用于多无人机系统协同移动接入的量子多智能体演员-评论家网络

Chanyoung Park、Won Joon Yun、Jae Pyoung Kim、Tiago Koketsu Rodrigues、
Soohyun Park、Soyi Jung，*IEEE*会员，以及Joongheon Kim，*IEEE*高级会员

arXiv:2302.04445v2 [cs.MA] 7 Jun 2023

摘要——**本文提出一种名为量子多智能体演员-评论家网络（QMACN）的新算法，用于自主构建采用多架无人机（UAV）的鲁棒移动接入系统。为促进多架无人机之间的协作，多智能体强化学习（MARL）技术被视为一种有前景的方法。这些方法使无人机能够协同学习，在共享环境中优化各自的动作，最终实现更高效的协同行为。此外，本研究采用量子计算（QC）原理，以增强相关无人机的训练过程和推理能力。通过利用量子计算独特的计算优势，我们的方法旨在提升无人机系统的整体效能。然而，由于量子比特使用所带来的含噪中等规模量子（NISQ）限制，采用量子计算会面临可扩展性挑战。所提出的算法通过实现量子集中式评论家，有效缓解了NISQ限制带来的约束。此外，通过多项数据密集型评估，验证了QMACN在训练速度和无线服务质量方面的性能提升优势。本文还验证了噪声注入方案可用于应对环境不确定性，从而实现鲁棒移动接入。**

索引词—**非地面网络（NTN）、自主移动连接系统、空中基站、多智能体系统、去中心化部分可观测马尔可夫决策过程（Dec-POMDP）、量子神经网络。**

I. 引言

A. 背景与动机

高机动性、灵活性和可达性是无人机 (UAVs) 的代表性优势，有助于为 5G 或 6G 以外的广泛服务提供支持 [2]–[6]。无人机部署持续增长，据美国联邦航空管理局 (FAA) 预测，到 2024 年，无人机机队数量将达到 160 万架 [7]。无人机应用包括实时监控 [8],包裹递送 [9]，智能工厂 [10], [11],以及灾害监测/管理 [12]。值得注意的是，无人机提供无线通信服务，无需固定的地面基础设施即可构建移动接入网络，从而提供更广泛的无线覆盖和实时服务 [13], [14]。无人机能够在军事问题或灾害等紧急情况下，以及因技术或经济原因难以安装地面基站的偏远地区，迅速与地面用户建立无线连接 [15]–[17]。借助无人机开发移动接入系统是一项具有挑战性的任务，因为实际环境中存在大量难以预料的变量和不确定因素。这些变量可能包括障碍物、强风、碰撞、有限的能源储备以及潜在故障。因此，设计一种能够自主且自适应地确定最佳无人机轨迹的机器学习算法，是预测并应对上述环境不确定性的另一种方法。

在众多机器学习（ML）和深度学习（DL）算法中，强化学习（RL）在需要序贯决策的任务中展现出最为出色的性能 [18]。强化学习能够在动态网络系统中有效提供资源管理服务，包括无人机辅助移动接入 [19]。此外，强化学习具有可扩展性，因此可以使用多智能体强化学习（MARL）来控制多架无人机。然而，在 MARL 架构中，无人机之间会发生各种交互，这些交互有助于同时训练其策略，并会影响彼此的策略训练表现。此外，由于物理限制，无人机只能观察到动态环境的部分信息。这些因素使 MARL 环境具有非平稳性。

利用量子纠缠的量子多智能体强化学习（QMARL）能够处理上述非平稳性问题 [20]。利用量子计算（QC）构建量子神经网络（QNN），与传统神经网络相比，可以使用更少的模型参数，高效利用计算资源，从而实现更可靠、更快速的学习 [1],[21], [22]。现有研究采用一种称为*CommNet* 的结构，执行基于集中式训练与分散式执行（CTDE）的计算，以实现多个智能体之间的协作 [23]。*CommNet* 是一种将所有智能体的状态信息作为输入，并在单个神经网络中学习所有智能体动作的结构。如果 m 个智能体将状态表示为 n维向量，则 CommNet 的输入将增加至 $m \times n$。因此，随着智能体数量增加，输入规模呈线性增长。然而，由于使用量子比特时受到近中等规模量子（NISQ）限制，这种线性扩展在量子计算环境中非常繁重，因为其使用

本研究的部分成果曾发表于意大利博洛尼亚举行的*IEEE*国际分布式计算系统会议（*ICDCS*），时间为7月 2022 [1]。本研究得到JSPS/NRF/NSFC A3前瞻计划的资助。（通讯作者：*Soohyun Park*、*Soyi Jung*）C. Park、W.J. Yun、J.P. Kim、S. Park和J. Kim任职于韩国首尔02841高丽大学电气工程学院（电子邮箱：{cosdeneb,ywjoon95,paulkim436,soohyun828,joongheon}@korea.ac.kr）。T.K. Rodrigues任职于日本仙台东北大学信息科学研究生院（电子邮箱：tiago.gama.rodrigues@it.is.tohoku.ac.jp）。S. Jung任职于韩国水原16499亚洲大学电气与计算机工程系（电子邮箱：sjung@ajou.ac.kr）。

量子比特，从而导致可扩展性问题 [24]。

IBM 和 Google 等公司制造的量子计算机体积庞大。因此，量子无人机的设想或许显得不切实际、难以实现。然而，随着微型量子计算机的发展，这一设想正迅速成为现实。迄今为止，已有多种成功研发的小型量子计算机。例如，Quantum Brilliance 开发的 Gen1 微型量子计算机采用 5 个量子比特，尺寸适合安装在机架式设备中 [25]，[26]。它的体积明显小于传统量子计算机，适合放置在桌面或台面上。此外，SpinQ 也成功研发了更小型的量子计算机 Gemini mini [27]。虽然它只能处理 2 个量子比特，但体积甚至比 Quantum Brilliance 的产品还小，因此与无人机的兼容性更高。最后，Alpine Quantum Technologies 研发了一台机架尺寸、具有 20 个量子比特的量子计算机 [28]。尽管上述量子计算机无法安装在无人机上，但它的重要意义在于证明了量子计算机的体积能够大幅缩小，同时仍保留可用量子比特的数量。从上述案例可以看出，可用量子比特的数量与量子计算机的体积直接相关，而微型量子计算机的性能不可避免地受到限制。然而，这一领域的研究仍在持续推进，随着技术进步，量子比特数量缩减的幅度也在不断减小。因此，在不久的将来，采用高性能小型量子计算机的量子无人机完全有望实现，这也进一步证明了本研究的实用性。基于量子计算和量子多智能体强化学习的发展，相关的深度学习和人工智能技术也可用于提升多无人机自主网络在移动性控制、网络调度等方面的性能。然而，无人机的运行极易受到环境变化的影响（例如天气会影响无人机的移动），这可能导致相关数据中出现大量噪声。因此，面向无人机系统的量子计算和量子多智能体强化学习解决方案在设计时应考虑这些噪声，并针对无人机的具体特性进行定制。

B. 算法设计原理

针对量子领域中随着所用量子比特数量增加而产生的量子误差可扩展性问题（如前所述），我们采用一种结构：由一个集中式评论家网络迭代训练多个行动者 网络 [1]。本文提出一种基于集中训练分散执行（CTDE）的训练算法，称为量子多智能体行动者-评论家网络（QMACN）算法，该算法利用量子计算（QC）的优势即量子纠缠，以及以更少参数高效利用计算资源）。如图 1 所示，本研究利用所提出的算法，构建了一种采用无人机的自主移动连接基础设施。在这一范式中，位于主服务器中的集中式评论家 在整个训练过程中评估每架无人机的决策能力。随后，

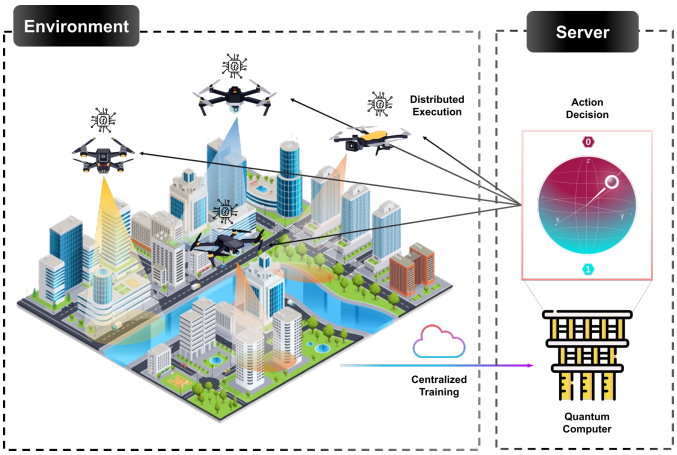


图 1. 所提出的量子多无人机系统移动接入架构。

集中式训练阶段完成后，无人机群通过采用已学习的参数化策略，协同向地面用户提供无线通信服务。借助所提出的以集中训练分散执行为核心的量子多智能体强化学习（QMARL）算法，由于集中式评论家实现了信息共享，因此无需进行无人机间通信。

最后，本文通过考虑影响无人机的环境信息噪声来设计真实环境，而这种噪声通常会阻碍无人机按预期运行。考虑到环境因素，本文通过性能评估证实，无人机能够更具适应性地应对环境噪声。

C. 贡献

我们的工作贡献如下。

- 本文提出了一种名为 QMACN 的新型算法，通过训练多架无人机在不确定且动态的环境中寻找最优轨迹，使其能够协作自主提供毫米波通信服务。

- 本文提出了一种新颖的集中训练分散执行框架，用于实现量子多智能体强化学习，同时缓解量子计算的局限性，即含噪中等规模量子时代量子比特数量增加所引发的量子误差。

- 本文通过噪声注入设计了一个逼真的环境，以构建多无人机系统中的鲁棒移动接入，其中的噪声是基于真实环境考虑的。强化学习的一个典型特征是训练智能体，使其能够自适应地响应动态要素。

- 本文通过在策略训练及其推理过程中开展各种数据密集型评估，证实了量子多智能体强化学习和真实噪声反映在训练策略方面的优势。

- 虽然 [1]，[10] 中的先前研究也提出了 QMACN 架构，但它们的应用分别主要面向单跳调度和智能工厂。相比之下，本研究着重开发自主移动接入系统，同时考虑具体的无人机规格和真实场景。此外，我们以更

全面的方式展示了量子优越性，采用通过噪声注入或数学计算复杂度分析获得的实验结果。

D. 论文组织结构

本文其余部分的组织如下：第II节介绍几篇相关工作的文章及相关技术。第III节回顾预备知识。第IV节描述基线模型的详细信息。第V节提出一种用于多无人机移动接入的新型量子多智能体强化学习算法。第VI节通过数据密集型仿真评估其性能。最后，第VII节对全文进行总结。

II. 相关工作

A. 通过无人机网络实现移动蜂窝接入

大量研究致力于蜂窝服务中无人机部署、轨迹和路径规划的优化，旨在改善无线通信系统。 [29] 中的研究旨在最小化无人机部署过程中的传输功率，同时考虑用户需求波动等各种不确定性。 [30], 通过结合K均值和Q学习来提升系统总容量； [31]则在地面基站与空中基站共存的场景下，同时考虑上行链路和下行链路干扰。在 [32], 中，提出了一种三维多无人机部署策略，在考虑同信道干扰的同时，确保不同用户分布下的服务质量（QoS）要求。 [33]的目标是在尽可能少地使用无人机基站的情况下最大化覆盖范围。此外， [34] 采用凸松弛技术，研究了多架无人机为小区边缘用户提供无线服务时的最优路径规划。为适应移动用户， [35]分析了飞艇式飞行基站（FlyBSs）的位置部署，并研究总容量与能耗之间的平衡。最后， [36]研究了一种基于圆填充的算法，用于优化准静态部署系统，旨在最大化整体吞吐量。尽管以往研究在各自目标方面表现出令人满意的性能，但其解决方法主要针对集中式优化问题。因此，上述方法无法为高度动态且大规模的无人机辅助网络提供实时解决方案。为应对这些挑战，本文提出了一种基于集中训练分散执行（CTDE）的多智能体强化学习（MARL）框架的方法，能够以分布式方式有效解决上述挑战。

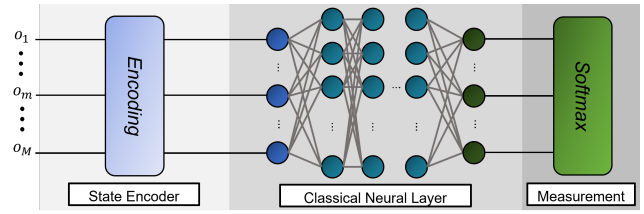
随后，有研究采用量子计算提出利用量子无人机构建通信网络，从而展示了量子无人机的可行性。在最初的研究中，量子无人机生成并传输纠缠光子对以传递信息；这种做法切实可行，因为它不需要高度先进的量子计算机。研究首先构建了一个仅由一对量子无人机构成的简单网络，以验证这一概念 [37]。之后，该研究通过增加量子无人机的数量和量子通信网络的规模进一步拓展 [38]。尽管网络规模扩大，且纠缠光子不稳定，该研究仍成功地将数据从一个地点准确传输到另一个地点。最后，也有研究探讨了在大都市环境中使用量子无人机 [39]。该研究的场景中有一支量子无人机群，它们充当信号中继节点来构建量子通信网络。因此，当地面量子服务器通过光子传输数据时，量子无人机群会放大数据信号，使其能够传输较长距离。由此，数据损失得以最大限度地降低，并实现了高效的量子通信网络。通过这些研究，我们确认量子无人机切实可行，并且能够投入实际应用。此外，随着微型量子计算机不断发展，未来不久还将实现复杂得多的任务。

B. 无人机与多智能体强化学习的应用

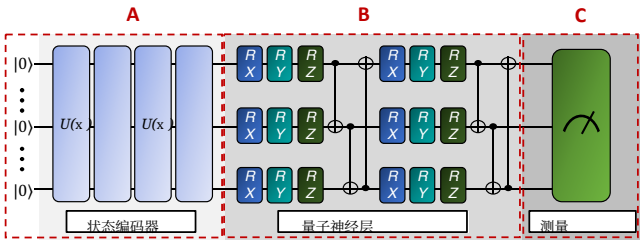
无人机能够提供空中视角、覆盖范围和机动性，从而有望在各种应用中增强多智能体强化学习系统的能力。通过多智能体强化学习训练的一支无人机机队可以执行监视和监测任务，例如出于安全目的进行目标跟踪或区域监测 [40]–[42]。配备传感器和摄像机的无人机可以采集空气质量、温度和湿度等环境因素的数据 [43]–[45]。可以使用多智能体强化学习训练一支无人机队伍，使其协调行动，覆盖广阔区域以采集数据。此外，还可以通过多智能体强化学习训练一组无人机，使其在灾区协同执行搜救任务，识别幸存者、评估损毁情况并提供援助 [46], [47]。最后，无人机可以在偏远地区或受灾地区建立临时通信网络。通过多智能体强化学习训练的一组无人机可以充当中继节点，彼此传递消息，直至消息到达预定目的地 [48]–[50]。在我们的研究中，目标是在无人机应用领域利用量子计算的量子优越性。与传统的神经网络驱动的多智能体强化学习相比，量子优越性表现为参数数量更少、计算复杂度更低。

C. 既有工作

我们之前将 QMARL 应用于网络与移动系统的工作见 [1], [10]。 [1], 设计并展示了一种并行单跳网络调度算法。将 QMARL 用于这种单跳网络调度可以显著提升性能。显然，这种网络调度的建模方式与我们考虑的多无人机移动控制有根本区别。 [10],讨论了基于 QMARL 的自主移动车辆移动控制。尽管 [10] 中提出的算法也考虑了移动控制，但它与本文提出的算法并不相同，因为本文包括：(i) 考虑动作和状态中的实际噪声因素，而这些因素给 MARL 计算带来了挑战；(ii) 计算复杂度分析；(iii) 考虑无人机能量模型，而 [10], 并未进行这一研究；最后，(iv) 构建空中环境与任务，这与智能工厂环境及任务 [10] 完全不同。



(a) 经典神经网络。



（b）量子神经网络。

图 2. 用于强化学习的、基于经典计算和量子计算的神经网络架构。

III. 量子机器学习

A. 量子比特

量子比特是量子计算中的基本信息单位。由于量子比特由两个基的组合表示—— $|0\rangle$ 和 $|1\rangle$，因此其取值可以是 0 到 1 之间的任意值。由这两个基共同构成的量子比特处于叠加 态，也可用于表示量子态 [51]。此外，量子比特之间还可以发生纠缠，从而显著增强两个独立量子比特之间的相关性。量子比特的这一特性使量子计算相比经典计算能够容纳并控制更多信息。假设有一个包含 q 个量子比特的系统，该系统希尔伯特态空间中的量子态可表示为 $|\psi\rangle = \alpha_1|0\cdots0\rangle + \cdots + \alpha_{2^q}|1\cdots1\rangle$，其中 α 表示各基的概率振幅，即，$\sum_{i=1}^{2^q}\alpha_i^2 = 1$ 和ψ 表示量子态。由于 α 是复数，因此量子态可表示为布洛赫球上的一个点。

B. 量子神经网络

为了使用量子比特设计和计算 QNN，量子比特应当具备可控性，以便训练神经网络。这种控制是通过利用基本量子门来实现的，用于控制量子比特在布洛赫球上的位置。基本量子门的代表性示例是旋转门，分别表示为 R_x、R_y 和 R_z，用于绕 x 轴、 y 轴和 z 轴旋转。更具体地说，这些门函数作为作用于单个量子比特的幺正运算执行，使其沿 x 轴、 y 轴和 z 轴的给定方向旋转特定角度。这些门不仅能够控制量子比特，还能对经典比特级数据进行编码。虽然基本量子旋转门是单量子比特门只能同时应用于单个量子比特，但也存在能够同时作用于两个或更多量子比特的多量子比特门。例如，*CNOT* 门通过对两个量子比特执行*XOR*运算，使多个量子比特之间产生纠缠 [52]。

基于上述理论和概念，通过组合各种类型的量子门来构建量子神经网络模型。传统的量子神经网络模型由以下三个组件构成：*i)* 状态编码电路（参见图 2(b) 中的 A）、*ii)* 参数化量子电路（PQC）（参见图 2(b) 中的 B）以及*iii)*量子测量（参见图 2(b) 中的 C）层。关于这三个连续组件的进一步说明将在后续章节中给出。

1）状态编码电路：首先，编码层的作用是将经典数据编码为量子态，因为量子电路无法将经典比特作为输入。因此，状态编码器将 q 个 $|0\rangle$ 传入一组旋转门，并使用作为参数的经典数据（记为 θ_{enc}），从而将比特转换为量子比特。此外，输入数据X 被拆分为 $[x_1 \cdots x_N]$，以便分别用作参数，其中 N 是输入 X的拆分数据数量。随后，编码层的输出量子态将包含经典数据的信息。

2）参数化量子电路：其次，参数化量子电路（PQC）负责执行所需的计算，其功能等同于经典神经网络（NN），尤其是图 2(a) 所示的累积隐藏层乘法。在参数化量子电路层中，输入量子态通过量子门按特定角度旋转，从而产生所需的值，例如动作值和状态值。在本文中，量子比特通过受控通用（CU）门进行计算，该门能够灵活控制旋转方向、纠缠和解纠缠。图 2(b) 展示了量子神经网络（QNN）模型的结构，可以看到编码层与 CU3 层依次重复多次。这种特定结构采用了数据重上传技术 [53], ，可同时对量子比特进行编码和旋转。因此，每个量子比特的计算效率都得到最大化，即所需的量子比特数量减少，从而生成多智能体强化学习（MARL）所需的值。

3）量子测量：最后，由PQC产生的量子态成为测量层的输入。在这一阶段，对输入进行测量，从而将量子数据解码回经典数据以进行优化。测量操作等价于相对于 z 轴的投影矩阵乘法。尽管 z轴最常用于测量，但也可以是任何其他适当定义的方向。对量子态进行测量后，量子态会坍缩，并成为一个可观测量。在解码过程后，该可观测量用于最小化损失函数。随后，应对其进行微分以用于反向传播。然而，量子数据无法进行微分，因为应用链式法则会使量子比特的状态完全坍缩。因此，量子神经网络训练采用通过可观测量损失函数的对称差商获取损失梯度的技术。

C. 量子强化学习

量子电路最终的可观测量值表示多智能体强化学习计算中一个智能体的动作概率。如图2(a)所示，它与*Softmax*函数相一致。利用经验回放缓冲区中的信息，可以计算环境中智能体的动作分布。随后，当智能体基于该动作进行决策时，将产生新的状态、观测和奖励数据，以进行下一次重复学习迭代。先前研究已验证，量子电路的性能优于由相同参数构成的基于深度神经网络的强化学习方法 [21], [54]。 [1]中的研究是开发经典多智能体强化学习算法纯量子版本的早期尝试之一。通过在单跳数据卸载环境中进行多次仿真，该研究确认量子多智能体强化学习利用量子计算的优势超越了其经典对应方法。随后， [10]中的研究通过在智能工厂场景中实施上述算法，提出了一项新颖贡献。尽管算法本身并未获得实质性改进，该工作仍通过严谨的仿真验证了该量子算法的实际有效性。此外，还探索了利用一种称为极点记忆的独特量子属性来开展元学习的潜力 [55]。结果表明，量子强化学习通过少样本学习执行多项任务，使量子强化学习具有很高的效率。本文算法利用量子计算实现弹性移动接入，并将量子多智能体强化学习与传统多智能体强化学习进行性能评估比较，针对本文所研究的问题展示了量子优越性。尽管 [1], [10] 中的先前研究也采用了量子多智能体强化学习框架，但其应用主要分别侧重于单跳调度和智能工厂。此外，与所述研究相比，本文通过采用第VI节中噪声注入获得的实验结果以及第V-C节中的数学计算复杂度分析，对量子优越性进行了更为全面的论证。

IV. 模型

A. 基于多无人机网络的移动接入模型

我们提出了一种名为 QMACN 的算法，旨在构建可靠且稳健的自主多无人机网络，以便在动态环境中应用 QMARL，如图 3 所示。如第 I 节所述，该算法应用于无人机支持的自主移动接入系统，旨在应对含噪中等规模量子时代因过度使用量子比特而带来的可扩展性挑战。在图 3 所示的示意图中，众多无人机作为执行者，在为地面用户提供无线通信服务的同时，收集包括自身位置、剩余能量水平、执行的动作以及获得的奖励在内的环境经验。随后，这些无人机将经验传送至中央服务器。在本文中，假设服务器的队列状态理想，因此不会产生任何延迟。本文还考虑了这样一种情况：若发送某架无人机的经验时通信失败，则该无人机会像标准传输控制协议（TCP）那样不断重传，直至发送成功。在中央服务器中，集中式评论家通过评估每架无人机在特定状态下的决策能力，为多个执行者 学习参数，并在整个训练过程中发挥作用。集中式训练阶段结束后，多架无人机依据各自学习到的策略，以分散方式协作，为地面用户提供无线通信服务。在集中训练分散执行环境中，无人机在向中央服务器发送经验用于训练时，可能会遇到各种不确定性，这些不确定性表现为环境噪声。无人机之间的交互会影响其他智能体的训练策略，各种噪声因素也会产生影响。这正是使用强化学习的原因。为在本文中考虑真实环境，我们将两类噪声输入多智能体强化学习系统：*i*）状态噪声和*ii*）动作噪声。

B. 无人机模型

由于无人机配备了机载电池，其运行受到电池容量有限的制约 [56]。当无人机因电池电量耗尽而发生故障时，可能会坠落至地面。这种情况可能导致其他无人机或地面用户受到身体伤害，最终损害所提供服务的可信度。因此，监测和管理无人机的剩余能量，对于确保服务可靠性至关重要 [57]。通常，无人机在移动接入中的能耗主要来自航空和无线通信。然而，本文仅考虑航空能耗，因为通信相关能耗（几瓦）远低于航空相关能耗（几百瓦） [13]。第 m 架无人机悬停时的能量消耗模型可用数学形式表示如下 [58]:

$$\mu_m(t) = \underbrace{\frac{\delta}{8}\rho s A \Omega^3 R^3}_{blade profile,\ P_o} + \underbrace{(1+k)\frac{W^{3/2}}{\sqrt{2\rho A}}}_{induced,\ P_i}, \qquad \text{（悬停）}, \quad (1)$$

其中， δ、 ρ、 s、 A、Ω、 R、 k和 W 分别表示剖面阻力系数、空气密度、旋翼实度、旋翼盘面积、桨叶角速度、旋翼半径、诱导功率增量修正系数，以及包含电池和螺旋桨的飞行器重量。此外，第 m架无人机往返飞行时的能量消耗模型可用数学形式表示如下 [58]:

$$\mu_m(t) = \underbrace{P_o\left(1+\frac{3v^2}{U_{tip}}\right)}_{blade profile} + \underbrace{P_i\left(\sqrt{1+\frac{v^4}{4v_0^4}}-\frac{v^2}{2v_0^2}\right)^{0.5}}_{induced} + \underbrace{\frac{1}{2}d_0\rho s A v^3}_{parasite}, \qquad \textit{(round-trip traveling)}, \quad (2)$$

其中， P_o、 P_i、 v、 U_{tip}、 v_0和 d_0 分别表示桨叶型阻功率、诱导功率、飞行速度、旋翼桨叶尖端速度、平均旋翼诱导速度和机身阻力比。

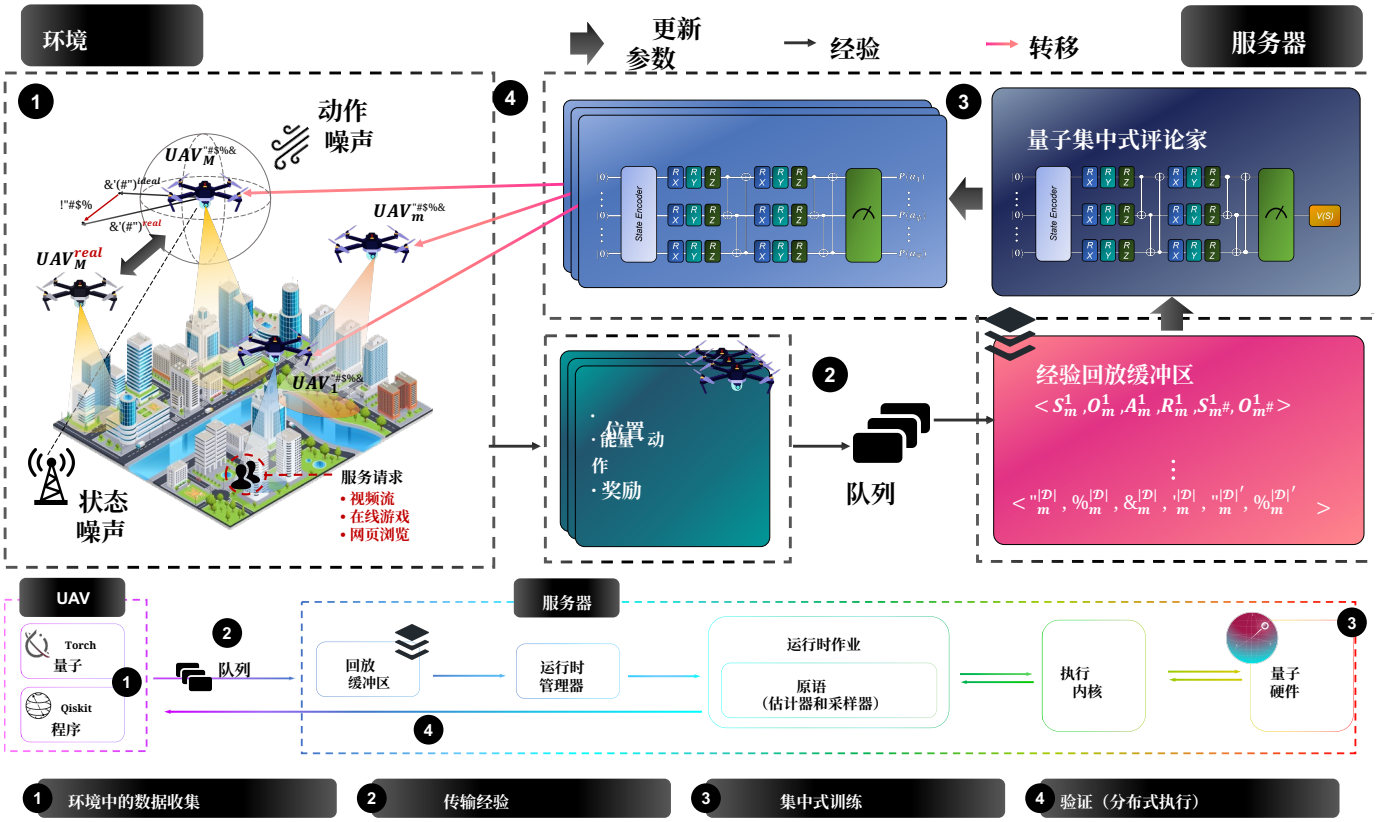


图 3. 用于构建自主量子多无人机网络的基于集中训练分散执行的训练流程。

C. 无人机定位中的噪声分布模型

状态噪声。由 QMACN 训练的无人机使用全球定位系统（GPS）确定其位置；许多移动系统和平台都可使用该系统 [59]。然而，GPS传感器存在噪声，即干扰、阻塞干扰和延迟锁定环（DLL）噪声，这些噪声可用高斯或非高斯分布建模。下文使用一种广泛采用的模型，即广义柯西噪声 [60]，来定义GPS传感器中的状态噪声。

$$p_{cy}(z,\sigma_z)=\frac{Y}{\left(1+v^{-1}[|z|/X]^k\right)^{v+1/k}}, \tag{3}$$

其中 $X=\left[\sigma_z^2\frac{\Gamma(1/k)\,(}{3/k}\right.$ $\left.\frac{}{2X\Gamma\ v\ \Gamma\ 1/k}=\right]^{1/2}$，$\int_0^\infty x^{t-1}e^{-x}\,dx$

$Y=\frac{kv^{-1/k}\Gamma_{(v+1/k)\,()()}}{}$，以及 $\Gamma(t)$（伽马函数）。在广义柯西密度的方差中，k 和 v 分别表示 GPS 噪声 [61]的冲击性和方差。我们在图 4(a) 中将状态噪声的概率分布函数（PDF）表示为广义柯西密度，其参数为 σ_z，其中设定 $k=0.20$、$v=40$和 $\sigma_z^2=0.22$ [60]。

动作噪声。该噪声采用 Weibull 分布建模，这是一种广泛用于风速建模的概率分布形式。根据 [62]，其表示如下：

$$p_{wb}(v,\sigma_v)=(c/A)\,(v/A)^{c-1}\exp(-vc/A), \tag{4}$$

其中，v 表示风速，$A=\left(\frac{\sigma_v}{\bar{v}}\right)^{-1.086}$和 $c=\frac{}{\Gamma\ \bar{v}\ (1+1/A)}$ 分别表示尺度参数和形状参数，风速的均值为 $\bar{v}$，标准差为 σ_v，这些数据取自数据 [62]。三年间（1999年6月1日至2002年5月31日）共进行了145，048次气象测量。图4(b)为威布尔分布的概率分布函数，其中 $A=10.97\,m/s$ 和 $c=2.29$ 用于表示风速的概率。此外，图4(c)表示风可能吹来的12个方向的概率。注意，N、S、E和W分别表示北、南、东和西方向。

现实与理想情况之间差距的建模。考虑到上述状态噪声，我们重新配置演员-评论家模型。本文同时考虑理想情况和现实状态/动作情形。理想情况不考虑环境中的噪声，而另一种情况则考虑噪声。这两种情况之间的关系可建模如下：

$$s_{\text{real}}=s_{\text{ideal}}+n_s, \tag{5}$$

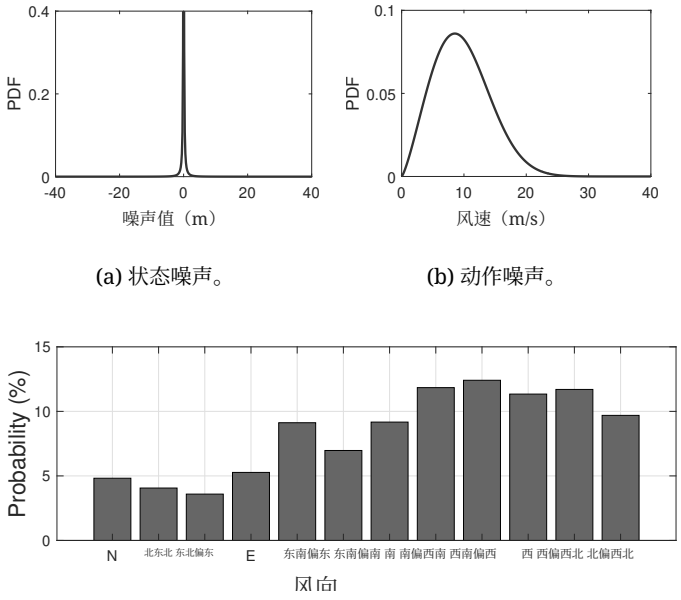
$$a_{\text{real}}=a_{\text{ideal}}+n_a, \tag{6}$$

其中，n_s 和 n_a 分别表示状态和动作中的噪声。请注意，$n_s\sim p_{cy}(z,\ \sigma_z)$ $\{x,y\}\in z$和 $n_a\sim p_{wb}(v,\ \sigma_v)$，其中 x 和 y 是无人机位置二维坐标中的基点。

V. MARL Algorithm for Multi-UAV Cooperation

A. 多智能体强化学习形式化

本文考虑的多无人机系统在数学上定义为去中心化部分可观测马尔可夫决策过程（Dec-POMDP），它广泛应用于集中训练分散执行的多智能体强化学习框架 [63]，，因为每架无人机受物理限制，只能依据



(a) 状态噪声。 (b) 动作噪声。

(c) 风向。

图 4. 状态噪声和动作噪声的概率分布函数。

部分环境信息进行分布式序贯决策。 M 架无人机的 Dec-POMDP 表示为 $\langle \mathcal{S}, \mathcal{A}, P, r, \mathcal{O}, \mathcal{Z}, \gamma \rangle$，其中 $\mathcal{S}$ 是状态集合，$s \in \mathcal{S}$； $\mathcal{A}$ 是动作集合， $a_m \in A_m \subset \mathcal{A}$，这里， a_m 表示第 m 架无人机的动作，它们构成无人机联合动作，即一个 $\in \mathcal{A}$，并且注意 A_m 表示第 m 架无人机的动作集合； P 是状态转移概率 P: $\Pr(s' \mid s, \mathbf{a}) = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{S}$，其中 a 为无人机联合动作； r 是给予每架无人机的奖励，满足 $r(s,\mathbf{a}\ s') = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$。此外， $\mathcal{O}$ 是观测集合，其中第 m 架无人机的观测记为 $o_m \in O_m \subset \mathcal{O}$，并且注意 O_m 表示第 m 架无人机的观测集合。这里，用于训练的无人机联合观测记为 o $\in \mathcal{O}$。 $\mathcal{Z}$ 是条件观测概率函数空间，即 $\mathcal{Z}(s', \mathbf{a}, \mathbf{o})$，而 γ 是折扣因子。

当第 m 架无人机在基于 $\mathcal{Z}$（s'，**a**，**o**）$= \mathcal{S} \times \mathcal{A} \to \mathcal{O}$进行观测的同时采取动作 a_m 时，状态将基于 P进行更新。完成此计算后，将基于 r（s，**a** s'）生成奖励。以下将详细介绍观测、状态、动作、奖励和目标。

观测。 第 m 架无人机根据其位置 $p_m \in \{x_m, y_m\}$ 部分观测自身的环境信息，其中 x_m 和 y_m 表示笛卡尔坐标。第 m 架无人机观测自身位置与其可观测范围内其他无人机之间的距离，即，

$$d_{mm'} = \begin{cases} \|p_m - p_{m'}\|_2, & \textit{if.} \quad \|p_m - p_{m'}\|_2 \le D_{th}, \\ -1, & \textit{(otherwise)}, \end{cases} \tag{7}$$

其中， D_{th} 和 $\|\cdot\|_2$ 分别表示观测范围和 L2 范数。此外，每架无人机都必须监测自身的能量水平 e_m，以避免因电池放电而导致的任何故障。观测第 m 架无人机的观测定义为 $\mathcal{O}_m \triangleq \{p_m, \bigcup_{m'=1}^{M}\{d_{mm'}\}, e_m\}$、$\forall m \in M$ 和 $\forall m' \in M$。

状态。 该状态由两个服务详情组成：*i)* 服务可用性 $c_{mn} \triangleq \{0,1\}$ 和 *ii)* 服务质量（QoS）q_{mn}。其中， c_{mn} 是一个变量，用于表示第 n 个用户是否由第 m 架无人机提供服务，具体通过其是否为零来区分。所提出网络中的总支持率可表示为 $\sum_{m=1}^{M}\sum_{n=1}^{N} c_{mn}$。然而，用户接收到的无线通信服务质量会因所请求的服务而异。因此，本文采用质量函数 [64] ，在假设用户随机请求不同数据速率等级（包括视频流、在线游戏或网页浏览）的情况下，评估最大可支持数据速率下的总服务质量 [65]。实际数据速率的详细计算方法见第 VI-A2 节。第 m 架无人机为第 n 个用户提供服务时，该用户的服务质量可表示如下:

$$q_{mn} = \begin{cases} \left(1 + \exp^{-w_a(\kappa_{mn} - w_b)}\right)^{-1}, & \textit{(video traffic)}, \\ \log\left(w_c \cdot \kappa_{mn} + w_d\right), & \textit{(otherwise)}, \end{cases} \tag{8}$$

其中， κ_{mn} 表示由第 m架无人机服务的第 n个用户的数据速率。此外，权重参数的取值为$w_a = 0.01$、 $w_b = 1024$、$w_c = 1$和 $w_d = 1$ [64]。这里，所有无人机同时具有相同的状态信息，因为总服务量会受到所有无人机的影响。因此，状态信息表示为$s \triangleq \bigcup_{m=1}^{M}\bigcup_{n=1}^{N}\{c_{mn}, q_{mn}\}$。

动作。 所有无人机在 t根据各自的策略依次采取动作。由于无人机处于二维笛卡尔坐标系（x， y） $\in \mathbb{R}^2$中，因此它们可以沿四个基点方向移动。于是，无人机可采取的动作集为 $\triangleq \{x_m \pm (v_m \times t)\ \ y_m \pm (v_m \times t)\}_{m=1}^{M}$，其中 v_m 表示无人机的速度。

奖励。 奖励 r（s， **a**）根据当前状态 s 和所有无人机所选的动作**a**生成。奖励函数如下:

$$r(s, \mathbf{a}, s') = w_c \times \sum_{m=1}^{M}\sum_{n=1}^{N}(c_{mn} \times q_{mn}) \times \tau \times \mathbb{1}(e_m), \tag{9}$$

其中， w_c 是用于使学习过程更加稳定的奖励权重； τ^t 是 t 处所有无人机的重叠率，必须降低该重叠率以减少它们之间的干扰；$\mathbb{1}(\cdot)$是用于区分其是否为零的指示函数。因此，可以看出，无人机根据奖励函数，试图在有限能量内最大化地面用户的支持速率和服务质量。

目标。 我们在多智能体强化学习中的主要目标表述如下:

$$\pi_{\boldsymbol{\theta}}^{*} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{s_{\text{real}} \sim E,\, \mathrm{a}_{\text{real}} \sim \pi_{\boldsymbol{\theta}}} \left[\sum_{t=1}^{T} \gamma^{t-1} \cdot r\left(s_{\text{real}}, \mathbf{a}_{\text{real}}, s'_{\text{real}}\right)\right], \tag{10}$$

其中， E、 T 和 $\gamma \in [0,1)$ 分别表示无人机所处的环境、回合长度和折扣因子。通过重新定义强化学习中的目标函数，无人机将考虑现实环境中的噪声（即 s_{real}、一个$_{\text{real}}$ 和 s'_{real}），从而使其策略能够更好地应对各种噪声。关于无人机如何利用重新配置的目标函数实现最优决策，详见第 V-B 节。

B. 用于协作式多无人机移动接入的 *QMACN*

QMARL 设计。第 I 节明确指出，量子比特数量有限会导致可扩展性问题。然而，在行动者-评论家训练方法 [66]，中，随着多智能体强化学习中智能体数量的增加，量子比特数量也必须增加。这一需求会引发量子误差，进而影响系统稳定性 [1], [24]。因此，我们提出一种新颖的 QMACN 算法，该算法采用集中训练、分散执行架构，其方法基于多智能体行动者-评论家强化学习框架 [1], [67]。在集中训练、分散执行架构中，设有一个集中式评论家和多个行动者网络，其中行动者网络的数量与去中心化部分可观测马尔可夫决策过程中智能体的数量相对应 [68]，如图 3 中的服务器所示。基于集中训练、分散执行架构的智能体分散地进行序贯决策，并通过评估集中式评论家网络的值，训练与策略对应的行动者网络，其表达式如下：

$$Q(o,a;\boldsymbol{\theta}) = \beta_a \langle O_a \rangle_{o,\boldsymbol{\theta}} = \beta_a \mathrm{Tr}(U^{a\dagger}(o;\boldsymbol{\theta}) M_a U^a(o;\boldsymbol{\theta})) \quad (11)$$

$$V(s;\boldsymbol{\phi}) = \beta_c \langle O \rangle_{s,\boldsymbol{\phi}} \quad = \beta_c \mathrm{Tr}(U^{c\dagger}(s;\boldsymbol{\phi}) M_c U^c(s;\boldsymbol{\phi})) \quad (12)$$

其中，算子 $\mathrm{Tr}(\cdot)$、$U(\cdot)$ 和 $(\cdot)^\dagger$ 分别表示迹算子、用于量子比特旋转的幺正操作，以及多个量子比特的纠缠和复共轭。测量量子态时，输出（称为观测值）介于 -1 和 1 之间，即 $\forall \langle O \rangle \in [-1,1]$；我们使用超参数（$\beta_a$，$\beta_c$）对演员网络和评论家网络进行充分训练。请注意，$M_a$ 和 M_c 是厄米矩阵。结合式（11）–（12）和超参数，演员-评论家网络可近似价值函数。

量子演员。在每个时间步 t，第 m 个量子演员根据其状态和观测信息，从当前可选动作中选择概率最大的动作，表示如下：

$$\mathrm{a}_{\mathrm{m,real}} = \underset{\mathrm{a}}{\mathrm{argmax}}\, \pi_{\boldsymbol{\theta}_m}(\mathrm{a}_{\mathrm{ideal}} | s_{\mathrm{ideal}} + n_s, o_m) + n_a, \quad (13)$$

满足约束

$$\pi_{\boldsymbol{\theta}_m}(\mathrm{a}_{\mathrm{ideal}} | s_{\mathrm{ideal}} + n_s, o_m) \triangleq \mathit{softmax}(Q(o,a;\boldsymbol{\theta}_m)), \quad (14)$$

$$\mathit{softmax}(\mathbf{x}) \triangleq \left[\frac{e^{x_1}}{\sum_{i=1}^N e^{x_i}}, \cdots, \frac{e^{x_N}}{\sum_{i=1}^N e^{x_i}} \right], \quad (15)$$

其中，$\mathit{Softmax}(\cdot)$ 是一种用于归一化输入的激活函数。通过使用它，我们可以提取式（11）中演员网络在可观测量$\langle O_a \rangle_{o,\,\boldsymbol{\theta}}$ 下所有动作的概率。

量子集中式评论家。集中训练分散执行（CTDE）包含一个集中式评论家，负责使用状态价值函数对当前状态进行评估，如下所示：

$$V_{\boldsymbol{\phi}}(s) = \langle O_c \rangle_{s,\boldsymbol{\phi}} \simeq \mathbb{E}_{s_{\mathrm{real}} \sim E,\, \mathrm{a}_{\mathrm{real}} \sim \pi_{\boldsymbol{\theta}}} \left[\sum_{t'=t}^{T} \gamma^{t'-t} \cdot r(s_{\mathrm{real}}, \mathbf{a}_{\mathrm{real}}, s'_{\mathrm{real}}) \right], \quad (16)$$

其中，s_t 是在 t 时测得的当前状态。我们还使用评论家网络的可观测量来评估当前状态的价值。

1）训练与推理：如式（10）所述，多智能体强化学习中的智能体试图最大化期望回报。我们利用集中式评论家网络的相应状态价值函数 V_ϕ，通过最大化共同目标来求取梯度。使用演员和评论家网络的参数（分别对应 $\boldsymbol{\theta}$ 和 $\boldsymbol{\phi}$），我们基于时序差分演员-评论家模型和贝尔曼最优方程构建多智能体策略梯度（MAPG），如下所示：

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{s_{\mathrm{real}}, \mathrm{o} \sim E} \left[\sum_{t=1}^{T} \sum_{m=1}^{M} \delta_{\boldsymbol{\phi}}^t \cdot \nabla_{\boldsymbol{\theta}_m} \log \pi_{\boldsymbol{\theta}_m}(\mathrm{a}_m^t | s^t, o_m^t) \right], \quad (17)$$

and

$$\nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}) = \sum_{t=1}^{T} \nabla_{\boldsymbol{\phi}} \left\| \delta_{\boldsymbol{\phi}}^t \right\|^2, \quad (18)$$

满足约束

$$\delta_{\boldsymbol{\phi}}^t = r\left(s^t, \mathbf{a}^t, s^{t+1}\right) + \gamma V_{\boldsymbol{\phi}}(s^{t+1}) - V_{\boldsymbol{\phi}}(s^t). \quad (19)$$

在上文给出的三个方程中，式（17）是演员网络的目标函数，方程中使用的神经网络参数通过梯度上升进行更新，以使其最大化，如下所示：

$$\boldsymbol{\theta}_m^{t+1} \approx \boldsymbol{\theta}_m^t + \alpha_{\mathrm{actor}} \times [\, \delta_{\boldsymbol{\phi}}^t \cdot \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}_m}(\mathrm{a}_m^t \,|\, s^t, \mathrm{o}_m^t)\,], \quad (20)$$

其中，α_{actor} 表示演员网络的学习率。式（18）对应集中式评论家网络的损失函数，应通过梯度下降将其最小化，如下所示：

$$\boldsymbol{\phi}^{t+1} \approx \boldsymbol{\phi}^t + \alpha_{\mathrm{critic}} \times [\, \delta_{\boldsymbol{\phi}}^t \cdot \nabla_{\boldsymbol{\phi}} V_{\boldsymbol{\phi}}(s^t)\,], \quad (21)$$

其中，α_{critic} 是集中式评论家网络的学习率。

我们介绍如何利用量子计算和经典计算获得损失梯度。为便于数学推导，以下将演员网络和评论家 网络分别记为 $\boldsymbol{\theta}$ 和 $\boldsymbol{\phi}$。它们的损失值通过式（19）中集中式评论家 $\delta_{\boldsymbol{\phi}}$ 的时序差分误差计算，其中演员/评论家的第 i个参数的导数表示如下：

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial J(\boldsymbol{\theta})}{\partial \pi_{\boldsymbol{\theta}}} \cdot \frac{\partial \pi_{\boldsymbol{\theta}}}{\partial \langle O \rangle_{o,\boldsymbol{\theta}}} \cdot \frac{\partial \langle O \rangle_{o,\boldsymbol{\theta}}}{\partial \theta_i}, \quad (22)$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\phi})}{\partial \phi_i} = \frac{\partial \mathcal{L}(\boldsymbol{\phi})}{\partial V_{\boldsymbol{\phi}}} \cdot \frac{\partial V_{\boldsymbol{\phi}}}{\partial \langle O \rangle_{s,\boldsymbol{\phi}}} \cdot \frac{\partial \langle O \rangle_{s,\boldsymbol{\phi}}}{\partial \phi_i}, \quad (23)$$

其中，式（22）–（23）右侧的一阶和二阶导数可以通过经典计算求得。然而，由于测量前量子态未知，无法计算后一个导数。因此，我们采用参数移位规则 [69]。通过将经典计算和量子计算的偏导数相乘，该规则使我们能够衔接经典计算与量子计算。参数移位规则应用于式（17）–（18），其中第 i个演员 参数的导数通过零阶导数得到，如下所示：

$$\frac{\partial \langle O \rangle_{o,\boldsymbol{\theta}}}{\partial \theta_i} = \langle O \rangle_{o,\boldsymbol{\theta} + \frac{\pi}{2}\mathbf{e}_i} - \langle O \rangle_{o,\boldsymbol{\theta} - \frac{\pi}{2}\mathbf{e}_i}, \quad (24)$$

其中，$\mathbf{e}_i$ 分别对应于 $\boldsymbol{\theta}$ 的第 i个基。类似地，公式 (23) 左侧的表达式由公式 (24) 得出。

最后，我们按照公式 (17)–(18) 中的详细说明计算目标函数的梯度。

2) 算法伪代码：所提出的基于集中训练分散执行的训练与推理过程详见算法 1，相关说明如下：

1) Initialize th参 t数 f *act*或 an d *centralized criti c*网络，分别为 $\boldsymbol{\theta}$ 和 $\boldsymbol{\phi}$（第 *1* 行）。2) 所有智能体通过重复执行以下过程（第 *2–19* 行）来学习其策略，直至所有训练轮数达到最大训练轮数：(i) 在每个训练轮次开始时，初始化环境以设置初始状态 s_0 （第 *3* 行）。(ii) 每架无人机在每个时间步根据其策略选择一个动作。执行动作后，环境转移到下一个时间步 s^{t+1} （第 *5–8* 行）。计算总奖励$\mathbf{r}^t$ 以及由无人机产生的转移对（即经验） $\xi = \{s^t, \mathbf{o}^t, \mathbf{r}^t, s^{t+1}, \mathbf{o}^{t+1}\}$，并将其存储在经验回放缓冲区中 $\mathcal{D}$ （第 *8–10* 行）。(iii) 无人机从 $\mathcal{D}$ 中随机采样小批量，以获取$V_{\boldsymbol{\phi}}$。通过这种方式，可以减少用于训练的数据的连续性，从而提升学习性能 [70]。需要注意的是，我们从无人机网络开始训练，以防止训练方向偏向初始数据（第 *12–18* 行）。(iv) 通过梯度下降更新集中式评论家 网络 $\boldsymbol{\phi}$ 的参数，使其损失函数趋向于式（18）中的值，从而减小损失函数值（第 *14–15* 行）。(v) 通过梯度上升更新所有 *actor* 网络 $\boldsymbol{\theta}$ 的参数，使其沿着目标函数值增大的方向趋向于式（17）中的目标函数值，该目标函数值由集中式评论家网络评估（第 *16* 行）。完成所有无人机的策略训练后，它们在环境中执行多次推理过程（第 *20–28* 行）。

算法1： QMACN 多无人机协作的训练与推理过程

多无人机协作

1 初始化演员 和集中式评论家网络的权重 分别记为 $\boldsymbol{\theta}$ 和 $\boldsymbol{\phi}$, $\forall m \in [1, M]$
2 **对于**训练轮次 = *1*, 最大训练轮次 **执行**
3 ▷ **初始化多无人机环境**，设置 s_0
4 **对于**时间步 = *1*, T **执行**
5 **对于**每架无人机 m**执行**
6 ▷ 根据其策略选择动作 a_m $\pi_{\boldsymbol{\theta}_m}(\mathrm{a}_m^t \mid s_{\mathrm{real}}^t, o_m^t)$ 在时间步 t
7 **end**
8 ▷ $s^t \rightarrow s^{t+1}$，$\mathbf{o}^t \rightarrow \mathbf{o}^{t+1}$，奖励为 $\mathbf{r}^t$
9 ▷设置 $\xi = \{s^t, \mathbf{o}^t, \mathbf{a}^t, \mathbf{r}^t, s^{t+1}\mathbf{o}^{t+1}\}$
10 ▷ 更新经验回放缓冲区 $\mathcal{D}$： $Enqueue(\mathcal{D}, \xi)$
11 **end**
12 **如果** $\mathcal{D}$ 已足够满，可以进行训练：**则**
13 **对于**每个无人机 m**执行**
14 ▷ 通过从 $\mathcal{D}$ 中采样小批量 $\mathcal{B}$，获取 $V_{\boldsymbol{\phi}}$
15 ▷ 更新 $\boldsymbol{\phi}$，通过**梯度下降**以优化损失函数的集中式评论家网络： $\nabla_{\boldsymbol{\phi}}\mathcal{L}$（$\boldsymbol{\phi}$）
16 ▷ 更新 $\boldsymbol{\theta}_m$，通过**梯度上升**优化目标函数，演员网络： $\nabla_{\boldsymbol{\theta}_m} J(\boldsymbol{\theta}_m)$
17 **end**
18 **end**
19 **结束**
20 **对于**回合 = *1*, *MaxEpisode***执行**
21 ▷ **初始化多无人机环境**，设置 s_0
22 **对于**时间步 = *1*, T **执行**
23 **对于**每架无人机 m**执行**
24 ▷ 根据其策略选择动作 a_m $\pi_{\boldsymbol{\theta}_m}(\mathrm{a}_m^t \mid s_{\mathrm{real}}^t, o_m^t)$ 在时间步 t
25 **end**
26 ▷ $s^t \rightarrow s^{t+1}$，$\mathbf{o}^t \rightarrow \mathbf{o}^{t+1}$，奖励为 $\mathbf{r}^t$
27 **end**
28 **结束**

表 I 无人机参数规格 [13]。

符号	数值
飞行速度，v	20 $[m/s]$
平均最长飞行时间	30 $[min]$
飞行电池容量	5,870 $[mAh]$
电压	15.2 $[V]$
包括电池和螺旋桨在内的飞行器重量，W	1375 $[g]$
旋翼半径，R	0.4 $[m]$
旋翼盘面积，$A = \pi R^2$	0.503 $[m^2]$
桨叶数量，b	4
旋翼实度，s，$\frac{0.0157b}{\pi R}$	0.05
桨叶角速度，Ω	300 $[radius/s]$
旋翼桨叶尖端速度，$U_{tip} = \Omega R^2$	120
机身阻力比，$d_0 = \frac{0.0151}{sA}$	0.6
空气密度，ρ	1.225 $[kg/m^3]$
悬停时旋翼平均诱导速度，$v_0 = \sqrt{\frac{W}{s\rho A}}$	4.03
剖面阻力系数，δ	0.012
诱导功率增量修正系数，k	0.1
最大实用升限	6000 m

C. 计算复杂度分析

本节将详细分析神经网络的计算复杂度[1]。首先，计算执行量子演员所需的操作次数。假设输入维度为 $|\mathcal{O}|$，输出维度为 $|\mathcal{A}|$，参数数量为 $|\boldsymbol{\theta}|$。那么，状态编码器中的操作次数与输入变量的数量成正比，即 $|\mathcal{O}|$。此外，参数化门操作的数量也与门的数量成正比，即 $|\boldsymbol{\theta}|$。另外，还需要进行 $|\mathcal{A}|$ 次量子测量和 Softmax 操作，以生成动作决策。最后，综合考虑所有操作，量子演员执行的计算复杂度可概括为 $\mathcal{C}_{QA} = \Omega(|\mathcal{O}| + |\boldsymbol{\theta}| + |\mathcal{A}|)$。类似地，将输入维度、输出维度和参数数量分别替换为 $|\mathcal{S}|$、1 和 $|\boldsymbol{\phi}|$，即可计算量子集中式评论家的计算复杂度。因此，量子集中式评论家的计算复杂度为 $\mathcal{C}_{QC}^E = \Omega(|\mathcal{S}| + |\boldsymbol{\phi}| + 1)$。

接下来，我们将详细说明每个训练周期的训练成本。为计算公式 (19) 中的目标差值，需要进行 $2{\cdot}T{\cdot}\mathcal{C}_{QC}^E$次运算。公式 (17) 右侧的后一项可通过 $2{\cdot}T{\cdot}M{\cdot}(|\mathcal{A}|+\mathcal{C}_{QA}^E)$ 次操作。总之，QMACN 每个训练轮次的训练成本可表示为 $\mathcal{C}_{QMACN} = T \cdot \mathcal{C}_{QC}^E + M \cdot \Omega((|\mathcal{A}| + C_{QA}^E))$。与 QMACN 相比，经典多智能体演员-评论家的演员执行计算复杂度为 $\mathcal{C}_{CA}^E = \Omega(|\mathcal{O}| \cdot |\boldsymbol{\theta}| \cdot |\mathcal{A}|)$，评论家执行计算复杂度为 $\mathcal{C}_{CC}^E = \Omega(|\mathcal{S}| \cdot |\boldsymbol{\phi}|)$，训练成本计算复杂度为 $\mathcal{C}_{MACN} = \Omega(T \cdot (M \cdot |\mathcal{O}| \cdot |\boldsymbol{\theta}| \cdot |\mathcal{A}| + |\mathcal{S}| \cdot |\boldsymbol{\phi}|))$。这些公式表明，QMACN 的计算复杂度低于

[1]下文中，我们使用算子 $|\cdot|$ 表示向量的维度。

表 II 实验设置参数

参数		取值
M	无人机数量	4
N	用户数量	25
T	回合长度	30 分钟
$\|\mathcal{S}\|$	状态维度	179
$\|\mathcal{A}\|$	动作维度	5
$\|\mathcal{D}\|$	经验回放缓冲区大小	50k
$\|\mathcal{B}\|$	小批量大小	32
γ	折扣因子	0.98
w_c	奖励系数	0.01
ϵ_{init}	初始 epsilon	0.275
-	退火 epsilon	0.00005
$\epsilon_{\min}$	最小 epsilon	0.01
-	训练轮数	10k
α_{actor}	学习率行为者	0.001
α_{critic}	学习率集中式评论家	0.00025
-	激活函数	ReLU
-	优化器	Adam

表 III 计算平台规格

系统	规格
CPU	Intel(R) Core(TM) i7-9700K CPU @3.60 GHz 随机存取存储器：64.0 GB
GPU	NVIDIA GeForce RTX 3090 核心数：10 ,496 内存：24 GB GDDR6X
平台 (PC)	内存：DDR4 64 GB 固态硬盘：500 GB（非易失性内存 express） 机械硬盘：1 TB 显卡：RTX 3090
软件库	Conda：v 22.11.1 Python：v 3.9.15 PyTorch：v 1.13.1 TorchQuantum：v 0.1.0 Numpy：v 1.23.4

经典多智能体演员-评论家的复杂度 .

VI. 性能评估

A. 系统设置

1）环境：我们的环境设置包含两个主要对象：地面用户和无人机。用户随机分布在 $6\,km \times 6\,km$ 区域/地图上；无人机则在 2500 米的高度飞行。我们假设每个回合开始时，所有无人机都位于地图中心。在整个训练阶段，我们采用 ϵ-贪心策略来控制无人机的探索与利用。此外，使用 Adam 优化器更新神经网络的权重。为切实呈现无人机的运行机制，本文采用*DJI Phantom4 Pro v2.0* 模型 [13]。所选无人机模型的能量相关参数如式（1）–（2）所示，汇总于表 I [13]。其他实验参数和计算平台的规格汇总于表 II–III。最后请注意，我们所使用的学习参数和超参数是在经过大量实验、尝试优化性能后确定的；为简洁起见，这些实验不再赘述。

2) 通信方法： 此外，我们采用 60 GHz 毫米波无线技术进行通信。考虑 60 GHz 无线网络的原因在于它具有 *i)* 大信道带宽、*ii)*低延迟传输、*iii)* 高波束方向性、*iv)* 高衍射能力，以及*v)* 高散射能力 [71]。这些特性使其受附近移动接入干扰的影响较小（空间复用），但容易受到阻挡影响 [71]。例如，在城市环境中，高层建筑的遮挡可能会影响毫米波无线系统。然而，我们提出的算法通过实现无人机的最优协同定位，最大限度地减少了这一问题。因此，借助高方向性和低延迟，可以为用户提供更好的服务质量 [71]。有关城市环境中通信信道的更多详细信息，请参见 [72]，相关内容也总结于附录A中。

3) 基准测试： 我们将把所提出方法的性能与多种基准进行比较，以验证下文所述的 QMARL。

- **‘QMACN双噪声’**使用我们的QMARL，在具有状态噪声和动作噪声的逼真环境中训练每架无人机的策略。因此，无人机的位置不仅受其动作决策影响。我们开展消融研究，探究各噪声成分对模型抗噪性能和鲁棒性的影响。

- **‘QMACN状态噪声’**和**‘QMACN动作噪声’**分别使用QMARL训练策略，其中仅包含来自GPS传感器的状态噪声和来自风的动作噪声。

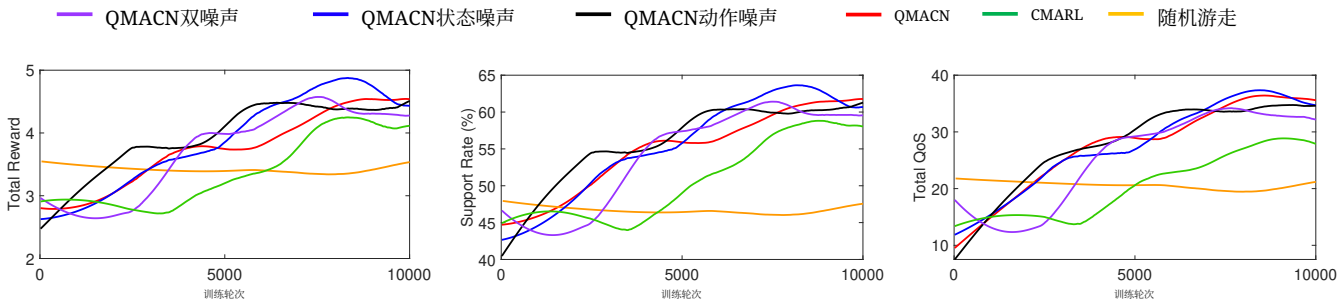
- **‘QMACN’**在无噪声环境中使用QMARL训练策略。这意味着无人机的位置仅取决于其动作决策。

- **‘CMARL’**是最新的、使用经典神经网络的多智能体强化学习方法，用于与我们提出的QMARL比较训练性能。在该训练方案中，多架无人机通过经典的集中训练分散执行方式，使用反向传播来训练各自的策略 [73]。请注意，我们将模型参数数量设置为与所提出的QMARL训练基准相同。

- **‘随机游走’**算法是指所有无人机不考虑给定的状态和观测信息，随机决定其动作。通过与之比较，我们验证了多智能体强化学习的优越性。

B. 奖励函数设计的可行性

在本小节中，我们将结合图5仔细考察奖励函数设计的可行性。从结果中可以看出，由于环境噪声效应，反映噪声影响的基准——“QMACN双噪声”、“QMACN状态噪声”和“QMACN动作噪声”——分别呈现出较大的波动，例如数据方差较高，相较于理想模型尤为明显。然而，基于QMACN的训练算法的总奖励最终收敛到更高的数值



(a) 平均总奖励。 (b) 平均支持率。 (c) 平均总服务质量。

图5. 所有训练基准中每个时间步的训练结果平均值。图5(a)展示了无人机在整个训练轮次中的总奖励，图5(b)展示了无人机在整个训练轮次中的平均总支持率，图5(c)展示了地面用户获得的无人机服务支持率和服务质量。

表 IV 基于 QMACN 的基准测试相较于基于经典神经网络的集中式多智能体强化学习的相对性能（单位：%）

❶：QMARL 算法相较于 CMARL 的训练优势			
基准	图5(a)	图5(b)	图5(c)
QMACN	21.27 %	27.44 %	6.390 %
含状态噪声	15。98 %	24.22 %	4。513 %
带动作噪声	19.36 %	23.83%	5.460 %
带双重噪声	8.150 %	15.28 %	2.566 %

表 V QMACN 对比分析

方法/规格	QC	NN	状态噪声	动作噪声
QMACN（双重噪声）	✓	✓	✓	✓
QMACN（状态噪声）	✓	✓	✓	✗
QMACN（动作噪声）	✓	✓	✗	✓
QMACN（理想情况）	✓	✓	✗	✗
CMARL	✗	✓	✗	✗
随机游走	✗	✗	✗	✗

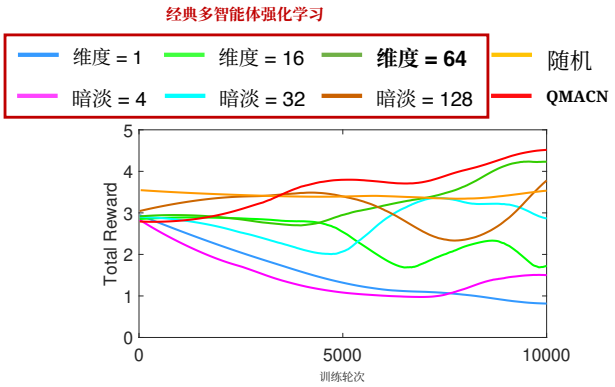


图 6. 无噪声（即 理想）环境下经典神经网络与量子计算的奖励收敛情况。维度 表示每个经典神经网络的隐藏层维度。

在训练过程结束时，该数值高于“CMARL”和“随机游走”。尽管参数较少，但如表 IV 所示，我们强调，所提出的QMACN在策略训练方面的表现优于CMARL。通过QMACN训练的无人机所获得的最终奖励值相近，依次为“QMACN”、“QMACN动作噪声”、“QMACN状态噪声”和“QMACN双噪声”。接下来，地面用户的平均支持率和服务质量也趋于相似的收敛状态，如图 5(b)–5(c)所示。这些结果表明，我们对奖励函数的设计合理，能够使多无人机实现可靠无线通信服务提供的预期目标。总之，我们确认，所设计的奖励函数不仅适用于理想条件，也适用于噪声环境。

C. 学习参数影响研究

为了通过实证评估超参数设置所导致的性能差异，我们使用不同的学习率值对所提出的算法进行评估，其中学习率是影响性能的主要超参数之一 [74],[75]。在强化学习中，学习率过高的模型训练可能会导致过冲，在使用策略梯度寻找最优权重时，阻碍模型收敛。相反，学习率过低的训练则可能导致损失函数最小化迭代次数过多。此外，还可能带来陷入局部极小值的风险。图 7 展示了学习率分别设为 0.01、0.001 和 0.0001 时，多无人机学习过程的实证评估结果。结果证实，当学习率为 0.001 时，奖励稳定收敛至最高值。这表明参数化量子电路内部旋转门的权重得到了最优训练。另一方面，可以观察到，当学习率为

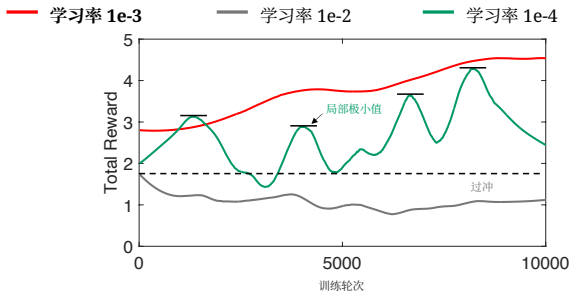
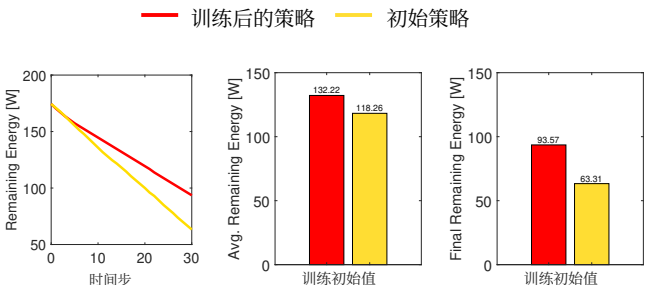


图 7. 智能体在所选学习率（0.001、0.01、0.0001）下的策略训练结果。



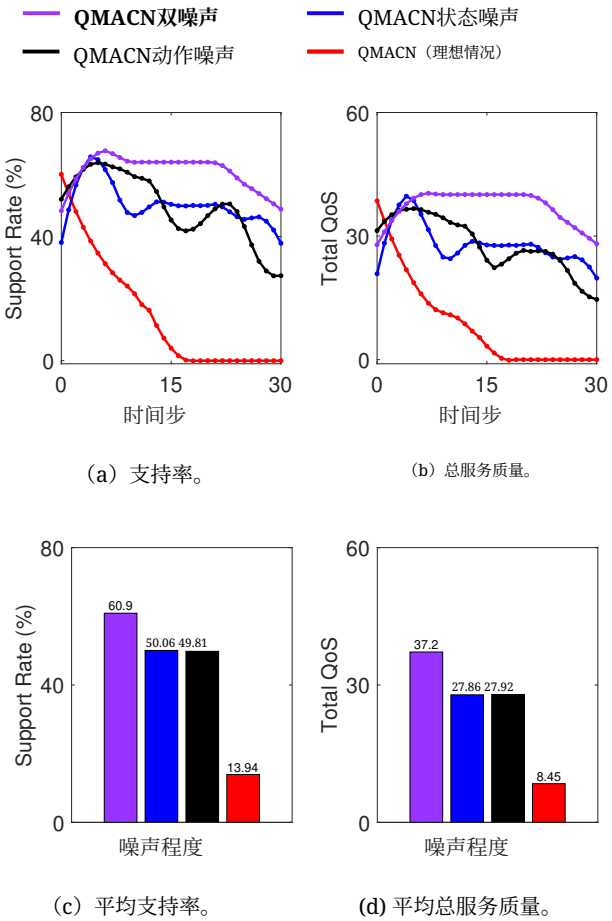
(a) 总能量。 （b）平均能量。 （c）最终能量。

图 8. 所有无人机智能体训练前后的能耗。

0.01 时会发生过冲。最后，当学习率为 0.0001 时，训练期间的方差较高。此外，还会出现多个局部极小值。

D. 与经典神经网络的比较

本小节将确定“CMARL”的隐藏层维度，并研究量子性能与现有神经网络相比的表现。图 6 展示了在理想环境中，经典神经网络（规模从 1 到 128）、理想模型和随机游走的奖励收敛趋势。在隐藏层中设置适当数量的神经元，可以在保持高准确率的同时缩短训练时间；但若不必要地增加神经元或层数，则会出现过拟合问题。这意味着，神经网络及其神经元数量越多，训练性能并不一定越好。在所有训练基准算法中，我们的 QMACN 在奖励收敛值和收敛速度方面表现最佳。除 QMACN 外，只有隐藏维度为 64 和 128 的经典神经网络在训练轮次结束时获得的奖励高于随机游走。然而，隐藏维度为 128 的经典神经网络训练表现比随机游走更不稳定，奖励也低于理想情况。因此，将 CMARL 的隐藏层设为 64 是合理的。图5(a)展示了所有训练基准在整个训练轮次中的奖励收敛趋势。与基于 QMACN 的训练基准相比，不基于 QMACN 的 CMARL 和随机游走在策略训练方面表现较差，总奖励值较低。总之，在无噪声（即“理想”）环境中，我们提出的 QMACN 训练算法能够使无人机学到比 CMARL 更接近最优的策略。



（a）支持率。 （b）总服务质量。

（c）平均支持率。 (d) 平均总服务质量。

图9. 存在状态和动作噪声时推理过程中的总体支持率和总服务质量。

E. 可靠移动蜂窝接入的能量管理

能量管理在无人机网络中至关重要，因为无人机是受能量限制的设备。因此，本节研究了基于 QMACN 的策略训练后无人机能耗的变化。图 8 展示了给定时间步进程中所有无人机的电池状态。可以观察到，采用 QMACN 训练策略的无人机比采用未训练策略的无人机消耗的能量更少。此外，如图所示，在整个时间步内，采用训练后策略的无人机平均剩余能量多出 11。8%。

图8(b)和图8(c)显示，在回合结束时，其能量比策略未经训练的无人机多47.8%。总而言之，第V-A节中设计的奖励函数成功训练了无人机的策略，使其朝着最大化服务质量和增强网络可靠性的方向发展。然而，无人机因能量不足而发生故障，会导致移动蜂窝接入中的服务效率低下。为解决这一问题， [40]中的先前研究表明，基于多智能体强化学习的无人机能够自适应地应对无人机电池完全放电的情况，方法是移动到故障无人机所覆盖的区域。

表 VI 噪声环境中注入双重噪声、状态噪声、动作噪声的 QMACN 相对于理想情况 QMACN 的性能（单位：%）

2：基于QMACN的无人机网络的相对鲁棒性		
算法	图9(c)	图9(d)
双重噪声	46.96 %	28.75 %
带状态噪声	36。12 %	19。41 %
带动作噪声	35。87 %	19。47 %

F. 噪声对真实环境中鲁棒性的影响

学习过程结束后，所有无人机在推理过程中依次依据学到的策略做出决策，以验证该策略的可行性。我们将通过100次迭代的推理，比较QMA CN在不同环境中的训练结果（即，消融研究）。如图9所示，在整个推理过程中，“双重噪声”在支持率（平均60.9%）和服务质量（平均37.2）方面的数值最高。“状态噪声”和“动作噪声”构成了次优的移动接入方案，其支持率和服务质量分别比“双重噪声”低18%和25%。在所有基于QMACN的训练算法中，“QMACN”受环境噪声影响，提供的无线通信服务最差。如表VI所示，注入噪声的QMACN，包括“QMACN双噪声”、“QMACN状态噪声”和“QMACN动作噪声”，在噪声环境中提供的无线通信服务优于未考虑环境噪声、仅训练无人机策略的“QMACN”。总之，纳入越来越贴近现实的噪声因素，能够使智能体学习到更稳健的策略，从而促进真实场景中有效移动接入系统的发展。

G. 讨论

本节将结合第V节中的实验结果，讨论我们提出的基于QMACN的训练算法的性能。

*1) MARL*环境中的量子优势：图5–6表明，与“CMARL”[55]相比，基于PQC的QMACN在参数更少的情况下，具有相当的策略训练性能。这一实证优势体现在记忆能力方面，这是PQC的代表性特征 [76]。在学习新标签时保留关于现有标签的知识，有助于以近乎最优的方式训练策略 [55]。基于量子的算法在这一特性上具有足够的灵活性，但采用深度神经网络策略的传统神经网络则不具备这一能力 [76]。正如我们的数值性能评估所观察到的，这一差异会影响训练速度或性能。此外，利用量子特征（即，叠加、纠缠）的量子门，相较于传统神经网络，还具有指数级计算增益的优势 [77]。

2) 噪声注入对逼真环境设计的影响：本文通过噪声反射开展了消融研究。一般而言，环境噪声会影响强化学习中智能体的行为。然而，某些高斯噪声能够改善策略的鲁棒性 [78]，并通过选择性噪声注入（SNI） [79]减轻其对梯度质量的不利影响。这种方法不仅能够使智能体更有效地实现预期目标，还能增强其抵御自然产生噪声的能力。我们在式（3）–（4）中的噪声模型遵循高斯分布，使无人机能够在环境中获得丰富的经验，并防止因图5(a)所示的受限训练环境而发生过拟合。即使考虑到一些波动，与“QMACN”相比，标记为“QMACN双噪声”“QMACN状态噪声”和“QMACN动作噪声”的基准测试也具有更快的收敛速度和更高的奖励值。此外，我们在图9中研究了含噪声的逼真环境中的模型鲁棒性。

VII. 总结

本文提出了一种基于集中训练分散执行的量子多智能体演员-评论家网络（QMACN）训练算法，用于构建基于多架无人机的鲁棒移动接入。为实现量子计算的实际应用，我们采用集中训练分散执行，以克服含噪中等规模量子时代量子误差带来的可扩展性问题，从而实现量子优越性。所提出的 QMACN 算法通过在训练速度、无线服务质量和网络可靠性方面的性能提升，验证了量子多智能体强化学习的优势，并在多种数据密集型评估中展现出这一优势。此外，我们验证了噪声注入方案有助于多架无人机应对环境不确定性，从而提高移动接入的鲁棒性。总而言之，与传统训练方法相比，所提出的 QMACN 算法以更少的模型参数，实现了更优的协作式多无人机移动接入性能，从而促进了计算资源的高效管理。

参考文献

[1] W. J. Yun、Y. Kwak、J. P. Kim、H. Cho、S. Jung、J. Park 和 J. Kim，“通过变分量子电路设计实现量子多智能体强化学习”，载于*IEEE ICDCS*会刊，意大利博洛尼亚，2022年7月。[2] D. S. Lakew、U. Sa’ad、N.-N. Dao、W. Na 和 S. Cho，“飞行自组网络中的路由：综合性综述”，*IEEE Communications Surveys &Tutorials*，第22卷，第2期，第1071–1120页，2020年第二季度。[3] D. S. Lakew、W. Na、N.-N. Dao 和 S. Cho，“异构无人机辅助无线通信的空中能量编排”，*IEEE Systems Journal*，第16卷，第2期，第2483–2494页，2022年6月。[4] T. K. Rodrigues、J. Liu 和 N. Kato，“多层6G网络中移动多接入边缘计算的卸载决策”，*IEEE Transactions on Emerging Topics in Computing*，第10卷，第3期，第1414–1427页，2022年7月/9月。[5] G. Gui、M. Liu、F. Tang、N. Kato 和 F. Adachi，“6G：为舒适、安全与智能的融合开启新视野”，*IEEE Wireless Communications*，第27卷，第5期，第126–132页，2020年10月。[6] W. Saad、M. Bennis 和 M. Chen，“6G无线系统愿景：应用、趋势、技术与开放研究问题”，*IEEE*网络，第34卷，第3期，第134–142页，2019年5月/6月。[7] M. Mozaffari、X. Lin 和 S. Hayes，“迈向拥有互联天空的6G：无人机及更广领域”，*IEEE Communications Magazine*，第59卷，第12期，第74–80页，2021年12月。[8] W. J. Yun、S. Park、J. Kim、M. Shin、S. Jung、D. A. Mohaisen 和 J.-H. Kim，“通过自主多无人机控制实现可靠监控的协作式多智能体深度强化学习”，*IEEETransactions on Industrial Informatics*，第18卷，第10期，第7086–7096页，2022年10月。[9] D. N. Das、R. Sewani、J. Wang 和 M. K. Tiwari，“包裹配送物流中的卡车与无人机协同路径规划”，*IEEE Transactions on Intelligent Transportation Systems*，第22卷，第9期，第5772–5782页，2021年9月。

[10] W. J. Yun、J. P. Kim、S. Jung、J.-H. Kim 和 J. Kim，“面向智能工厂管理中互联网连接的多机器人协同的量子多智能体演员-评论家神经网络”，IEEE物联网期刊，第10卷，第11期，第9942–9952页，2023年6月。[11] G. Lee、W. Saad 和 M. Bennis，“工业4.0智能工厂中无人机辅助分布式雾计算的在线优化”，载于IEEE全球通信大会，阿布扎比，阿拉伯联合酋长国，2018年12月。[12] Y. Kawamoto、T. Mitsuhashi 和 N. Kato，“灾害场景中用于车对车（V2V）通信的无人机辅助信息扩散”，IEEE Transactions on Emerging Topics in Computing，第10卷，第4期，第1909–1917页，2022年10月/12月。[13] S. Jung、W. J. Yun、M. Shin、J. Kim 和 J.-H. Kim，“面向云辅助多无人机充电系统的编排调度与多智能体深度强化学习”，IEEE车辆技术汇刊，第70卷，第6期，第5362–5377页，2021年6月。[14] M. Mozaffari、W. Saad、M. Bennis、Y.-H. Nam 和 M. Debbah，“无线网络中的无人机教程：应用、挑战与开放问题”，IEEE Communications Surveys & Tutorials，第21卷，第3期，第2334–2360页，2019年第三季度。[15] M. Zhang、S. Fu 和 Q. Fan，“无人机基站的联合三维部署与功率分配：一种深度强化学习方法”，IEEE无线通信快报，第10卷，第10期，第2309–2312页，2021年10月。[16] I. Bor-Yaliniz 和 H. Yanikomeroglu，“无线接入网异构性的新前沿：多层无人机蜂窝”，IEEE Communications Magazine，第54卷，第11期，第48–55页，2016年11月。[17] Y. Zeng、R. Zhang 和 T. J. Lim，“采用无人机的无线通信：机遇与挑战”，IEEE Communications Magazine，第54卷，第5期，第36–42页，2016年5月。[18] J. Park、S. Samarakoon、A. Elgabli、J. Kim、M. Bennis、S.-L. Kim 和 M. Debbah，“无线网络上的高通信效率分布式学习：原理与应用”，IEEE会刊，第109卷，第5期，第796–819页，2021年5月。[19] M.-A. Lahmeri、M. A. Kishk 和 M.-S. Alouini，“无人机赋能无线网络中的人工智能：综述”，IEEE通信学会开放期刊，第2卷，第1015–1040页，2021年4月。[20] S. Oh、J. Choi 和 J. Kim，“量子卷积神经网络（QCNN）教程”，载于IEEE ICTC会刊，济州岛，韩国，2020年10月。[21] S. Y.-C. Chen、C.-H. H. Yang、J. Qi、P.-Y. Chen、X. Ma 和 H.-S. Goan，“用于深度强化学习的变分量子电路”，IEEE Access，第8卷，第141 007–141 024页，2020年7月。[22] T. Ohyama、Y. Kawamoto 和 N. Kato，“利用量子计算通过组合优化实现智能反射面（IRS）分配调度方法”，IEEE Transactions on Emerging Topics in Computing，第10卷，第3期，第1633–1644页，2022年7月/9月。[23] S. Sukhbaatar、R. Fergus 等，“通过反向传播学习多智能体通信”，载于神经信息处理系统大会会刊，第29卷，西班牙巴塞罗那，2016年12月，第2244–2252页。[24] P. W. Shor，“减少量子计算机存储器退相干的方案”，物理评论A辑，第52卷，第4期，第R2493页，1995年。[25] A. Horsley，“量子计算硬件”，2022年1月。 [在线]。可访问：https://quantumbrilliance.com/quantum-brilliance-hardware [26] Y.-C. Liu、Y.-C. Dzeng 和 C.-C. Ting，“氮空位中心金刚石量子比特：制造、设计及其在量子计算中的应用”，IEEE纳米技术杂志，第16卷，第4期，第37–43页，2022年8月。[27] J. Xiang，“Gemini-mini·2量子比特便携式核磁共振量子计算机”，2022年1月。 [在线]。可访问：https://www.spinquanta.com/products-solutions/geminiMini [28] R. Blatt、T. Monz 和 P. Zoller，“全球领先的19英寸机架式量子计算机”，2022年5月。 [在线]。可访问：https://www.aqt.eu/pine-system-19-rack-mounted-quantum-computer/ [29] M. Mozaffari、W. Saad、M. Bennis 和 M. Debbah，“用于高能效无人机部署的最优传输理论”，载于IEEE国际通信大会会刊，吉隆坡，马来西亚，2016年5月，第1–6页。[30] Z. Kaleem、W. Khalid、A. Muqaibel、A. A. Nasir、C. Yuen 和 G. K. Karagiannidis，“变化高度下用于总容量提升的学习辅助无人机三维部署与功率分配”，IEEE通信快报，第26卷，第7期，第1633–1637页，2022年7月。[31] M. A. Ali 和 A. Jamalipour，“蜂窝网络上下行操作中的无人机部署与功率分配”，IEEE《通信汇刊》，第68卷，第7期，第4383–4393页，2020年7月。[32] I. Valiulahi 和 C. Masouros，“存在同信道干扰时用于吞吐量最大化的多无人机部署”，IEEE物联网期刊，第8卷，第5期，第3605–3618页，2021年3月。[33] E. Kalantari、H. Yanikomeroglu 和 A. Yongacoglu，“无线蜂窝网络中无人机基站的数量与三维部署”，载于IEEE车辆技术大会会刊，蒙特利尔，加拿大魁北克省，2016年9月，第1–6页。

[34] F. Cheng、S. Zhang、Z. Li、Y. Chen、N. Zhao、F. R. Yu 和 V. C. Leung，《多小区边缘数据卸载的无人机轨迹优化》，IEEE车辆技术汇刊，第67卷，第7期，第6732–6736页，2018年7月。[35] Z. Becvar、M. Nikooroo 和 P. Mach，《为移动用户提供服务的飞艇式飞行基站的能耗》，IEEE Transactions on Communications，第70卷，第10期，第7006–7022页，2022年10月。[36] N. Babu、C. B. Papadias 和 P. Popovski，《无人机通信系统中空中接入点的能效型三维部署》，IEEE《通信快报》，第24卷，第12期，第2883–2887页，2020年12月。[37] H.-Y. Liu、X.-H. Tian、C. Gu、P. Fan、X. Ni、R. Yang、J.-N. Zhang、M. Hu、J. Guo、X. Cao 等，《面向移动量子网络的基于无人机的纠缠分发》，《国家科学评论》，第7卷，第5期，第921–928页，2020年5月。[38] H.-Y. Liu、X.-H. Tian、C. Gu 和 P. Fan，《使用无人机作为移动节点的光中继纠缠分发》，《物理评论快报》，第126卷，第2期，文章编号020503，2021年1月。[39] F. Chiti、R. Picchi 和 L. Pierucci，《都市量子无人机网络与计算：软件定义视角》，IEEE Access，第10卷，第126 062–126 073页，2022年12月。[40] W. J. Yun、S. Park、J. Kim、M. Shin、S. Jung、D. A. Mohaisen 和 J.-H. Kim，《通过自主多无人机控制实现可靠监视的协作式多智能体深度强化学习》，IEEE Transactions on Industrial Informatics，第18卷，第10期，第7086–7096页，2022年10月。[41] Z. Xia、J. Du、J. Wang、C. Jiang、Y. Ren、G. Li 和 Z. Han，《多智能体强化学习辅助的智能无人机集群目标跟踪》，IEEE车辆技术汇刊，第71卷，第1期，第931–945页，2021年1月。[42] D. Baldazo、J. Parras 和 S. Zazo，《用于洪水监测的无人机集群分散式多智能体深度强化学习》，载于欧洲信号处理会议论文集，西班牙 A Coru˜na，2019年9月，第1–5页。[43] T. Cai、Z. Yang、Y. Chen、W. Chen、Z. Zheng、Y. Yu 和 H.-N. Dai，《基于多智能体深度强化学习的无人机辅助群智感知协作式数据感知与计算卸载》，IEEE《网络科学与工程汇刊》，第9卷，第5期，第3197–3211页，2021年9月/10月。[44] O. S. Oubbati、M. Atiquzzaman、A. Lakas、A. Baz、H. Alhakami 和 W. Alhakami，《多无人机赋能的信息新鲜度感知无线供能通信网络：一种多智能体强化学习策略》，载于 IEEE信息通信会议研讨会，加拿大不列颠哥伦比亚省温哥华，2021年5月，第1–6页。[45] X. Tang、X. Lan、L. Li、Y. Zhang 和 Z. Han，《通过多智能体强化学习激励权益证明区块链，以实现无人机辅助物联网中的安全数据收集》，IEEE《通信领域精选期刊》，第40卷，第12期，第3470–3484页，2022年12月。[46] V. Sadhu、C. Sun、A. Karimian、R. Tron 和 D. Pompili，《Aerial-DeepSearch：用于搜索任务的分布式多智能体深度强化学习》，载于 IEEE移动自组织系统与传感器网络会议，印度德里，2020年12月，第165–173页。[47] N. G´omez、N. Pe˜na、S. Rinc´on、S. Amaya 和 J. Calderon，《基于粒子群优化方法的多智能体系统搜救领航者–跟随者行为》，载于东南地区会议论文集，美国阿拉巴马州莫比尔，2022年5月，第413–420页。[48] X. Tan、L. Zhou、H. Wang、Y. Sun、H. Zhao、B.-C. Seet、J. Wei 和 V. C. Leung，《认知无线电网络中基于协作式多智能体强化学习的分布式动态频谱接入》，IEEE物联网期刊，第9卷，第19期，第19 477–19 488页，2022年10月。[49] H. Albinsaid、K. Singh、S. Biswas 和 C.-P. Li，《基于多智能体强化学习的分布式动态频谱接入》，IEEE《认知通信与网络汇刊》，第8卷，第2期，第1174–1185页，2021年6月。[50] Z. Wang、H. Yao、T. Mai、Z. Xiong、X. Wu、D. Wu 和 S. Guo，《学习无人机集群网络中的路由：一种多智能体强化学习方法》，IEEE车辆技术汇刊，第1–14页，2022年12月。[51] D. Bouwmeester 和 A. Zeilinger，《量子信息物理学：基本概念》。施普林格，2000年，第1–14页。[52] C. P. Williams、S. H. Clearwater 等，《量子计算探索》。施普林格，1998年。[53] A. P´erez-Salinas、A. Cervera-Lierta、E. Gil-Fuster 和 J. I. Latorre，《用于通用量子分类器的数据重上传》，Quantum，第4卷，文章编号226，2020年2月。

[54] O. Lockwood 和 M. Si，《使用量子变分电路进行强化学习》，载于 AAAI论文集，第16卷，第1期，线上，2020年10月，第245–251页。[55] W. J. Yun、J. Park 和 J. Kim，《量子多智能体元强化学习》，载于 AAAI 论文集，美国华盛顿特区，2023 年 2 月。[56] A. P. Cohen 、 S. A. Shaheen 和 E. M. Farrar，《城市空中交通：历史、生态系统、市场潜力与挑战》，IEEE Transactions on Intelligent Transportation Systems，第22卷，第9期，第6074–6087页，2021年9月。[57] H. Mei、K. Yang、Q. Liu 和 K. Wang，《无人机赋能的边缘云系统中虚拟化移动克隆的联合轨迹与资源优化》，IEEE物联网期刊，第7卷，第7期，第5906–5921页，2019年7月。[58] Y. Zeng 和 R. Zhang，《通过轨迹优化实现能效型无人机通信》，IEEE无线通信汇刊，第16卷，第6期，第3747–3760页，2017年6月。[59] J. Paek、J. Kim 和 R. Govindan，《面向智能手机的能效型自适应速率GPS定位》，载于 ACM MobiSys论文集，美国加利福尼亚州旧金山，2010年6月，第299–314页。[60] L. Liu 和 M. G. Amin，《采用预相关滤波器和采样率的非高斯噪声环境下GPS接收机性能分析》，IEEE信号处理汇刊，第56卷，第3期，第990–1004页，2008年3月。[61] S. A. Kassam，《非高斯噪声中的信号检测》。施普林格科学与商业媒体，2012年。[62] J. Feng 和 W. Z. Shen，《利用风速和风向的联合分布对风电场布局优化中的风况进行建模》，Energies，第8卷，第4期，第3075–3092页，2015年4月。[63] T. Rashid、M. Samvelyan、C. S. De Witt、G. Farquhar、J. Foerster 和 S. Whiteson，《用于深度多智能体强化学习的单调价值函数分解》，机器学习研究杂志，第21卷，第1期，第7234–7284页，2020年1月。[64] S. Jung、J. Kim、M. Levorato、C. Cordeiro 和 J.-H. Kim，《基础设施辅助的联网车辆行驶中体验共享》，IEEE车辆技术汇刊，第70卷，第8期，第7307–7321页，2021年8月。[65] C.-C. Lai、C.-T. Chen 和 L.-C. Wang，《面向任意分布用户并保证数据速率的按需密度感知无人机基站三维部署》，IEEE无线通信快报，第8卷，第3期，第913–916页，2019年6月。[66] V. Konda 和 J. Tsitsiklis，《行动者-评论家算法》，载于神经信息处理系统大会论文集，第12卷，美国科罗拉多州丹佛，1999年11月/12月。[67] R. Lowe、Y. Wu、A. Tamar、J. Harb、O. Pieter Abbeel 和 I. Mordatch，《混合协作–竞争环境中的多智能体行动者-评论家》，载于神经信息处理系统大会论文集，第30卷，2017年。[68] F. A. Oliehoek 和 C. Amato，《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。[69] G. E. Crooks，《利用参数移位规则和量子门分解计算参数化量子门的梯度》，CoRR，第abs/1905.13311卷，2019年5月。[70] V. Mnih、K. Kavukcuoglu、D. Silver、A. Graves、I. Antonoglou、D. Wier-stra 和 M. Riedmiller，《利用深度强化学习玩雅达利游戏》，arXiv预印本 arXiv:1312.5602，2013年。[71] S. Singh、R. Mudumbai 和 U. Madhow，《高度定向 60 GHz网状网络的干扰分析：重新思考介质访问控制的理由》，IEEE/ACM网络汇刊，第5期，第1513–1527页，2011年10月。[72] C. Park、H. Lee、W. J. Yun、S. Jung、C. Cordeiro 和 J. Kim，《通过多无人机控制实现可靠且能效型移动接入的协作式多智能体深度强化学习》，arXiv预印本 arXiv:2210.00945，2022年。[73] R. Hecht-Nielsen，《反向传播神经网络理论》，载于《用于感知的神经网络》。爱思唯尔，1992年，第65–93页。[74] Z. Lv、L. Qiao、J. Li 和 H. Song，《物联网中的深度学习赋能安全问题》，IEEE物联网期刊，第8卷，第12期，第9531–9538页，2021年6月。[75] S. Ray，《机器学习算法简述》，载于COMITCon论文集，印度法里达巴德，2019年2月，第35–39页。[76] S. Jerbi、C. Gyurik、S. Marshall、H. Briegel 和 V. Dunjko，《用于强化学习的参数化量子策略》，载于神经信息处理系统大会论文集，线上，2021年12月。[77] P. W. Shor，《量子计算机上用于素数分解和离散对数的多项式时间算法》，SIAM评论，第41卷，第2期，第303–332页，1999年。[78] Z. He、A. S. Rakin 和 D. Fan，《参数化噪声注入：可训练随机性，以提升深度神经网络对抗对抗攻击”，载于*IEEE*计算机视觉与模式识别会议论文集，美国加利福尼亚州长滩，2019年6月，第588–597页。[79] M. Igl、K. Ciosek、Y. Li、S. Tschiatschek、C. Zhang、S. Devlin和K. Hofmann，“通过选择性噪声注入和信息瓶颈实现强化学习中的泛化”，载于神经信息处理系统大会论文集，第32卷，加拿大不列颠哥伦比亚省温哥华，2019年12月。[80] A. Maltsev、E. Perahia、R. Maslennikov、A. Lomayev、A. Khoryaev和A. Sevastyanov，“TGad信道模型的路径损耗模型开发”，*IEEE 802.11–09/0553r1*，2009年5月。[81] I. R. Sector，“用于400 MHz至约70 GHz频段共用研究的固定和移动业务全向、扇区及其他天线的参考辐射方向图”，*ITU-R F*建议书，第1336–5页，2019年。[82] J. Kim和W. Lee，“面向超互联雾计算应用的60 GHz毫米波技术可行性研究”，*IEEE*物联网期刊，第4卷，第5期，第1165–1173页，2017年10月。[83] J. Kim、S.-C. Kwon和G. Choi，“采用60 GHz辐射和IEEE 802.11ad基带的基础设施到车辆远程信息处理平台中的视频流性能”，*IEEE*车辆技术汇刊，第65卷，第12期，第10 111–10 115页，2016年12月。

附录A
60 GHz毫米波模型

干扰分析。在分析无线网络性能时，必须考虑干扰的影响。为对 60 GHz 无线网络进行干扰分析，需要考虑其路径损耗和天线方向图。首先，IEEE 802.11ad [80]中定义的 60 GHz 路径损耗模型如下：

$$L(d) = 32.5 + 20\log_{10}(f_{\mathrm{GHz}}) + 10 \cdot n \cdot \log_{10}(d), \quad (25)$$

其中， f_{GHz}、 n和 d 分别表示以 GHz 为单位的载波频率（即， $f_{\mathrm{GHz}} = 60$）、路径损耗系数（$n = 2$），以及以米为单位的发射机（Tx）与接收机（Rx）之间的距离。接下来，国际电信联盟（ITU）标准文档 [81] 中定义的高斯 60 GHz 天线方向图如下：

$$G_{\mathrm{dBi}}(\varphi,\theta) = \begin{cases} G_{\mathrm{dBi}}^{\max} - 12\Delta^2, & 0 \le \Delta < 1 \\ G_{\mathrm{dBi}}^{\max} - 12 - 15\ln\Delta, & 1 \le \Delta \end{cases} \quad (26)$$

其中， $G_{\mathrm{dBi}}^{\max}$ 为最大天线增益， φ 和 θ 为方位角和仰角，其中 $-180^\circ \le \varphi \le 180^\circ$和 $-90^\circ \le \theta \le 90^\circ$。此外， Δ 可通过$\Delta \triangleq |\Sigma^{\ddagger} \cdot \Sigma^{\dagger}|$ 求得，其中 $\Sigma^{\ddagger} \triangleq \cos^{-1}(\cos\varphi\cos\theta)$, 且 $\Sigma^{\dagger} \triangleq \sqrt{\left[\cos\left\{\tan^{-1}\left(\frac{\tan\theta}{\varphi}\right)\right\}/\varphi_3\right]^2 + \left[\sin\left\{\tan^{-1}\left(\frac{\tan\theta}{\varphi}\right)\right\}/\theta_3\right]^2}\sin$其中， φ_3和 θ_3分别为方位面和仰角面上的 3 dB半功率波束宽度；并假设 $\varphi_3 = \theta_3 = 10^\circ$ [82]。然后，无线链路 i 上的传输所产生的干扰（其中 $i \neq l$， $i \in \mathcal{I}$， $\mathcal{I}$ 为无线链路集合）对无线链路 l 造成的干扰可表示如下：

$$I_{\mathrm{mW}}^{i,l} = f_{\mathrm{dBm\text{-}mW}}\left(\mathcal{G}_{\mathrm{dBi}}^{i,\mathrm{Tx}}(\varphi_1,\varphi_2) + \mathcal{P}_{\mathrm{dBm}}^{i,\mathrm{Tx}} - L\left(d_{(i,l)}\right)\right), \quad (27)$$

其中， $\mathcal{P}^{i,\,\mathrm{Tx\ dBm}}$ 是链路 i 上传输的发射功率； $L\left(d_{(i,l)}\right)$ 是路径损耗，如公式 (25) 所示，其中 $d_{(i,l)}$ 是第 i 个发射端与第 l 个接收端之间的距离； $\mathcal{G}^{i,\mathrm{Tx\ dBi}}$ 是第 i 个发射端与第 l 个接收端之间的天线增益，取决于 φ_1 和 φ_2，它们是这两个发射端与接收端之间的角度差，如公式 (26) 所示； $f_{\mathrm{dBm\text{-}mW}}$ 是将分贝毫瓦转换为毫瓦 (mW) 的函数，即 例如 $f_{\mathrm{dBm\text{-}mW}} \triangleq 10^{(x/10)}$。

表 VII 60 GHz IEEE 802.11ad 调制与编码方案表 [82]。

Rx 灵敏度	可用 MCS	最大数据速率	容量估计与香农
-78 分贝毫瓦	MCS0	27.5 兆比特每秒	1.43 吉比特每秒
-68 分贝毫瓦	MCS1	385 兆比特每秒	2.04 吉比特每秒
-66 分贝毫瓦	MCS2	770 兆比特每秒	2.40 吉比特每秒
-65 分贝毫瓦	MCS3	962.5 兆比特每秒	2.81 吉比特每秒
-64 分贝毫瓦	MCS4	1155 兆比特每秒	3.25 吉比特每秒
-63 分贝毫瓦	MCS6	1540 兆比特每秒	3.74 吉比特每秒
-62 分贝毫瓦	MCS7	1925 兆比特每秒	4.25 吉比特每秒
-61 分贝毫瓦	MCS8	2310 兆比特每秒	5.38 吉比特每秒
-59 分贝毫瓦	MCS9	2502.5 兆比特每秒	7.90 吉比特每秒
-55 分贝毫瓦	MCS10	3080 兆比特每秒	8.57 吉比特每秒
-54 分贝毫瓦	MCS11	3850 兆比特每秒	9.23 吉比特每秒
-53 分贝毫瓦	MCS12	4620 兆比特每秒	43.48 吉比特每秒

使用香农公式计算容量。本文采用香农容量来计算 60 GHz 无线网络中的理论数据速率。香农容量公式给出了公式 (28) 中链路 l 上发射端与接收端之间、距离为 d 时的信道容量 $\mathcal{C}_l(d)$，如下所示：

$$\mathcal{C}_l\left(d\right) = \mathbf{BW} \cdot \log_2\left(\frac{\mathcal{P}_{\mathrm{mW}}^{l,\mathrm{Rx}}\left(d\right)}{n_{\mathrm{mW}} + \sum_{i\in\mathcal{I},\, i\neq l} I_{\mathrm{mW}}^{i,l}} + 1\right), \quad (28)$$

其中，BW 表示 60 GHz 信道带宽（IEEE 802.11ad [82]中为 2.16 GHz）；$\mathcal{P}^{l}$, ^{Rx} mW（d）表示第 l 个接收端链路上以 mW 为单位的接收信号强度；I^{il}, mW 表示式 (27) 中的干扰；n_{mW} 表示以 mW 为单位的背景噪声。此处，$\mathcal{P}^{l}$, ^{Rx} dBi（d）可计算为 $\mathcal{P}^{l,\mathrm{Rx}}$ dBi（d）$= \mathcal{G}^{l\ \mathrm{Tx}} + \mathcal{P}^{l\ \mathrm{Tx}} - L\ d\ + \mathcal{G}^{l\ \mathrm{Rx}}$，$\mathcal{G}^{l\ \mathrm{Tx}}$，$\mathcal{P}^{l\ \mathrm{Tx}}$，dBi，dBm（），dBi，其中，dBi 和 dBm 均为等效全向辐射功率

链路 l 的发射端的发射天线增益和发射功率；L（d）表示式 (25) 中的路径损耗；$\mathcal{G}^{l,\ \mathrm{Rx}}$ dBi 表示接收天线增益（ [82] 中为 3 dBi）。在 [82]，中，60 GHz 信道上的等效全向辐射功率（EIRP）最大值设为 43 dBm，其中 $\mathcal{G}^{l,\ \mathrm{Tx\,dBi}} = 19\,\mathrm{dBi}$（以 dBi 为单位的发射天线增益），$\mathcal{P}^{l,\ Tx\,\mathrm{dBm}} = 24\,\mathrm{dBm}$（链路 l 上以 dBm 为单位的发射功率）；n_{mW} 表示背景噪声，且$n_{\mathrm{mW}} = f_{\text{dBm-mW}}\left(k_{\mathrm{B}}T_e + 10\log_{10}\mathbf{BW} + \sigma\right)$，其中 $k_{\mathrm{B}}T_e$ 表示噪声功率谱密度（ [82] 中为 $-174\mathrm{dBm/Hz}$），σ 表示额外损耗，即实现损耗（ [82] 中为 10 dB）与噪声系数（ [82] 中为 5 dB）之和。

采用 IEEE 802.11ad 调制与编码方案（MCS）计算数据速率。由于收发器和射频实现限制，实际数据速率低于根据香农公式得出的速率。因此，实际数据速率是通过采用 IEEE 802.11ad [83] 中的调制与编码方案（ MCS ）计算得出的。首先，计算考虑干扰时的接收端信号强度。然后，通过匹配表 VII 中的 MCS 索引，得出链路 l 实际能够支持的最大数据速率。

Chanyoung Park 自2022年9月起在韩国首尔的高丽大学攻读电气与计算机工程博士学位。他于2022年在韩国水原的亚洲大学获得电气与计算机工程理学学士学位，并以荣誉成绩提前毕业。

Won Joon Yun 自2021年3月起在大韩民国首尔的高丽大学攻读电气与计算机工程博士学位，并于该校获得电气工程理学学士学位。

Jae Pyoung Kim 自2023年3月起在大韩民国首尔的高丽大学攻读电气与计算机工程硕士学位，并于该校获得电气工程理学学士学位。

Tiago Koketsu Rodrigues 现任日本东北大学助理教授。他的研究兴趣包括人工智能、机器学习、网络建模与仿真以及云系统。2017年至2020年，他担任《IEEE车辆技术汇刊》首席系统管理员，负责监督所有投稿的审稿流程以及整个投稿系统。他现任《IEEE车辆技术汇刊》和《IEEE网络》编辑。

Soohyun Park 目前正在大韩民国首尔的高丽大学攻读电气与计算机工程博士学位。她于2019年获得大韩民国首尔中央大学计算机科学与工程理学学士学位。她的研究重点包括深度学习算法及其应用。

她于2019年荣获IEEE车辆技术学会（VTS）首尔分会奖。

Soyi Jung 自2022年9月起担任大韩民国水原亚洲大学电气与计算机工程系助理教授。在加入亚洲大学之前，她于2021年至2022年担任大韩民国春川翰林大学助理教授；于2021年至2022年担任美国加利福尼亚州尔湾加利福尼亚大学欧文分校唐纳德·布伦信息与计算机科学学院访问学者；于2021年担任大韩民国首尔高丽大学研究教授；并于2015年至2016年在大韩民国果川韩国测试研究院（KTR）担任研究员。她于2013年、2015年和2021年分别获得大韩民国水原亚洲大学电气与计算机工程学士、硕士和博士学位。

Joongheon Kim （M'06–SM'18）自2019年起任职于韩国首尔的高丽大学，目前担任电气工程学院副教授。他于2004年和2006年分别获得韩国首尔高丽大学计算机科学与工程专业的学士和硕士学位，并于2014年获得美国加利福尼亚州洛杉矶南加州大学（USC）计算机科学博士学位。在加入高丽大学之前，他曾担任LG电子（韩国首尔，2006年–2009年）的研究工程师、英特尔公司（美国加利福尼亚州圣克拉拉，2013年–2016年）的系统工程师，以及中央大学（韩国首尔，2016年–2019年）计算机科学与工程专业的助理教授。

他担任IEEE《车辆技术汇刊》、《机器学习、通信与网络汇刊》和IEEE《通信标准杂志》的编辑。他还是*IEEE*通信学会（*ComSoc*）和*IEEE*系统理事会的杰出讲师。他是韩国通信与信息科学研究院（KICS）的执行董事。