面向多目标救援的通信受限无人机集群分布式策略

俞汉清[1]，林艳[1,2]，贾林琼[1]，李强[3]，张一晋[1]
（1. 南京理工大学电子工程与光电技术学院，江苏 南京 210094；
2. 东南大学移动通信国家重点实验室，江苏 南京 210096；
3. 鹏城实验室，广东 深圳 518000）

摘 要：现有无人机集群的协同决策设计所依据的信息共享缺乏对无人机之间通信能力的合理假设。针对电量、载荷和路线约束下的无人机集群多目标救援问题，结合无人机飞行路线，考虑通信能力对无人机之间信息共享的限制。首先，将问题建模成部分可观测马尔可夫决策过程；然后，利用循环神经网络提出基于深度强化学习的能够适应通信拓扑结构不断变化的分布式救援策略。仿真结果表明，所提策略相较于其他策略在通信受限的情况下具有更佳的分布式救援性能，无人机数量和无人机通信能力需要依据救援场景进行联合设置方能达到无人机集群救援性能和使用成本的最佳折中。

关键词：无人机；多目标救援；马尔可夫决策过程；分布式策略；强化学习
中图分类号：TN911
文献标志码：A
doi: 10.11959/j.issn.2096−3750.2022.00284

A distributed strategy for the multi-target rescue using a UAV swarm under communication constraints

YU Hanqing[1], LIN Yan[1,2], JIA Linqiong[1], LI Qiang[3], Zhang Yijin[1]
1. School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
3. Peng Cheng Laboratory, Shenzhen 518000, China

Abstract: The current designs of the cooperative decision-making of an unmanned aerial vehicle (UAV) swarm usually adopt unreasonable assumptions on the communication ability between UAVs. Focusing on a multi-target rescue problem of a UAV swarm under constraints of energy, load and path, the limitation on the information sharing due to the communication constraints and the flight path of UAVs were taken into account. Firstly, the problem was formulated as a partially observable Markov decision process (POMDP). Then, a recurrent neural network was used to propose a deep-reinforcement-learning-based distributed rescue strategy, which is able to adapt to the changeable communication topology. Simulation results show that the proposed strategy outperforms other strategies under communication constraints, and further show that a careful joint setting of the size and communication ability of a UAV swarm is needed to achieve the best compromise between the UAV swarm rescue performance and the cost.

Key words: unmanned aerial vehicle, multi-target rescue, Markov decision process, distributed strategy, reinforcement learning

收稿日期：2021−10−13；**修回日期：**2022−06−15
通信作者：张一晋，yijin.zhang@gmail.com
基金项目：国家自然科学基金资助项目（No.62071236，No.62001225）；中央高校基本科研业务费资助项目（No.30920021127）；江苏省自然科学基金资助项目（No.BK20190454）；鹏城实验室重大攻关项目（No.PCL2021A15）；东南大学移动通信国家重点实验室开放研究基金资助项目（No.2022D07）

Foundation Items: The National Natural Science Foundation of China (No.62071236, No.62001225), The Fundamental Research Funds for the Central Universities of China (No.30920021127), The Natural Science Foundation of Jiangsu Province (No.BK20190454), The Major Key Project of PCL (No.PCL2021A15), The Open Research Fund of National Mobile Communications Research Laboratory, Southeast University (No.2022D07)

0 引言

相较于单无人机，无人机集群使用大量各司其职的无人机共同完成任务，有效弥补了单无人机能力有限的短板[1-2]。近年来，随着无人机飞控、通信以及协同决策技术[3-5]的快速发展，无人机集群协作正逐步替代单无人机应用于电力线检查、精准农业、海洋监测及抢险救灾等场景[6-8]。显而易见，协同决策需要借助无线通信技术提供的信息共享完成决策，因此，通信受限下的无人机集群协同决策成为工业界和学术界的研究热点[9-11]。

无人机集群集中式决策借助无线通信为某个控制单元收集无人机集群全局信息，并指派此控制单元对无人机集群进行统一的任务规划[12]，是最常用的无人机集群协同决策方式之一。文献[13]将部分可观察马尔可夫决策过程与顺次分配技术结合，设计了一种近似最优在线规划算法，最大化单个智能体对团队任务目标的边际贡献。文献[14]提出了一种改进模拟退火 K 均值算法，解决多无人机协同侦察任务分配问题。文献[15]基于满意决策设计了一种多无人机协同攻击目标分配算法。文献[16]基于马尔可夫决策过程和贝尔曼方程提出了一种无人机辅助充电算法。文献[17]基于深度强化学习设计了一种无人机集群轨迹优化算法。然而，集中式决策方式往往要求无人机装载成本较高的无线通信模块，要求控制单元具有强大的存储和计算能力，且一旦控制单元失效，整个无人机集群将面临无法完成任务的风险[18]。

与集中式决策不同，无人机集群分布式决策仅要求每个无人机基于本地信息分布式地做出决策，具有实时性强、灵活性高、鲁棒性好的优点[18-19]。文献[20]设计了一种基于深度强化学习的无人机集群洪水监测算法，使用无人机对洪水区域的局部观测做出决策。文献[21]基于遗传算法设计了一种多无人机航路规划算法，有效提高了无人机的监控效率。文献[22]针对无人机集群在动态环境下的任务分配问题，提出了一种动态蚁群分工模型。

尽管如此，以上研究均未考虑无人机通信能力对无人机集群分布式策略设计的影响。鉴于此，文献[13]基于蒙特卡洛树搜索设计了一种分布式侦察监视算法，但需要预先为每个无人机分配指定的侦察区域，从而基于无人机之间相对位置变化不大的事实假设无人机通信拓扑固定。文献[23]设计了一种基于虚拟长机状态估计的编队控制方法，但仍然假设无人机集群通信拓扑固定。上述文献理想化地假设某个无人机仅可与特定的另外一个无人机通信，未考虑由多种因素导致的通信中断、通信拓扑结构变化等情况。文献[24]针对无人机集群对多个目标进行跟踪的问题，基于多智能体部分可观测马尔可夫决策模型提出了一种基于最大共识协议的分布式决策方法，考虑了通信半径对通信拓扑的动态影响，但未同时考虑无人机电量和路线约束。

针对以往无人机集群协同决策研究对无人机之间通信情况假设过于理想化的问题，本文考虑无人机电量、载荷、路线以及通信能力约束下的无人机集群多目标协同救援场景，使用部分可观测马尔可夫决策过程（POMDP, partially observable Markov decision process）对此救援问题进行建模，并根据此 POMDP 建模和循环神经网络（RNN, recurrent neural network）提出基于集中式训练、分布式执行的深度强化学习算法的救援策略。仿真结果表明，所提策略相较于其他策略在通信受限情况下具有更佳的分布式救援性能，并显示了无人机数量和无人机通信能力需要依据救援场景进行联合设置方能达到无人机集群救援性能和使用成本的最佳折中。

本文的主要贡献如下：1) 相较于集中式的决策方法，本文基于 D3QN（D3QN, double dueling deep Q-network）技术提出集中式训练、分布式执行的救援策略，能够更好地应用于救援场景下常见的缺乏集中式决策者的情况。本文所考虑问题的联合动作空间很大，集中式算法难以求解，而本文算法在智能体较多时仍具有较低的复杂度。2) 相较于传统分布式的决策方法，本文综合考虑了在复杂环境中无人机通信拓扑改变的问题以及无人机自身的约束，提出了能够满足无人机电量、载荷和路线约束，适应通信拓扑结构不断变化的分布式救援策略。因此，本文所提策略能更好地应用于复杂救援环境。

1 系统模型

无人机集群多目标救援场景如图 1 所示，本文聚焦于具有众多不同救援紧迫程度的不同待救援位置的无人机集群救援场景。在此场景中，具有一定通信能力的各无人机携带有限的用于飞行的电量和有限的救援物资由固定起始位置出发。由于地形复杂、环境恶劣等因素（如存在墙、碎片或其他

障碍物），无人机必须经由一系列中间位置到达各待救援位置，在各位置均会遭遇某自然灾害威胁自身安全，到达任一待救援位置后即可投放物资，并且必须在自身电量值耗尽前回到起始位置。

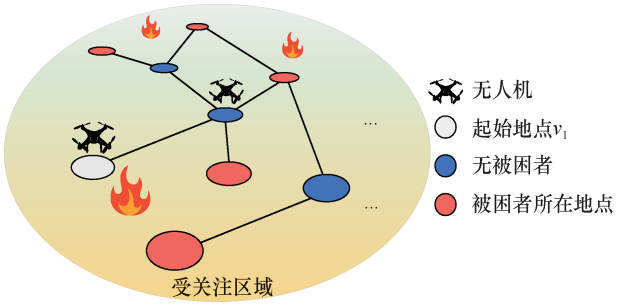

图 1　无人机集群多目标救援场景

考虑受限的无人机飞行路径，将无人机集群救援场景的物理区域定义为一个顶点个数为N的无向图$G \triangleq (V,E)$，其中，V表示顶点的集合，E表示边的集合。每个顶点表示一个地理位置，可能为待救援位置，亦可能为中间位置；图 1 的每一条边表示无人机可在其所连接的顶点之间进行往返。将被困者表示为$g \in \mathcal{G}$，将待救援位置表示为$v^g \in V^{\mathcal{G}} \subseteq V$，无人机起始位置表示为$v_1$。

由于不同地点的灾情不同，各待救援位置具有不同的救援紧迫程度，由函数$f: V^{\mathcal{G}} \to \mathbf{R}^+$确定。并且无人机成功投放一次物资后，被困者不再需要救援，此时$f(v^g)=0$。将时间划分为若干个时隙$t \in \mathcal{T} = \{1,2,\cdots,T\}$。无人机$z \in \mathcal{Z}$在时隙 1 从顶点$v_1$出发，携带电量为$D_z^{\mathrm{nrg}}$（单位：格），救援物资数量为$D_z^{\mathrm{sup}}$（单位：个）。在每个时隙初可选择停留于上一时隙初所在位置$v_n \in V$或者飞行至相邻位置$v_n' \in \mathrm{adj}_G(v_n)$，其中，$\mathrm{adj}_G(v_n)$表示顶点$v_n$的相邻顶点集合。如果无人机$z$位于待救援位置则可以选择是否投放 1 个物资，但投放成功率为p^{sup}，最后在电量耗尽前需要回到顶点v_1结束救援任务。在任一时隙，无人机停留不投放物资则消耗电量$c^{\mathrm{nrg,stay}}$（单位：格），停留并投放物资则消耗电量$c^{\mathrm{nrg,sup}}$（单位：格），移动则消耗电量$c^{\mathrm{nrg,mov}}$（单位：格）。需要指出，每个顶点均表示一定范围的地理空间，因此多个无人机可同时位于相同的顶点。

将除起点之外顶点$v_n \in V\backslash\{v_1\}$的$K_R^n$个威胁状态[13]表示为$e_R^{v_n} \in R^n \triangleq \{R_1^n, R_2^n, \cdots, R_{K_R^n}^n\}$，其取值在相邻时隙之间的变化相互独立且遵循有限状态马尔可夫链（FSMC, finite-state Markov chain）。无人机飞行至顶点$v_n \in V$时，一个时隙受到的伤害值由函数$h^n: R^n \to \mathbf{R}^+$确定。特别地，顶点$v_1$仅有一个威胁状态$R_1^1$，且$h^1(R_1^1) \equiv 0$。

在每个时隙初，无人机$z \in \mathcal{Z}$可观测其所处顶点的威胁状态，如果是待救援位置可额外得知该位置是否已成功救援。另外，无人机始终准确知道自己的位置信息，然而，由于救援区域环境恶劣，无人机的传感器可能因浓烟、大雨等因素的干扰而无法获得当前位置的威胁状态和救援状态信息。定义无人机在每时隙初能获得当前位置的威胁状态和救援状态信息的概率为p^{ans}。无人机尝试取得信息后，可以与离其距离小于或等于l^{com}的无人机进行无线通信[25]，告知它们对各顶点的最新观测信息、自身位置信息、自身剩余电量信息、自身剩余物资数量信息以及最近执行的任务，从而达到优化救援策略的效果。

基于上述模型，将无人机z在时隙t的位置表示为$v_{z,t}^{\mathrm{UAV}}$，剩余电量表示为$d_{z,t}^{\mathrm{nrg}}$，剩余物资数量表示为$d_{z,t}^{\mathrm{sup}}$，将时隙t位于顶点v_n的无人机个数表示为$N_t(v_n)$。$\beta_{z,t,g}=1$表示无人机z在时隙t向被困者g投放救援物资，$\beta_{z,t,g}=0$表示未投放。令$\boldsymbol{1}_X$为指示函数，当且仅当条件X为真时$\boldsymbol{1}_X=1$，否则$\boldsymbol{1}_X=0$。无人机集群救援问题可以描述为

$$
\begin{aligned}
&\max_{\{\beta_{z,t,g}\}_{\forall z\in\mathcal{Z},\forall t\in\mathcal{T},\forall g\in\mathcal{G}},\{v_{z,t}^{\mathrm{UAV}}\}_{\forall t,\forall z\in\mathcal{Z}}} \mathbb{E}\\
&\left[\sum_{t\in\mathcal{T}}\sum_{z\in\mathcal{Z}}\sum_{g\in\mathcal{G}}(1-\alpha)p^{\mathrm{sup}}\beta_{z,t,g}f(v^g)-\sum_{t\in\mathcal{T}}\sum_{v_n\in V}\alpha h^n(e_R^{v_n})N_t(v_n)\right]\\
&\text{s.t.}\quad \mathrm{C1}:\ d_{z,T}^{\mathrm{sup}}\geqslant 0,\quad \forall z\in\mathcal{Z}\\
&\qquad \mathrm{C2}:\ d_{z,T}^{\mathrm{nrg}}\geqslant 0,\quad \forall z\in\mathcal{Z}\\
&\qquad \mathrm{C3}:\ v_{z,1}^{\mathrm{UAV}}=v_{z,T}^{\mathrm{UAV}}=v_1,\quad \forall z\in\mathcal{Z}\\
&\qquad \mathrm{C4}:\ v_{z,t+1}^{\mathrm{UAV}}\in\mathrm{adj}_G(v_{z,t}^{\mathrm{UAV}})\bigcup\{v_{z,t}^{\mathrm{UAV}}\},\quad \forall z\in\mathcal{Z},\forall t\in\mathcal{T}\\
&\qquad \mathrm{C5}:\ \beta_{z,t,g}\in\{0,1\},\quad \forall z\in\mathcal{Z},\forall t\in\mathcal{T},\forall g\in\mathcal{G}\\
&\qquad \mathrm{C6}:\ \boldsymbol{1}_{v^g=v_{z,t}^{\mathrm{UAV}}}-\beta_{z,t,g}\geqslant 0,\quad \forall z\in\mathcal{Z},\forall t\in\mathcal{T},\forall g\in\mathcal{G}
\end{aligned}
\tag{1}
$$

其中，α是权重系数，目标函数是无人机集群的救援效用，表示无人机集群要在尽可能多地救援目标的同时减少自身受到的伤害，约束 C1 和 C2 分别表示无人机的载荷和电量约束，约束 C3 表示无人机要在电量耗尽前回到起点，约束 C4 表示无人机的

路线约束，约束 C5 表示无人机可选择是否向某个被困者空投物资，约束 C6 表示无人机仅能向当前所在位置空投物资。需要指出，各无人机仅能根据本地信息以及通信所获得的信息决定其自身行为，因此，此优化问题具有分布式部分观测性特点。

如式(1)所示，本文的主要目标是在无人机部分可观测性条件下以最大化长期救援效用为目标寻求电量、载荷和路线约束下的最优分布式救援策略。此研究目标具有以下挑战：1) 需要考虑多种约束的互相耦合对策略设计的影响；2) 需要考虑通信能力和各无人机实时位置对各无人机部分观测的影响；3) 需要考虑未知动态环境分布式设计带来的高计算复杂度问题。

相较于文献[13-17]提出的基于全局信息的集中式协同决策方案，分布式决策要求各无人机仅能根据自身所知信息以及环境反馈制定执行动作，因此在相同性能要求下具有更大的挑战。另外，尽管已有文献研究无人机集群分布式决策，但未综合考虑无人机自身约束和无人机间通信拓扑变化的影响。例如，文献[23]假设无人机集群的通信拓扑固定不变，无法适用于无人机通信拓扑频繁改变的救援场景；文献[24]考虑在已知无人机动态通信拓扑变化规律以最大化费希尔信息为优化目标，基于分布式 POMDP 理论设计了分布式协同决策方案，但该方案建模需要环境模型，因此不适用于求解未知动态环境下长期性能的优化问题，并且其最优策略求解复杂度高，因此不适用于求解本文所考虑的未知动态环境下联合状态及联合动作空间维度较高的多目标救援问题。

2 POMDP 建模

将上述无人机救援问题建模为如下 POMDP $\langle \mathcal{S},\mathcal{A}_s,\mathcal{P},\mathcal{O},\Omega,\mathcal{D},r\rangle$，其中，$\mathcal{S}$ 为联合状态空间，$\mathcal{A}_s$ 为联合动作空间，$\mathcal{P}$ 为状态转移概率集合，$\mathcal{O}$ 为联合观测空间，Ω 为观测概率集合，$\mathcal{D}$ 为约束集合，r 为奖励函数。

1) 状态：定义联合状态 s 为

$$\begin{aligned} s \triangleq [&(v_1^{\mathrm{UAV}},\cdots,v_{|\mathcal{Z}|}^{\mathrm{UAV}}),(d_1^{\mathrm{nrg}},\cdots,d_{|\mathcal{Z}|}^{\mathrm{nrg}}),(d_1^{\mathrm{sup}},\cdots,d_{|\mathcal{Z}|}^{\mathrm{sup}}),\\ &(e_s^{v^1},\cdots,e_s^{v^{|\mathcal{G}|}}),(e_R^{v_1},\cdots,e_R^{v_N})]\in\mathcal{S} \end{aligned} \tag{2}$$

其中，$e_s^{v^g}$ 是被困者 g 的救援状态，$e_s^{v^g}=0$ 表示被困者 g 已收到应急救援物资，$e_s^{v^g}=1$ 表示未收到救援物资。

2) 动作：定义无人机 z 的动作，a_z^{stay} 表示停留在当前位置不动；a_z^{sup} 表示向当前位置空投救援物资；$a_{v_n,z}^{\mathrm{mov}}$ 表示移动到顶点 v_n。其动作空间为

$$\mathcal{A}_{s,z} \triangleq \{a_z^{\mathrm{stay}},a_z^{\mathrm{sup}}\}\bigcup\{a_{v_n,z}^{\mathrm{mov}}:v_n\in \mathrm{adj}_G(v_z^{\mathrm{UAV}})\} \tag{3}$$

联合动作空间为 $\mathcal{A}_s \triangleq \prod_{z\in\mathcal{Z}}\mathcal{A}_{s,z}$。

3) 状态转移概率：由第 1 节系统模型可知，无人机当前位置 v_z^{UAV}、剩余电量 d_z^{nrg} 以及剩余物资 d_z^{sup} 随时间的变化均是确定性的、不受环境影响的，且仅由无人机 z 在时隙 t 的动作 $a_{z,t}$ 决定，可以得到

$$v_{z,t+1}^{\mathrm{UAV}}=\begin{cases} v_n, & a_{z,t}=a_{v_n}^{\mathrm{mov}} \\ v_{z,t}^{\mathrm{UAV}}, & \text{其他} \end{cases} \tag{4}$$

$$d_{z,t+1}^{\mathrm{nrg}}=d_{z,t}^{\mathrm{nrg}}-c^{\mathrm{nrg}}(a_{z,t}) \tag{5}$$

$$d_{z,t+1}^{\mathrm{sup}}=d_{z,t}^{\mathrm{sup}}-c^{\mathrm{sup}}(a_{z,t}) \tag{6}$$

其中，$v_{z,t}^{\mathrm{UAV}}$ 是无人机 z 在时隙 t 的位置，$d_{z,t}^{\mathrm{nrg}}$ 是无人机 z 在时隙 t 的剩余电量，$d_{z,t}^{\mathrm{sup}}$ 是无人机 z 在时隙 t 的剩余物资，$c^{\mathrm{nrg}}(a_{z,t})$ 是无人机执行动作 $a_{z,t}$ 消耗的电量，$c^{\mathrm{sup}}(a_{z,t})$ 是无人机执行动作 $a_{z,t}$ 消耗的物资数量。顶点 v_n 的威胁状态根据上文所述 FSMC 进行转移，定义转移概率矩阵 P_R^n 为

$$P_R^n \triangleq \begin{bmatrix} p_{11}^n & \cdots & p_{1K_R^n}^n \\ p_{21}^n & \cdots & p_{1K_R^n}^n \\ \vdots & \ddots & \vdots \\ p_{K_R^n1}^n & \cdots & p_{K_R^nK_R^n}^n \end{bmatrix} \tag{7}$$

其中，p_{ij}^n 表示顶点在当前时隙的威胁状态为 R_i^n 时，下一时隙的威胁状态为 R_j^n 的概率。考虑无人机投放的物资可能由于救援环境中的自然灾害遭遇破坏，以及被困者由于地形复杂等原因无法获得无人机投放的物资，定义被困者成功接收无人机投放物资的概率为 p^{sup}，即

$$p^{\mathrm{sup}} \triangleq \mathbb{P}(e_s^{v_g}=0,t=t_0+1\,|\,e_s^{v_g}=1,a_t=a^{\mathrm{sup}},t=t_0) \tag{8}$$

4) 观测：定义无人机 z 在时隙 t 直接收集的信息为随机变量 $I_{z,t}$，用 I^{null} 表示将无人机未收集到信息。根据系统模型可以得到，$\mathbb{P}(I_{z,t}=I^{\mathrm{null}})=1-p^{\mathrm{ans}}$，$\mathbb{P}(I_{z,t}\neq I^{\mathrm{null}})=p^{\mathrm{ans}}$。当 $I_{z,t}\neq I^{\mathrm{null}}$ 时，定义 $I_{z,t}$ 为

$$I_{m,t} \triangleq \begin{cases} \left(e_s^{v_g},e_R^{v_z^{\mathrm{UAV}}}\right), & v_{z,t}^{\mathrm{UAV}}=v^g \\ e_R^{v_z^{\mathrm{UAV}}}, & \text{其他} \end{cases} \tag{9}$$

考虑无人机之间的互相通信，定义无人机 $i \in \mathcal{Z}$ 在时隙 t 的观测为

$$o_{i,t} \triangleq I_{i,t} \cup \left\{ I_{j,t}, v_j^{\mathrm{UAV}}, d_j^{\mathrm{nrg}}, d_j^{\mathrm{sup}}, a_j^{\mathrm{last}} \mid l_{ij} < l^{\mathrm{com}} \right\} \in \mathcal{O}_i \quad (10)$$

其中，l_{ij} 为无人机 i 与无人机 $j \in \mathcal{Z}$ 的距离，a_j^{last} 为无人机 j 执行的上一个动作。定义联合观测空间为 $\mathcal{O} \triangleq \prod_{z \in \mathcal{Z}} \mathcal{O}_z$ 。

5) 观测概率：当观测 o 与当前状态 s 中相应部分一致时 $\Omega(o \mid s,a)=1$，否则 $\Omega(o \mid s,a)=0$ 。

6) 约束：由于无人机 z 在执行任务中消耗的电量和物资数量分别不能超过 D_z^{nrg} 和 D_z^{sup}，可以得到

$$\begin{cases} \sum_{t=1}^{T} c^{\mathrm{nrg}}(a_{z,t}) \leqslant D_z^{\mathrm{nrg}}, \quad \forall z \in \mathcal{Z} \\ \sum_{t=1}^{T} c^{\mathrm{sup}}(a_{z,t}) \leqslant D_z^{\mathrm{sup}}, \quad \forall z \in \mathcal{Z} \end{cases} \quad (11)$$

另外，由于无人机 z 需要在电量耗尽前返回顶点 v_1，所以无人机应时刻保证剩余的能量能够返回起始顶点 v_1，可以得到

$$c^{\mathrm{nrg}}(a_{z,t}) + c^{\mathrm{nrg,mov}} \mathrm{ShortestPath}(v_{z,t+1}^{\mathrm{UAV}}, v_1) \leqslant d_z^{\mathrm{nrg}}, \quad \forall z \in \mathcal{Z} \quad (12)$$

其中，$\mathrm{ShortestPath}(v_{z,t+1}^{\mathrm{UAV}}, v_1)$ 为从顶点 $v_{z,t+1}^{\mathrm{UAV}}$ 到顶点 v_1 的最小跳数。

7) 奖励：$r_z : \mathcal{S} \times \mathcal{A}_{s,z} \to \mathbb{R}$ 为无人机 z 的奖励函数。容易得到，当 i 架无人机同时救援被困者 g 时，被困者 g 的救援成功率为 $1-(1-p^{\mathrm{sup}})^i$ 。可见，每增加一架无人机带来的边际成功率递减。鉴于此，定义位于 v_n 的无人机 z 在联合状态 s 下执行动作 a_z 获得的奖励 $r_z(s,a_z)$ 为

$$r_z(s,a_z) = \begin{cases} (1-\alpha)(1-p^{\mathrm{sup}})^{n_z^{\mathrm{sup}}} p^{\mathrm{sup}} e_s^{v^g} f(v^g) - \alpha h^n(e_R^{v_n}), \\ \qquad\qquad v_z^{\mathrm{UAV}} = v^g, a_z = a_z^{\mathrm{sup}} \\ -\alpha h^n(e_R^{v_n}), \qquad \text{其他} \end{cases} \quad (13)$$

其中，n_z^{sup} 是同一时隙在顶点 v_n 执行空投任务，且编号属于 $\{1,2,\cdots,z-1\}$ 的无人机数量。

3　基于深度强化学习的分布式救援策略

定义一个智能体为一个移动物理实体，即一个无人机。根据上文所述 POMDP 模型，各智能体要进行决策不仅需要考虑当前的观测，而且要考虑各智能体已经执行的所有动作和历史的观测，从而各智能体能获得的未来奖励也取决于所有历史。传统的前馈神经网络（FNN, feedforward neural network）要求当前的输出仅与当前输入有关，是无记忆的，不具备处理历史信息的能力。不同于 FNN，RNN 引入了隐藏状态而具有记忆功能，从而使得当前的输出不仅与当前输入有关，而且与历史的输入有关，因此能够有效处理该部分可观测问题。下面基于深度循环 Q 学习设计通信受限下无人机集群分布式多目标救援策略。

3.1　算法架构

采用的深度强化学习算法架构[26]如图 2 所示。其中，RNN 以智能体的观测为输入，以智能体下一步

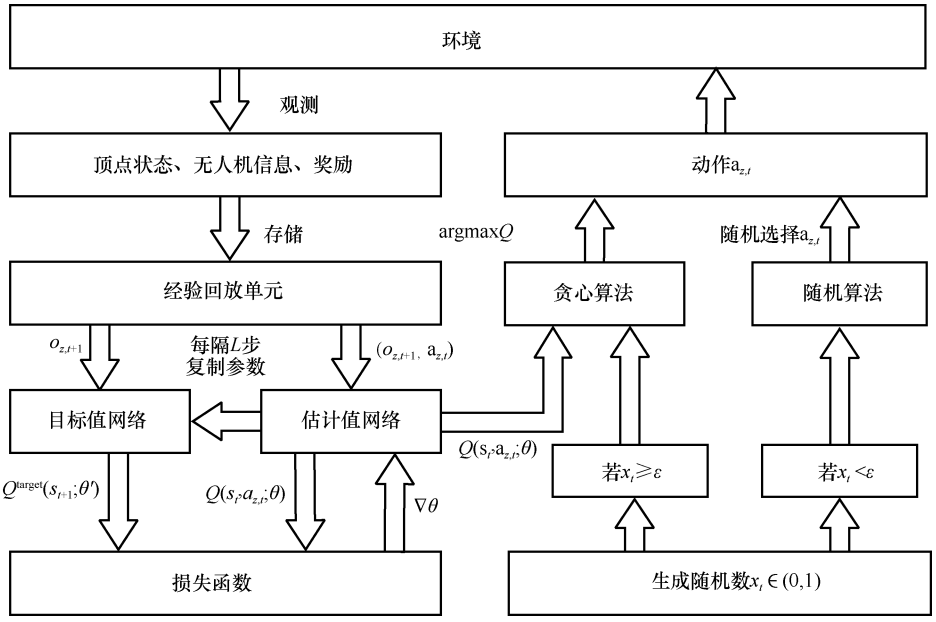

图 2　深度强化学习算法架构

动作的 Q 值为输出，并采用时间差分算法进行训练。尽管如此，由于神经网络的参数及其输出的 Q 值在每次训练后都将改变，并且联合状态空间亦很大，若直接使用时间差分算法进行训练将导致算法的性能在训练过程中难以稳定。鉴于此，本文算法包含了两个神经网络，即估计值网络和目标值网络[27]，而估计值网络在每训练 L 步后将复制目标值网络的参数。神经网络的训练目标是最小化以下损失函数

$$\mathcal{L}=\left(r_{z,t}+\gamma Q^{\text{target}}(s_{t+1})-Q(s_t,a_{z,t})\right)^2 \tag{14}$$

其中，γ 是折损因子，$r_{z,t}$ 是智能体 z 在时隙 t 获得的奖励，s_t 是时隙 t 的状态，$Q(s_t,a_{z,t})$ 和 $Q^{\text{target}}(s_{t+1})$ 分别是估计值网络和目标值网络的输出。

进一步，使用经验回放单元进行训练。当智能体与环境交互时，先将交互过程中产生的观测、动作和奖励存放在经验回放单元中，然后在训练时从中随机抽取多个样本进行训练。此种方式不仅可以重复利用过去的经验、提高样本的利用率、节省智能体与环境交互的时间，而且可以打破样本之间的相关性，让样本更趋近于独立同分布，从而更加有利于神经网络的训练。

在多智能体强化学习问题中，每个智能体能获得的奖励不仅取决于其自身的动作，也与其他智能体有关。若一个智能体的策略改变，将会影响其他智能体策略的选取，因此将给算法收敛带来困难。由于深度强化学习算法通常在仿真环境中训练，集中式训练—分布式执行的方法可以有效解决该问题[28]。此方法假设所有智能体在训练时能通过无限制通信共享各自的信息，从而能够使用联合观测和联合动作进行训练，进而充分考虑了智能体之间耦合关系对算法的影响；但在执行时只能根据通信受限下的自身局部观测做出决策。

为防止神经网络陷入局部最优解，本文采用 ε-greedy 算法寻求最优策略，其中 $0<\varepsilon<1$。智能体根据网络输出的 Q 值，每次以 ε 的概率从动作空间随机地选择一个动作；以 $1-\varepsilon$ 的概率用贪心算法选择最大化 Q 值的动作，即

$$a_{z,t}\in\underset{a_z\in\mathcal{A}_{s,z}}{\arg\max}\,Q(s_t,a_{z,t}) \tag{15}$$

3.2 网络结构

采用的网络结构如图 3 所示，门控循环单元（GRU, gated recurrent unit）作为 RNN 的一种形式，具有结构简单、训练效率高以及性能较好等优点，能够解决传统 RNN 梯度消失的问题。首先，将智能体的观测视为网络输入，并将所有输入归一化到[0, 1]或[−1, 1]；其次，全连接网络通过修正线性单元连接到 GRU；然后，GRU 的输出分为两个分支，一个分支为状态值函数 $u(s)$，另一个分支为动作优势值函数 $A(s,a)$ [29]；最后，网络输出各个动作的 Q 值 $Q(s,a)=u(s)+A(s,a)$。

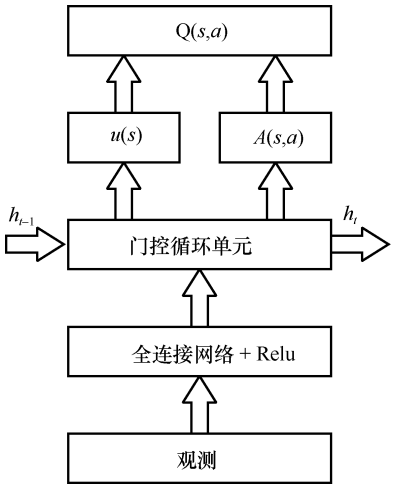

图 3 网络结构

4 仿真与分析

考虑救援场景及救援配置参数（携带电量和物资数）对无人机集群救援策略有重要影响，本节将分别对不同规模救援场景及救援配置进行所提策略性能的仿真验证，并探索无人机数量和通信距离对策略性能的影响。

设置每个无人机执行移动任务时每个时隙均消耗电量 $c^{\text{nrg,mov}}=2$，而执行其余任务时均消耗电量 $c^{\text{nrg,sup}}=c^{\text{nrg,stay}}=1$。设置每个无人机空投物资成功概率均为 $p^{\text{sup}}=0.8$，获得所在位置威胁状态和救援状态信息的概率均为 $p^{\text{ans}}=0.8$。设置权重系数 $\alpha=0.5$，被困者 g 的救援紧迫程度 $f(v^g)\in[30,70]$，且在区间内均匀分布。除顶点 v_1 外，设置每个顶点具有 R_1、R_2、R_3 3 种可能威胁状态，其对应的威胁值为 0、3、10。其中，设置初始威胁状态为等概率随机生成的，而设置威胁状态转移矩阵为

$$\boldsymbol{P}_R=\begin{bmatrix}0.8 & 0.2 & 0\\ 0.4 & 0.4 & 0.2\\ 0 & 0.2 & 0.8\end{bmatrix} \tag{16}$$

在本文策略所采用的深度强化学习算法中，设置学习率为 5×10^{-4}，折损因子 $\gamma=0.95$；为了让算

法在训练初期更侧重于探索，同时防止训练后期陷入局部最优解，设置 ε-greedy 算法参数为

$$\varepsilon = \max(0.02, 0.5 - 0.005 \times n^{\text{episodes}}) \tag{17}$$

其中，n^{episodes} 是已经经过的回合数。算法每一个回合训练一次，训练 L=200 次更新一次目标值网络。进一步，对深度强化学习算法进行 8 次独立训练，并且每训练 100 次进行 32 次仿真以获得某一种情况下的集群平均总奖励仿真值。

另外，本节将考虑以下 3 种策略作为对照策略：随机救援策略，在每个时隙等概率随机挑选一个可行动作；基于独立 Q 学习（IQL, independent Q-Learning）[30]的救援策略，直接将单智能体深度 Q 学习扩展到多智能体强化学习环境中，使得单智能体将其他智能体当作环境的一部分，从而能够独立进行学习；基于集中式强化学习的策略[17]，将整个无人机集群视为一个智能体，集中式地做出决策。

4.1　大规模救援场景仿真

本节考虑一种大规模救援区域的拓扑结构如图 4 所示，并且针对任意无人机 z 设置携带的最大电量 $D_z^{\text{nrg}} = 50$，最大物资数量 $D_z^{\text{sup}} = 10$。

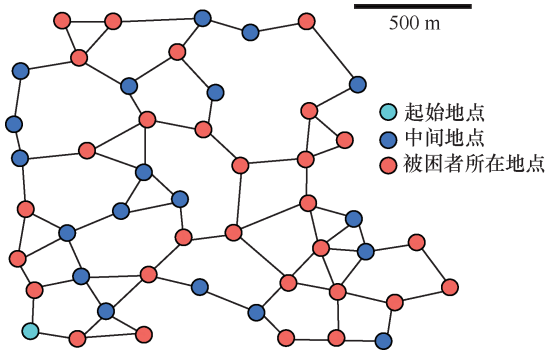

图 4　一种大规模救援区域的拓扑结构

大规模场景中不同无人机数量下的救援性能如图 5 所示，在 $l^{\text{com}} = 1\,000$ 的情况下比较了所提策略与对照策略按照 90%置信区间获得的大规模救援场景集群平均总奖励。正如预期，受益于图 2 所示的算法架构，本文策略能有效处理智能体之间的耦合关系，性能显著优于基于 IQL 的策略和随机策略。可以观察到，本文策略的救援性能随训练次数的增加逐渐增加，最后趋于收敛，在 $|\mathcal{Z}| = 3$ 时均获得了约 500 的奖励，较基于 IQL 的策略和随机策略奖励分别提升约 320 和 600。进一步可以发现，在本文策略中，无人机集群获得的总奖励随无人机数量的增加先上升后下降。出现此现象的原因是：给定救援场景，无人机数量适当增加可救援更多被困者，从而获得更高的奖励，但无人机数量过多将会使无人机集群受到的总伤害增加。此现象表明无人机数量优势在某些通信受限场景下并不能带来更多奖励。

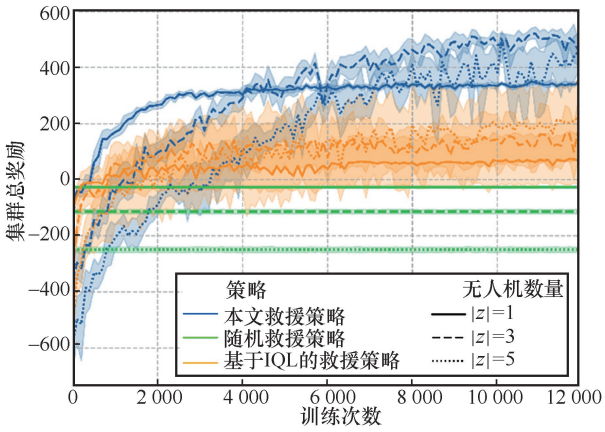

图 5　大规模场景中不同无人机数量下的救援性能（$l^{\text{com}} = 1\,000$ m）

大规模场景中不同通信距离下的救援性能如图 6 所示，进一步显示了无人机数量 $|\mathcal{Z}| = 3$ 时无人机通信距离对所提策略的大规模场景救援性能的影响。可以看出，所提策略在任何通信距离下都优于基于 IQL 的策略。进一步可以发现，当通信距离 $l^{\text{com}} = 0$ 时，无人机完全无法通信，只能得知自身信息，救援性能最差；当 l^{com} 增加时，无人机集群受益于通信获得的信息共享能够获得更高的集群总奖励；当 $l^{\text{com}} = 1\,000$ 时性能最好，获得的奖励约为 600。出现此现象的主要原因是：无法通信使无人机无法有效协作，只能通过有限的线索猜测其他无人机的行动，从而使得各无人机同一时刻的救援对象产生了较大重合，进而未能充分发挥各无人机的救援能力，反而增加了无人机集群遭遇威胁的风险，再次验证了通信距离对无人机集群决策的重要性。

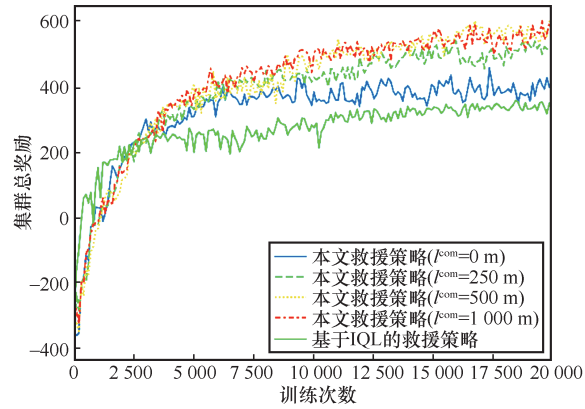

图 6　大规模场景中不同通信距离下的救援性能（$|\mathcal{Z}| = 3$）

4.2　小规模救援场景仿真

为了进一步验证无人机数量和无人机通信距离对集群的影响，一种小规模救援区域的拓扑结构如图 7 所示，设置任意无人机 z 携带的最大电量 $D_z^{\mathrm{nrg}}=30$，最大物资数量为 $D_z^{\mathrm{sup}}=5$。

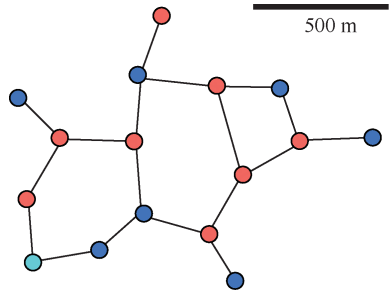

图 7　一种小规模救援区域的拓扑结构

小规模场景中不同无人机数量下的救援性能如图 8 所示，在 $l^{\mathrm{com}}=1\,000$ 的情况下比较了所提策略与对照策略获得的小规模救援场景集群平均总奖励。可以看出，本文策略仍然表现最佳，且在小规模场景中的收敛速度更快。相较于集中式强化学习策略[17]，本文所提策略的性能具有更好的稳定性，且收敛更快。其原因是高维联合动作空间会造成集中式强化学习策略难以训练收敛，而本文所提的分布式强化学习策略一方面利用了 D3QN，减少过估计的优势，使得性能更加稳定，另一方面采用分布式执行框架，使得各智能体只需要学习局部观测到局部动作空间的映射，显著提高收敛速度。其次，与图 5 类似，无人机集群获得的总奖励随无人机数量的增加先上升后下降，但此时最佳的无人机数量为 2。由此可见，执行的任务规模越小，所需要的最佳无人机数量就越少。在小规模场景中，仅需要少量无人机就能较好地完成救援任务，使用过多无人机反而会使集群受到的总伤害增加。

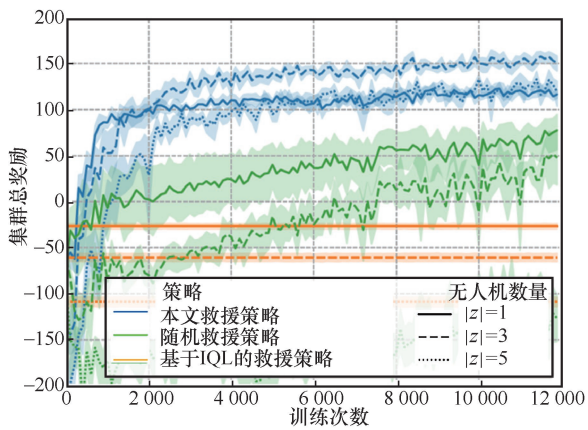

图 8　小规模场景中不同无人机数量下的救援性能（$l^{\mathrm{com}}=1\,000$ m）

小规模场景中不同通信距离下的救援性能如图 9 所示，显示了无人机数量 $|\mathcal{Z}|=3$ 时无人机通信距离对所提策略的小规模场景救援性能的影响。可以看出，本文策略仍在各种通信距离情况下都优于基于 IQL 的策略，性能随通信距离的增加有提高趋势，但并不明显。相较于路径多样、复杂的大规模场景而言，小规模救援场景的拓扑结构较为简单，可通行的路径少。无人机在这种情况下仅需要各自救援不同的区域并返回，与其他无人机的耦合小，因此即使不与其他无人机通信也能获得较好的救援效果。

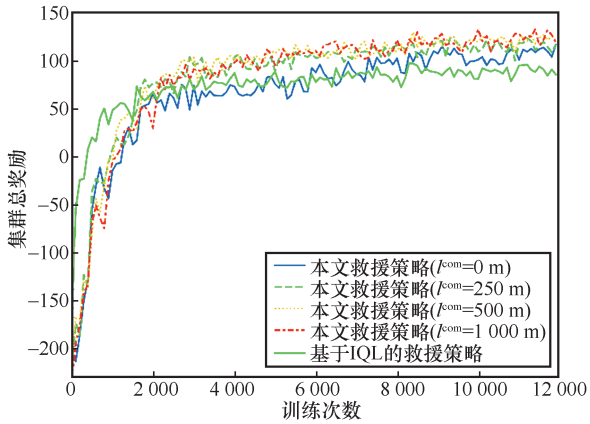

图 9　小规模场景中不同通信距离下的救援性能（$|\mathcal{Z}|=3$）

不同通信距离下使用本文所提策略的各无人机之间的平均距离如图 10 所示，显示了通信距离对所提策略中无人机之间平均距离的影响。可以看出，无人机之间的平均距离在小规模场景中随着通信距离变化几乎保持不变。可见，通信距离对集群决策的影响不大，进一步验证了图 9 的结论。在大规模场景中，无人机之间的平均距离随通信距离的增加而增加。当无人机完全不能通信时，平均距离最小。此时各无人机倾向于救援与其他无人机相同或相近的地点，容易产生重复救援。而当无人机通信距离较大时，无人机之间的平均距离也较大，此时无人机倾向于分别探索不同的区域，既避免了重复救援，也能通过通信使无人机集群获得更多环境的信息。

以上仿真结果显示了所提出的基于深度强化学习算法的分布式无人机救援策略在通信受限情况下的性能优势，并且表明无人机数量和无人机通信距离需要依据救援场景进行联合的合理设置方能达到最佳救援性能。

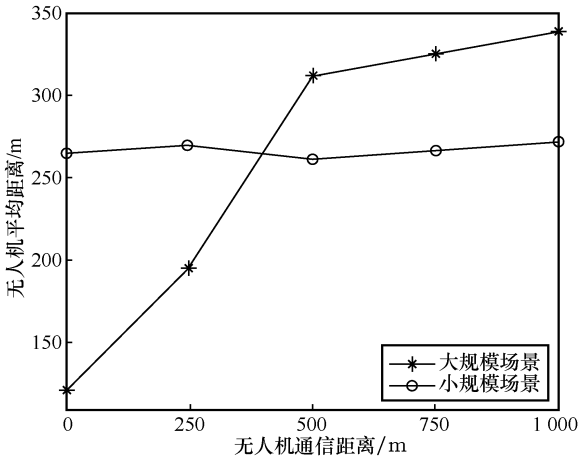

图 10　不同通信距离下使用本文所提策略的各无人机之间的平均距离（$|\mathcal{Z}|=3$）

5　结束语

针对电量、载荷、路线和通信能力约束下的无人机集群多目标救援问题，本文首先建立了 POMDP 模型，其次基于集中式训练—分布式执行的深度强化学习算法提出了能够适应通信拓扑结构不断变化的分布式救援策略。仿真结果表明，所提策略相较于其他策略在通信受限的情况下具有更佳的分布式救援性能，并显示了无人机数量和无人机通信能力需要依据救援场景进行联合设置方能达到无人机集群救援性能和使用成本的最佳折中。

参考文献：

[1] 韩亮, 任章, 董希旺, 等. 多无人机协同控制方法及应用研究[J]. 导航定位与授时, 2018, 5(4): 1-7.
HAN L, REN Z, DONG X W, et al. Research on cooperative control method and application for multiple unmanned aerial vehicles[J]. Navigation Positioning and Timing, 2018, 5(4): 1-7.

[2] HILDMANN H, KOVACS E. Review: using unmanned aerial vehicles (UAVs) as mobile sensing platforms (MSPs) for disaster response, civil security and public safety[J]. Drones, 2019, 3(3): 59.

[3] FAN B K, LI Y, ZHANG R Y, et al. Review on the technological development and application of UAV systems[J]. Chinese Journal of Electronics, 2020, 29(2): 199-207.

[4] CAMPION M, RANGANATHAN P, FARUQUE S. UAV swarm communication and control architectures: a review[J]. Journal of Unmanned Vehicle Systems, 2019, 7(2): 93-106.

[5] ZHOU Y K, RAO B, WANG W. UAV swarm intelligence: recent advances and future trends[J]. IEEE Access, 8: 183856-183878.

[6] BACCO M, CHESSA S, BENEDETTO M, et al. UAVs and UAV swarms for civilian applications: communications and image processing in the SCIADRO project[C]//Wireless and Satellite Systems, 2018: 115-124.

[7] RUETTEN L, REGIS P A, FEIL-SEIFER D, et al. Area-optimized UAV swarm network for search and rescue operations[C]//Proceedings of 2020 10th Annual Computing and Communication Workshop and Conference (CCWC). Piscataway: IEEE Press, 2020: 613-618.

[8] MATESE A, TOSCANO P, DI GENNARO S, et al. Intercomparison of UAV, aircraft and satellite remote sensing platforms for precision viticulture[J]. Remote Sensing, 2015, 7(3): 2971-2990.

[9] KIM K S, KIM H Y, CHOI H L. A bid-based grouping method for communication-efficient decentralized multi-UAV task allocation[J]. International Journal of Aeronautical and Space Sciences, 2020, 21(1): 290-302.

[10] LADOSZ P, OH H, ZHENG G, et al. Gaussian process based channel prediction for communication-relay UAV in urban environments[J]. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(1): 313-325.

[11] 宗群, 王丹丹, 邵士凯, 等. 多无人机协同编队飞行控制研究现状及发展[J]. 哈尔滨工业大学学报, 2017, 49(3): 1-14.
ZONG Q, WANG D D, SHAO S K, et al. Research status and development of multi UAV coordinated formation flight control[J]. Journal of Harbin Institute of Technology, 2017, 49(3): 1-14.

[12] 张可为, 赵晓林, 李宗哲, 等. 多无人机侦察任务分配方法研究综述[J]. 电光与控制, 2021, 28(7): 68-72, 82.
ZHANG K W, ZHAO X L, LI Z Z, et al. A review of multi-UAV reconnaissance mission assignment methods[J]. Electronics Optics & Control, 2021, 28(7): 68-72, 82.

[13] 陈少飞. 无人机集群系统侦察监视任务规划方法[D]. 长沙: 国防科学技术大学, 2016.
CHEN S F. Planning for reconnaissance and monitoring using UAV swarms[D]. Changsha: National University of Defense Technology, 2016.

[14] ZHAO J W, ZHAO J J. Study on multi-UAV task clustering and task planning in cooperative reconnaissance[C]//Proceedings of 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics. Piscataway: IEEE Press, 2014: 392-395.

[15] 黄捷, 陈谋, 姜长生. 无人机空对地多目标攻击的满意分配决策技术[J]. 电光与控制, 2014, 21(7): 10-13, 30.
HUANG J, CHEN M, JIANG C S. Satisficing decision-making on task allocation for UAVs in air-to-ground attacking[J]. Electronics Optics & Control, 2014, 21(7): 10-13, 30.

[16] SU C X, YE F, WANG L C, et al. UAV-assisted wireless charging for energy-constrained IoT devices using dynamic matching[J]. IEEE Internet of Things Journal, 2020, 7(6): 4789-4800.

[17] ABEDIN S F, MUNIR M S, TRAN N H, et al. Data freshness and energy-efficient UAV navigation optimization: a deep reinforcement learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(9): 5994-6006.

[18] WHITBROOK A, MENG Q G, CHUNG P W H. Reliable, distributed scheduling and rescheduling for time-critical, multiagent systems[J]. IEEE Transactions on Automation Science and Engineering, 2018, 15(2): 732-747.

[19] 杜永浩, 邢立宁, 蔡昭权. 无人飞行器集群智能调度技术综述[J]. 自动化学报, 2020, 46(2): 222-241.

DU Y H, XING L N, CAI Z Q. Survey on intelligent scheduling technologies for unmanned flying craft clusters[J]. Acta Automatica Sinica, 2020, 46(2): 222-241.

[20] BALDAZO D, PARRAS J, ZAZO S. Decentralized multi-agent deep reinforcement learning in swarms of drones for flood monitoring[C]//Proceedings of 2019 27th European Signal Processing Conference (EUSIPCO). Piscataway: IEEE Press, 2019: 1-5.

[21] 左益宏, 柳长安, 罗昌行, 等. 多无人机监控航路规划[J]. 飞行力学, 2004, 22(3): 31-34.
ZUO Y H, LIU C G, LUO C X, et al. Path planning for surveillance of multiple unmanned air vehicles[J]. Flight Dynamics, 2004,22(3): 31-34.

[22] WU H S, LI H, XIAO R B, et al. Modeling and simulation of dynamic ant colony's labor division for task allocation of UAV swarm[J]. Physica A: Statistical Mechanics and Its Applications, 2018, 491: 127-141.

[23] 成成, 张跃, 储海荣, 等. 分布式多无人机协同编队队形控制仿真[J]. 计算机仿真, 2019, 36(5): 31-37.
CHENG C, ZHANG Y, CHU H R, et al. Simulation of distributed cooperative formation control for multi-UAVs[J]. Computer Simulation, 2019, 36(5): 31-37.

[24] ZHAO Y Y, WANG X K, WANG C, et al. Systemic design of distributed multi-UAV cooperative decision-making for multi-target tracking[J]. Autonomous Agents and Multi-Agent Systems, 2019, 33(1/2): 132-158.

[25] FU X W, PAN J, WANG H X, et al. A formation maintenance and reconstruction method of UAV swarm based on distributed control[J]. Aerospace Science and Technology, 2020, 104: 105981.

[26] SAMVELYAN M, RASHID T, DE WITT C S, et al. The StarCraft multi-agent challenge[C]//Proceedings of AAMAS '19: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. 2019: 2186-2188.

[27] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[28] KRAEMER L, BANERJEE B. Multi-agent reinforcement learning as a rehearsal for decentralized planning[J]. Neurocomputing, 2016, 190: 82-94.

[29] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2016: 1995-2003.

[30] TAMPUU A, MATIISEN T, KODELJA D, et al. Multiagent cooperation and competition with deep reinforcement learning[J]. PLoS One, 2017, 12(4): e0172395.

[作者简介]

俞汉清（1999– ），男，南京理工大学电子工程与光电技术学院在读，CCF 学生会员，主要研究方向为深度学习、强化学习在无线网络的应用。

林艳（1990– ），女，博士，南京理工大学副教授，主要研究方向为 6G 无线资源分配、强化学习等。

贾林琼（1989– ），女，博士，南京理工大学讲师，主要研究方向为可见光通信与移动通信。

李强（1973– ），男，博士，鹏城实验室正高级工程师，主要研究方向为物联网与 5G/B5G。

张一晋（1982– ），博士，南京理工大学教授，主要研究方向为序列设计、无线网络与人工智能。