多无人机协同数据采集：面向物联网的信息新鲜度

Xijun Wang、Mengjie Yi、Juan Liu、Yan Zhang、Meng Wang 和 Bo Bai

arXiv:2303.01381v1 [cs.IT] 2 Mar 2023

摘要

在物联网（IoT）中保持信息新鲜度是一项至关重要且极具挑战性的问题。本文研究多无人机协同数据采集，旨在最小化总平均信息年龄（AoI）。我们考虑无人机的多种约束，包括运动学、能量、轨迹和碰撞避免，以优化数据采集过程。具体而言，每架无人机的机载能量有限，它们从各自的初始位置起飞，飞越传感器节点，并与其他无人机协作收集更新数据包。在规定的持续时间结束后，无人机必须在剩余能量非负的情况下抵达各自的最终目的地，以确保有足够的能量完成任务。合理设计无人机的轨迹以及传感器节点的传输调度，对于提升信息新鲜度至关重要。我们将多无人机数据收集问题建模为分散式部分可观测马尔可夫决策过程（Dec-POMDP），因为每架无人机都不了解环境的动态，且只能观测部分传感器。为应对这一问题带来的挑战，我们提出了一种多智能体深度强化学习（DRL）算法，采用集中式学习和分散式执行。除了奖励塑形，我们还使用动作掩码过滤无效动作，确保满足约束。仿真结果表明，与基线算法相比，所提出的算法能够显著降低总平均信息年龄；此外，采用动作掩码方法还可以提高所提出算法的收敛速度。

本研究的部分成果发表于 IEEE INFOCOM 研讨会， 2020 [1] 年 7 月。

X. Wang 任职于中国广州 510006 的中山大学电子与信息技术学院（电子邮箱：wangxijun@mail.sysu.edu.cn）。M. Yi 任职于中国西安 710071 的西安电子科技大学网络工程学院（电子邮箱：mjyi@stu.xidian.edu.cn）。J. Liu 任职于中国浙江 315211 的宁波大学电气工程与计算机科学学院（电子邮箱：eeliujuan@gmail.com）。Y. Zhang 任职于中国陕西西安 710071 的西安电子科技大学综合业务网国家重点实验室、信息科学研究所（电子邮箱：yanzhang@xidian.edu.cn）。M. Wang 和 B. Bai 任职于中国香港特别行政区香港科学园华为技术有限公司中央研究院2012实验室理论实验室（电子邮箱：wang.meng7@huawei.com；ee.bobbai@gmail.com）。

@ 2023年IEEE。允许个人使用本材料。所有其他用途均须获得 IEEE 的许可，包括为广告或促销目的重印或再版本材料、收集新编作品以转售或重新分发至服务器或列表，或在其他作品中重复使用本作品中任何受版权保护的组成部分。

索引词

A信息时效性、深度强化学习、物联网、无人机。

I. 引言

在地面网络基础设施有限的偏远地区，人们越来越需要支持物联网（IoT）应用的泛在连接，例如监测农业生长、追踪海洋生物以及监视边境地区等。为满足这一需求，无人机（UAV）凭借高度的机动性和灵活性，已成为一种很有前景的技术，可为服务不足的地区提供及时、灵活且弹性的服务 [2] [3]。具体而言，配备微处理器和无线收发器的无人机能够动态移动至物联网网络中的各个传感器，以收集或传播信息。凭借完全可控的移动性和较高的高度，无人机能够利用空地信道上的视距（LoS）无线通信，提高吞吐量并降低传输能耗。此外，无人机可按需部署，并可根据动态通信环境迅速调整位置，从而实现快速、灵活的重新配置。

近年来，无人机赋能的物联网网络凭借其诸多优势，受到了广泛关注并引发了大量研究兴趣。然而，现有的大多数研究都聚焦于优化传统性能指标，例如系统吞吐量、覆盖范围和时延 [4]–[8]。这些性能指标无法量化无人机赋能的物联网网络中信息的新鲜度，而信息新鲜度对于任务关键型应用至关重要 [9]，[10]，，例如森林火灾控制与扑灭，以及灾害救援监测与搜救。在这些应用中，根据从物联网设备收集的信息所做决策的准确性，很大程度上取决于信息的新鲜度。因此，人们采用信息年龄（AoI）来衡量接收端视角下信息的新鲜程度[11]–[13]。

由于无人机的机载能量有限，使用单架无人机很难，甚至不可能在严格的时间要求内收集大范围内所有传感器的数据。因此，有必要研究多无人机协同数据采集。然而，与单无人机赋能的物联网网络相比，多架无人机的存在带来了若干技术挑战。首先，由于无人机之间可能发生碰撞，在设计无人机轨迹时必须确保额外的避碰约束。其次，由于同时进行状态更新时存在相互干扰，

还必须仔细设计物联网设备的调度，以减少传输失败。第三，由于环境的动态变化受到所有无人机动作的影响，每架无人机所面临的环境都是非平稳的。最后，动作空间会随无人机数量呈指数增长，从而使寻找最优决策变得更加困难。该领域以往的研究依赖全局状态信息进行协调 [14], [15], 这可能导致显著的通信开销。因此，有必要进一步研究如何以去中心化方式利用多架无人机从传感器协作收集状态更新数据包，从而保持信息新鲜度。

受这些事实的启发，我们考虑没有地面网络基础设施的物联网网络，并研究多无人机协同数据采集以提升信息新鲜度的问题。在该模型中，每架无人机都作为一个独立的智能体。具体而言，每架无人机携带有限的机载能量，从基地出发，飞向传感器以收集状态更新数据包，并最终在固定时段内抵达目的地。这些无人机必须在飞行过程中避免碰撞，并在能量耗尽前抵达目的地。每个传感器节点都配备了电池，并且能够从周围环境中收集能量。传感器只有在拥有足够能量时才可用。我们联合设计无人机的轨迹和传感器调度，以最小化总平均信息年龄。本文的主要贡献总结如下。

- 鉴于缺乏地面网络基础设施、每架无人机的观测受限以及环境动态未知（包括传感器的能量收集速率和空地信道的视距概率），我们将多无人机数据收集问题建模为有限时域分散式部分可观测马尔可夫决策过程（Dec-POMDP）。

- 我们提出了一种基于多智能体深度强化学习（DRL）的集中训练、分散执行（CTDE）算法，以应对非平稳环境。具体而言，在训练阶段，无人机利用全局信息进行联合训练；而在执行阶段，每架无人机根据自身观测独立做出决策。此外，为确保满足运动学、能量和轨迹约束，我们使用动作掩码在训练和执行阶段过滤无效动作。

- 我们进行了大量仿真，以评估所提出算法的性能。结果表明，所提出的算法能够有效协调无人机的轨迹和传感器调度。与基线算法相比，总

平均信息年龄显著降低。

本文组织如下：第二节总结相关工作。第三节介绍系统模型与问题建模。第四节提出多智能体深度强化学习方法。第五节给出仿真结果并进行讨论。最后，第六节总结全文。

II. 相关工作

A. 单无人机数据收集

利用单架无人机（UAV）在物联网网络中收集新鲜数据已得到广泛研究。在 [16], 中，基于Q学习设计无人机的飞行轨迹，以最小化过期数据包的数量，同时考虑施加于每个传感器的信息年龄截止期限约束。在 [17], 中，提出了两个年龄最优数据收集问题，分别以最小化平均信息年龄或峰值信息年龄为目标；其中，传感器被划分为若干不重叠簇，无人机沿收集点飞行，从一组传感器中收集数据。在 [18], 中，针对无人机辅助无线供电物联网网络，同时设计无人机的充电时间和轨迹，以最小化平均信息年龄；无人机会在收集传感器的状态更新之前为其充电。在 [19], 中，考虑了无人机的三种数据采集策略，即悬停、飞行和混合策略，并采用动态规划算法优化传感器节点的访问顺序和数据采集策略，以最小化平均信息年龄，同时满足传感器节点的能量约束。上述研究 [16]–[19] 均涉及无人机仅从每个传感器收集一次状态更新数据包，而在 [20], [21], 中，无人机可以在给定时间段内多次访问每个传感器。在 [20], 中，提出了一种深度强化学习算法，以最小化归一化信息年龄加权和为目标，优化无人机的轨迹和传感器调度；在 [21] 中，又将深度强化学习与凸优化相结合，对该方法进行了扩展。

然而，这些先前的研究忽略了无人机的能耗。实际上，由于无人机的机载电池容量有限，在设计其飞行轨迹时，考虑无人机的能耗至关重要。在我们先前的工作 [1], 中，我们同时考虑能量和时间约束，联合设计了信息年龄最优的飞行轨迹和传输调度。在 [22], 中，无人机被用作移动中继，并通过联合优化无人机的飞行轨迹、传感器和无人机两端的能量分配及传输时长，使平均峰值信息年龄最小化。在 [23], 中，我们针对蜂窝连接无人机，从时间和能耗的角度研究了感知与通信权衡。在 [24], 中，我们联合优化无人机的飞行速度、悬停位置和带宽分配，以

最小化预期平均信息年龄、无人机推进能量以及物联网设备传输能量的加权和。

B. 多无人机数据收集

已有多项研究探讨了利用多架无人机进行协同数据收集。一项研究 [25], 使用无人机收集智能交通系统中车辆产生的数据，并采用深度确定性策略梯度算法优化无人机轨迹和车辆的传输调度，以最小化期望加权信息年龄总和。另一项研究 [26] 采用深度Q网络设计无人机的协同轨迹，以在能量和信息年龄约束下最大化总能量效率。尽管这两项研究都涉及多架无人机，但所有无人机的动作都由单个智能体控制，这可能导致动作空间非常庞大。

相比之下，研究 [14] 提出了一种分布式感知与发送协议，其中每架无人机都作为独立的智能体，并基于复合动作演员-评论家算法自主决策，以最小化信息年龄。研究 [15], 进一步扩展了这一方法，提出了一种基于深度确定性策略梯度的多无人机轨迹设计算法，以最小化无人机到设备通信的信息年龄。然而，在这些研究中，每架无人机都必须与基站通信，才能获取其他所有无人机状态的完整信息，从而产生大量信令开销。

在 [27], 中，使用了两架无人机，一架作为能量发射机，另一架作为数据收集器，并采用独立深度Q网络（IDQN）算法来设计无人机轨迹，以最小化信息年龄、增强向设备的能量传输并降低无人机能耗。使用 IDQN 时，每架无人机根据自身观测做出决策，但在非平稳的多智能体环境中，IDQN 不保证收敛。为解决这一问题，本文将协同数据收集问题建模为分散式部分可观测马尔可夫决策过程，并采用集中训练、分散执行框架。具体而言，通过利用全局状态信息对多架无人机进行集中训练以确保收敛；随后，每架无人机根据局部观测做出决策，无需中央实体收集并分发全局状态信息。

III. 系统模型与问题建模

本节首先介绍由多无人机支持的物联网网络。然后，给出能耗模型、空地信道模型以及信息年龄的演变。

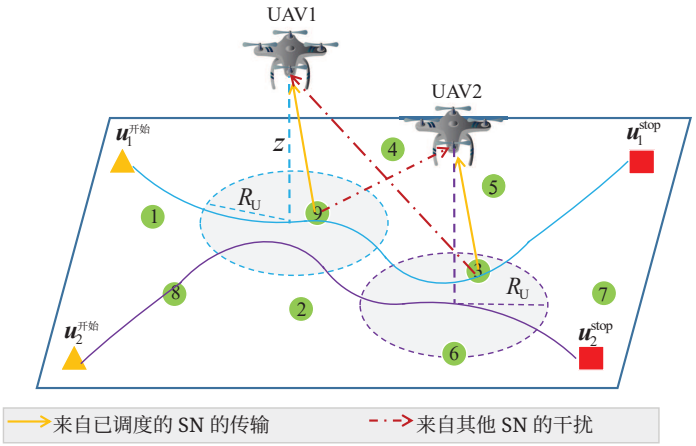


图 1. 多无人机支持的数据收集示意图。

最后，通过优化无人机的飞行轨迹和传感器的传输调度，构建数据收集问题。

A. 场景描述

如图 1 所示，我们考虑一个由 N 个随机分布的传感器节点（SN）组成的多无人机物联网网络。所有 SN 的集合记为 $\mathcal{N} = \{1, 2, \ldots, N\}$，SN $n \in \mathcal{N}$ 的位置在三维笛卡尔坐标系中表示为 $\boldsymbol{w}_n = (x_n, y_n, 0)$。我们部署了由 M 架旋翼无人机组成的集群，以协同收集 SN 的状态更新；这些状态更新将由无人机直接用于执行特定任务。所有无人机的集合记为 $\mathcal{M} = \{1, 2, \ldots, M\}$。我们假设系统为离散时间系统，时间被划分为长度相等的时隙，每个时隙持续 τ_0 秒。无人机需要在总计 T 个时隙内支持协同数据收集。[1] 具体而言，无人机 m（$m \in \mathcal{M}$）从初始位置 $\boldsymbol{u}_m^{\text{start}}$ 起飞，飞越各个 SN 并收集其状态更新。在第 T 个时隙结束时，无人机需要抵达最终目的地 $\boldsymbol{u}_m^{\text{stop}}$。

在时隙 t 开始时，无人机 $m \in \mathcal{M}$ 在三维笛卡尔坐标系中的位置记为 $\boldsymbol{u}_m(t) = (x_m(t)\ y_m(t)\ z_m(t))$，其中 $(x_m(t)\ y_m(t))$ 是无人机 m 位置在地面上的投影，$z_m(t)$ 是无人机 m 的高度。假设所有无人机均以相同的固定高度飞行，即 $z_m(t) = z, \forall t$。时隙 t 开始时，无人机 m 的速度用极坐标表示，记为 $\boldsymbol{v}_m(t) = (v_m^s(t)\ \varphi_m(t))$，其中

[1] T 的值取决于无人机的具体任务。

$v_m^s(t) = \|\boldsymbol{v}_m(t)\|$ 表示无人机 m 的速度，$\varphi_m(t)$ 表示速度方向，且 $0 \le \varphi_m(t) \le 2\pi$。假设在一个时隙内加速度恒定，即 $\boldsymbol{a}_{c,m}(t) = \dot{\boldsymbol{v}}_m(t)$，因此速度可按 $\boldsymbol{v}_m(t+1) = \boldsymbol{v}_m(t) + \boldsymbol{a}_{c,m}(t)\tau_0$ 更新。由于四旋翼无人机只需调节四个旋翼的转速便可轻松转向，我们假设无人机可以在时隙开始时立即改变方向，且在该时隙剩余时间内飞行方向保持不变。实际上，无人机受到运动学约束。具体而言，无人机 m 在时隙 t 的速度不得超过其最大值 $v_{\max}^s$，即 $v_m^s(t) \le v_{\max}^s$；无人机 m 在时隙 t 的转向角 $\triangle\varphi_m(t) = \varphi_m(t) - \varphi_m(t-1)$ 不得超过其最大值 $\triangle\varphi_{\max}$，即 $|\triangle\varphi_m(t)| \le \triangle\varphi_{\max}$。[2] 所有无人机的速度记为 $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M)$，其中 $\boldsymbol{v}_m = (\boldsymbol{v}_m(1)\ \boldsymbol{v}_m(2) \ldots, \boldsymbol{v}_m(T))$ 表示无人机 m 的速度序列。无人机 m 的飞行轨迹定义为其飞过的位置序列，即 $\boldsymbol{p}_m = (\boldsymbol{u}_m(1)\ \boldsymbol{u}_m(2) \ldots, \boldsymbol{u}_m(T))$，其中 $\boldsymbol{u}_m(1) = \boldsymbol{u}_m^{\text{start}}$ 且 $\boldsymbol{u}_m(T) = \boldsymbol{u}_m^{\text{stop}}$。为避免无人机之间发生碰撞，任意两架无人机在时隙 t 的距离不得小于安全距离，即对于任意 $m \ne m'(m, m' \in \mathcal{M})$，$d_{m,m'}(t) \ge d_{\text{safe}}$。

每架无人机在地面上都有一个半径为 $R_U, m \in \mathcal{M}$ 的覆盖区域，该半径固定不变，由传感器节点的发射功率、无人机的高度、天线增益和信道路径损耗决定。具体而言，我们将半径设为

$$R_U = \sqrt{d^2 - z^2} \qquad d =, \text{其}$$

$$\frac{c}{4\pi f_c}\left(\frac{P_c G_s G_u}{\xi_{\text{th}}\sigma^2 \eta_{\text{NLoS}}}\right)^{1/\varsigma} \ {}^{3} \qquad \overline{d}_{n\,m}\ t$$

中 。具体符号将在 III-C 小节中解释。令 $,(\,)$ 表示在时隙 t 开始时，传感器节点 n 与无人机 m 在地面上的投影之间的距离。如果 $\overline{d}_{n,m}(t) \le R_U$，则在时隙 t，传感器节点 n 位于无人机 m 的覆盖区域内。每个时隙中，每架无人机都必须决定调度其覆盖区域内的哪个传感器节点来更新其状态。令 $\boldsymbol{b}_m = (b_m(1)\ b_m(2) \ldots, b_m(T))$ 表示无人机 m 的传感节点调度向量，其中$b_m(t) = n$ 表示传感器节点 n 在时隙 t 向无人机 m 发送状态更新数据包，而$b_m(t) = 0$ 表示无人机 m 在时隙 t 未调度任何传感器节点。所有无人机的传感节点调度向量表示为 $\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M)$。

[2]这项工作还可以扩展到无人机的三维移动性。具体而言，无人机的速度表示为$\boldsymbol{v}_m(t) = (v_m^s(t)\varphi_{p,m}(t)\varphi_{a,m}(t))$，其中 $v_m^s(t) = \|\boldsymbol{v}_m(t)\|$ 表示无人机的速度，m 且 $(\varphi_{p,m}(t)\,\varphi_{a,m}(t))$ 表示无人机的方向。$\varphi_{p,m}(t)$ 是相对于正 z 轴的极角，其中 $0 \le \varphi_{p,m}(t) \le \pi$；$\varphi_{a,m}(t)$ 是在 xy 平面内相对于正 x 轴的方位角，其中 $-\pi \le \varphi_{a,m}(t) \le \pi$。速度和转向角的运动学约束与二维移动性中的约束类似。此外，无人机的高度限制在 $z_{\min}$ 至 $z_{\max}$ 之间。

[3]为简单起见，在本研究中，我们将覆盖区域视为半径固定为 R_U 的圆形，该半径由非视距路径损耗系数 η_{NLoS} 确定。这是因为非视距条件通常会导致更高的路径损耗，而在此圆形区域内具有视距信道的任何次级用户也会被无人机覆盖。然而，我们承认，同时考虑视距和非视距路径损耗对于更准确地表示覆盖区域至关重要。虽然我们的方法适用于基础评估，但还可以进一步扩展为覆盖区域可变的更一般模型。

B. 能量模型

1）无人机能量模型： 我们假设所有无人机具有相同数量的初始机载能量，记为 $E_{\max}$。无人机的主要能耗包括通信能量和推进能量。由于每架无人机都充当状态更新数据包的接收端，其通信能耗远小于推进能耗，因此可以忽略。 无人机 m E_m（t） 在时隙 t中的推进能耗可表示如下 [7]:

$$E_m(t) = \tau_0 n_r \left[\frac{\sigma}{8} \left(\frac{T_m^{\mathrm{h}}(t)}{c_T \rho A} + 3\left(v_m^s(t)\right)^2 \right) \sqrt{\frac{T_m^{\mathrm{h}}(t)\rho c_s^2 A}{c_T}} + \frac{1}{2} d_0 \rho c_s A \left(v_m^s(t)\right)^3 + (1+c_f) T_m^{\mathrm{h}}(t) \left(\sqrt{\frac{\left(T_m^{\mathrm{h}}(t)\right)^2}{4\rho^2 A^2} + \frac{\left(v_m^s(t)\right)^4}{4}} - \frac{\left(v_m^s(t)\right)^2}{2} \right)^{\frac{1}{2}} \right], \tag{1}$$

其中， n_r 为旋翼数量， σ 为局部桨叶剖面阻力系数， c_T 表示基于桨盘面积的推力系数， ρ 表示空气密度， A 和 c_s 分别表示每个旋翼的桨盘面积和旋翼实度， d_0 为每个旋翼的机身阻力比， c_f为诱导功率增量修正因子， T_m^{h}（t）为每个旋翼的推力。为便于说明，我们仅考虑速度方向上的直线加速度，并忽略垂直于速度方向的加速度分量 [7]。因此，每个旋翼的推力可表示为

$$T_m^{\mathrm{h}}(t) = \frac{1}{n_r} \left[\left(W a_{c,m}(t) + \frac{1}{2} \rho \left(v_m^s(t)\right)^2 S_{FP} \right)^2 + (Wg)^2 \right]^{1/2}, \tag{2}$$

其中，$a_{c,\ m}$（t） $=$ （v_m^s（$t+1$） $-v_m^s$（t）） $/\tau_0$， S_{FP}表示机身等效平板面积，W 表示无人机的重量， g 为重力加速度。

2) 传感器节点能量模型： 我们假设每个传感器节点都配有一块有限容量的电池 $E_{\max}^{\mathrm{sn}}$，并且能够从周围环境中采集能量。令 $E_n^{\mathrm{sn}}(t)$ 表示时隙 t开始时传感器节点 n 的电池电量。不失一般性，我们假设所有传感器节点的电池在初始时均已充满，即 $E_n^{\mathrm{sn}}(1) = E_{\max}^{\mathrm{sn}}$ 对任意 $n \in \mathcal{N}$ 均成立。我们还假设传感器节点能够持续从太阳能和风能等来源采集能量 [28]–[30]。然而，由于周围环境中的能量受天气影响且不稳定，我们将每个传感器节点采集到的能量建模为参数为 λ_n 的独立伯努利过程。这意味着在每个时隙内，能量到达传感器节点 n的概率为 λ_n 。此外，由于传感器节点用于采集能量的可再生能源发电装置（例如太阳能电池板）与其收发器相互独立，传感器节点可以同时

采集能量并传输信息。我们用指示变量 $\kappa_n(t)$ 表示能量到达，其中 $\kappa_n(t) = 1$ 表示传感器节点 n 在时隙 t 采集到 E_n^{har}能量，κ_n而$t(= 0)$ 则表示未采集到能量。我们假设已调度的传感器节点在一个时隙内以固定功率 P_c发送状态更新数据包，因此传感器节点 n 进行状态更新所需的能量为$E_n^{\text{c}} = P_c\tau_0$。令 $\zeta_n(t)$ 为一个指示变量，用于表示传感器节点 n 是否在时隙t被调度发送，其中 $\zeta_n(t) = 1$ 表示其被任意无人机调度，而 $\zeta_n(t) = 0$ 则表示未被调度。具体而言，有$\zeta_n(t) = 1 - \prod_{m=1}^{M} \mathbf{1}(b_m(t) \neq n)$。值得注意的是，只有当传感器节点的电池电量充足时，才会对其进行调度，即 $E_n^{\text{sn}}(t) \geq E_n^c$。因此，传感器节点 n 的电池电量动态变化如下:

$$E_n^{\text{sn}}(t+1) = \begin{cases} \min\{E_n^{\text{sn}}(t) + \kappa_n(t)E_n^{\text{har}} - E_n^{\text{c}}, E_{\max}^{\text{sn}}\}, & \text{if } \zeta_n(t) = 1, \\ \min\{E_n^{\text{sn}}(t) + \kappa_n(t)E_n^{\text{har}}, E_{\max}^{\text{sn}}\}, & \text{otherwise.} \end{cases} \tag{3}$$

C. 空地信道

环境中的障碍物会影响空地（A2G）信号传播。根据具体的传播环境，空地信道可以是视距（LoS）或非视距（NLoS）信道。然而，在实际场景中，通常无法获知障碍物的确切位置、高度和数量。无人机与特定 SN 之间存在视距的概率是给定的。令 $d_{n,\ m}(t)$ 表示时隙 t 中 SNn 与无人机 m 之间的欧几里得距离。那么，SN n 与无人机 m 之间的视距概率为 [31]

$$p_{n,m}^{\text{LoS}}(t) = \frac{1}{1 + \beta_0 \exp\left(-\beta_1\left(\frac{180}{\pi}\arcsin\left(\frac{z}{d_{n,m}(t)}\right) - \beta_0\right)\right)}, \tag{4}$$

其中，β_0 和 β_1 是由环境决定的常数。那么，时隙 t 中 SNn 与无人机 m 之间的路径损耗可表示为

$$PL_{n,m}(t) = \begin{cases} \left(\frac{4\pi f_c d_{n,m}(t)}{c}\right)^{\varsigma} \eta_{\text{LoS}}, & \text{w.p. } p_{n,m}^{\text{LoS}}(t), \\ \left(\frac{4\pi f_c d_{n,m}(t)}{c}\right)^{\varsigma} \eta_{\text{NLoS}}, & \text{otherwise.} \end{cases} \tag{5}$$

其中，ς 是路径损耗指数，f_c 是载波频率，c 表示真空中的光速，η_{LoS} 和 η_{NLoS} ($\eta_{\text{NLoS}} > \eta_{\text{LoS}} > 1$) 分别是视距链路和非视距链路的额外路径损耗系数。我们假设无人机在调度之前不知道信道状态是视距还是非视距，也不知道视距概率。

假设每个状态更新数据包都在一个时隙内传输。与 [32], 中一样，我们选择时隙长度 τ_0，使得无人机与传感器节点之间的距离在每个时隙内近似保持不变。例如，τ_0 可以选择为满足

$\tau_0 v_{\max}^s \ll z$。由于多架无人机在同一频段收集传感器节点的状态更新，在同一时隙内传输的传感器节点可能会相互干扰。我们假设传感器节点和无人机均配备全向天线，其天线增益分别记为 G_s 和 G_u。不失一般性，设 $G_s = G_u = 0$ dB。值得注意的是，本研究可扩展至采用定向天线的场景。于是，传感器节点 n 与无人机 m 之间空地信道的信干噪比为

$$\xi_{n,m}(t) = \frac{P_c G_s G_u \gamma_{n,m}(t)}{\sigma^2 + I_{n,m}(t)}, \tag{6}$$

其中，P_c 是每个传感器节点的传输功率，$\gamma_{n,m}(t) = \frac{1}{PL_{n,m}(t)}$ 是信道增益，σ^2 是噪声功率。在式（6）中，$I_{n,m}(t) = \sum_{n' \in \mathcal{N}_{-m}(t)/n} P_c \gamma_{n',m}(t)$ 是总干扰功率，其中 $\mathcal{N}_{-m}(t)$ 是时隙 t 中由其他无人机 m'（即 $m' \in \mathcal{M}$、$m' \neq m$）调度的所有传感器节点的集合。若信干噪比 $\xi_{n,m}(t)$ 不低于阈值 ξ_{th}，则传感器节点 n 的数据包在时隙 t 被无人机 m 成功接收；否则，接收失败。

D. 信息年龄

我们假设无人机直接利用收集到的状态更新进行后续决策，如 [21] 所示，并使用信息年龄衡量无人机处信息的新鲜度。具体而言，传感器节点 n $\delta_n(t)$ 的信息年龄定义为任一无人机最近一次接收到的状态更新自生成以来经过的时间。我们考虑按需生成策略，即传感器节点每当被调度时便生成一个状态更新数据包。状态更新数据包包含有关所关注物理过程状态的信息，以及对应的时刻（即样本生成的时间）。如果无人机 m 在时隙 t 调度传感器节点 n 更新其状态，且传输成功（即 $\xi_{n,m}(t) \geq \xi_{\text{th}}$），则传感器节点 n 的信息年龄将减小至一（因为数据包传输占用一个时隙）；否则，信息年龄将增加一。于是，传感器节点 n 的信息年龄动态可表示为

$$\delta_n(t+1) = \begin{cases} 1, & \text{if } \zeta_n(t) = 1 \text{ and } \xi_{n,m}(t) \geq \xi_{\text{th}}, \\ \min(\delta_n(t) + 1, \delta_{\max}), & \text{otherwise}, \end{cases} \tag{7}$$

其中 $\delta_{\max}$ 是信息年龄的最大值。我们设置信息年龄上限的原因在于，过时严重的状态更新对于时间关键型物联网应用毫无用处 [9],[33]。此外，$\delta_{\max}$ 的取值取决于具体应用，可以较小，也可以较大。我们还假设每个传感器节点的信息年龄由该传感器节点自身维护，因为无人机会在成功传输后立即确认。

E. 问题构建

在这个多无人机数据采集问题中，我们的目标是通过联合优化无人机的轨迹和传感器节点的调度，最小化时间平均总期望信息年龄，即各传感器节点对应的期望信息年龄之和的时间平均值。该优化问题可表示如下：

$$\text{P1:} \min_{\boldsymbol{V},\boldsymbol{B}} \quad \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{n=1}^{N}\delta_n(t)\right], \tag{8a}$$

$$\text{s.t.} \quad d_{m,m'}(t) \geq d_{\text{safe}}, \forall m, m' \in \mathcal{M}, m \neq m', \tag{8b}$$

$$0 \leq v_m^s(t) \leq v_{\max}^s, \forall m \in \mathcal{M}, \tag{8c}$$

$$0 \leq \varphi_m(t) \leq 2\pi, \forall m \in \mathcal{M}, \tag{8d}$$

$$|\triangle\varphi_m(t)| \leq \triangle\varphi_{\max}, \forall m \in \mathcal{M}, t \geq 2, \tag{8e}$$

$$\zeta_n(t) \leq \mathbf{1}(E_n(t) \geq E_n^c, \overline{d}_{n,m}(t) \leq R_{\text{U}}), \forall n \in \mathcal{N}, \exists m \in \mathcal{M}, \tag{8f}$$

$$\sum_{t=1}^{T} E_m(t) \leq E_{\max}, \forall m \in \mathcal{M}, \tag{8g}$$

$$\boldsymbol{u}_m(1) = \boldsymbol{u}_m^{\text{start}}, \boldsymbol{u}_m(T) = \boldsymbol{u}_m^{\text{stop}}, \forall m \in \mathcal{M}. \tag{8h}$$

避碰约束由式（8b）给出。速度约束（8c）以及方向约束（8d）和（8e）确保无人机满足运动学约束。传感节点调度约束（8f）确保只有当电池剩余能量足以进行传输，且该传感器节点位于至少一架无人机的覆盖区域内时，才对其进行调度。无人机能量约束（8g）确保每架无人机在 T个时隙结束之前不会耗尽能量。轨迹约束（8h）保证每架无人机在第一个时隙从其初始位置出发，并在 T个时隙结束时到达最终位置。由于环境动态未知、能耗具有因果性，且无人机的观测能力有限，求解上述随机优化问题极具挑战性。需要掌握所有环境动态信息的传统优化方法并不适用于该问题。因此，我们提出一种多智能体深度强化学习方法，联合设计轨迹规划和传输调度。

IV. 多智能体深度强化学习方法

在本节中，我们考虑多无人机数据采集问题，其中一组无人机协同从传感器节点收集状态更新。每架无人机作为一个独立智能体运行，

在不与其他无人机共享信息的情况下做出自己的决策。由于每架无人机都不知道其他无人机的动作，因此只能观测其附近传感器节点的信息年龄和电池电量。为此，我们首先将多无人机数据采集问题建模为分散式部分可观测马尔可夫决策过程，使多个在不同智能体上运行的独立部分可观测马尔可夫决策过程能够彼此独立地运行，同时共同实现依赖于所有智能体行为的目标函数 [34]。然后，我们提出一种基于多智能体深度强化学习的算法，以获得每架无人机的策略。

A. 分散式部分可观测马尔可夫决策过程的建模

在分散式部分可观测马尔可夫决策过程中，多个无人机与由状态集合 $\mathcal{S}$ 表示的环境进行交互。每架无人机均由专属智能体控制，目标函数取决于所有智能体的动作。在每个时隙，每架无人机都会接收自身的观测并采取相应的动作。随后，每架无人机都会获得相应的代价，环境则转移到新的状态。下面将更详细地定义分散式部分可观测马尔可夫决策过程的状态、观测、动作、状态转移和代价。

1) 状态： 时隙 t 中的状态定义为 $\boldsymbol{s}(t) = (\boldsymbol{u}(t)\ \boldsymbol{\delta}(t)\ \boldsymbol{V}(t)\ \boldsymbol{\varphi}(t-1)\ \boldsymbol{E}^{\text{sn}}(t)\ \boldsymbol{\phi}(t)\ \boldsymbol{\psi}(t)) \in \mathcal{S}$，由以下七个部分组成：

- $\boldsymbol{u}(t) = (\boldsymbol{u}_1(t)\ \boldsymbol{u}_2(t) \ldots, \boldsymbol{u}_M(t))$ 表示所有无人机在时隙 t 开始时的位置。
- $\boldsymbol{\delta}(t) = (\delta_1(t)\ \delta_2(t) \ldots, \delta_N(t))$ 是时隙 t 所有传感器节点的信息年龄。
- $\boldsymbol{V}(t) = (v_1^s(t)\ v_2^s(t) \ldots, v_M^s(t))$ 是时隙 t 开始时所有无人机的速度。
- $\boldsymbol{\varphi}(t-1) = (\varphi_1(t-1)\ \varphi_2(t-1) \ldots, \varphi_M(t-1))$ 是时隙 $t-1$ 所有无人机的速度方向。
- $\boldsymbol{E}_{\text{传感器节点}}(t) = (E_{\text{传感器节点1}}(t)\ E_{\text{传感器节点2}}(t) \ldots, E_{\text{传感器节点}N}(t))$ 是时隙 t 所有传感器节点的电池电量。
- $\boldsymbol{\phi}(t) = (\phi_1(t)\ \phi_2(t) \ldots, \phi_M(t))$ 表示所有无人机在时隙t开始时的时间差。具体而言，$\phi_m(t) = T - t + 1 - T\text{req}_m(t)$ 是飞行航程剩余时间（即 $T-t+1$）与无人机 m 到达最终目的地所需时间（即 $T\text{req}_m(t)$）之间的差值。我们假设，当无人机 m 的速度为零时，其速度方向可以是 $(0, 2\pi)$ 之间的任意值。计算 $T\text{req}_m(t)$ 有两种情况。第一种情况下，如果无人机 m 可以立即将速度方向转向目的地，即$\triangle\varphi\text{stop}_m(t) \leq \triangle\varphi\ \text{max}$ 或 $v_m^s(t) = 0$，那么 $T\text{req}_m(t)$ 由加速至最大速度所需的时隙数，以及随后以最大速度飞往目的地所需的时隙数组成。另一种情况下，如果无人机 m 无法立即将速度方向转向目的地，我们规定它先将速度降至零，然后将速度方向转向目的地并加速至最大速度，

之后以最大速度飞往目的地。因此， $T_m^{\text{req}}(t)$ 由将速度减至零、加速至最大速度，以及随后以最大速度飞往目的地所需的时隙数组成。为简化计算和表述，当无人机 m 无法立即将速度方向转向目的地时，我们假设其速度方向与最终位置的方向相反。这样，我们可以得到 $T_m^{\text{req}}(t)$ 的上界。因此， $T_m^{\text{req}}(t)$ 可表示为

$$T_m^{\text{req}}(t)=\begin{cases}1+\lceil\frac{\left\|\boldsymbol{u}_m(t)-\boldsymbol{u}_m^{\text{stop}}\right\|-\triangle_{m,1}(t)}{v_{\max}^s\tau_0}\rceil, & \text{if } \triangle\varphi_m^{\text{stop}}(t)\le\triangle\varphi_{\max} \text{ or } v_m^s(t)=0,\\ 2+\lceil\frac{\left\|\boldsymbol{u}_m(t)-\boldsymbol{u}_m^{\text{stop}}\right\|+\triangle_{m,2}(t)-\triangle_{m,3}}{v_{\max}^s\tau_0}\rceil, & \text{otherwise,}\end{cases} \tag{9}$$

其中， $\lceil\cdot\rceil$ 表示向上取整， $\triangle\varphi_m^{\text{stop}}(t)=|\varphi_m(t)-\varphi_m^{\text{stop}}(t)|$ 表示无人机 m $\varphi_m(t)$ 的速度方向与从无人机 m 当前位置指向目的地 $\varphi_m^{\text{stop}}(t)$ 的向量方向之间的角度差的绝对值。 $\triangle_{m,1}(t)=\frac{v_{\max}^s+v_m^s(t)}{2}\tau_0$ 表示无人机 m 从 $v_m^s(t)$ 到最大速度 $v_{\max}^s$ 的飞行距离， $\triangle_{m,2}(t)=\frac{v_m^s(t)}{2}\tau_0$ 表示无人机 m 从 $v_m^s(t)$ 到零速度的飞行距离，而 $\triangle_{m,3}=\frac{v_{\max}^s}{2}\tau_0$ 表示无人机 m 从零速度到 $v_{\max}^s$ 的飞行距离。

- $\boldsymbol{\psi}(t)=(\psi_1(t)\ \psi_2(t)\ldots,\psi_M(t))$ 表示所有无人机在时隙 t开始时的能量差。特别地， $\psi_m(t)=E_{\text{最大}}-\sum_{i=1}^{t-1}E_m(i)-E_{\text{需求}m}(t)$ 表示无人机 m的剩余能量与其所需能量之间的差值；其中，剩余能量为 $E_{\text{最大}}-\sum_{i=1}^{t-1}E_m(i)$，而所需能量是指无人机$m$ 从当前位置到达目的地所需的能量，即

$$E_m^{\text{req}}(t)=\begin{cases}E_m^{a_{c,m}^1(t)}+\sum_{i=1}^{T_m^{\text{req}}(t)-1}E_m^{v_{\max}^s,0}, & \text{if } \triangle\varphi_m^{\text{stop}}(t)\le\triangle\varphi_{\max} \text{ or } v_m^s(t)=0,\\ E_m^{a_{c,m}^2(t)}+E_m^{0,a_{c,m}^3}+\sum_{i=1}^{T_m^{\text{req}}(t)-2}E_m^{v_{\max}^s,0}, & \text{otherwise,}\end{cases} \tag{10}$$

其中， $E^{a_c^1}$、$_m(t)$ $_m$ 表示在a_c、$_m(t)=\frac{v_{\max}^s-v_m^s(t)}{\tau_0}$时，时隙 t 中消耗的能量； $E^{a_c^2}$、$_m(t)$ $_m$ 表示在a_c、$_m(t)=\frac{-v_m^s(t)}{\tau_0}$时，时隙 t 中消耗的能量； E^0、a_c^3、$_m$ ^{m}表示在 $v_m^s(t)=0$ 和a_c、$_m(t)=\frac{v_{\max}^s}{\tau_0}$时，一个时隙内消耗的能量； $E^{v_{\max}^s}$、$_m$ ^{0}表示在 $v_m^s(t)=v_{\max}^s$ 和a_c、$_m(t)=0$时，一个时隙内消耗的能量。

2）观测：我们假设传感器节点会在每个时隙开始时通过正交接入报告其信息年龄和电池电量，并且每架无人机都能从位于其覆盖区域内的传感器节点接收这些信息。时隙 t 时无人机 m 的观测表示为 $\boldsymbol{o}_m(t)=(\boldsymbol{u}_m(t)\ \boldsymbol{\delta}_m(t)\ v_m^s(t)\ \varphi_m(t-1)\ \boldsymbol{E}_m^{\text{sn}}(t)\ \phi_m(t)\ \psi_m(t))$， $\forall m\in\mathcal{M}$，其中，$\hat{}\boldsymbol{\delta}_m(t)=(\hat{}\delta_m,1(t),\hat{}\delta_m,2(t)\ldots,\hat{}\delta_{m,N}(t))$ 和 $\hat{}\boldsymbol{E}_m^{\text{sn}}(t)=(\hat{}E_m^{\text{sn}},1(t),\hat{}E_m^{\text{sn}},2(t)\ldots,\hat{}E_{m,N}^{\text{sn}}(t))$ 分别表示无人机 m 在时

隙 t时能够观测到的传感器节点的信息年龄和电池电量。特别地，如果传感器节点 n 在时隙 t时位于无人机 m 的覆盖范围内，则有 $\delta_{m,n}(t)=\delta_n(t)$ ^ ^ ^ 和 $E^{\text{sn}}_{m,n}(t)=E^{\text{sn}}_n(t)$；否则，$\delta_{m,n}(t)$ 为 none，且 $E^{\text{sn}}_{m,n}(t)$ 为 none。

3）动作：无人机 m 在时隙 t 中的动作 $\boldsymbol{a}_m(t)=(v^s_m(t+1)\ \varphi_m(t)\ b_m(t))$，由其在时隙 $t+1$ 开始时的速度 $v^s_m(t+1)$、速度方向 $\varphi_m(t)$ 以及在时隙 t 中的传感器节点调度 $b_m(t)$ 所表征。我们将 $v^s_m(t+1)$ 和 $\varphi_m(t)$ 的取值分别离散化为 $v^s_m(t+1)\in\{0,\frac{1}{N_1}v^s_{\max},\frac{2}{N_1}v^s_{\max},\dots,v^s_{\max}\}$ 和 $\varphi_m(t)\in\{0,\frac{1}{N_2}2\pi,\frac{2}{N_2}2\pi,\dots,2\pi\}$，其中 N_1 和 N_2 为正整数。因此，每架无人机的动作空间定义为 $\mathcal{A}=\mathcal{V}^s\times\Phi\times\mathcal{B}$，其中 $\mathcal{V}^s=\{0,\frac{1}{N_1}v^s_{\max},\frac{2}{N_1}v^s_{\max},\dots,v^s_{\max}\}$ 是无人机速度的集合，$\Phi=\{0,\frac{1}{N_2}2\pi,\frac{2}{N_2}2\pi,\dots,2\pi\}$ 是无人机方向的集合，$\mathcal{B}=\{0,1,\dots,N\}$ 是调度动作的集合。然后，所有无人机的联合动作由 $\tilde{\boldsymbol{a}}(t)=(\boldsymbol{a}_1(t)\dots,\boldsymbol{a}_M(t))$ 给出。

由于运动学约束，在每个 $\mathcal{V}^s\times\Phi$ 状态下并非所有移动方式都有效，并且可用动作会因状态而异。为筛除无效动作并避免不必要的探索，我们将无人机 m 的可用移动方式集合定义为 $\mathcal{V}'_m(t)\ \{v^s_m(t+1)\ \varphi_m(t)|v^s_m(t+1)\in\mathcal{V}^s,\varphi_m(t)\in\Phi,(\varphi_m(t-1)-\triangle\varphi_{\max})\bmod 2\pi\le\varphi_m(t)\le(\varphi_m(t-1)+\triangle\varphi_{\max})\bmod 2\pi\}$。然而，由于式（8g）—（8h）中的长期约束，无人机可能无法在整个 T 时隙期间从 $\mathcal{V}'_m(t)$中选择动作。如果 $\phi_m(t)<0$ 或 $\psi_m(t)<0$，由于时间不足或能量不足，无人机将无法在第 T 个时隙内到达目的地。在这些情况下，当时间差或能量差较小时，无人机可能需要停止探索并遵循预定策略。令 $\overline{E}$ 表示一个时隙内可消耗的最大能量。由于 $\phi_m(t+1)\ge\phi_m(t)-4$ 且 $\psi_m(t+1)\ge\psi_m(t)-4\overline{E}$（如附录中所证明），当$\phi_m(t)>4$ 且 $\psi_m(t)>4\overline{E}$时，无人机可以自由地从 $\mathcal{V}'_m(t)$中选择移动动作。另一方面，当 $\phi_m(t)\le 4$ 或 $\psi_m(t)\le 4\overline{E}$时，存在两种预定策略：其一，如果 $\triangle\varphi^{\text{stop}}_m(t)\le\triangle\varphi_{\max}$ 或 $v^s_m(t)=0$，无人机 m 必须调整方向朝向目的地，并以最大速度飞行。其二，如果 $\triangle\varphi^{\text{stop}}_m(t)>\triangle\varphi_{\max}$ 且 $v^s_m(t)\neq 0$，无人机 m 必须先将速度减至零，再加速至最大速度，然后以最大速度飞往目的地。

同时，无人机 m 可以调度其覆盖区域内的某个传感器节点进行更新，也可以不进行数据收集而飞行。然而，由于能量因果性，无人机覆盖区域内并非所有传感器节点都能够被调度进行传输。因此，无人机 m 在时隙 t 中的可用调度集合为

$$\mathcal{B}'_m(t)=\left\{n\mid n\in\mathcal{B},\overline{d}_{n,m}(t)\le R_{\text{U}},E_n(t)\ge E^c_n\right\}. \tag{11}$$

4) 状态转移：我们详细描述 $\boldsymbol{s}(t)$ 中每个元素的状态转移。每个传感器节点的电池电量和信息年龄分别按照式（3）和（7）进行更新。无人机 m 的位置变化可表示为

$$\boldsymbol{u}_m(t+1) = \boldsymbol{u}_m(t) + (\frac{v_m^s(t) + v_m^s(t+1)}{2}\cos\varphi_m(t)\tau_0, \frac{v_m^s(t) + v_m^s(t+1)}{2}\sin\varphi_m(t)\tau_0). \quad (12)$$

时间差 $\phi_m(t)$ 根据无人机 m 的位置和速度进行更新。根据时间差的定义，其更新方式如下：

$$\phi_m(t+1) = \phi_m(t) - 1 + T_m^{\text{req}}(t) - T_m^{\text{req}}(t+1). \quad (13)$$

同样，能量差按如下方式更新：

$$\psi_m(t+1) = \psi_m(t) - E_m(t) + E_m^{\text{req}}(t) - E_m^{\text{req}}(t+1). \quad (14)$$

5）代价：代价取决于所有无人机的状态和联合动作。总体目标是最小化总信息年龄。此外，如果任意两架无人机发生碰撞，违反约束 (8b)，则会对代价施加惩罚项。因此，代价定义如下：

$$c(t) = \begin{cases} \sum_{n=1}^{N}\delta_n(t) + k_1, & \text{if } d_{m,m'}(t) < d_{\text{safe}}, m \neq m', m, m' \in \mathcal{M}, \\ \sum_{n=1}^{N}\delta_n(t), & \text{otherwise.} \end{cases} \quad (15)$$

其中，k_1 是一个足够大的正数。

我们的目标是找到一个最优策略 π^*，以确定长度为 T 的有限时域内的连续动作。由于无人机只能部分观测系统状态，因此每架无人机的策略都是从其观测空间到动作空间的映射。具体而言，给定初始状态 $\boldsymbol{s}(1)$，可通过最小化时间平均期望代价来获得最优策略，如下所示：

$$\pi^* = \arg\min_{\pi}\frac{1}{T}\mathbb{E}_{\pi}\left[\sum_{t=1}^{T} c(t)|s(1)\right]. \quad (16)$$

此处，期望是相对于 π 所诱导的轨迹分布以及状态转移而言的。然而，当状态空间和动作空间较大时，经典强化学习方法（如 Q 学习）并不可行。这是因为 Q 学习收敛到最优解的条件之一是每个状态-动作对都必须被访问无限次，而当状态-动作对的维度极大时，这是不可能的 [35]。因此，在下一小节中，我们将采用多智能体深度强化学习方法来解决这个问题。

B. 多智能体深度强化学习方法

如果中央控制器能够完全观测环境状态，那么可以通过集中式学习解决多架无人机的数据收集问题。具体而言，所有无人机均由单个智能体控制，其输入为环境状态，输出为所有无人机动作的组合。在集中式学习中，基于状态 $\boldsymbol{s}$ 和联合动作 $\tilde{\boldsymbol{a}}$ 的全局动作价值函数 $Q_{\pi}^{G}(\boldsymbol{s}(t)\ \tilde{\boldsymbol{a}}(t))$ 可表示如下： [36]

$$Q_{\pi}^{G}(\boldsymbol{s}(t),\tilde{\boldsymbol{a}}(t))=\mathbb{E}\left[\sum_{i=t}^{T}c(i)|\boldsymbol{s}(t)=\boldsymbol{s},\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}\right], \tag{17}$$

该函数表示在状态 $\boldsymbol{s}$ 中选择联合动作 $\tilde{\boldsymbol{a}}$ ，随后从时隙 t 开始遵循策略 π 所产生的累积期望成本。最优策略 $Q^{G*}(\boldsymbol{s}(t)\ \tilde{\boldsymbol{a}}(t))$ 的动作价值函数可通过 [36] 进行估计

$$Q^{G}(\boldsymbol{s}(t),\tilde{\boldsymbol{a}}(t))=(1-\alpha)Q^{G}(\boldsymbol{s}(t),\tilde{\boldsymbol{a}}(t))+\alpha\left[c(t)+\min_{\tilde{\boldsymbol{a}}'}Q^{G}(\boldsymbol{s}(t+1),\tilde{\boldsymbol{a}}')\right]. \tag{18}$$

最优策略是在每个时刻采取使动作价值函数最小的动作。

然而，无人机无法获取其他无人机的动作和观测信息，因此无法估计全局动作价值函数。因此，在这种分布式设置中无法采用集中式学习。实现去中心化学习的一种自然方法是使用 IDQN [37]，独立训练每架无人机，其中每架无人机采用深度Q网络方法，并通过深度神经网络（DNN）近似其动作价值函数。然而，由于同时进行学习和探索，IDQN 可能无法在非平稳多智能体环境中收敛。

为了解决上述困难，我们提出了一种基于QMIX的算法，该算法可以采用集中式方式进行训练，但以去中心化方式执行 [38]。与IDQN类似，QMIX中包含多个智能体网络，每个网络对应一架无人机。此外，QMIX中还包含一个混合网络。在训练过程中，如图2所示，每架无人机利用一个智能体网络，根据自身的观测和动作估计局部动作价值函数，而混合网络则结合系统状态和所有局部动作价值函数，生成全局动作价值函数。QMIX的本质在于保证全局动作价值函数与局部动作价值函数之间的单调性。在测试过程中，如图3所示，每个智能体只需根据自身的观测和前一动作执行其动作，进而最小化全局

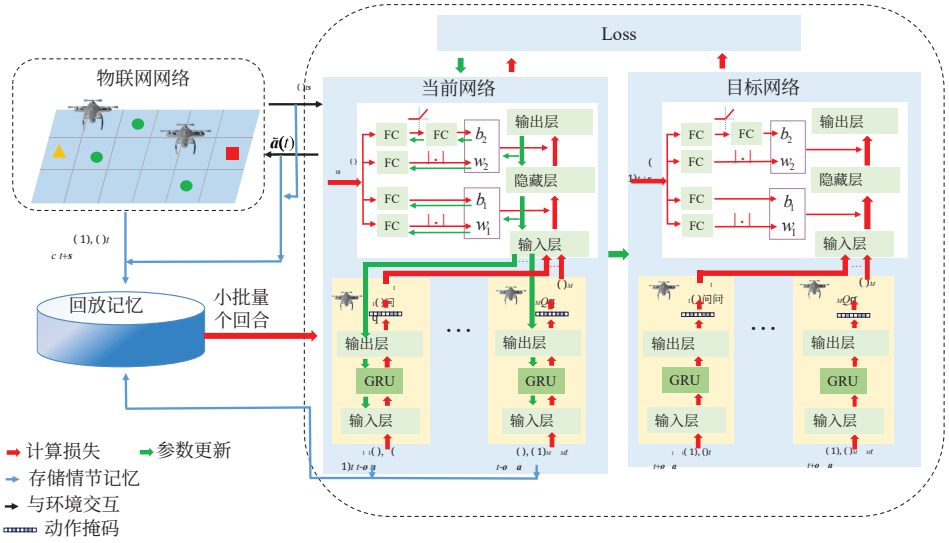


图2. QMIX算法在训练阶段的框架。前向传播过程用红色实线箭头表示，用于计算损失；反向传播过程用绿色实线箭头表示，用于更新神经网络的参数。顶部的白色框表示混合网络，底部的黄色框表示智能体网络。蓝色箭头表示经验的收集，黑色箭头表示与环境的交互。黑色网格表示动作掩码。

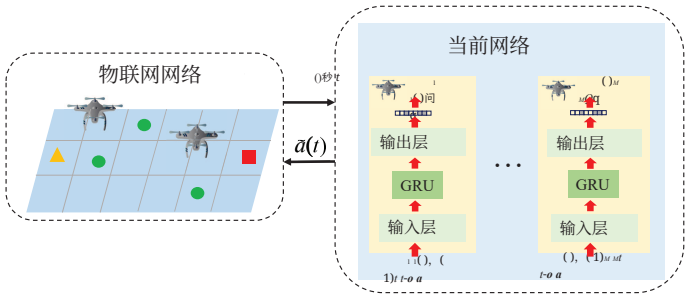


图 3. QMIX算法在测试阶段的框架。

动作价值函数。下面将解释QMIX如何利用智能体网络和混合网络实现集中训练、分散执行。

1）智能体网络：对于每架无人机，都有一个智能体网络来评估其局部动作价值函数 $Q_m(\boldsymbol{\tau}_m(t)\ \boldsymbol{a}_m(t);\ \theta_m)$，其中 θ_m 是智能体网络的权重。每个智能体网络都包含一个输入层、一个门控循环单元（GRU）层和一个输出层。智能体网络的输入层收集当前观测 $\boldsymbol{o}_m(t)$以及前一动作 $\boldsymbol{a}_m(t-1)$

无人机 m 的，然后将联合观测-动作历史 $\boldsymbol{\tau}_m(t) = (\boldsymbol{o}_m(t)\ \boldsymbol{a}_m(t-1))$输入GRU层。作为标准循环神经网络的改进版本，GRU层包含更新门和重置门，分别有助于捕获序列中的长期依赖关系和短期依赖关系 [39]。因此，GRU层能够克服底层分散式部分可观测马尔可夫决策过程的部分可观测性。随后，GRU层输出其隐藏状态，并将其输入输出层。输出层基于GRU层的隐藏状态生成局部动作价值函数。

注1. 训练智能体网络的计算复杂度取决于输入层 $h_{\mathrm{I}}^{\mathrm{a}}$、GRU层 $h_{\mathrm{G}}^{\mathrm{a}}$ 和输出层 $h_{\mathrm{O}}^{\mathrm{a}}$ 中神经元的数量。与 [40], 类似，我们可以得到 GRU层的计算复杂度为 $\mathcal{O}(h_{\mathrm{G}}^{\mathrm{a}}(3h_{\mathrm{I}}^{\mathrm{a}} + 3h_{\mathrm{G}}^{\mathrm{a}} + h_{\mathrm{O}}^{\mathrm{a}}))$。此外，具有 h_{I} 个输入神经元和 h_{O} 个输出神经元的全连接层的计算复杂度为 $\mathcal{O}(h_{\mathrm{I}}h_{\mathrm{O}})$ [41]。总之，训练智能体网络的计算复杂度为 $\mathcal{O}(h_{\mathrm{I}}^{\mathrm{a}}h_{\mathrm{G}}^{\mathrm{a}}+h_{\mathrm{G}}^{\mathrm{a}}(3h_{\mathrm{I}}^{\mathrm{a}}+3h_{\mathrm{G}}^{\mathrm{a}}+h_{\mathrm{O}}^{\mathrm{a}})+h_{\mathrm{G}}^{\mathrm{a}}h_{\mathrm{O}}^{\mathrm{a}})$，其中 $h_{\mathrm{I}}^{\mathrm{a}}$ 和 $h_{\mathrm{O}}^{\mathrm{a}}$ 分别表示观测空间和动作空间的维度。

2）混合网络：混合网络由一个包含一个隐藏层的全连接神经网络（FNN）和一组超网络组成。混合网络的权重记为 θ_0。我们使用混合网络来评估全局动作价值函数 $Q^G(\tilde{\boldsymbol{\tau}}(t)\ \tilde{\boldsymbol{a}}(t),\boldsymbol{s}(t);\ \theta)$，其中 $\tilde{\boldsymbol{\tau}}(t) = (\boldsymbol{\tau}_1(t)\ldots,\boldsymbol{\tau}_M(t))$，且 $\theta = (\theta_0,\theta_1,\ldots,\theta_M)$。混合网络的输入包括每个智能体网络的输出，即 $Q_m(\boldsymbol{\tau}_m,\boldsymbol{a}_m)\ \forall m \in \mathcal{M}$，以及环境状态。具体而言，FNN 将每个智能体网络的输出作为输入，并通过一个隐藏层以单调方式将它们混合。输出层仅包含一个神经元，用于生成全局动作价值函数。与传统 FNN 不同，混合网络中 FNN 的权重和偏置由一组超网络生成。所有超网络均以状态 $\boldsymbol{s}(t)$ 为输入，因此可以根据状态以非单调方式调整 $Q^G(\tilde{\boldsymbol{\tau}}(t)\ \tilde{\boldsymbol{a}}(t)\ \boldsymbol{s}(t);\ \theta)$。其中有两个超网络分别生成输入层与隐藏层之间、隐藏层与输出层之间的权重。这两个超网络各由一个全连接层（FC）组成。FC 层的输出采用绝对值激活函数，以获得 FNN 的非负权重。此外，FNN 隐藏层和输出层的偏置也由另一个超网络以类似方式生成。具体而言，我们采用一个包含一个 FC 层的超网络来获得隐藏层的偏置，并采用一个包含两个 FC 层且使用修正线性单元（ReLU）非线性的超网络来获得输出层的偏置，

分别如此。[4] 通过这种结构，混合网络保证全局动作价值函数随着局部动作价值函数 [38]，单调递增，即，

$$\frac{\partial Q^G(\tilde{\boldsymbol{\tau}}(t), \tilde{\boldsymbol{a}}(t), \boldsymbol{s}(t); \theta)}{\partial Q_m(\boldsymbol{\tau}_m(t), \boldsymbol{a}_m(t); \theta_m)} \geq 0, \forall m \in \mathcal{M}. \tag{19}$$

因此，使全局动作价值函数最小化的联合动作，等价于使各个局部动作价值函数最小化的单个动作。具体而言，

$$\arg\min_{\tilde{\boldsymbol{a}}} Q^G(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{a}}, \boldsymbol{s}; \theta) = \begin{pmatrix} \arg\min_{\boldsymbol{a}_1} Q_1(\boldsymbol{\tau}_1, \boldsymbol{a}_1; \theta_1) \\ \arg\min_{\boldsymbol{a}_2} Q_2(\boldsymbol{\tau}_2, \boldsymbol{a}_2; \theta_2) \\ \vdots \\ \arg\min_{\boldsymbol{a}_M} Q_M(\boldsymbol{\tau}_M, \boldsymbol{a}_M; \theta_M) \end{pmatrix}. \tag{20}$$

这样，每个智能体都可以通过去中心化方式，贪心地选择使其局部动作价值函数最小化的动作。

注 2. 混合网络的训练计算复杂度取决于全连接神经网络和超网络中的神经元数量。具体而言，具有一个隐藏层的全连接神经网络的计算复杂度为 $\mathcal{O}(h_{\mathrm{I}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}} + h_{\mathrm{H}}^{\mathrm{m}} h_{\mathrm{O}}^{\mathrm{m}})$，其中 $h_{\mathrm{I}}^{\mathrm{m}}$为输入层的神经元数量， $h_{\mathrm{H}}^{\mathrm{m}}$ 为隐藏层的神经元数量， $h_{\mathrm{O}}^{\mathrm{m}}$ 为输出层的神经元数量。设 $h_{\mathrm{HI}}^{\mathrm{m}}$ 表示超网络输入层的神经元数量。两个超网络分别生成输入层与隐藏层之间的权重和偏置，其计算复杂度分别为 $\mathcal{O}(h_{\mathrm{HI}}^{\mathrm{m}} M h_{\mathrm{H}}^{\mathrm{m}})$ 和 $\mathcal{O}(h_{\mathrm{HI}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}})$。同样，两个超网络分别生成隐藏层与输出层之间的权重和偏置，其计算复杂度分别为 $\mathcal{O}(h_{\mathrm{HI}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}} h_{\mathrm{O}}^{\mathrm{m}})$ 和 $\mathcal{O}(h_{\mathrm{HI}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}} + h_{\mathrm{H}}^{\mathrm{m}} h_{\mathrm{O}}^{\mathrm{m}})$。总而言之，混合网络的计算复杂度为

$$\begin{aligned} \mathcal{O}^{\mathrm{m}} =& \mathcal{O}(h_{\mathrm{I}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}} + h_{\mathrm{H}}^{\mathrm{m}} h_{\mathrm{O}}^{\mathrm{m}} + h_{\mathrm{HI}}^{\mathrm{m}} M h_{\mathrm{H}}^{\mathrm{m}} + h_{\mathrm{HI}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}} + h_{\mathrm{HI}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}} h_{\mathrm{O}}^{\mathrm{m}} + h_{\mathrm{HI}}^{\mathrm{m}} h_{\mathrm{H}}^{\mathrm{m}} + h_{\mathrm{H}}^{\mathrm{m}} h_{\mathrm{O}}^{\mathrm{m}}) \\ =& \mathcal{O}(h_{\mathrm{H}}^{\mathrm{m}} (h_{\mathrm{I}}^{\mathrm{m}} + 2h_{\mathrm{O}}^{\mathrm{m}} + (M + 2 + h_{\mathrm{O}}^{\mathrm{m}}) h_{\mathrm{HI}}^{\mathrm{m}})). \end{aligned} \tag{21}$$

注意， $h_{\mathrm{HI}}^{\mathrm{m}}$、 $h_{\mathrm{I}}^{\mathrm{m}}$和 $h_{\mathrm{O}}^{\mathrm{m}}$ 分别表示状态空间的维度、智能体的数量 M和 1。

3）集中式训练：智能体网络和混合网络采用离线集中式训练。与深度Q网络类似，为稳定训练过程，使用两组神经网络，如图2所示。一组是权重为 θ 的当前网络，另一组是权重为 θ^- 的目标网络。这里，权重 $\theta =$（θ_0、 θ_1、.....、 θ_M）包括

[4]注意，偏置 不受非负条件的限制，因为单调性不取决于偏置的符号。

混合网络的参数 θ_0 和所有智能体网络的参数 θ_m（对应于 $m \in \mathcal{M}$）。当前网络用作函数近似器，其权重在每个时隙更新。目标网络用于计算目标动作价值函数，其权重在一段时间内保持固定，然后每隔 O 步替换为当前网络的最新权重。此外，还使用经验回放。为高效训练每个智能体网络中的 GRU 层，我们对一个回合中的完整经验进行采样，而不是像深度Q网络那样在回放记忆中随机采样转移元组。这样，GRU层的隐藏状态便能从回合 [42] 中的经验学习时间相关性。

QMIX 的当前网络通过在每个回合最小化损失函数进行训练。具体而言，对于从回放记忆中采样的回合中的经验，其损失函数为

$$L(\theta) = \frac{1}{Z}\sum_{i=1}^{Z}(y^G(i) - Q^G(\tilde{\boldsymbol{\tau}}(i), \tilde{\boldsymbol{a}}(i), \boldsymbol{s}(i); \theta))^2, \tag{22}$$

其中， Z 是小批量中的转移数量， $y^G(i) = c(i) + \min_{\tilde{a} \in \mathcal{A}'(i+1)} Q^G(\tilde{\boldsymbol{\tau}}(i+1), \tilde{\boldsymbol{a}}, \boldsymbol{s}(i+1); \theta^-)$。这里， $\mathcal{A}'(i+1) = (\mathcal{A}'_1(i+1)\text{.....}, \mathcal{A}'_M(i+1))$ 是时隙 $i+1$ 中所有无人机的联合可用动作， $\mathcal{A}'_m(i+1)$ 是根据观测 $\boldsymbol{o}_m(i+1)$，通过动作掩码方法获得的时隙 $i+1$ 中无人机 m 的可用动作。注意， $Q^G(\tilde{\boldsymbol{\tau}}(i), \tilde{\boldsymbol{a}}(i), \boldsymbol{s}(i); \theta)$ 和 $Q^G(\tilde{\boldsymbol{\tau}}(i+1), \tilde{\boldsymbol{a}}, \boldsymbol{s}(i+1); \theta^-)$ 分别由当前网络和目标网络获得。随后，当前网络的权重通过半梯度算法更新如下：

$$\theta = \theta + \frac{\alpha}{Z}\sum_{i=1}^{Z}(y^G(i) - Q^G(\tilde{\boldsymbol{\tau}}(i), \tilde{\boldsymbol{a}}(i), \boldsymbol{s}(i); \theta))\nabla_\theta Q^G(\tilde{\boldsymbol{\tau}}(i), \tilde{\boldsymbol{a}}(i), \boldsymbol{s}(i); \theta). \tag{23}$$

集中式训练的详细过程总结于算法1中。首先，初始化当前网络和目标网络，并清空回放记忆。由于当前网络是基于回合经验进行更新的，因此在执行梯度下降步骤（第25行）之前，我们会收集整个回合的经验（第 5~14行）。由于动作掩码会筛除违反约束（8c）—（8h）的无效动作，因此回合在两种情况下终止：1）两架无人机发生碰撞，$d_{m,\ m'}(t) < d_{\text{safe}}, m \neq m', m, m' \in \mathcal{M}$；2）所有无人机在时隙 T到达最终位置。在终止状态之前，每个智能体同时利用 ε-贪心策略选择动作，其中 ε 会随着时隙数量的增加而减小 t （第 6~10行）。所有智能体执行动作后，即可获得代价 $c(t)$和下一观测 $\tilde{\boldsymbol{o}}(t+1)$。环境也会转移到下一状态 $\boldsymbol{s}(t+1)$（第11行）。然后，我们从以下内容中随机抽取一个回合小批量：

算法1 多无人机数据采集算法的训练过程。1：初始化当前网络 $Q^G(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{a}}, \boldsymbol{s}; \theta)$和目标网络 $Q^G(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{a}}, \boldsymbol{s}; \theta^-)$；2：初始化回放记忆 D；3：对于 $episode = 1 : EP$，执行 4：环境初始化初始状态 $\boldsymbol{s}(1)$，智能体初始化其观测$\boldsymbol{o}_m(1)$；5：当 $\boldsymbol{s}(t)$不是终止状态时，执行 6：对于智能体 $m = 1 : M$，执行 7：计算可用动作 $\mathcal{A}'_m(t) \triangleq (\mathcal{V}'_m(t)\ \mathcal{B}'_m(t))$；8：以概率 ϵ选择随机动作 $\boldsymbol{a}_m(t) \in \mathcal{A}'_m(t)$；9：否则选择 $\boldsymbol{a}_m(t) = \arg\min_{\boldsymbol{a}_m \in \mathcal{A}'_m(t)} Q_m(\boldsymbol{\tau}_m(t)\ \boldsymbol{a}_m; \theta_m(t))$；10：结束循环 11：执行联合动作 $\tilde{\boldsymbol{a}}(t)$，并获取代价 $c(t)$、下一状态 $\boldsymbol{s}(t+1)$和下一联合观测 $\tilde{\boldsymbol{o}}(t+1)$；~ ~12：将（$\boldsymbol{s}(t)\ \tilde{\boldsymbol{\tau}}(t)\ \mathcal{A}'(t)\ \tilde{\boldsymbol{a}}(t)\ c(t)\ \boldsymbol{s}(t+1)\ \tilde{\boldsymbol{\tau}}(t+1)\ \mathcal{A}'(t+1)$）存入回放记忆；13：$\boldsymbol{s}(t) = \boldsymbol{s}(t+1)$；14：结束循环 15：从 D中随机抽取一个回合小批量；16：计算回合经验 y^G的目标动作价值：17：对于 $i = 1 : Z$，执行 18：如果 $\boldsymbol{s}(i)$不是终止状态，则 19：$y^G(i) = c(i) + \min_{\tilde{\boldsymbol{a}} \in \tilde{\mathcal{A}}'(i+1)} Q^G(\tilde{\boldsymbol{\tau}}(i+1)\ \tilde{\boldsymbol{a}}, \boldsymbol{s}(i+1); \theta^-)$；20：否则 21：$y^G(i) = c(i)$；22：结束条件判断 23：结束循环 24：$y^G = \sum_{i=1}^{Z} y^G(i)$；25：通过执行式（23）中的梯度下降来更新当前网络；26：如果已经过了 O 个更新间隔步数，则 27：通过设置 $\theta^- = \theta$同步目标网络；28：结束条件判断 29：结束循环

回放记忆 D 以更新当前网络（第 15~25行）。每经过固定数量的 O 个回合，便通过复制当前网络参数来更新目标网络（第 26~28行）。请注意，在所有回合中，传感器节点的数量和位置始终保持不变。

4）分布式执行：离线集中式训练完成后，智能体网络便可在线以去中心化方式执行。具体而言，每架无人机都利用预训练智能体网络，根据自身观测以分布式方式做出决策。QMIX 的测试框架如图3所示。具体来说，在每个时隙中，每架无人机都会选择使其局部动作价值函数最小化的动作，即 $\boldsymbol{a}_m(t) = \arg\min_{\boldsymbol{a}_m \in \mathcal{A}'_m(t)} Q_m(\boldsymbol{\tau}_m, \boldsymbol{a}_m;\ \theta_m)$，并据此更新观测。每架无人机持续飞行，直到在第 T 个时隙到达最终位置。分布式执行的具体步骤总结于算法2中。

算法2 多无人机数据采集算法的测试过程。

1：初始化环境；

2: 对于 $episode = 1:\ EPT$ 执行

3: 环境初始化初始状态 $\boldsymbol{s}(1)$，智能体初始化其观测 $\boldsymbol{o}_m(1)$；

4: 当 $\boldsymbol{s}(t)$ 不是终止状态时执行

5: 对于智能体 $m = 1:\ M$ 执行

6: 计算可用动作 $\mathcal{A}'_m(t)$;

7: 选择 $\boldsymbol{a}_m(t) =$ 参数 $\min_{\boldsymbol{a}_m \in \mathcal{A}'_m(t)} Q_m(\boldsymbol{\tau}_m(t)\ \boldsymbol{a}_m;\ \theta_m(t))$;

8: 结束循环

9: 执行联合动作 $\tilde{\boldsymbol{a}}(t)$，并获得代价 $c(t)$、下一状态 $\boldsymbol{s}(t+1)$ 以及下一联合观测 $\tilde{\boldsymbol{o}}(t+1)$;

10: $\boldsymbol{s}(t) = \boldsymbol{s}(t+1)$;

11: 结束 while 循环

12: 结束 for 循环

V. 仿真结果

在本节中，我们进行了大量仿真，以评估所提出算法的性能。首先，我们介绍仿真设置和三种基线算法。然后，我们评估所提出算法在不同系统参数下的收敛性和有效性。

表 I 系统参数

参数	值	参数	值
飞行高度 z	100 m [15]	空气密度 ρ	1.225 kg/m^3
持续时间 T	100 个时隙	每个旋翼的盘面面积 A	0.0314 m^2
时隙长度 τ_0	0.5 s	每个传感器节点的发射功率 P_c	5 mW [43]
安全距离 d_{safe}	10 m	噪声功率 σ^2	-110dBm [44]
无人机最大速度 $v^s_{\max}$	20 m/s [7]	β_0, β_1	11.95, 0.14 [45]
无人机最大转弯角 $\triangle\varphi_{\max}$	$\pi/3$	η_{LoS}, η_{NLoS}	1.6 dB, 23 dB
电池容量 $E_{\max}$,$E^{\text{sn}}_{\max}$	2.4×10^4 J, 5 mJ	载波频率f_c	2 GHz [45]
$E^{\text{har}}_{\text{n}}$, λ_{n}	0.42 mJ, 0.9	真空中的光速 c	3×10^8 米/秒
无人机质量 W	2 千克 [7]	旋翼数量 n_r	4
重力加速度 g	9.8 米/秒^2	c_f, c_T	0.131, 0.302
局部桨叶剖面阻力系数 σ	0.012	c_s, d_0	0.0955, 0.834

A. 仿真设置

在仿真中，我们在一个 800 m $\times$ 800 m 的正方形区域内均匀随机部署传感器节点。为便于说明，我们将无人机的初始和最终位置分别均匀设置在该区域的底部和顶部。例如，当 $M = 3$ 时，无人机的初始位置为 (0 m, 0 m, 100 m)、(360 m, 0 m, 100 m) 和 (760 m, 0 m, 100 m)，最终位置为 (0 m, 760 m, 100 m)、(360 m, 760 m, 100 m) 和 (760 m, 760 m, 100 m)。值得注意的是，每架无人机的初始位置和最终位置可以相同，不同无人机的初始位置或最终位置也可以相同。每个回合开始时，所有无人机和传感器节点均已充满电。除非另有说明，物联网网络的仿真参数见表 I。

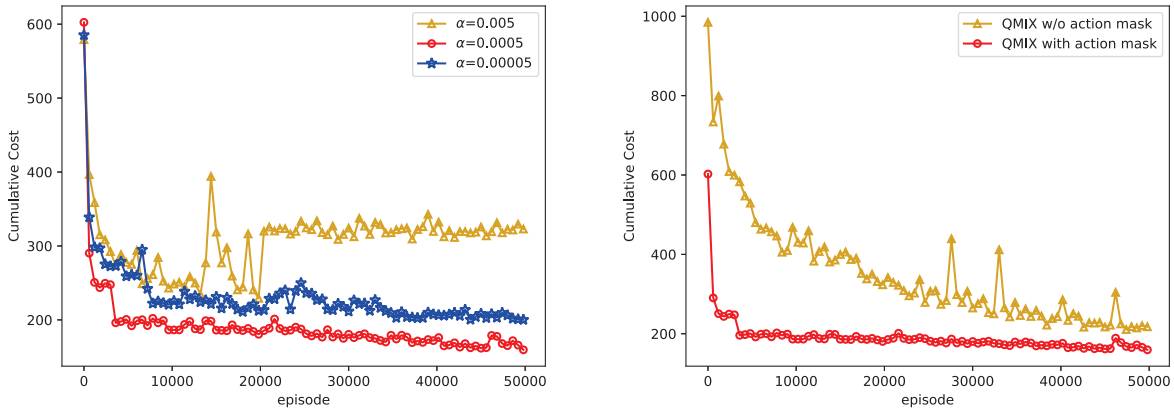
基于QMIX的算法通过 PyTorch 库实现。具体而言，智能体网络包含一个具有 256 个隐藏神经元的 GRU层。由于每架无人机具有相同的状态空间和动作空间，因此每架无人机的智能体网络都相同 [46]。混合网络中的隐藏层包含 256 个神经元。基于QMIX的算法的超参数列于表 II 中。

接下来，我们将提出的基于QMIX的算法与三种基线算法进行比较：

- 最近调度算法：这是基于QMIX的算法的简化版本，其轨迹采用与多智能体深度强化学习方法相同的方法进行设计，但每架无人机都调度与自身距离最近的传感器节点。

表 II 基于QMIX的算法超参数

参数	值	参数	值
训练回合数 EP	50000	最小值 ε	0.01
D, O	1000, 200	学习率 α	0.0005
小批量大小	32	优化器	Adam
初始值 ε	0.99	激活函数	ReLU
ε-贪心递减	9.9e-6	N_1, N_2	1, 6



(a) (b)

图4. 所提出的基于QMIX的算法的收敛性评估（$\xi_{\text{th}} = 5$dB、 $M = 2$和 $N = 15$）。（a）不同学习率 α的影响。（b）动作掩码方法的影响。

- 基于聚类的算法：在该算法中，首先通过 K-均值将所有传感器节点聚为 M 个簇，簇中心的初始位置设为无人机的起始位置。然后，每架无人机从对应簇中的传感器节点收集数据。在每个时隙，无人机飞向其所在簇中信息年龄最大的传感器节点，并调度其覆盖范围内信息年龄最大的传感器节点。

- 基于IDQN的算法：在该算法中，每架无人机作为一个独立智能体，并采用深度Q网络处理轨迹设计和传感节点调度。

B. 性能评估

我们在图4(a)中展示了所提出的基于QMIX的算法的收敛性。具体而言，图中绘制了不同学习率α下累积成本随训练回合数的变化情况。如图所示，过高的学习率会导致不稳定，而过低的学习率则会导致算法收敛到局部最优。

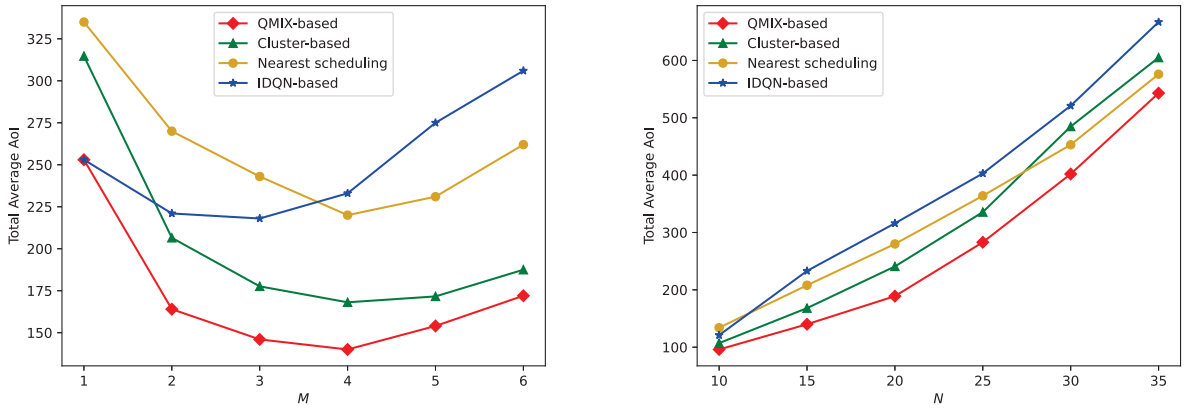
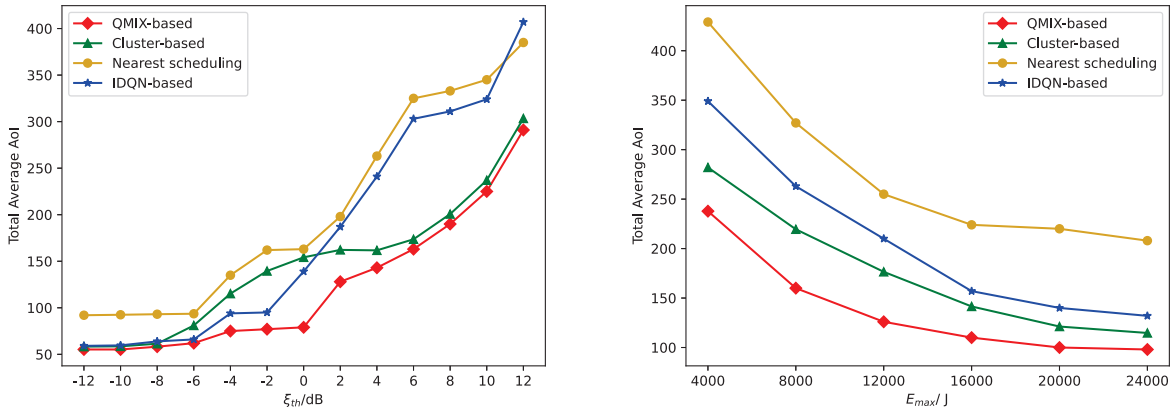


图5. 基于QMIX的算法与其他三种基线算法的比较（$\xi_{\text{th}} = 5$ dB）。（a）总平均信息年龄随无人机数量 M 的变化情况，$N = 15$。（b）总平均信息年龄随传感器节点数量 N 的变化情况，$M = 4$。

在这种情况下，我们发现学习率为0.0005时性能最佳，因为与0.005和0.00005的学习率相比，它产生的累积成本最低。因此，在后续分析中，我们选择该学习率用于QMIX算法。在图4(b)中，我们考察了动作掩码方法对QMIX收敛性的影响。通过比较采用和未采用该方法的QMIX可以看出，引入动作掩码能够实现更快的收敛和更低的累积成本。这是因为动作掩码能够隐藏违反约束的动作，从而防止无效探索，进而提升QMIX算法的性能。

图5(a)展示了在 $N = 15$ 和$\xi_{\text{th}} = 5$ dB条件下，总平均信息年龄随无人机数量 M 的变化情况。随着 M 增加，基于QMIX的算法的总平均信息年龄先降低后升高。这一趋势可以从两个方面解释。一方面，随着 M 增大，每架无人机需要调度的传感器节点平均数量会减少。由于每架无人机同一时刻最多只能调度一个传感器节点，因此无人机数量越多，每个传感器节点就能越频繁地被调度，从而降低总平均信息年龄。另一方面，随着 M 增加，干扰水平也会升高。因此，状态更新被无人机成功接收的概率会降低，进而导致总平均信息年龄上升。显然，当 $M = 1$时，基于QMIX的算法与基于IDQN的算法性能相同，因为二者都会退化为单智能体深度强化学习。图5(b)展示了在 $\xi_{\text{th}} = 5$ dB和 $M = 4$条件下，总平均信息年龄随传感器节点数量 N 的变化情况。可以看出，总平均信息年龄会增加



(a) (b)

图6. 基于QMIX的算法与其他三种基线算法的比较（$N = 15$）。(a) 在 $M = 4$条件下，总平均信息年龄随信干噪比阈值 ξ_{th} 的变化。(b) 在 $\xi_{\text{th}} = 5$ dB和 $M = 3$条件下，总平均信息年龄随总能量 $E_{\max}$ 的变化。

随着 N 增加。这是因为每架无人机同一时刻最多只能调度一个传感器节点进行状态更新。因此，传感器节点数量越多，每个传感器节点等待的时间就越长。此外，随着传感器节点数量持续增加，总平均信息年龄会迅速上升。这是因为，当 N较大时，更多传感器节点需要等待数量有限的无人机进行调度，导致总平均信息年龄快速增加。在两个子图中，所提出的基于QMIX的算法的总平均信息年龄均低于其他三种基线算法。这一优势是通过无人机协同训练，并利用全局价值函数联合优化无人机轨迹和传感器节点调度实现的。

图6(a)展示了在 $N = 15$ 和 $M = 4$条件下，总平均信息年龄随阈值 ξ_{th} 的变化情况。从图6(a)可以看出，总平均信息年龄会随着 ξ_{th} 的增加而上升。这是因为 ξ_{th} 越大，成功传输的概率越低，从而导致总平均信息年龄更高。还可以看出，尽管基于聚类的算法在 ξ_{th} 较低和较高时的表现与基于QMIX的算法相近，但在 ξ_{th}的所有范围内，基于QMIX的算法始终能实现最低的总平均信息年龄。图6(b)评估了在 $N = 15$、 $\xi_{\text{th}} = 5$ dB和 $M = 3$条件下，总平均信息年龄随无人机电池容量 $E_{\max}$ 的变化情况。初始位置和最终位置设为相同，无人机分别从(0 m, 360 m, 100 m)、(360 m, 360 m, 100 m)和(760 m, 360 m, 100 m)出发。如式(1)所示，无人机在一个时隙内消耗的能量主要由其速度和加速度决定。从图6(b)可以看出，总平均信息年龄起初会随着 $E_{\max}$ 急剧下降

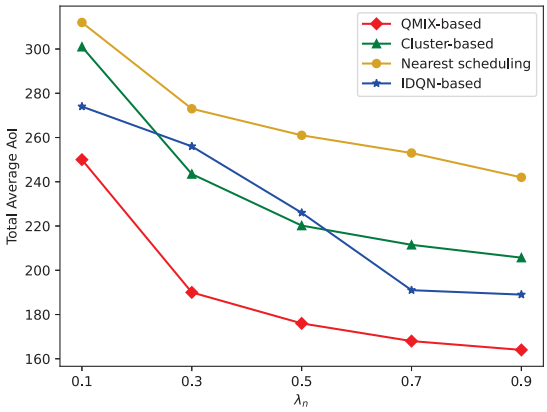
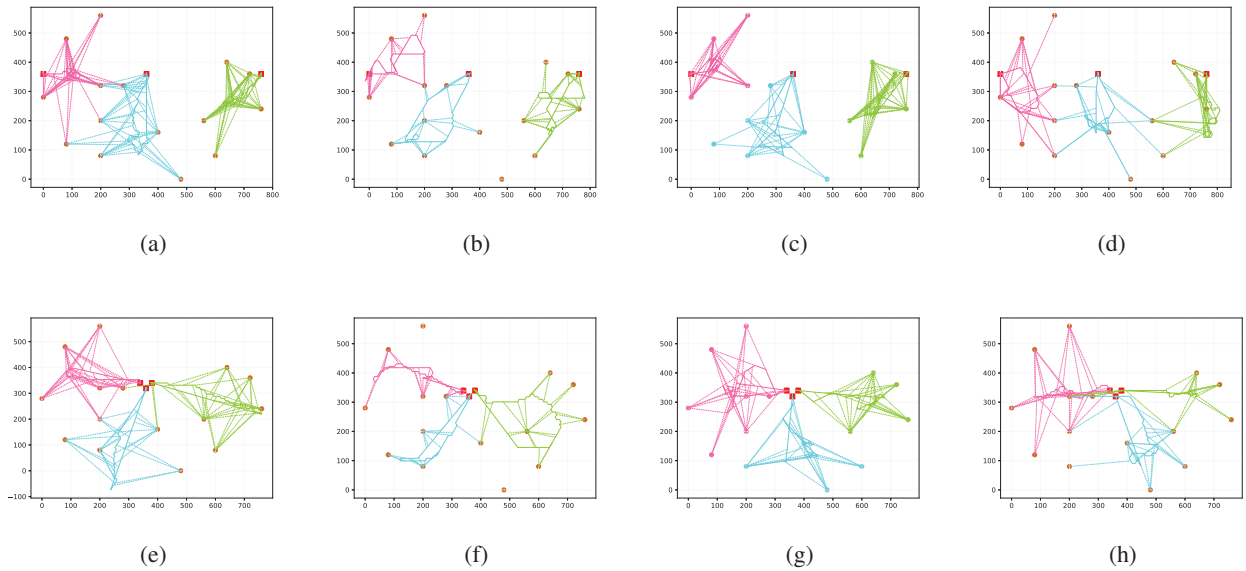


图7。总　　平均信息年龄与传感器节点的能量到达概率 λ_n 的关系，其中 $N = 15$、 $\xi_{\text{th}} = 5\,\text{dB}$，且 $M = 2$。

随后下降速度变慢。原因在于，当 $E_{\max}$ 较小时，能量约束较为严格，无人机需要在某些时隙悬停以节省能量。当 $E_{\max}$ 较大时，能量约束的限制较小，无人机可以更自由地移动以收集数据包。

图7展示了在$N = 15$、 $\xi_{\text{th}} = 5\,\text{dB}$，且 $M = 2$的条件下，传感器节点的能量到达概率 λ_n 对总平均信息年龄的影响。从图7可以看出，随着 λ_n 增加，总平均信息年龄会降低。这是因为当 λ_n 较小时，传感器节点的能量不足，导致其因缺乏能量而无法及时进行状态更新。然而，随着 λ_n 增加，当无人机飞至传感器节点附近时，传感器节点更有可能拥有足够的能量进行更新。当 $\lambda_n \geq 0.7$时，总平均信息年龄的降幅变得很小，因为传感器节点能够及时充能。我们还可以看到，与其他基线算法相比，基于QMIX的算法实现了最低的总平均信息年龄。这是因为基于QMIX的算法可以联合优化无人机轨迹和调度策略，即使在 λ_n 较小时，也能有效调整它们以降低总平均信息年龄。

图8展示了不同算法下无人机的轨迹和传感器节点的调度。实线表示无人机的轨迹，虚线表示传感器节点的调度。此外，我们用不同颜色表示不同的无人机及其调度的传感器节点。对于每架无人机，其初始位置和最终位置相同，并在图中用正方形表示。在图8(a)–(d)中，无人机的初始位置和最终位置相距较远；而在图8(e)–(h)中，两者相距较近。在图8(a)中，多架无人机能够协同飞行并调度传感器节点，从而降低干扰和总平均信息年龄。然而，在图8(b)中，最近调度算法中的传感器节点调度并未协调，



(a) (b) (c) (d)

(e) (f) (g) (h)

图8. 一个回合中无人机轨迹和传感器节点调度的示意图（$\xi_{\text{th}} = 7$ dB、 $N = 15$、 $p_h = 0.9$ 和 $M = 3$）。(a) 初始和最终位置相距较远时的基于QMIX的算法。(b) 初始和最终位置相距较远时的最近调度算法。(c) 初始和最终位置相距较远时的基于聚类的算法。(d) 初始和最终位置相距较远时的基于IDQN的算法。(e) 初始和最终位置较近时的基于QMIX的算法。(f) 初始和最终位置较近时的最近调度算法。(g) 初始和最终位置较近时的基于聚类的算法。(h) 初始和最终位置较近时的基于IDQN的算法。

导致干扰较大。在图8(c)中，基于聚类的算法中的无人机轨迹并非最优，因为传感器节点一旦完成聚类，其状态更新便只能由特定无人机收集。在图8(d)中，基于IDQN的算法只能分别优化各架无人机的轨迹和传感器节点调度，从而导致局部最优解。如图8(e)–(h)所示，我们的方法在无人机初始位置和最终位置相距较近的场景中也能表现良好。通过对比可以看出，基于QMIX的算法能够让无人机分担工作负载，并从不同的传感器节点收集数据。此外，它们还能够协同收集部分传感器节点的数据，从而降低总平均信息年龄。

表 III 总结了我们所提出的算法与其他基线算法在复杂度和性能方面的比较。基于聚类的算法的复杂度主要由 K均值算法决定，其复杂度为 $\mathcal{O}(MN)$。另一方面，基于QMIX的算法、最近邻调度算法和基于IDQN的算法的复杂度主要取决于其神经网络的计算。基于QMIX的算法和最近邻调度算法具有相同的智能体网络和混合网络，但最近邻调度算法的智能体网络输出

表III 计算复杂度与性能比较

方法	计算复杂度	总平均信息年龄
基于QMIX	$M\mathcal{O}^{\mathrm{a}}((N+1)(N_1+1)(N_2+1)) + \mathcal{O}^{\mathrm{m}}$	140
基于聚类	$\mathcal{O}(MN)$	168.1
最近邻调度	$M\mathcal{O}^{\mathrm{a}}((N_1+1)(N_2+1)) + \mathcal{O}^{\mathrm{m}}$	208
基于IDQN	$M\mathcal{O}^{\mathrm{a}}((N+1)(N_1+1)(N_2+1))$	233

层较小。为了清楚地说明这一差异，我们将智能体网络 $\mathcal{O}^{\mathrm{a}}$ 的计算复杂度表示为智能体网络 h_{O}输出维度的函数。基于IDQN的算法与基于QMIX的算法具有相同的智能体网络，但缺少混合网络。因此，基于QMIX的算法的复杂度高于其他基线算法。表中还列出了所有基线算法和我们所提出的算法的总平均信息年龄，该结果是在 $N = 15$、$M = 4$、$E_{\max} = 2.4 \times 10^4\,\mathrm{J}$ 和 $\xi_{\mathrm{th}} = 5\,\mathrm{dB}$ 条件下获得的。可以看出，基于QMIX的算法性能最佳。

VI. 结论

本研究探讨了如何利用多架协作无人机，在多无人机物联网网络中实现数据收集最优化的问题。我们考虑了运动学、能量、轨迹和碰撞避免约束，并以最小化总平均信息年龄为目标。为此，我们提出了基于QMIX的算法，以联合优化无人机的轨迹和传感器节点的调度。具体而言，我们利用全局价值函数对无人机进行集中式训练，随后各无人机根据自身的局部观测，以分布式方式执行数据收集。仿真结果表明，所提出的基于QMIX的算法优于其他基线方法，且动作掩码方法有助于加快基于QMIX的算法的收敛。我们观察到，采用基于QMIX的算法时，无人机能够分担工作负载，分别从不同的传感器节点收集数据，有时甚至会协作从同一个传感器节点收集数据。我们还发现，由于传输机会增加与相互干扰加剧之间存在权衡，因此存在一个能够使总平均信息年龄最小化的无人机数量。未来，我们将考虑基于NR帧结构优化轨迹规划和传输调度，其中无人机轨迹可在大时间尺度上优化，而传输调度则在小时间尺度上优化。

附录 A

PROOF FOR THE PROPERTY OF $\phi_m(t+1) \geq \phi_m(t) - 4$ AND $\psi_m(t+1) \geq \psi_m(t) - 4\overline{E}$

当无人机能够在时隙 t开始时立即将速度方向转向目的地，即 $\triangle\varphi_m^{\text{stop}}(t) \leq \triangle\varphi_{\max}$ 或 $v_m^s(t) = 0$，且时隙 $t+1$ 开始时的速度方向与当前位置指向目的地的方向相反，即 $\varphi_m(t+1) = \varphi_m^{\text{stop}}(t) + \pi$时，可得到 $(\phi_m(t+1) - \phi_m(t))$ 和 $(\psi_m(t+1) - \psi_m(t))$ 的最小值。

在这种情况下，$\boldsymbol{u}_m(t+1)$、$\boldsymbol{u}_m(t)$ 和 $\boldsymbol{u}_m^{\text{stop}}$位于同一直线上。因此，我们有 $\|\boldsymbol{u}_m(t) - \boldsymbol{u}_m^{\text{stop}}\| - \|\boldsymbol{u}_m(t+1) - \boldsymbol{u}_m^{\text{stop}}\| = \frac{v_m^s(t)+v_m^s(t+1)}{2}\tau_0$。然后，我们可以得到

$$
\begin{aligned}
& T_m^{\text{req}}(t) - T_m^{\text{req}}(t+1) \\
&= \left\lceil \frac{\|\boldsymbol{u}_m(t) - \boldsymbol{u}_m^{\text{stop}}\| - \frac{v_m^s(t)+v_{\max}^s}{2}\tau_0}{v_{\max}^s \tau_0} \right\rceil - \left\lceil \frac{\|\boldsymbol{u}_m(t+1) - \boldsymbol{u}_m^{\text{stop}}\| + \frac{v_m^s(t+1)-v_{\max}^s}{2}\tau_0}{v_{\max}^s \tau_0} \right\rceil - 1 \\
&\overset{(a)}{\geq} \left\lfloor \frac{-\frac{v_m^s(t)+v_m^s(t+1)}{2} - \frac{v_m^s(t)+v_{\max}^s}{2} - \frac{v_m^s(t+1)-v_{\max}^s}{2}}{v_{\max}^s} \right\rfloor - 1 \\
&= \left\lfloor \frac{-v_m^s(t) - v_m^s(t+1)}{v_{\max}^s} \right\rfloor - 1 \\
&\geq -3,
\end{aligned} \tag{24}
$$

其中，(a) 是因为 $\lceil i \rceil - \lceil j \rceil \geq \lfloor i - j \rfloor$。将 (24) 代入 (13)，可得 $\phi_m(t+1) \geq \phi_m(t) - 4$.

假设每个时隙的能耗达到最大值 $\overline{E}$，我们可以得到$\psi_m(t) - \psi_m(t+1) \leq 4\overline{E}$，这是因为 $\phi_m(t) - \phi_m(t+1) \leq 4$。因此，我们有 $\psi_m(t+1) \geq \psi_m(t) - 4\overline{E}$。

参考文献

[1] M. Yi、X. Wang、J. Liu、Y. Zhang 和 B. Bai，《面向无人机辅助物联网网络中新鲜数据采集的深度强化学习》，载于 IEEE INFOCOM 研讨会论文集，2020年7月。[2] M. Mozaffari、W. Saad、M. Bennis、Y. Nam 和 M. Debbah，《无线网络中的无人机：应用、挑战与开放问题教程》，IEEE 通信调查与教程，第 21 卷，第 3 期，第 2334–2360 页，2019 年。[3] Y. Zeng 和 R. Zhang，《从空中接入：面向 5G 及后续技术的无人机通信教程》，《IEEE 会刊》，第 107 卷，第 12 期，第 2327–2375 页，2019 年 12 月。[4] J. Gong、T. Chang、C. Shen 和 X. Chen，《无线传感器网络数据收集的无人机飞行时间最小化》，《IEEE 选定通信领域期刊》，第 36 卷，第 9 期，第 1942–1954 页，2018 年。[5] Y. Zeng 和 R. Zhang，《采用轨迹优化的节能型无人机通信》，《IEEE 无线通信汇刊》，第 16 卷，第 6 期，第 3747–3760 页，2017 年。

[6] C. H. Liu、Z. Chen、J. Tang、J. Xu 和 C. Piao，《面向高效、公平通信覆盖的节能无人机控制：一种深度强化学习方法》，《IEEE 选定通信领域期刊》，第 36 卷，第 9 期，第 2059–2070 页，2018 年。[7] R. Ding、F. Gao 和 X. S. Shen，《面向节能、公平通信的三维无人机轨迹设计与频段分配：一种深度强化学习方法》，《IEEE 无线通信汇刊》，第 19 卷，第 12 期，第 7796–7809 页，2020 年。[8] M. Pourghasemian、M. R. Abedi、S. S. Hosseini、N. Mokari、M. R. Javan 和 E. A. Jorswieck，《面向支持 NFV 的空中网络的基于人工智能、具备移动性感知能力的节能资源分配与轨迹设计》，《IEEE 绿色通信与网络汇刊》，第 1–1 页，2022 年。[9] M. A. Abd-Elmagid、N. Pappas 和 H. S. Dhillon，《信息年龄在物联网中的作用》，《IEEE 通信杂志》，第 57 卷，第 12 期，第 72–77 页，2019 年。[10] O. M. Bushnaq、A. Chaaban 和 T. Y. Al-Naffouri，《无人机物联网网络在未来野火探测中的作用》，《IEEE 物联网期刊》，第 8 卷，第 23 期，第 16 984–16 999 页，2021 年。[11] S. Kaul、M. Gruteser、V. Rai 和 J. Kenney，《车联网中信息年龄最小化》，载于 IEEE 第八届年度通信学会传感器、网状网与自组织通信网络会议论文集，2011 年，第 350–358 页。[12] M. Akbari、M. R. Abedi、R. Joda、M. Pourghasemian、N. Mokari 和 M. Erol-Kantarci，《利用深度强化学习在工业物联网中进行信息年龄感知的虚拟网络功能调度》，《IEEE 选定通信领域期刊》，第 39 卷，第 8 期，第 2487–2500 页，2021 年。[13] C. Guo、X. Wang、L. Liang 和 G. Y. Li，《智能车联网中的信息年龄、时延与可靠性》，《IEEE 网络》，第 1–8 页，2022 年。[14] J. Hu、H. Zhang、L. Song、R. Schober 和 H. V. Poor，《协作式无人机互联网：基于多智能体深度强化学习的分布式轨迹设计》，《IEEE 通信汇刊》，第 68 卷，第 11 期，第 6807–6821 页，2020 年。[15] F. Wu、H. Zhang、J. Wu、Z. Han、H. V. Poor 和 L. Song，《无人机到设备的底层通信：基于多智能体深度强化学习的信息年龄最小化》，《IEEE 通信汇刊》，第 69 卷，第 7 期，第 4461–4475 页，2021 年。[16] W. Li、L. Wang 和 A. Fei，《通过路径规划减少无人机辅助数据感知中的数据包过期丢失》，《IEEE 无线通信快报》，第 8 卷，第 6 期，第 1520–1523 页，2019 年 12 月。[17] J. Liu、P. Tong、X. Wang、B. Bai 和 H. Dai，《无线传感器网络中面向信息新鲜度的无人机辅助数据收集》，《IEEE 无线通信汇刊》，第 1–1 页，2020 年。[18] H. Hu、K. Xiong、G. Qu、Q. Ni、P. Fan 和 K. B. Letaief，《无线供能物联网网络中的信息年龄最小化轨迹规划与数据收集》，《IEEE 物联网期刊》，第 8 卷，第 2 期，第 1211–1223 页，2021 年。[19] Z. Jia、X. Qin、Z. Wang 和 B. Liu，《无人机辅助物联网网络中的基于信息年龄的路径规划与数据采集》，载于 IEEE 国际通信会议研讨会论文集（ICC Workshops），2019 年，第 1–6 页。[20] M. A. Abd-Elmagid、A. Ferdowsi、H. S. Dhillon 和 W. Saad，《用于最小化无人机辅助网络信息年龄的深度强化学习》，载于 IEEE 全球通信会议论文集（GLOBECOM），普阿科，夏威夷州，美国，2019 年 5 月。[21] A. Ferdowsi、M. A. Abd-Elmagid、W. Saad 和 H. S. Dhillon，《用于无人机辅助网络中信息年龄最优联合轨迹与调度设计的神经组合深度强化学习》，《IEEE 选定通信领域期刊》，第 39 卷，第 5 期，第 1250–1265 页，2021 年。[22] M. A. Abd-Elmagid 和 H. S. Dhillon，《无人机辅助物联网网络中的平均峰值信息年龄最小化》，《IEEE 车辆技术汇刊》，第 68 卷，第 2 期，第 2003–2008 页，2019 年。[23] S. Zhang、H. Zhang、Z. Han、H. V. Poor 和 L. Song，《蜂窝式无人机互联网中的信息年龄：感知与通信权衡设计》，《IEEE 无线通信汇刊》，第 19 卷，第 10 期，第 6578–6592 页，2020 年。[24] M. Sun、X. Xu、X. Qin 和 P. Zhang，《面向物联网网络的信息年龄与能耗感知无人机辅助数据收集：一种深度强化学习方法》，《IEEE 物联网期刊》，第 8 卷，第 24 期，第 17 275–17 289 页，2021 年。[25] M. Samir、C. Assi、S. Sharafeddine、D. Ebrahimi 和 A. Ghrayeb，《信息年龄感知的轨迹规划

智能交通系统中的无人机：一种深度学习方法，”IEEE Trans. Veh. Technol., vol. 69, no. 11, pp. 12382–12 395, 2020.[26] S. F. Abedin, M. S. Munir, N. H. Tran, Z. Han, and C. S. Hong, “数据新鲜度与节能无人机导航优化：一种深度强化学习方法，”IEEE Trans. Intell. Transp. Syst., pp. 1–13, 2020.[27] O. S. Oubbati, M. Atiquzzaman, A. Lakas, A. Baz, H. Alhakami, and W. Alhakami, “多无人机赋能的感知信息年龄无线供能通信网络：一种多智能体强化学习策略，”in Proc. IEEE Int. Conf. Comput. Commun. Workshops (INFOCOM研讨会), 2021, pp. 1–6.[28] A. Baknina and S. Ulukus, “能量采集广播信道的最优及近最优在线策略，”IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3696–3708, 2016.[29] D. Shaviv and A. Ozgur, “能量采集节点的普遍近最优在线功率控制，”IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3620–3631, 2016.[30] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, “有限容量电池能量采集传感器的最小信息年龄传输：在线策略，”IEEE Trans. Inf. Theory, vol. 66, no. 1, pp. 534–556, 2020.[31] A. Al-Hourani, S. Kandeepan, and S. Lardner, “实现最大覆盖范围的低空平台最优高度，”IEEE Wireless Commun. Lett., vol. 3, no. 6, pp. 569–572, 2014.[32] Y. Zeng, X. Xu, and R. Zhang, “无人机辅助组播中用于最小化完成时间的轨迹设计，”IEEE Trans. Wireless Commun., vol. 17, no. 4, pp. 2233–2246, 2018.[33] L. Liu, K. Xiong, J. Cao, Y. Lu, P. Fan, and K. B. Letaief, “射频无线能量传输下无人机辅助数据收集的平均信息年龄最小化：一种深度强化学习方案，”IEEE Internet Things J., vol. 9, no. 7, pp. 5216–5228, 2022.[34] F. A. Oliehoek, C. Amato et al.,《去中心化部分可观测马尔可夫决策过程简明导论》。施普林格，2016年，第1卷。[35] C. J. Watkins and P. Dayan, “Q学习，”机器学习，vol. 8, no. 3, pp. 279–292, 1992.[36] R. S. Sutton and A. G. Barto，《强化学习：导论》。麻省理工学院出版社，2018年。[37] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, “深度强化学习中的多智能体协作与竞争，”PLOS ONE，vol. 12, no. 4, p. e0172395, 2017.[38] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, “QMIX：用于深度多智能体强化学习的单调价值函数分解，”arXiv:1803.11485, 2018.[39] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “门控循环神经网络在序列建模中的实证评估，”arXiv:1412.3555, 2014.[40] H. Sak, A. Senior, and F. Beaufays, “基于长短期记忆的大词汇量语音识别循环神经网络架构，”arXiv:1402.1128, p. arXiv:1402.1128, Feb. 2014.[41] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, “面向移动设备的量化卷积神经网络，”in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2016.[42] M. Hausknecht and P. Stone, “面向部分可观测马尔可夫决策过程的深度循环Q学习，”arXiv:1507.06527, 2015.[43] S. Leng and A. Yener, “无线自组织网络的信息年龄最小化：一种深度强化学习方法，”in Proc. IEEE Global Commun. Conf. (GLOBECOM). IEEE, 2019, pp. 1–6.[44] H. Hu, K. Xiong, G. Qu, Q. Ni, P. Fan, and K. B. Letaief, “无人机辅助无线供能物联网网络中的信息年龄最小化轨迹规划与数据收集，”IEEE Internet Things J., vol. 8, no. 2, pp. 1211–1223, 2021.[45] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “用于节能物联网通信的移动无人机，”IEEE Trans. Wireless Commun., vol. 16, no. 11, pp. 7574–7589, 2017.[46] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, “通过深度多智能体强化学习学习通信，”in Proc. 29th Adv. Neural Inf. Process. Syst. (NeurIPS), 2016, pp. 2137–2145.