the deadline. Our proposal and TGreedy achieve similar performance, followed by TS, and DGreedy causes severe latency since it does not take deadlines into consideration. As can be seen from Fig. 3(b), our proposal achieves the lowest total cost, especially when the deadline is tight. For T^s_{end}, the gain is 4.04%, 13.15% and 81.71% as compared to



(a) The percentage of on-time served ROIs.



(b) The total cost.

Fig. 3. The percentage of on-time served ROIs and the total cost under different ranges of deadline, $L = 4$ km, $\alpha = 2.34$.

TS, TGreedy and DGreedy, respectively. Results indicate that our proposal can balance the swiftness and timeliness.

Then, we investigate the bearing capacity of the system, i.e., the maximum region that could be covered potentially with at least 95% ROIs served on time by one UAV and one GV. Given the range of service deadline $T_{end} = 10$ hours, Fig. 4 shows that the maximum manageable region is 36 km^2. Although TGreedy has a slight advantage on timeliness when the region is relatively small (e.g., $L \leq 5$ km), it performs poorly as the region becomes larger. This is reasonable because, with less strict deadlines, the UAV will have abundant time for traveling to ROIs that far away from each other, which is advantageous for TGreedy as it always reaches the most urgent ROIs regardless of its location. While the design of trajectory plays an important role when the deadline becomes tighter, thus our proposal shows superior performance. For example, when $L = 6$ km, it outperforms TS, TGreedy and DGreedy by 3.06%, 0.32% and 21.31%, respectively. The

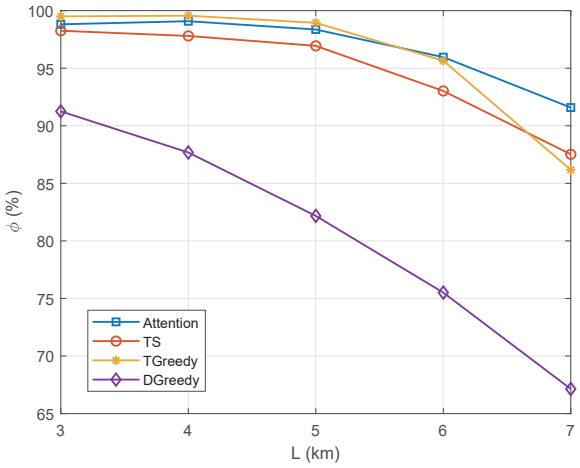


Fig. 4. The percentage of on-time served ROIs as a function of the region size with $T_{end} = 10$ hours.

TABLE I
TOTAL COST OF DIFFERENT METHODS

L (km)	3	4	5	6	7
α	1.48	2.34	3.41	4.68	6.18
Attention	1.64	2.61	3.88	6.35	12.70
TS	1.66	2.67	4.04	7.04	14.43
TGreedy	1.78	2.92	4.45	7.84	21.40
DGreedy	2.67	8.39	30.62	101.55	275.98

corresponding total cost of the four methods is given in Table I, which also indicates that our proposal outperforms others. Since our proposed attention-based scheme can exploit global information, the performance gap between our proposal and others becomes larger as the region extends.

V. CONCLUSION

In this paper, we investigated the UAV-aided post-disaster communications with deadline constraints, where a ground vehicle acting as a battery swap station is introduced to assist the UAV. Inspired by the promising development of deep reinforcement learning, we proposed a novel learning-based cooperative route planning scheme with the aim of minimizing the mission's total cost. Specifically, after determining the hovering locations based on the best coverage radius, we used a Seq2Seq neural network to learn the policy of the trajectory planning with time constraints, then a heuristic algorithm is employed to tackle the coordinated trajectory design of the aerial and ground vehicle. Numerical results indicated that our proposal outperforms the compared algorithms, especially when the deadline is tight, which offers an appealing balance between swiftness and timeliness.

REFERENCES

[1] H. Sallouha, M. M. Azari, and S. Pollin, "Energy-constrained UAV trajectory design for ground node localization," in *Proc. IEEE GLOBECOM'18*, Abu Dhabi, United Arab Emirates, Dec. 2018.

[2] Y. Sun, T. Wang, and S. Wang, "Location optimization and user association for unmanned aerial vehicles assisted mobile networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10 056–10 065, Oct. 2019.

[3] Y. Zhu and S. Wang, "Aerial data collection with coordinated UAV and truck route planning in wireless sensor network," in *Proc. IEEE GLOBECOM'21*, Madrid, Spain, Dec. 2021.

[4] N. Zhao *et al.*, "UAV-assisted emergency networks in disasters," *IEEE Wirel. Commun.*, vol. 26, no. Feb., pp. 45–51, 2019.

[5] F. Demiane, S. Sharafeddine, and O. Farhat, "An optimized UAV trajectory planning for localization in disaster scenarios," *Comput. Netw.*, vol. 179, p. 107378, Oct. 2020.

[6] Y. Lin, T. Wang, and S. Wang, "UAV-assisted emergency communications: An extended multi-armed bandit perspective," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 938–941, Mar. 2019.

[7] T. Zhang *et al.*, "Trajectory optimization for UAV emergency communication with limited user equipment energy: A Safe-DQN approach," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 3, pp. 1236–1247, 2021.

[8] X. Xu and Y. Zeng, "Time-weighted coverage of integrated aerial and ground networks for post-disaster communications," in *Proc. IEEE WCNCW'20*, Seoul, Korea (South), Jun. 2020.

[9] Y. Wang *et al.*, "Mobile wireless rechargeable UAV networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 60, no. 3, pp. 33–39, Mar. 2022.

[10] Y. Zhu and S. Wang, "Efficient aerial data collection with cooperative trajectory planning for large-scale wireless sensor networks," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 433–444, Jan. 2022.

[11] R. G. Ribeiro *et al.*, "Unmanned-aerial-vehicle routing problem with mobile charging stations for assisting search and rescue missions in postdisaster scenarios," *IEEE Trans. Syst. Man Cybern. Syst.*, 2021, 10.1109/TSMC.2021.3088776.

[12] M. T. Rashid, D. Y. Zhang, and D. Wang, "SocialDrone: An integrated social media and drone sensing system for reliable disaster response," in *Proc. IEEE INFOCOM'20*, Toronto, ON, Canada, Jul. 2020.

[13] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wirel. Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[14] Y. Dumas *et al.*, "An optimal algorithm for the traveling salesman problem with time windows," *Oper. Res.*, vol. 43, no. 2, pp. 367–371, 1995.

[15] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Proc. NeurIPS'15*, Montreal, Quebec, Canada, Dec. 2015.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS'17*, Long Beach, California, USA, Dec. 2017.

[17] W. Kool, H. van Hoof, and M. Welling, "Attention, Learn to solve routing problems!" in *Proc. ICLR'19*, New Orleans, LA, USA, May 2019.

[18] I. Bello *et al.*, "Neural combinatorial optimization with reinforcement learning," 2016. [Online]. Available: https://arxiv.org/abs/1611.09940

[19] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, no. 3, pp. 229–256, 1992.

[20] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[21] M. Zachariasen and M. Dam, *Tabu Search on the Geometric Traveling Salesman Problem*. Boston, MA: Springer US, 1996, pp. 571–587.