面向草原修复的多无人机协同方法研究

焦栋斌[1+]，袁雨辰[1]，王贤义[1]，陈鑫安[1]，林华康[1]

（1. 兰州大学 信息科学与工程学院，兰州 730000；+ 指导教师 $E-mail:jiaodb@lzu.edu.cn$）

摘　要：草原作为地球重要的生态系统之一，如何采取无人机修复等现代方法修复草原使其免受退化是环保领域的重要问题。考虑到无人机电池能量有限及草原退化程度各异等各项现实约束，我们对草原修复过程建立了详细的数学模型，继而提出了深度强化学习训练单无人机模型、多无人机协同算法调度进行草原修复以求解该问题。单无人机模型利用深度强化学习 Actor-Critic 算法在自身修复地图下的草原修复过程建模为马尔可夫决策过程，构建编码器-解码器架构的深度神经网络模型完成单无人机修复方案的求解。编码器通过经典的 Transformer 架构对各项输入的节点特征与无人机特征进行编码，借由自回归的指针网络解码器逐步得到修复方案。考虑到多无人机协同对草原修复效率的明显提高，我们设计了一种轻量级的多无人机协同策略，将人工设计的启发式策略与深度强化学习到的策略相结合，以期望提高多无人机情况下的草原修复能力。实验结果显示，在预设的问题规模中，我们提出的算法具有出色的寻优能力，能在有效的时间内获得比对比算法更优的草原修复方案。

关键词：草原修复；深度强化学习；多无人机协同

Grassland Restoration Method Based on Multi-Drone Collaboration

JIAO Dongbin[1+], YUAN Yuchen[1], WANG Xianyi[1], CHEN Xinan[1], LIN Huakang[1]

(1. College of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China)

Abstract: Grassland, as one of the crucial ecosystems on Earth, faces the significant challenge of degradation prevention through modern approaches such as drone-based restoration. Addressing the practical constraints of limited drone battery energy and varying degrees of grassland degradation, we have developed a detailed mathematical model for grassland restoration. Subsequently, we propose a solution to this issue by training a single drone model using deep reinforcement learning and coordinating multiple drones for collaborative scheduling in grassland restoration. The single drone model leverages the Actor-Critic algorithm of deep reinforcement learning to model the grassland restoration process under its own repair map as a Markov decision process, utilizing a deep neural network model with an encoder-decoder architecture to solve the single drone restoration strategy. The encoder encodes node features and drone features using the classical Transformer architecture, while the decoder, based on an autoregressive pointer network, progressively generates the restoration plan. Considering the substantial enhancement in grassland restoration efficiency achieved through multi-drone collaboration, we have devised a lightweight collaborative strategy for multiple drones by integrating heuristic strategies with deep reinforcement learning policies to enhance the grassland restoration capabilities under multi-drone scenarios. Experimental results demonstrate that, within the predefined problem scope, our proposed algorithm exhibits exceptional optimization capabilities, yielding superior grassland restoration solutions over comparative algorithms within a effective time.

Keywords: grassland restoration; deep reinforcement learning; multi-UAV collaboration

草原占据了地球陆地总面积的 26%-40%[1]，具有防风固沙、涵养水源、固碳释氧、调节气候、美化环境及维护生物多样性等重要功能[2]，在维护生态和保障民生等方面都发挥着不可或缺的作用[3]。长期以来，由于人类活动、气候变化、外来物种入侵等多种因素的影响，草原面临大规模退化困境[4]。

近年来，无人机 (unmanned aerial vehicles, UAV) 以其高度自动化、低成本低人力高效益的日益成为生态修复中数据收集、低空作业的重要工具[5]。然而，由于无人机自身的电池能量、飞行载荷等限制[6]，如何减少无人机作业过程中的无效能耗、尽可能增大无人机草原修复的实际效果是无人机在草原修复场景下得以成功应用的关键之一。

现有的大多数关于无人机路径规划工作，研究热点集中于通过高效的区域覆盖算法来延长无人机的工作时间，如通过将问题建模为旅行商问题 (Traveling Salesman Problem, TSP)、优化无人机的能量来提升无人机的飞行时间[7]，或将问题建模为特殊的定向问题 (Orienteering Problem, OP)，通过启发式算法求解[8]等。大多数研究都聚焦于单一的路径规划任务中，而未同时考虑路径规划和应用场景下的相关任务的双重优化。

因此，我们建立起了详尽自洽的多无人机协同的草原修复最大化模型，首先利用强化学习方法训练单无人机模型，进而提出多无人机协同的草原修复方法，以期望在无人机能量不足、单无人机修复能力有限等限制条件下，研究草原修复面积最大化问题。

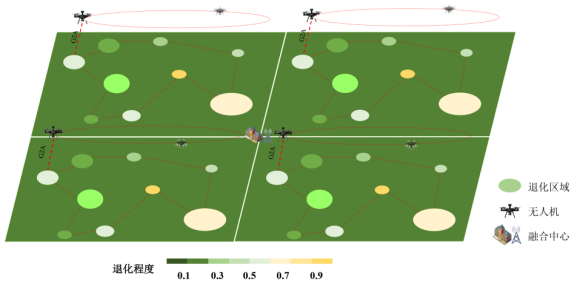

图 1 多无人机修复退化区域示例

1 多无人机协同的草原修复面积最大化模型

我们将待修复草原实例定义为一个完全无向图 $G=(V,E)$，其中 $V=\{v_0,v_1,...,v_N\}$ 表示所有带修复区域的集合，每个区域 v_i 具有位置坐标、退化程度和待修复区域面积等信息，其中区域 v_0 定义为地面信息融合中心，退化程度和待修复面积均为 0。$E=(e_{ij}|i,j\in V,i\neq j)$ 表示所有边的集合，每条边 e_{ij} 仅有边长这一属性。无人机 (多旋翼) 起飞时电池能量定义为 E_{max}，携带的草种重量为 Q。根据国际生态修复实践原则和标准[9]，我们将待修复草原的退化程度范围定为 $[0.3,0.8]$。

针对待修复草原实例，我们的目标是在无人机从基站出发并在为每个待修复区域提供服务前耗尽能量的情况下，最大化无人机修复的总面积 (假设仅当无人机飞行至待修复区域时才能进行播种)。由于无人机在播撒草籽修复过程中自身重量不断变化，其飞行能耗又与之直接相关，如何权衡无人机飞行能耗与修复能耗之间的关系时问题的核心所在。在修复过程中，无人机的能量消耗主要包括三个部分：无人机在待修复区域中播种的能量消耗、无人机进行航拍的能量消耗以及无人机在飞行过程中的能量消耗。相应地，这三部分能量消耗可以分别表示为 E_s,E_{ap} 和 E_f。此外，无人机的能量消耗恒定速度下的载荷重量成正比，Dorling[10] 等推导出 h 转子无人机的功耗方程，表示如下：

$$P(\overline{q}_{ij})=(M+\overline{q}_{ij})^{\frac{3}{2}}\sqrt{\frac{g^3}{2\rho\zeta h}} \tag{1}$$

其中，$M=W+m$，W 表示无人机框架重量，m 表示其电池重量。$\overline{q}_{ij}$ 表示无人机当前载荷重量（草籽重量），g 为标准重力加速度，ρ 表示空气密度，ζ 表示无人机旋转叶片盘面积，h 表示无人机旋翼数量，上述参数单位均参考国际单位制标准。

为简化问题，我们假设无人机在待修复区域间以固定高度、恒定速度飞行，同时忽略不同天气条件对无人机飞行的影响，如温度、风力、雨量和沙尘暴等。此外，考虑到无人机能量的限制，我们假设所有草原上的退化区域均可在一次修复过程中由一架无人机修复 (无论是否完全修复)。

1.1 各项能量约束

（1）无人机飞行能耗

无人机飞行期间的能耗，即单架无人机从基站出发，根据修复方案遍历带修复区域最后返回基站的飞行路径能耗。我们设 $x_{ij}\in 0,1$ 为 $0-1$ 决策变量，定义如下：

$$x_{ij}=\begin{cases}1 & e_{ij}\in uav\_path\\ 0 & otherwise\end{cases} \tag{2}$$

我们假定无人机在固定高度以恒定速度 v 飞行，每单位距离的能耗相同。因此修复过程中无人机飞行的总能耗可以表示为：

$$E_f=\sum_{i=0}^{N}\sum_{i\neq j}^{N}e_{ij}^f d_{ij}x_{ij} \tag{3}$$

其中，e_{ij}^f 表示无人机的飞行单位距离的能耗，与无人机在待修复区域 v_i 和 v_j 之间携带的种子重量 $\overline{q}_{ij}$ 相关，可由公式 (1) 实时计算得出。$\overline{q}_ij$ 表示从待修复区域 v_i 和 v_j 中无人机携带的草种重量，满足以下条件：

$$\sum_{j=0,i\neq j}^{N}\overline{q}_{ji}-\sum_{j=0,i\neq j}^{N}\overline{q}_{kj}=Q_i,\forall i\in V_a \tag{4}$$

$$\overline{q}_{ij}\leq Qx_{ij},\forall(i,j)\in A \tag{5}$$

（2）无人机播种能耗

我们将每个待修复区域离散为 $c_i(i=1,...,n)$ 个单位圆面积。无人机播种的能量消耗可视为与草地退化程度和其携带种子的重量相关的函数。无人机每单位播种面积的能耗可以定义为：$e_i=\eta q_i$。其中 η 是一个与能量消耗相关的正参数，q_i 是第 i 个待修复区域每单位圆修复所需种子的重量。进一步地，修复每个区域单位圆面积所需的种子重量可视为该区域的草地退化程度的函数[11]：$q_i=(1+l_i)^\gamma$。其中，$l_i\in[0.3,0.8]$ 表示第 i 个待修复区域的草地退化程度，γ 是与草地环境有关的正参数。综上所述，无人机 u 在第 i 个待修复区域中修复 σ_i 个单位圆所需的草种重量可以表示为 $Q_i^u=\sigma_i q_i$，σ_i 表示无人机修复的单位圆数量 ($1\leq\sigma_i\leq c_i$)。因此，无人机 u 在待修复区域内播种的总能量消耗可以表示为：

$$E_s=\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sigma_i e_i x_{ij} \tag{6}$$

其中，二元 $0-1$ 变量 x_{ij} 用于确定无人机是否将对待修复区域 v_j 进行修复。

（3）无人机拍照能耗

除飞行能耗以外，无人机还在各待修复区域收集修复信息。无人机搭载高光谱相机进行航拍的总能耗可表示为：

$$E_{ap} = e_{ap} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \sigma_i x_{ij} \tag{7}$$

其中，e_{ap} 表示无人机在 σ_i 个被修复的单位圆中航拍所消耗的能量。

1.2 优化目标

由于能量有限，无人机在一次修复过程中无法修复草原上所有的退化区域。因此，我们的目标是通过无人机技术尽可能多地修复草原的退化面积，即优化目标为：

$$C = \sum_{i=1}^{N} \sigma_i \tag{8}$$

1.3 问题描述

综上所述，我们考虑在无人机的最大能量限制下，对待修复区域的播种修复和航拍能耗、待修复区域间的飞行能耗以及无人机的飞行轨迹进行组合优化。无人机通过将自身决策与彼此信息实时交流从而实现协同 (修复地图再分配等)。多无人机协同的草原修复模型可被描述为：

$$\max_{x_{iju}, \sigma_{iu}} \quad C = \sum_{u=1}^{U} \sum_{i=1}^{N} \sum_{j \neq i}^{N} x_{iju} \sigma_{iu} \tag{9a}$$

$$\begin{aligned}
\text{s.t.} \ & \sum_{i=1}^{N} \sum_{j \neq i}^{N} \sigma_{iu} e_{iu} x_{iju} + e_{ap} \sum_{i=1}^{N} \sum_{j \neq i}^{N} x_{iju} \sigma_{iu} \\
& + \sum_{i=0}^{N} \sum_{j \neq i}^{N} (M + \overline{q}_{iju})^{\frac{3}{2}} \sqrt{\frac{g^3}{2\rho\zeta h}} x_{iju} d_{iju} \leq E_{max}^{u}, \quad (9b) \\
& \sum_{i=1, i \neq j}^{N} \sigma_{iu} q_{iu} x_{iju} \leq Q^u, \ \forall j \in V_u, \quad (9c) \\
& \sum_{j=0, i \neq j}^{N} \overline{q}_{jiu} - \sum_{j=0, i \neq j}^{N} \overline{q}_{iju} = \sigma_{iu} q_{iu}, \ \forall i \in V_u, \quad (9d) \\
& \overline{q}_{iju} \leq Q^u x_{iju}, \ \forall (i,j) \in V_u, \quad (9e) \\
& \sum_{i=0, i \neq j}^{N} x_{iju} = \sum_{i=0, i \neq j}^{N} x_{iju} = 1, \ \forall i, j \in V_u \quad (9f) \\
& \sum_{j=1}^{N} x_{0ju} = \sum_{j=1}^{N} x_{j0u} = 1, \quad (9g) \\
& x_{iju} \in \{0, 1\}, \quad (9h) \\
& 1 \leq \sigma_{iu} \leq c_i \quad (9i) \\
& \overline{q}_{iju} \geq 0. \quad (9j)
\end{aligned}$$

2 算法求解思路

对于序列决策问题，基于迭代搜索的启发式方法往往能取得能取得较优的效果[12]。然而，启发式方法受限于过长的求解时间，同时在一个待修复草原实例下的结果无法迁移，在面对多个待修复草原实例时求解时间往往难以接受。相较之下，神经网络方法以其离线训练、在线决策，泛化能力极强的特性，在面对规模不同、分布各异的序列决策问题时仍然能在极短时间得出较优解, 且无需重复训练，因而在组合优化领域广受关注[3]。因此，我们首先提出以单无人机作为智能体，构建深度神经网络模型，利用强化学习方法进行训练，试图学习出一个合理的草原修复策略，以大大节省问题的求解时间。

我们采用强化学习方法相关概念，将无人机视作智能体，其与待修复草原环境的交互过程建模为一个马尔可夫决策过程，如下图所示。智能体的策略建模为神经网络，在与环境的交互中不断收集数据，网络模型通过强化学习 Actor-Critic 算法进行训练，目标是学习到一个策略 $\pi_\theta(a|s)$(θ 为神经网络参数) 使得草原修复面积最大化。

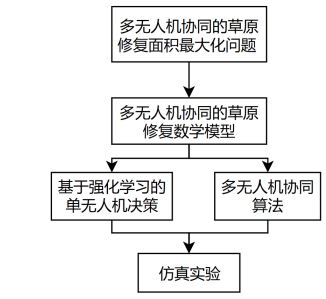

图 2 问题求解思路

3 基于强化学习的最优修复方案

3.1 马尔可夫过程建模

我们将修复过程中的每架无人机视为一个智能体，将待修复草原视为以欧几里得距离为边权值的无向全联通图。对单架无人机而言，其部分修复过程可视为从某点出发，遍历修复地图并完成相应修复任务，最后返回起点。

在无人机的修复过程中，由于其在两块相邻修复区域间的飞行能耗正比于此时无人机自身重量与路径长度，而自身重量又随着播散种子修复区域的过程动态变化，因而其总飞行能耗与总飞行长度、修复区域的顺序及修复面积均有复杂关系。换言之，作为优化目标的总修复面积由于与无人机耗能而与其飞行路径间接相关，因而二者之间存在复杂的耦合关系，如何解决修复面积的最优化，毫无疑问是一个困难的 NP-hard 问题。

针对该难题，我们拟采用强化学习方法学习这一复杂的函数关系，具体建模如下：

（1）状态

本文将一个待修复草原实例设为 $V=\{v_i\}_{i=0}^{n}$，其中包含了各个待修复区域的位置、退化程度等各项信息。各点以 v_0 为起始点按照某种策略 π，逐渐完善部分解 $\{(v_0,a_0),(v_i,a_i)\}_{i=1}^{uav\_now}$。其中，编号 i 表示无人机访问各点的次序，uav_now 为无人机已经过的点数量，a_i 代表按照策略 π 在点 v_i 修复的面积。我们将该部分解作为智能体在强化学习中的状态，记作 $s(<i), i\in[1,n+1]$。初始状态记作 $s(<1)=v_0$，表示无人机处于地面中心；结束状态记作 $s(<n+1)=s$，表示无人机已访问全部点。

（2）动作

如上所述，无人机状态是修复过程的部分解，因而我们将动作设置为无人机修复过程某一步的解，即将无人机在状态 $s(<i)$ 时的动作记作 $\pi_i=s(i), i\in[1,n+1]$，即可实现强化学习中动作-状态对的自然更新。

（3）状态转移函数

如上所述，在动作选取完毕后无人机新的状态也被唯一确定，因而状态转移函数具有完全确定性，记其为 $S(s(<i)|s(i))=s(<i+1), i\in[1,n]$。

（4）策略函数

如上所述，草原修复过程被创建为一个马尔可夫决策过程，因而其策略函数可被链式分解为：

$$p(\pi|s)=\prod_{i=1}^{n}p(s(i)|s(<i)) \tag{10}$$

其中，$p(s(i)|s(<i))$ 表示智能体在状态 $s(<i)$ 下选择动作 $s(i)$ 的概率。

（5）奖励函数

作为强化学习方法的核心之一，奖励函数的设计必须考虑全面，以防止模型难以收敛。经我们实验证明，单纯以最基本的修复面积作为模型的奖励函数会使得模型收敛速度大大减缓，同时考虑到无人机自身能量限制引入的惩罚项，我们最终将一组解的奖励函数设置为：

$$R(\pi|V)=\alpha_p * Pel+\alpha_r * \sum_{i=1}^{n}a_i \tag{11}$$

$$Pel=\begin{cases} d_n^0+\sum_{i=1}^{n}d_i^{i+1} & E_{rest}<0 \\ 0 & E_{rest}\geq 0 \end{cases} \tag{12}$$

上式中，α_p,α_r 为修正系数，Pel 为控制无人机能量约束的惩罚项。同时为了加快模型收敛速度，我们将不符合无人机能量约束时的惩罚项设置为其从起点到返回地面中心经历的路径长度，以确保在模型收敛的前期尽可能获得的解路径长度较短、符合能量约束，并在此前提下进行修复面积的决策。

3.2 神经网络模型构建

对于序列决策问题，编码器-解码器模型[13]作为最经典的神经网络模型结构之一，取得了优异的效果。我们采用稍加修改的经典的 Tranformer 作为编码器，指针网络[14]作为自回归解码器以实现构造式求解。

3.2.1 特征提取

我们对待求解的待修复草原模型提取两部分特征，即待修复区域的坐标、退化程度等静态特征，及无人机自身剩余能量、剩余重量等动态特征。静态、动态元素使用各自 Transformer 作为特征提取器。其中动态元素在构造解的过程中不断变化，需要结合部分解进行循环编码-解码，从而模拟无人机求解过程中负载动态变化的过程，更好地利用无人机当前信息作出更合理的决策。

3.2.2 编码器

参考了 Bresson[15]等人的方法。编码器结构如图 3 所示，由 L 层堆叠而成，采取 BatchNorm 归一化，公式可表示为：

$$h^{l=0}=h^{in}W^{in}\in\mathbb{R}^{n}\times d \tag{13a}$$

$$h_{rc}^{l+1}=BN(MHA^{l+1}(h^{l})+h^{l}) \tag{13b}$$

$$h^{l+1}=BN(Relu(h_{rc}^{l+1}W_1^{l+1})W_2^{l+1}+h_{rc}^{l+1}) \tag{13c}$$

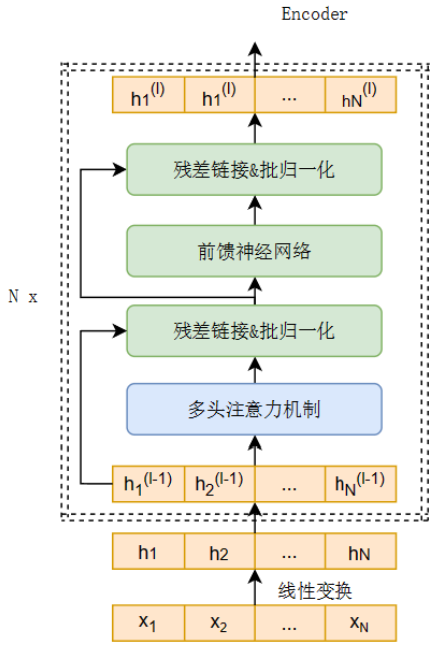

图 3 编码器结构

3.2.3 解码器

自回归解码器的作用是循环解码以逐步实现解的构造。首先输入静态、动态编码器的输出作为输入，通过注意力机制选择出第一个结点，然后利用循环神经网络学习已选择的结点信息更新解嵌入，再通过注意力机制选择下一个待修复区域及修复面积，以此类推直至所有结点[14]。

具体而言，解码器的输入有：静态特征、动态特征、访问的上一个结点的隐藏特征。开始时上一个结点的隐藏特征初始化为全零向量，之后每次决策出下一个待修复区域及修复面积便更新动态特征及上一个结点的隐藏特征。同时，采用循环神经网络学习已选择的结点信息，可表示为：

$$rnn^i, h^i_{gru} = \begin{cases} GRU(h^i_0, 0) & i = 0 \\ GRU(h^i_{i-1}, h^{i-1}_{gru}) & i > 0 \end{cases} \tag{14}$$

其中，0 为全零向量，$GRU(gatedrecurrentunit)$ 为循环神经网络更新解嵌入，h^i_{i-1} 为上一轮选择结点的结点嵌入。

自回归解码器接收上述输入后，先对所有输入做一次自注意力以进一步聚集所有结点的嵌入，而后通过两次线性聚合层得到当前修复地图的全局信息上下文并将其与当前局部信息相结合，以得出各个到达待修复区域 i 修复相应大小区域 a_i 的概率 $p(_i|s_i)$(二维)。同时，我们需要将解码器输出的概率进行一次掩码操作以过滤掉所有之前以到达过的待修复区域及各修复区域中不合法的修复面积方案。

$$h^{l=i}_p = Attention(static, dynamic^i, rnn^i) \tag{15a}$$

$$\mu_i = \begin{cases} C \times Tanh((W_q h^{l=i}_p)^T (W_k h^{l=i}_p)) & x_i \notin \pi_i \\ -\infty & otherwise \end{cases} \tag{15b}$$

$$p(_i|s_i) = Softmax(\mu_i) \tag{15c}$$

其中，$W_q, W_k \in \mathbb{R}_{d\times d}$ 为可学习参数矩阵。我们通过掩码设置 $\mu_i = -\infty$ 以避免重复选择结点，每次选择对应结点(即待修复区域) 及修复面积后，对解码器输入的动态元素及掩码进行更新以剔除不可行结点及修复方案。

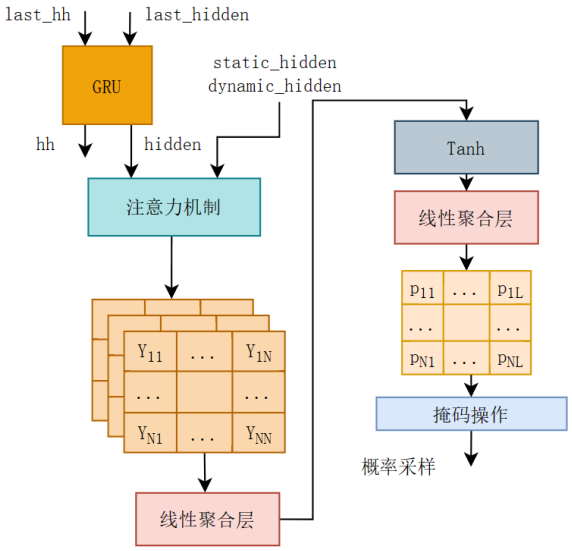

图 4 解码器结构

3.3 模型训练

对于建立的单无人机修复模型，我们采用强化学习 Actor-Critic 算法[16]进行梯度更新。模型中作为主体的 Actor 网络用于近似草原修复过程中的策略概率函数 $p(\pi|s)$，记其网络参数为 θ，则网络的待优化目标可表示为：

$$J(\theta \mid s) = \mathbb{E}_{t\sim p_{\theta(\cdot|V)}} R(\pi \mid s) \tag{16}$$

我们参考了 Williams 等的方法[17]，将上述策略梯度函数表示为：

$$\begin{aligned} \nabla_\theta J(\theta \mid s) &= \mathbb{E}_{t\sim p_{\theta(\cdot|s)}} \left[(R(\pi \mid s) - b(s)) \nabla_\theta \ln p_\theta(\pi|s)\right] \\ &= \mathbb{E}_{t\sim p_{\theta}(\cdot|s)} \left[A(\pi \mid s) \nabla_\theta \ln p_\theta(\pi|s)\right] \end{aligned} \tag{17}$$

其中，$R(\pi|s)$ 代表草原修复过程中计算得的奖励值；动作概率 $p_\theta(\pi|s)$ 代表 Actor 网络在草原修复过程中每次自回归时选择修复区域及修复面积大小的概率；$b(s)$ 为 REINFORCE 算法[17]中的基线函数，即 Critic 网络输出的估计值；$A(\pi|s) = R(\pi|s) - b(s)$ 为优势函数，表示在状态 s 下策略 π 的优劣，乘以 $\nabla_\theta \ln p_\theta(\pi|s)$，表示若优势函数 $A(\pi|s)$ 为正数则增大概率，若为负数则减小概率。

若设训练的图批次大小为 B，每张图中点集大小为 L，每次训练都按照策略函数 $p_\theta(\pi|s)$ 从各图中抽取一组可能的解，则 Actor 网络梯度可近似为：

$$\nabla_\theta J(\theta \mid s) \approx \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{L} [A(\pi_j \mid s_j) \nabla_\theta \ln p_\theta(\pi_j|s_j)] \tag{18}$$

将 Critic 网络参数设为 θ_c, 则 Critic 网络优化目标可近似为:

$$L(\theta_c \mid s) \approx \frac{1}{B}\sum_{i=1}^{B}\|b_{\theta_c}(V_i) - R(\pi|V_i))\|_2^2 \tag{19}$$

模型的具体训练过程如下算法 3 所示。算法通过在每个 epoch 内随机生成实例用于对网络进行训练与验证。通过指针网络的自回归性质构造性地得到当前策略下实例的解，并将其作为下一次网络的输入直至得到完整的解。根据公式计算得到相应的优势函数值后，通过 Adam（adaptive moment estimation）优化器[18]对 Actor-Critic 网络进行参数更新，其中基线函数 b(s) 为 Critic 网络对当前实例估计的奖励值。

算法 3 Actor-Critic 网络训练算法

输入：Actor-Crtic 网络所有参数 θ、θ_c，学习率 α，训练回合数 N_{epoch}，训练集大小 B_T，验证集大小 B_V, 训练批次 B，训练序列长度 L，训练序列面积上限 $area_{max}$ 及 $area_{min}$。

输出：收敛的神经网络参数 θ、θ_c

1. $i_{epoch} \leftarrow 0$

2. while $i_{epoch} < N_{epoch}$ do

3. 随机生成规模为 B_T 的训练集 $train\_set$ 及规模为 B_V 的验证集 $valid\_set$，将训练集分割为 $\lfloor \frac{B_T}{B} \rfloor$ 份

4. $t_e \leftarrow 0$

5. for t_e to $\lfloor \frac{B_T}{B} \rfloor$ do

6. 取训练集中第 t_e 份作为 T_{data}

7. $i_{steps} \leftarrow 0$

8. $Initial(mask)$ \\初始化 mask 掩码

9. $Initial(resolution)$ \\初始化解的存储空间

10. while $i_{steps} < L$ do

11. $(ptr, area) \leftarrow$ Actor(T_{data}) \\Actor 网络输出部分解

12. while Mask($(ptr, area), mask$):

13. $(ptr, area) \leftarrow$ Actor(T_{data}) \\屏蔽求解过程中非法的解

14. end while

15. Update($mask$)

16. Update($resolution, (ptr, area)$)

17. $i_{steps} \leftarrow i_{steps} + 1$

18. end while

19. R $\leftarrow$ Reward($resolution, T_{data}$) \\根据公式计算当前策略获得的奖励值

20. b $\leftarrow$ Critic(T_{data}) \\根据公式计算 Critic 网络的基线估计值

21. A $\leftarrow \|R - b\|_2^2$ \\计算优势函数值

22. $\nabla_\theta \leftarrow$ Loss($resolution, A$) \\计算网络梯度

23. $\theta, \theta_c \leftarrow$ Adam($\theta, \theta_c, \alpha, \nabla_\theta$)

24. end for

25. if $i_{epoch} == 0$ then

26. $\theta^\star \leftarrow \theta$

27. else

28. if Reward(Valid($\theta, valid\_set$)) 优于 Reward(Valid($\theta^\star, valid\_set$)) then

29. $\theta^\star \leftarrow \theta$

30. end if

31. end if

32. $i_{epoch} \leftarrow i_{epoch} + 1$

33. end while

4 多无人机协同调度算法

假设各无人机每轮巡航所携带的种子数量相同，在修复过程中无人机自身重量因种子播撒而减轻，因而导致其能飞行能耗变化，无人机已知自身修复地图的详细信息 (退化区域的数量及区域位置、退化程度及待修复面积等退化区域信息)，距离计算采用二维平面内的欧式距离公式。

本文考虑了一种多无人机协同调度算法，该算法通过单个无人机的局部信息和中心汇总的全局信息来实现多个无人机之间的协同。在算法的输入部分，除了对参数进行设定外，本文还需要为每架无人机初始化修复地图 M 及记录无人机信息的状态集 S。M 中涵盖了每架无人机的待修复任务点，每个任务点包含有任务点编号、坐标的位置、退化程度以及可修复面积；S 中则记录无人机当前的位置、剩余能量、已访问序列和已修复面积。同时，为了模拟无人机与地面中心交互的先后顺序，我们引入信号量这一机制以确保无人机与地面中心交流的通畅，该信号量定义为无人机在当前修复地图下最先完成首个待修复区域目标的时间，故其与无人机前往下一个区域的飞行时间与修复时间直接相关。

而在算法的输出部分，算法结束后应由中心对无人机状态集合 S 进行汇总，输出各无人机的访问序列、修复面积及剩余能量，以便对算法进行分析。

算法 4 多无人机协同调度算法

输入：参数序列 $Parms$，无人机修复地图集合 M_u，无人机状态集合 S_u

输出：无人机访问的节点序列 O_p，无人机修复的面积 O_a，无人机剩余能量 O_e

1. M_u^i = Initial(M_u) \\无人机根据初始化方法 (如 K-means) 等分配初始地图

2. P_u^i = Initial(P_u) \\初始化无人机信号量以决定决策优先级

3. while $M_u \neq \varnothing$ do

4. E_u^{r1} = PlaningPath(M_u, P_u^{self}) \\无人机群第一次路径规划

5. SendToCenter(S_u, M_u, E_u^{r1}) \\无人机第一次上报中心

6. M_u^{tmp} = updateMap(M^{global}, P_u^{self}) \\中心更新地图

7. RecvToUAV($M_u^{P_u^{tmp}}$)\\中心给无人机下发新地图

8. E_u^{r2} = PlaningPath(M_u^{tmp}, P_u^{self})\\无人机群第二次路径规划

9. SendToCenter($E_u^{r2}, Area_u^{r2}$)\\无人机第二次上报中心

10. if $\sum_{u=1}^{U} Area_u^{r2} \geq \sum_{u=1}^{U} Area_u^{r1}$ then

11. RecvToUAV(M_u^{tmp})\\无人机选择修复面积更多的地图

12. $M_u = M_u^{tmp}$

13. end if

14. $\sigma_u^{max\_p}$ = DecideArea(E_u^r, M_u)\\信号量最优先的无人机决策修复面积

15. ActionUAV($\sigma_u^{max\_p}, c_{max\_p}, P_u^{max\_p}$)\\无人机在当前点执行修复和信息采集

16. DropPointFromMap($M_u, P_u^{max\_p}$)

17. $P_u^{max\_p}$ = FlyToPoint($P_u^{benefit}$) \\无人机飞往下个最优点

18. Update($P_u^{max\_p}$) \\更新无人机信号量

19. end while

20. FlyToPoint(P_u^0)\\无人机返回起点

简而言之，该设计的多无人机协同算法主要思想是动态调整无人机修复地图，通过适应度函数 $\Theta(uav, M_u)$（如最短路径函数等）在修复过程中结合无人机自身位置等信息再分配地图以期望得到一个更好的结果。

无人机从地面控制中心出发，其初始修复地图可按照 K-means 等聚类方法划分，首先前往修复地图中适应度函数最小（如距离）的结点。随后，各无人机以当前位置为出发点，按照修复地图进行第一次最优路径规划；并计算该方案下，去除完成本路径及返航的飞行能耗、在剩余各点 (包含当前点) 进行信息收集的能耗及修复一单位面积所需要的能耗后，本机修复的面积。此处要特别说明，无人机在路径规划中通过循环确保得出的解一定符合能力约束要求，即按照某一路径规划方案，依次访问每个点，并完成信息采集与至少 1 单位面积的修复工作后，仍然剩余的能量大于 0。

无人机记录本次的规划方案，并将当前的修复地图、自身位置、修复面积等信息上报中心。中心汇总各无人机的修复地图得到全局修复地图，剔除其中每个无人机当前所在点；中心依次遍历地图中各点，根据已知信息，结合适应度函数 $\Theta(uav, M_u)$ 寻找与其最适合的无人机，并将该点加入对应无人机的新地图中。完成地图更新后，中心将新地图下发给各无人机。信号量最大的无人机以当前位置为出发点，按照调整过的新修复地图进行第二次最优路径规划，并计算该方案下的修复面积。对本次的规划方案做好记录后，并将第二次规划中的修复面积等信息上报中心。中心汇总上报信息，计算第二次的多机体系总修复面积；中心将两次修复面积进行比较，若第一次大于等于第二次，则新地图作废，反之，用新地图替代旧地图；中心将最终决策下达各无人机。无人机根据中心决策判断是否更新地图，并选择地图对应的剩余能量和规划方案。信号量最大的无人机在当前点完成信息采集和面积修复工作后，按照规划方案前往下一个点，并将当前点加入已访问序列，同时更新自身信号量。

重复上述决策过程，当某架无人机的修复地图为空时，该无人机返回中心，并在此后不再参与无人机群同中心的交互；当全局修复地图为空时 (全部无人机返航)，该算法结束。

5 实验与分析

本部分我们展示了一些仿真实验的具体配置与结果。

5.1 仿真配置

表 1 实验软/硬件配置

配置	描述
CPU	AMD Ryzen Threadripper 3970X 32-Core Processor, 3.79 GHz
GPU	NVIDIA GeForce RTX 3090 Ti
RAM	ADATA 192GB-DDR4
OS	Ubuntu Server 22.04.3 LTS
Python	Python 3.9
PyTorch	version 1.12.1+cu113

5.2 实例设置

为了验证模型和算法的有效性，本文在规模大小为 700×700 的草原区域实例上进行了模拟，同时设置了具有不同数量退化区域的草原场景，每个退化区域的退化程度 $li \in [0.3, 0.8]$, 退化区域随机生成。所有实例的基站都位于 $(0, 0)$ 处。对于所有草原场景中的每个退化区域，修复所需的单位圆数 σ_i 与草原区域的大小相对应，数量在 $[5, 20]$ 之间。对于所有场景，无人机的初始能量 E_{max} 设置为 3×10^7，以上设置中随机变量初始化均遵循均匀分布。

5.3 算法对比设置

为了与研究热点中的最优化路径规划策略[19]进行比较，我们将对比算法设置为一种双层的决策策略，即无人机首先通过最优路径规划给出一条自身修复地图内的最短路径，而后利用自身能量按照贪婪策略按照退化程度由低到高修复该路径内的待修复区域，直到达到无人机的能量约束上限。

我们称上述的策略为 SPGS(Shortest Path Greedy Strategy)，将其与强化学习方法进行比较。同时，我们还对二者分别搭配了相同的多无人机协同算法以对我们提出的两部分算法内容做进一步的验证。

5.4 参数设置

实验仿真具体参数如下。

表 2 实验软/硬件配置

参数	描述	值
M	无人机重量	1.5
g	重力加速度	9.8
ρ	空气流体密度	1.024
ζ	无人机旋转螺旋桨面积	0.2
h	无人机旋翼数	6
γ	修复能耗正系数	2
η	草原环境相关的正系数	1×10^4
e_{ap}	无人机收集一个单位圆信息所需的能耗	2×10^4
E_{max}	无人机所携带的最大能量	3×10^6
B	强化学习模型训练时的图批次大小	512
N_{epoch}	强化学习模型训练时的回合数	512

5.5 仿真结果

我们提出的单无人机强化学习模型用于同时优化路径规划和进行修复决策，主要涉及的待修复区域的位置、退化程度信息和无人机剩余能量等信息。因此，本文将每个待修复区域视为一个三维坐标点，包含了区域的水平、垂直坐标以及退化程度；将无人机自身信息视为变化的三维信息，包含无人机当前草籽负载、当前剩余能量和折合后的剩余待修复区域耗能。

模型的具体训练考虑了单架无人机负责 20 个退化区域的情况。得益于指针网络对输入规模的延展性，本文综合考虑强化模型的训练时长后，将行动者网络在训练时的输入规模设置为 20 个坐标点 (含出发点)。在该规模下训练完成的模型，在无人机待修复地图大小为 20 左右时均能获得较好的效果。我我们将强化学习模型每批次训练的图样本数量设置为 512，进行了 100 轮训练，Actor-Critic 网络的模型训练过程如图所示。

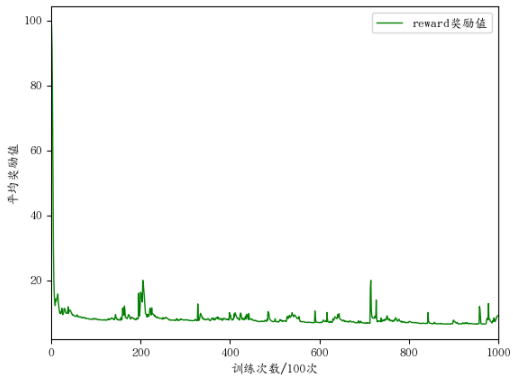

图 5 Actor 网络奖励值

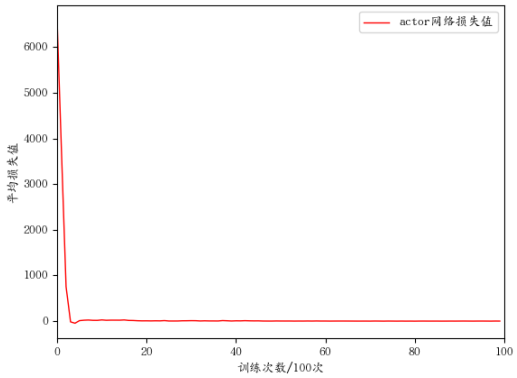

图 6 Actor 网络损失值

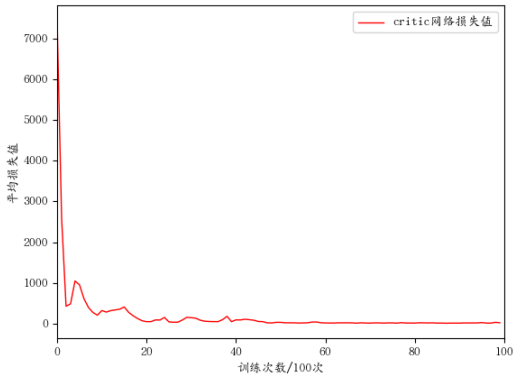

图 7 Critic 网络损失值

由图 5、6、7 可以看出，Actor-Critic 网络收敛迅速，同时 Actor 网络奖励值收敛良好，说明该智能体在训练过程中学得了良好的策略。

同时，我们还将训练的模型推广到多无人机的情形下，同时结合设计的多无人机协同算法，我们随机生成了不同无人机数量下的 100 个待修复草原实例进行仿真，实例仿真结果如表 3、4 所示：

表 3 多无人机协同算法

UAV 数量	RL			SPGS		
	Best	Avg	SD	Best	Avg	SD
2	518	474.6	24.8	345	288.7	24.8
3	796	721	35.1	524	484.6	21.7
4	1036	979.2	34	864	767.4	77.4
5	1201	1175.7	29.1	1081	872.1	53.1
6	1436	1398.4	36.4	1211	1184.2	43.8

表 4 非多无人机协同算法

UAV	RL			SPGS		
数量	Best	Avg	SD	Best	Avg	SD
2	505	449.2	37.1	281	185.3	48.4
3	727	697.4	19.1	411	317.5	67.4
4	965	910.7	49.5	652	585.6	57.7
5	1160	1121.1	25.3	862	792.1	45.3
6	1386	1333.5	42.4	1043	952.2	48.8

表 3、4 展示了在不同无人机数量情况下下协同和非协同算法的实验效果。从表格中可以清晰地看出，随着草原区域规模和无人机数量的增加，表格中多无人机协同算法在所有场景下的最优值和平均值明显优于非协同算法。同时，强化学习 RL 方法又明显由于贪婪的双层决策策略。具体而言，从表中可以看出，当无人机数量的大小从 1 增加到 6 时，多无人机协同的强化学习 Actor-Critic 方法相对于其他算法的解决方案结果改进百分比可以达到最大增长率分别为 17.02%、38.28% 和 35.98%。此外，多无人机协同的强化学习 Actor-Critic 方法在最优值和平均值方面均优于其他算法。因此，可以得出结论，多协同优化无人机轨迹设计和草原修复面积分配问题是解决多无人机最大化修复草原面积问题的一个良好解决方案。

6 结束语

我们提出了结合深度强化学习方法求解的多无人机协同的草原修复方法，利用深度强化学习的序列决策能力，构建编码器-解码器架构深度神经网络模型，使用强化学习 Actor-Critic 方法对模型进行训练，无需人为设计即可自动学习出优秀的策略。深度神经网络模型在编码器中的多头注意力层对待修复区域个体特征和待修复区域位置特征进行信息交互，通过自回归解码器输出下一步可能的修复方案。此外，本文还提出了多无人机协同的修复算法，在充分利用已有强化学习方法的基础上进一步优化求解效果。

实验结果表明，本文提出的算法优于基于贪婪策略的双层决策算法，且随着问题规模的增大优势更为明显；相较于贪婪策略，本文提出的单无人机训练、多无人机协同的策略具有更强寻优能力，在规定时间内进一步提升了求解效果，展示出神经网络算法强大的泛化与寻优能力，同时为求解大规模草原修复问题提供了新的思路和方法。

最后，在仿真结果与分析部分，由于项目时间限制，其中仍有众多方面可以值得扩展研究。比如：在给定的参数范围内，选择更多的启发式算法来作为对比算法，从而能对控制算法的性能有更公正的评估。或在本文提出的控制算法框架下，通过控制变量法改变草场规模、无人机初始能量等参数多次训练神经网络，而后通过大量的随机模拟取最佳、最差及平均等指标，来探讨算法的稳定性。

参考文献

[1] CHAPIN F S, SALA O E, HUBER-SANNWALD E. Global biodiversity in a changing environment: scenarios for the 21st century: Vol. 152[M]. Springer Science & Business Media, 2013.

[2] GIBSON D J. Grasses and grassland ecology[M]. Oxford University Press, 2009.

[3] REINERMANN S, ASAM S, KUENZER C. Remote sensing of grassland production and management—a review[J]. Remote Sensing, 2020, 12(12): 1949.

[4] STEFFEN W, RICHARDSON K, ROCKSTRÖM J, et al. Planetary boundaries: Guiding human development on a changing planet[J]. science, 2015, 347(6223): 1259855.

[5] ZENG Y, ZHANG R. Energy-efficient uav communication with trajectory optimization[J]. IEEE Transactions on wireless communications, 2017, 16(6): 3747-3760.

[6] NEX F, REMONDINO F. Uav for 3d mapping applications: a review[J]. Applied geomatics, 2014, 6: 1-15.

[7] SHIVGAN R, DONG Z. Energy-efficient drone coverage path planning using genetic algorithm[C]//2020 IEEE 21st International Conference on High Performance Switching and Routing (HPSR). IEEE, 2020: 1-6.

[8] ROSSELLO N B, CARPIO R F, GASPARRI A, et al. Information-driven path planning for uav with limited autonomy in large-scale field monitoring[J]. IEEE Transactions on Automation Science and Engineering, 2021, 19(3): 2450-2460.

[9] GANN G D, MCDONALD T, WALDER B, et al. International principles and standards for the practice of ecological restoration[J]. Restoration ecology, 2019, 27(S1): S1-S46.

[10] DORLING K, HEINRICHS J, MESSIER G G, et al. Vehicle routing problems for drone delivery[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016, 47(1): 70-85.

[11] KLAUS V H, SCHÄFER D, KLEINEBECKER T, et al. Enriching plant diversity in grasslands by large-scale experimental sward disturbance and seed addition along gradients of land-use intensity[J]. Journal of Plant Ecology, 2017, 10 (4): 581-591.

[12] FENG L, ONG Y S, TAN A H, et al. Memes as building blocks: a case study on evolutionary optimization+ transfer learning for routing problems[J]. Memetic Computing, 2015, 7: 159-180.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[14] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[J]. Advances in neural information processing systems, 2015, 28.

[15] BRESSON X, LAURENT T. The transformer network for the traveling salesman problem[A]. 2021. arXiv: 2103.03012.

[16] SUTTON R S, MCALLESTER D, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation[J]. Advances in neural information processing systems, 1999, 12.

[17] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine learning, 1992, 8: 229-256.

[18] KINGMA D P, BA J. Adam: A method for stochastic optimization[A]. 2014.

[19] AGGARWAL S, KUMAR N. Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges[J]. Computer communications, 2020, 149: 270-299.

[20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[A]. 2014.

[21] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[22] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[23] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[24] LIN M, CHEN Q, YAN S. Network in network[A]. 2013.

[25] EBERHART R C, SHI Y, KENNEDY J. Swarm intelligence[M]. Elsevier, 2001.

[26] HOLLAND J H. Genetic algorithms[J]. Scientific american, 1992, 267(1): 66-73.

[27] SAHU A, PANIGRAHI S K, PATTNAIK S. Fast convergence particle swarm optimization for functions optimization[J]. Procedia Technology, 2012, 4: 319-324.

[28] HU W, YEN G G. Adaptive multiobjective particle swarm optimization based on parallel cell coordinate system[J]. IEEE Transactions on Evolutionary Computation, 2013, 19 (1): 1-18.

[29] MAZYAVKINA N, SVIRIDOV S, IVANOV S, et al. Reinforcement learning for combinatorial optimization: A survey[J]. Computers & Operations Research, 2021, 134: 105400.