面向第六代移动通信技术的整体网络虚拟化与泛在网络智能

Xuemin (Sherman) Shen，*IEEE* 会士，Jie Gao，*IEEE* 高级会员，Wen Wu，*IEEE* 会员，Mushu Li，*IEEE* 会员，Conghao Zhou，*IEEE* 学生会员，以及 Weihua Zhuang，*IEEE* 会士

（特邀论文）

arXiv:2301.00519v1 [cs.NI] 2 Jan 2023

摘要—**本文是一篇教程论文，探讨网络架构的演进与发展前景，并提出一种面向第六代（6G）网络的新型概念架构。所提出的架构包含两个关键要素，即整体网络虚拟化与泛在人工智能（AI）。整体网络虚拟化包括网络切片和数字孪生，分别从服务提供和服务需求的角度出发，以实现以服务为中心和以用户为中心的网络。泛在网络智能分别从面向人工智能的网络和面向网络的人工智能两个角度，将人工智能融入未来网络。基于整体网络虚拟化与泛在网络智能，所提出的架构能够促进三种相互作用：数字孪生与网络切片范式之间的相互作用、模型驱动方法与数据驱动方法在网络管理中的相互作用，以及虚拟化与人工智能之间的相互作用，从而最大限度地提升6G网络的灵活性、可扩展性、适应性和智能化水平。我们还指出了与所提出架构相关的挑战与开放问题。通过阐述我们的愿景，我们旨在激发对6G潜在架构的进一步讨论与发展。**

索引词—**第六代移动通信技术、网络架构、网络虚拟化、数字孪生、面向网络的人工智能、面向人工智能的网络。**

一、引言

A. 背景

随着第五代移动通信技术网络在全球范围内持续部署，无线通信与网络领域的技术界已开始着眼于面向2030年及以后的第六代移动通信技术网络。尽管定义第六代移动通信技术的确切概念和技术尚未确定，但越来越多的研究工作开始讨论相关愿景、需求、用例和候选技术，例如 [1]–[3]。在这些讨论中，关于第六代移动通信技术的一些初步共识逐渐形成。例如，就第六代移动通信技术的主要需求而言，提升安全性 [4]和能源效率 [5] 的紧迫性已得到一致认同。就第六代移动通信技术的用例而言，增强型移动宽带（eMBB）的组合，超可靠低时延通信（uRLLC）和海量机器类通信（mMTC）也已被提出，尽管不同研究工作使用的术语各不相同 [6], [7]。至于候选技术，常被提及的例子包括卫星、空中、地面和水下网络的融合 [8], [9]、（亚）太赫兹与可见光通信 [10]、人工智能（AI）赋能的网络 [11]–[13]，等等。

一个值得特别关注的共识是，第六代移动通信技术可能需要全新的网络架构。在成本效益和效率的驱动下，网络架构的演进会随着网络所提供服务的变化而发展。例如，为了引入数据服务，第三代移动通信技术架构中出现了分组交换核心网组件，作为电路交换语音服务的补充 [14]。随后，为了应对数据流量的指数级增长，第四代移动通信技术引入了重新设计并简化的网络架构，采用扁平化的全互联网协议（IP）网络，从而提高数据速率并降低时延 [15]。在第五代移动通信技术时代，随着网络变得前所未有的异构化、服务日趋多样化，人们提出了多种网络架构创新，以实现灵活的面向服务网络，包括软件定义网络（SDN） [16]、云无线接入网（C-RAN） [17] 和网络切片 [18],[19]。因此，为了支持空前多样的服务，并满足极为严格的服务质量（QoS）或体验质量（QoE）需求，面向未来的第六代移动通信技术很可能需要在网络架构方面实现突破性创新。

在构想第六代移动通信技术的架构时，很难忽视两个关键要素，即虚拟化和人工智能。网络虚拟化已在第五代移动通信技术的架构中发挥重要作用 [20]。资源、功能和网络的虚拟化分别实现了资源共享、网络功能的软件化实现以及面向服务的网络，从而提高资源利用率，同时降低网络部署和运营成本。虚拟化体现了网络管理向软件化发展的趋势，使网络管理更加灵活、可扩展且适应性更强 [21]。因此，可以预见，虚拟化在第六代移动通信技术架构中仍将至关重要。至于第二个关键要素，即人工智能，全球越来越多的研究团队正在探索人工智能驱动的网络，人们对人工智能赋能第六代移动通信技术寄予厚望 [1], [22]。与基于启发式方法或数学模型的通信与网络方案相比，只要有足够的数据用于训练，基于人工智能的方法就能处理复杂的网络问题并获得准确结果，前提是

本研究工作得到了加拿大自然科学与工程研究理事会（NSERC）的研究经费支持。Xuemin (Sherman) Shen、Mushu Li、Conghao Zhou 和 Weihua Zhuang 均任职于加拿大安大略省滑铁卢市滑铁卢大学电气与计算机工程系，邮编 N2L 3G1（电子邮箱： {sshen, m475li, c89zhou, wzhuang}@uwaterloo.ca）。Jie Gao 任职于美国威斯康星州密尔沃基市马凯特大学电气与计算机工程系，邮编 53233（电子邮箱：j.gao@marquette.edu）。Wen Wu 任职于中国广东省深圳市鹏城实验室前沿研究中心，邮编 518055（电子邮箱：wuw02@pcl.ac.cn）。本文的通讯作者为 Weihua Zhuang。

有足够的数据可用于训练。这一优势适用于日益异构且动态变化的网络，在这类网络中，数学模型可能并不存在，或者无法准确描述所研究的问题。因此，不难预见人工智能在第六代移动通信技术中的重要性。

B. 第六代移动通信技术所需的架构创新

认识到虚拟化和人工智能的重要性后，我们进一步探究了当前最先进技术的局限性，以了解第六代移动通信技术所需的架构创新。现有虚拟化技术大多涉及通信网络中的服务提供 。例如，网络切片着重于可用的网络资源、服务提供能力以及各类服务的 QoS满意度 [23]。尽管这种侧重于服务提供的虚拟化使第五代移动通信技术能够处理多种共存服务，但对于第六代移动通信技术而言可能还不够，因为终端用户服务需求 的特征可能是实现以用户为中心的网络的关键。因此，未来的虚拟化应同时关注网络的服务提供能力以及网络中终端用户的服务需求。这将使终端用户也能实现虚拟化，而不再仅限于网络虚拟化。至于人工智能，现有研究大多针对网络的特定功能（例如路由 [24]）层（例如物理层 [25]）、网络片段（例如接入网络 [26]）或应用（例如自动驾驶 [27]）。与此同时，如何将人工智能融入跨越不同层或网络片段的网络架构，仍需进一步研究。人工智能驱动的网络的范围和程度尚待确定。

随着虚拟化扩展到涵盖服务提供和服务需求的同时，人工智能渗透到网络的每个角落，这两个要素之间的紧密联系是可以预见的，并可能主导第六代移动通信技术的架构需求。第一个联系通过网络和最终用户数据 [28]实现。虚拟化有助于描述网络服务提供能力、服务性能、资源利用率，以及在未来网络中的终端用户服务需求。因此，将产生海量数据，这些数据可用于描述网络和终端用户。如果得到妥善管理，此类数据能够同时赋能人工智能驱动的网络和人工智能应用（例如目标检测） [29]。第二个联系通过网络控制实现。人工智能可用于做出与虚拟化相关的决策，包括服务准入、切片建立、动态虚拟网络功能编排和资源调度。未来，人工智能还可以帮助控制终端用户虚拟化的数据收集，并提取虚拟化终端用户的特征。因此，人工智能有望提升虚拟化的功效和适应性。第三个联系通过网络资源实现。网络虚拟化的主要动机是协调不同服务之间的资源共享，从而提高网络资源利用率和服务满意度。人工智能驱动的网络可以致力于高效利用网络资源。由于虚拟化和人工智能都会消耗计算、通信和存储资源，因此二者可能会争夺网络资源。然而，人工智能有可能通过智能网络规划和操作提高虚拟化的效率，而虚拟化则可以通过适当的数据提供和管理提高人工智能的效率。因此，二者应当协同工作，以提升网络资源利用率和服务质量。

通过数据和控制实现上述连接，在现有第五代移动通信技术架构中已可见雏形。例如，第三代合作伙伴计划（3GPP）在 Release 15 [30] 中为第五代移动通信技术引入了网络数据分析功能（NWDAF），并在 Release 16 [31]中引入了网络自动化使能器（eNA）。该架构设计为 NWDAF 提供了一个框架，使其能够从其他网络功能（例如策略控制和网络切片选择功能）收集数据，并将分析结果（例如数据流量统计和预测）反馈给这些网络功能。在第六代移动通信技术中，数据收集和分析的范围与层级都将大幅扩展。网络数据分析很可能不再局限于一两个特定功能，而是由人工智能驱动，并可在整个网络中随处获得。同样，用于网络管理的数据不应再受类型、内容或格式的限制，而应按需提供有关网络和终端用户的信息。通过拓展虚拟化和人工智能在网络架构中的作用，可以满足这些期望。

C. 我们的愿景

我们对第六代移动通信技术网络架构的愿景，基于虚拟化和人工智能的重要性、它们在现有网络中的局限性，以及二者之间不可或缺的联系。具体而言，我们旨在设计一种网络架构：i）分别从服务提供和服务需求的角度支持网络和终端用户的虚拟化；ii）在统一架构下，将人工智能融入各种网络功能、层、部分和应用；更重要的是，iii）促进虚拟化与人工智能之间的相互作用，使二者能够共存、融合并相互增强。为进一步明确这一愿景，我们提出以下三个关键问题：

- 如何在网络切片之外进一步推进虚拟化?
- 如何让人工智能融入网络的方方面面?
- 如何通过网络架构设计有效整合虚拟化与人工智能?

为回答上述问题，我们提出了面向第六代移动通信技术网络架构的整体网络虚拟化和泛在网络智能理念。整体网络虚拟化通过融入网络切片和数字孪生范式，推动虚拟化迈向第六代移动通信技术。前者实现以服务为中心的网络管理，后者则为未来网络的虚拟化增添了以用户为中心的视角。泛在网络智能从面向网络的人工智能和面向人工智能的网络两个视角，实现人工智能在网络中的通用集成。前者强调人工智能在网络管理中的作用，后者则利用网络设计来支持人工智能应用。在这篇教程论文中，我们针对整体网络虚拟化和泛在网络智能，综述了现有研究，

介绍了我们的网络架构设计，并阐述了其优势。通过整合这两个组成部分，我们进一步提出了一种整体概念网络架构，以实现我们的愿景：构建面向第六代移动通信技术、具有前所未有的灵活性、可扩展性、适应性和智能性的网络。

本文可从三个方面为读者提供有用信息并带来裨益。首先，对于关注虚拟化和人工智能技术历史发展与当前进展的读者，我们梳理了相关文献，并回顾了这两者在通信网络中的应用背景。其次，对于探索虚拟化和人工智能未来发展方向的读者，我们提出了推动其迈向第六代移动通信技术的原创构想。具体而言，本文介绍了一系列设计与构想，例如将数字孪生融入整体网络虚拟化、利用互联人工智能进行网络管理、采用训练与推理分离的人工智能切片，以及数据-模型混合驱动方法。最后，在介绍整体网络虚拟化和泛在网络智能的愿景后，我们提出开放问题和挑战，以启发后续研究。

文献中已有少数关于虚拟化和人工智能的综述 [21], [32]–[34]。关于虚拟化，Minerva 等人介绍了物联网背景下基于数字孪生的现有应用 [32], ，另有一篇综述介绍了网络切片的关键使能技术和设计原则 [21]。关于人工智能，Boutaba 等人 对网络领域多个方面的人工智能应用进行了全面综述 [33],，另有一篇综述聚焦于基于深度学习（DL）的无线组网应用 [34]。相比之下，本文着眼于第六代移动通信技术的愿景。具体而言，在介绍最先进的虚拟化和人工智能技术后，我们提出了整体网络虚拟化和泛在网络智能等原创设计，以构建面向6G网络的新型概念架构。

D. 论文结构

本文教程的结构如图1所示。

第二节从整体网络虚拟化的角度阐述我们对第六代移动通信技术网络的愿景。我们在第 II-A 小节回顾现有的网络虚拟化概念和技术。然后，在第 II-B 小节中，我们介绍终端用户虚拟化，重点关注数字孪生。最后，在第 II-C 小节中，我们提出整体网络虚拟化的构想，重点介绍一种六层虚拟化架构。

第三节从泛在网络智能的角度阐述我们对第六代移动通信技术网络的愿景。第 III-A 小节概述了可能对第六代移动通信技术网络有用的代表性人工智能技术。第 III-B 小节介绍泛在网络智能的动机，并提出一种四级人工智能架构。第 III-C 和 III-D 小节分别总结现有研究，并提出我们对面向网络的人工智能和面向人工智能的网络的构想。

第四节整合了整体网络虚拟化与泛在网络智能，并提出了我们对第六代移动通信技术的整体愿景。第四节 A 小节回顾了有关第六代移动通信技术架构的相关研究。第四节 B 小节介绍了一种适用于 6G网络的概念架构，该架构融合了整体网络虚拟化与泛在网络智能。第四节 C 和 D 小节讨论了所提出架构的组件、子系统及其潜在实现。第四节 E 至 G 小节详细阐述了所提出架构所支持的三种相互作用，即数字孪生与网络切片之间、数据驱动方法与模型驱动方法之间，以及虚拟化与人工智能之间的相互作用。

表I：缩略语列表

3GPP	第三代合作伙伴计划
5G	第五代
6G	第六代
AI	人工智能
AL	人工智能级别
AP	接入点
API	应用程序编程接口
ARQ	自动重传请求
BS	基站
云无线接入网	云无线接入网
DL	深度学习
DNN	深度神经网络
DRL	深度强化学习
eMBB	增强型移动宽带
FL	联邦学习
IoT	物联网
IP	互联网协议
ITU	国际电信联盟
LSTM	长短期记忆
LTE	长期演进
mMTC	海量机器类通信
MIMO	多输入多输出
ML	机器学习
NFV	网络功能虚拟化
NN	神经网络
NWDAF	网络数据分析功能
QoE	体验质量
QoS	服务质量
RAN	无线接入网
SBS	小基站
SDN	软件定义网络
SNR	信噪比
UAV	无人驾驶飞行器
超可靠低时延通信	超可靠低时延通信
VL	虚拟化层
VM	虚拟机
WSN	无线传感器网络

第五节指出了与所提出的网络架构相关的关键挑战和未解决的问题，第六节总结了这项研究。

表I列出了本文中使用的缩略语。

II. 整体网络虚拟化

本节首先回顾现有网络中的虚拟化技术及其优势，然后介绍整体网络虚拟化的概念。

A. 网络虚拟化

网络虚拟化的概念和技术经过三十多年的发展， [35]。早期有关网络虚拟化的研究包括虚拟局域网，其旨在促进分布式系统中不同类型的操作（服务） [36], ，并提供灵活的网络控制、提高链路利用率 [37]。网络虚拟化的另一个例子是虚拟专用

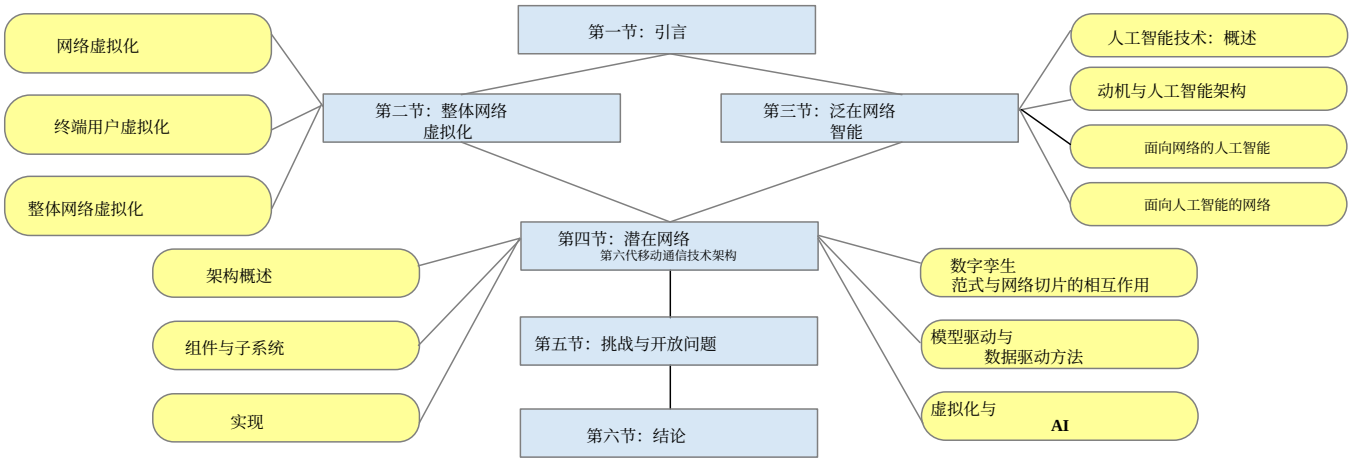


图1：本文的结构。

网络，它能够建立高效、安全的通信链路，连接地理位置分散的终端用户。随着时间推移，对可编程网络管理的需求扩展到增强网络架构这一目标。

云计算的发展推动了网络虚拟化的近期进展，包括网络功能虚拟化（NFV）和网络切片。借助 NFV，运行在通用计算服务器虚拟机上的软件实例取代了用于各种网络功能的定制专有硬件 [38]。在网络核心网中，NFV 应用于交换、防火墙、深度包检测和会话边界控制器等功能 [39]。在无线接入网中，NFV 应用于帧生成、调制、载波分配等 [40]。NFV 的实现为网络切片提供了基础，而网络切片是第五代移动通信技术（5G）中的一项关键网络架构创新。网络切片采用面向服务的网络管理视角，在共享的物理网络基础设施之上，为不同服务创建多个端到端虚拟网络，即切片。借助网络切片，网络资源会先在网络规划阶段预留给相应服务，之后在网络运行阶段分配给各个用户 [11]。切片的创建、调整和终止依据服务需求时空分布的变化进行，从而为网络管理提供高度的灵活性和适应性 [23]。[1]

虚拟化可以在网络中的不同层级和规模上应用。现有技术包括节点、链路、资源和网络层级的虚拟化。虚拟节点是对网络中底层节点的抽象，例如服务器、路由器和交换机；节点虚拟化的典型例子包括存储和计算服务器虚拟化 [41], [42]。虚拟链路是互连虚拟节点的逻辑信道。虚拟资源是对网络中计算、内存、存储和通信资源的抽象 [43],，而位于不同位置的物理资源可以构成虚拟资源池 [38]。例如，网络功能虚拟化是通过运行软件来执行网络控制或服务功能，并为其提供必要资源。虚拟网络是由虚拟节点和链路组成的组合，通过适当的虚拟资源分配来满足服务请求的服务质量要求，并由必要的网络协议提供支持。除上述研究工作外，表 II 总结了节点、链路、资源和网络虚拟化领域更具代表性的研究工作。

无论级别和规模如何，网络领域中的虚拟化通常都具备以下特征：

- 抽象——抽象提供网络的高层概览，同时隐藏底层物理网络实体（节点、链路或网络）或资源的细节 [63]。这简化了网络管理，并有助于灵活地提供服务；

- 共存——对应同一物理实体的多个虚拟实体可以共存，或者多个虚拟资源池可以在同一物理资源池上共存 [35]。这使面向服务的虚拟网络成为可能，并提高网络资源利用效率；
- 隔离——对应同一物理实体的共存虚拟实体应独立运行 [64]。这对于保证服务可靠性、安全性、可扩展性和 QoS 满意度至关重要。

学术界和业界都投入了大量精力研究网络虚拟化。在核心网络虚拟化方面，一些研究利用软件定义网络技术，通过不同的协议或应用程序编程接口（API）分离控制平面和数据平面，例如 OpenFlow [65]。此外，网络虚拟化已扩展至无线接入网，并提出了多种无线接入网虚拟化框架。SoftRAN 框架根据控制决策对时间的敏感程度，实现集中式和分布式无线接入网控制 [66]。另一个框架 FlexRAN 提供了用于实时无线接入网控制的分层架构，并采用灵活的应用程序接口来分离无线接入网中的控制平面和数据平面 [67]。在 AT&T 等业界企业的推动下，例如

[1]软件定义网络和云无线接入网也与网络虚拟化密切相关，因为虚拟化显著简化并加快了它们在现代无线网络中的实现。

表 II：节点、链路、资源和网络虚拟化的代表性研究工作

Type	Work	应用场景	研究重点	目标
Node 虚拟化	[44]	边缘计算	虚拟边缘节点放置	低成本放置并快速响应用户请求
	[45]	云计算	虚拟机（VM）部署-ment	可靠的虚拟机部署与路由
	[46]	IP 网络	作为互联网协议覆盖层的虚拟节点/路由器 lay	针对链路故障的实用互联网协议层弹性
	[47]	无线传感器网络（无线传感器网络）	用于传感器虚拟化的无线传感器网络架构	多个应用共享同一个无线传感器网络
	[48]	云无线接入网	接入点聚类	根据服务质量要求构建面向特定用户的虚拟基站
Link 虚拟化	[49]	WSN	虚拟骨干网构建	以高性能实现低复杂度骨干网构建保证
	[50]	通用	虚拟链路嵌入	在给定带宽需求的情况下降低拥塞概率
	[51]	互联网服务提供商 带有软件定义网络的互联网服务提供商网络	虚拟链路提供	在满足服务质量和鲁棒性约束条件
资源 虚拟化	[52]	云计算	复合 虚拟资源 映射	将计算与网络资源高效映射到底层联网云中的资源
	[53]	云计算	虚拟机迁移	虚拟机迁移期间低成本传输虚拟机存储和内存通过广域网
	[54]	无线 接入 网络（无线接入网）	无线资源虚拟化	在多个移动网络运营商
	[55]	RAN	无线资源虚拟化	通过无线资源虚拟化实现时延受限的 QoS 保障 tion
	[56]	车载网络	资源 共享 之间 切片	重复利用通信与缓存资源以支持应用程序满足不同的服务质量要求
网络 虚拟化	[57]	5G核心网 with SDN	网络功能链嵌 嵌入	在满足网络资源约束条件的前提下最小化嵌入成本
	[58]	采用软件定义网络的核心网络	网络功能链嵌-入	在满足网络资源约束的前提下，最小化网络中的总流量约束条件
	[59]	云无线接 入网	切片请求接纳	在网络约束下，最大化云无线接入网运营商的收入资源约束
	[60]	异构 无线 网络	动态无线电资源切片 ing	通过最优带宽切片和最大化网络效用用户关联
	[61]	5G 无线接入网	无线接入网中的无线资源分配 无线接入网切片	满足 通过适当的资源映射和调度来满足服务质量要求 调度
	[62]	IoT	面向服务 身份验证	保护隐私的切片选择与服务数据安全访问

以及中国移动，开放式无线接入网（O-RAN）被提出作为一个开源、开放接口的平台，以支持无线接入网虚拟化 [68], ，该平台能够整合人工智能，并提供用于数据驱动网络的应用程序接口 [69]–[71]。

虚拟化技术的采用使现代网络具备可编程性、灵活性和可扩展性，从而显著提高了网络部署与操作的成本效益。鉴于这些优势，可以预见，先进的虚拟化技术将成为第六代移动通信技术的必需。与此同时，现有的网络虚拟化范围具有局限性，因为虚拟化技术主要聚焦于网络基础设施和资源，而对终端用户的关注相对较少。在第六代移动通信技术中，终端用户虚拟化将因以下两个原因而变得必要。首先，随着终端用户设备、资源密集型服务以及异构动态网络日益多样化，在第六代移动通信技术时代，为终端用户提供服务质量体验保证将更具挑战性。对终端用户进行准确表征和抽象是实现终端用户虚拟化的必要条件，也可能是满足体验质量的前提。其次，由于人工智能将成为第六代移动通信技术的一大亮点，人工智能服务和基于人工智能的网络管理需要大量用户数据作为支撑。鉴于对数据的这种需求，终端用户虚拟化可以成为收集、管理和处理终端用户数据的一种具有竞争力的方法。

B. 终端用户虚拟化

直到最近，只有少数研究探讨了网络环境中的终端用户虚拟化。一个较早且与终端用户虚拟化相关的例子是网络托管化身，即终端用户的虚拟代理，可用于用户离线时的文件下载等应用 [72]。另一个例子是虚拟对象，它被提出作为物联网（IoT）平台的一个组成部分 [73]。其动机是通过虚拟化处理物理对象（终端用户）的异质性，并促进向终端用户提供服务。

作为实现终端用户虚拟化的一种潜在范式，数字孪生近来备受关注。数字孪生这一概念最初由 Michael Grieves 于 2003 [74], [75] 提出，用于工业领域的产品生命周期管理。随后，美国国家航空航天局和美国空军飞行器部门开发了一种适用于车辆的数字孪生范式，用于预测车辆的剩余使用寿命和任务成功概率 [76]。数字孪生的特点是对物理对象或流程进行完整的数字化表示，并在物理对象或流程与其对应的数字副本之间实现实时同步。数字孪生可以包含大量来自物理对象或流程的数据，以供高级分析使用；分析结果则可用于提升相应物理对象或流程的性能。文献 [77] 讨论了一般应用场景中的数字孪生示例，以及数字孪生支持大数据分析所需的潜在需求。文献 [78] 提出了在工业系统中实现代表物联网设备的数字孪生的潜在方案。其他具有代表性的数字孪生研究工作总结见表 III。

目前，大多数关于网络领域数字孪生的研究都聚焦于应用，例如分布式时钟同步 [79] 和计算卸载 [80]。相比之下，从网络

架构和网络管理视角研究数字孪生的工作目前还很有限。[2] [83], 文献 [83], 提出了一种以云为中心、基于数字孪生的网络架构，其中部署在网络边缘的终端用户数字孪生充当通信助手或网络数据记录器。

数字孪生似乎是实现终端用户虚拟化的一种直观解答。然而，鉴于灵活高效的网络管理与服务提供这一目标，将现有的、以网络切片为代表的网络虚拟化扩展到终端用户并非易事。例如，简单地使用节点级虚拟化，并在虚拟网络中将终端用户表示为虚拟数据源或数据汇，是很容易实现的。此外，尽管终端用户可能拥有通信和计算资源，资源级虚拟化却无法刻画终端用户特有的属性（例如位置和移动性），也无法刻画终端用户特定于服务的属性（例如数据流量变化）。有必要了解基于数字孪生的终端用户虚拟化的潜在优势、需求和实现，尤其应关注数字孪生与现有网络虚拟化框架的集成。

基于数字孪生的终端用户虚拟化可能具有双重优势，即丰富的终端用户数据和强大的网络仿真能力。网络基础设施和资源的虚拟化刻画了网络状态以及服务提供能力，而终端用户的数字孪生则可以提供有关服务需求和用户QoS/QoE满意度的大量数据。这些数据能够通过促进充分知情的网络规划与运营决策，在网络管理中发挥重要作用。此外，终端用户与其数字孪生之间的实时或近实时同步，能够实现强大的网络仿真。例如，可以创建同一虚拟网络的多个实例，并通过同步的终端用户数字孪生，将实时终端用户信息（例如位置和数据流量）提供给所有实例。[3] 不同的网络规划或运营策略可以在不同实例中应用和仿真，同时，每个实例都通过终端用户数字孪生提供的信息，与现实网络环境保持同步。

为了参与网络虚拟化，终端用户的数字孪生应满足以下需求：

- 灵活性：将终端用户抽象为数字孪生时，必须具备足够的灵活性，以表示异构的物理设备（如智能手机、车辆和工业传感器），并服务于各种应用（如虚拟现实游戏、自动驾驶和工业自动化）；
- 兼容性：基于数字孪生的终端用户虚拟化应补充并增强最先进的网络虚拟化，即网络切片。例如，终端用户的数字孪生应提供数据，以支持各种网络切片，而每个切片可能只能访问与该切片相关的一部分数据；
- 可定制性：数字孪生的属性应根据相应的服务、网络流量、资源利用率等进行定制和更新。例如，数字孪生所包含的数据数量和类型应能够调整，而非固定不变。此外，虽然数字孪生主要面向终端用户，但也应能够表示其他网络实体，例如搭载于无人机（UAV）上的移动基站（BS）。

此外，还应考虑创建和维护数字孪生所产生的网络资源消耗。

鉴于上述优势和需求，我们旨在围绕数字孪生的实现回答以下关键问题：

- 位置：数字孪生应托管在网络中的何处?
- 归属：数字孪生应存在于网络切片内部还是外部?
- 数据：数字孪生应包含哪些与网络相关的数据属性? 对于特定属性，应包含多少历史数据？是否应包含预测用户信息?
- 同步：如何通过从物理对象获取新数据，确定更新数字孪生中各项数据的频率?
- 控制：谁应根据哪些信息确定并更新数字孪生模型?

在下一小节中，我们提出一种用于整体网络虚拟化的新型概念架构，该架构融合了数字孪生与网络切片，并深入探讨上述问题。

C. 整体网络虚拟化

我们提出一种新型虚拟化架构，即整体网络虚拟化，用于将数字孪生融入网络虚拟化，从而提升网络管理和服务提供能力。所提出的虚拟化架构由六层组成，如图 2 所示，其中虚拟化层（VL）1 是负责数据收集的底层，VL 6 是负责数字孪生模型控制的顶层。各层概述如下：

VL 1 – 数据收集：根据规定的数据精度、上传方式、收集频率等要求，从相应的物理实体收集选定终端用户的数字孪生表示所需的数据。数据通过接入点收集，数据收集由部署在网络边缘的本地控制器控制；

VL 2 – 一级抽象：根据来自数字孪生模型控制层（即 VL 6 ）的当前数字孪生模型，确定每个数字孪生所包含数据的内容和格式。数字孪生利用 VL 1 收集的数据生成并更新。抽象可能包括聚合来自不同来源的数据、更新历史数据，以及为新的或额外的最终用户创建数字孪生。此层创建的数字孪生

[2]一些研究工作关注分布式网络（例如车联网），并采用数字孪生作为网络虚拟化而非终端用户虚拟化的一种方法 [81], [82]。[3]该仿真可应用于虚拟网络片段，例如网络边缘。

表 III：数字孪生相关研究工作

Work	应用	物理对象类型-ject	数字孪生的作用	数字孪生的使用目标
[84]	水下网络用于海洋观测	水下传感器/执行器	监测和测试观测系-tem	可视化海洋观测系统并增强仿真
[85]	车辆互联网的边缘计算车辆互联网	车辆	收集and共享信息关于车辆和周围环境	通过促进数据分析，赋能计算卸载
[86]	5G网络切片	网络切片	预测和监测切片性能能	辅助自主网络切片
[87]	智能工厂	输送系统中的工作站送系统	评估和验证控制策略gies	实施智能输送系统
[88]	智慧城市	道路基础设施	监测道路并检测车辆—辆/人员	通过收集和处理数据来支持智慧城市应用data
[89]	IoT	带有传感功能的对象能力	存储数据以检测事件并识别行为	通过态势感知和可解释性促进合成感知可解释性
[90]	工业4.0	工业机械	生成训练数据集和模拟结果	利用有限的标注数据实现准确的异常检测
[91]	智慧医疗	患者	处理数据以供分析，并开发人工智能模型	改善医疗运营
[92]	工业4.0	技术资产（例如，机器、环境）	整合模型和用于仿真的数据	增强基于仿真的系统工程
[93]	移动边缘计算ing	现实世界网络环境	训练学习算法并监测网络环境	实现学习，以优化用户关联、资源分配和卸载
[94]	工业4.0	产品、工作站、传送带	安全性关键流程的数据共享与控制安全性关键流程	构建基于状态复制与同步的安全架构
[95]	信息物理系统	通用物理设备	监测、诊断与预测tics	支持情境感知交互等应用以及驾驶辅助
[96]	IoT	通用物理系统tems	管理上下文信息并自我适应	通过自主机制提升自主性并加强协作数字孪生
[97]	焊接制造ing	人机交互系统	监测焊接机器人并实现仿真	可视化焊工行为并训练焊工
[98]	智能制造	Job shop调度系统	获取调度数据并进行仿真的调度策略	实现及时响应并减少调度计划偏差
[99]	智能制造	制造系统	预测并验证系统性-能	提高自主性并增强故障诊断能力
[100]	智能建筑	光伏能源转换单元	估算光伏能量的状态能量转换单元	提高故障检测的准确性
[101]	车联网	车辆和路侧单元	监测车辆的实时状态辆和路侧单元	支持网络资源管理
[102]	移动边缘缓存	车辆	捕捉.....的社会特征车辆	提升缓存管理的有效性

是一级数字孪生，代表各个最终用户，并托管在连接到本地控制器的服务器上；

VL 3 – 局部处理与控制：一级数字孪生的数据在网络边缘进行处理，以预测个体用户的行为，例如用户数据流量和移动模式，并制定用户级服务决策，例如计算卸载、内容交付和链路层协议适配。局部处理还可包括基于一级数字孪生对边缘网络或其部分进行仿真。局部控制可包括进一步聚合来自一级数字孪生的数据，以供 VL 4 使用；根据用户移动性迁移数字孪生；以及选择用于数字孪生表示的终端用户。与 VL2 的情况类似，局部处理与控制在隶属于本地控制器的服务器上进行；

VL 4 – 二级抽象：VL 3 聚合的数据在 VL 4 中按服务类型整理为各网络切片对应的数据。随后，为每个切片生成描述切片配置、切片资源利用率、切片服务等级协议满足度等的附加数据。然后，对上述数据进行抽象，以构建或更新各种切片的数字孪生。在此层创建的数字孪生是二级数字孪生，与虚拟网络（切片）相关联。二级数字孪生托管在连接到网络集中式控制器的服务器上；

VL 5 – 切片级处理与控制：该 data 来自网络切片二级数字孪生的数据经过处理，用于特定服务预测（例如时空服务需求分布预测）或切片级决策（例如规划与运营决策）。切片级处理可能包括基于二级数字孪生对端到端切片或其部分进行仿真。切片级控制可能包括切片准入、资源预留和切片服务覆盖控制。与 VL 4 的情况类似，服务级处理与控制在隶属于网络集中式控制器的服务器上进行；

VL 6 – 数字孪生模型控制：该层根据可用于数字孪生的网络资源、基于当前数字孪生得出的网络管理与服务提供决策的性能，以及动态时空服务需求，确定并更新一级和二级数字孪生的模型。例如，VL 6 会确定数据精度、不同数据属性的同步频率、每项属性的数字孪生中包含的历史数据量，以及是否纳入预测用户信息。此外，该层还会决定每个切片可访问的一级数字孪生数据子集。数字孪生模型控制也由隶属于网络集中式控制器的服务器执行；

由 VL 6 配置的一级数字孪生模型可能包含以下数据，这些数据应由本地控制器从 VL 1 的终端用户处收集：(1) 连接性和信道信息，例如终端用户连接的接入点以及每个连接的信道状态信息；

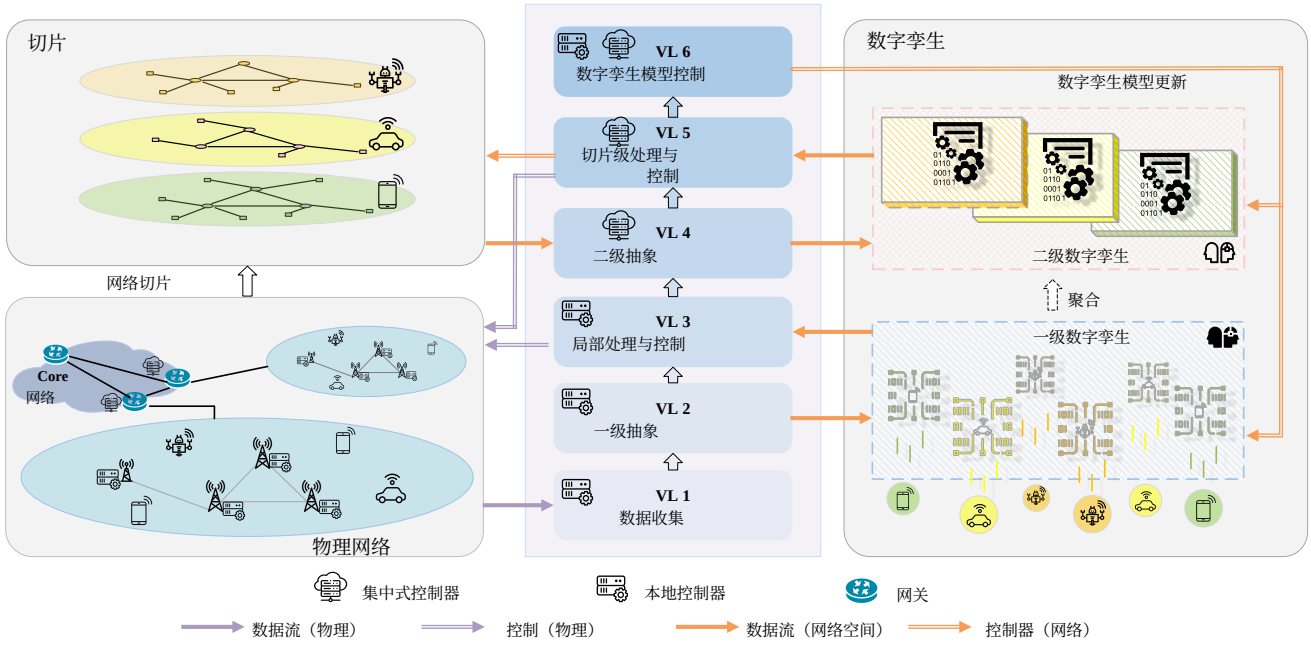


图 2：整体网络虚拟化的概念性六层虚拟化架构

n.

(2) 服务信息，例如活跃服务类型、每项服务的数据流量和服务质量满足度；(3) 用户信息，例如用户画像、用户位置和移动性、分配给用户的网络资源，以及用户本地的计算和缓存能力；(4) 其他特定于用例的信息，例如增强现实交互式游戏的运动传感器读数，或工业物联网设备的操作日志。由 VL 6 配置的二级数字孪生模型可能包含以下数据，这些数据应由集中式控制器收集或生成：(1) 切片服务需求，例如服务请求数量和时空服务请求分布；(2) 切片资源配置，例如为切片预留的通信、计算和缓存资源；(3) 切片性能，例如切片服务等级协议满足度、切片资源利用率和切片能耗；(4) 切片策略，例如用于网络功能部署、资源预留和资源调度的方法或算法；(5) 其他特定于用例的信息，例如无人机辅助网络的无人机轨迹配置。请注意，不同的终端用户数字孪生模型适用于不同类型的终端用户，并且每个网络切片都可以有其专门定义的切片数字孪生模型。例如，车辆和工业物联网设备的数字孪生很可能包含不同的数据；工业物联网切片与智能家居切片之间，或不同网络运营商的切片之间，其数字孪生模型也可能有所不同。因此，定制化需求使得 VL 6 中的数字孪生模型控制成为必要。

在概念性虚拟化架构中，VL 1 至 VL 3 与网络中的本地控制器接口，VL 4 和 VL 5 与网络的集中式控制器接口，而 VL 6 则同时与本地控制器和集中式控制器接口。该架构充分发挥了数字孪生的两大优势，即为网络管理提供丰富的数据，并支持强大的网络仿真。此外，该架构还满足了前文所述数字孪生在灵活性、兼容性和定制化方面的需求。最后但同样重要的是，它回答了 II-B 小节中提出的有关数字孪生实现的关键问题。

通过图 2 所示的架构设计，数字孪生与网络切片融入整体网络虚拟化的理念。网络切片采用了现有的网络虚拟化技术，例如网络功能虚拟化。数字孪生通过向切片提供经过组织和定制的终端用户数据，并进一步将切片抽象为二级数字孪生，从而增强网络切片。两级数字孪生的设计避免了针对同一用户为不同切片创建和维护多个数字孪生所造成的额外资源消耗，以及由此产生的同步负担。相反，每个切片都可以访问一级数字孪生中的数据子集，这些数据与相应服务相关，或属于位置、移动性等一般用户信息；随后，将相关数据进一步聚合到该切片的二级数字孪生中。在此架构中，网络切片遵循以服务为中心的网络管理，而数字孪生则为虚拟化加入了以用户为中心的视角。具体而言，一级数字孪生刻画终端用户及其服务需求，二级数字孪生刻画网络服务提供能力、网络性能和网络资源利用率。总体而言，数字孪生范式与网络切片共同支持网络管理和服务提供，而网络则根据网络动态的需要配置数字孪生和网络切片。

D. 整体网络虚拟化：总结

本节回顾了网络虚拟化的现有范围和技术，指出当前网络虚拟化的不足，介绍了将网络虚拟化和终端用户虚拟化相结合的整体网络虚拟化理念，并为整体网络虚拟化构建了六层虚拟化架构。

第五代移动通信技术中的资源、网络功能和网络虚拟化预计在第六代移动通信技术中仍将发挥重要作用，因为它们有助于实现灵活且自适应的网络管理。同时，以网络切片和网络功能虚拟化为代表的第五代移动通信技术虚拟化技术，主要从服务提供的角度关注网络虚拟化。在第六代移动通信技术中，扩展虚拟化的范围并纳入终端用户虚拟化将至关重要。

数字孪生范式是实现终端用户虚拟化的一种有前景的解答。在第六代移动通信技术中，数字孪生可用于描述各个终端用户的状态和服务需求。针对第六代移动通信技术网络的数字孪生研究仍处于起步阶段，目前存在多种定义或实现方式。在我们对整体网络虚拟化的构想中，数字孪生是可配置的数据组合，包含历史数据和实时数据，以及收集的数据和生成的数据，用于描述终端用户、基础设施或网络切片。此外，相应的数据收集和处理也可以配置。

为整合整体网络虚拟化，我们提出了面向第六代移动通信技术的六层虚拟化架构。该架构为系统地集成数字孪生和网络切片提供了参考设计，并回答了与6G网络中的数字孪生相关的重要问题，包括数字孪生部署在哪里、包含哪些数据，以及如何对其进行管理。

III. 泛在网络智能

泛在网络智能是我们对第六代移动通信技术愿景的第二个要素。本节首先概述现有的人工智能技术，然后介绍其动机，并提出泛在网络智能的四级架构。接着，我们从面向网络的人工智能和面向人工智能的网络两个角度阐述泛在网络智能的理念，并回顾相关研究工作。本节不对具体的人工智能技术进行综述，而是聚焦于泛在网络智能的架构和方法。

A. 人工智能技术：概述

人工智能旨在设计能够展现人类智能、像人类一样执行任务甚至表现得更好的智能机器或系统 [131]。机器学习（ML）的进步推动了人工智能在学术界和工业界的成功。机器学习技术支持的应用，例如计算机视觉和自然语言处理，其准确率能够超越人类水平。近年来，人工智能在构建智能网络方面的潜力，使其在无线网络研究领域受到广泛关注。

机器学习技术可分为三类：无监督学习、监督学习和强化学习。从学习结构来看，这些技术又可细分为集中式技术和分布式技术。表IV列出了无线网络中常用的机器学习技术。

无监督学习无需使用带标签数据集，即可评估数据中隐藏的特征和模式，用于数据分析，例如预测。无监督学习技术的一种常见应用是数据聚类，例如，k均值法 [103] 和混合模型 [105], 可用于解决网络规划问题，例如无线传感器网络中的簇形成 [104]和小蜂窝部署 [132]。神经网络可用于促进新型无监督学习算法的开发。例如，基于神经网络的自编码器能够利用数量有限的神经元学习输入数据的压缩特征，并可用于数据预测，例如流量预测 [107]。

监督学习利用给定的带标签数据集，挖掘输入数据与输出数据之间的映射关系。监督学习技术能够从数据集中的输入数据到带标签输出数据推导出映射函数，即训练模型。通过应用训练模型，可以评估新输入所对应的输出，并将其用于决策或预测。深度神经网络（DNN）是监督学习的一种典型方法。DNN 使用多层人工神经元来估计输入数据与输出数据之间的非线性相关性，并通过迭代不断提高估计精度。DNN 技术已在通信领域取得了许多成功应用。例如，卷积神经网络（CNN）利用卷积层和池化层识别多维输入数据之间的相关性，并已应用于调制分类 [112]；循环神经网络（ RNN ）挖掘序列数据中的相关性，已广泛用于流量预测[113] 和无线信道建模 [133]。

强化学习通过与环境交互、感知网络状态并评估反馈，迭代学习最优策略。其目标是在动态环境中最大化累积奖励。深度强化学习（DRL）结合了深度神经网络（DNN）与强化学习技术，广泛应用于资源管理，以解决复杂的决策问题。在深度强化学习中，神经网络充当近似器，用于存储高维状态或动作，使其能够高效地解决复杂问题。深度强化学习已广泛应用于无线网络的网络优化 [134],资源分配 [19], [118],[135],和用户关联 [116], [121], [123] 。

随着移动边缘计算的发展，分布式人工智能得以发展，用于利用网络边缘的计算资源，并减少因数据收集和交换而产生的通信开销 [136]。学习模型可以在网络边缘以半分布式或全分布式方式进行训练和评估。具体而言，联邦学习（FL）作为最受欢迎的分布式学习技术之一，利用分布在网络边缘的数据训练模型。一个集中式

表 IV：常见机器学习算法。

	无监督学习	监督学习	强化学习
集中式学习算法	•K均值 [103], [104] •混合模型 [105], [106] •自编码器 [107] •生成对抗网络 [108], [109]	•支持向量机 [110] •逻辑回归 [111] •深度神经网络 [107], [112], [113]	•深度Q学习 [114]–[116] •策略梯度 [117]–[119] •演员-评论家 [120], [121] •深度确定性策略梯度 (DDPG) [19], [122], [123]
分布式学习算法	•联邦学习 [124], [125] •分割学习 [126]		•多智能体强化学习 [127]–[130]

控制器汇总本地计算的学习模型，并更新学习模型中的参数。由于采用这种去中心化的模型训练方式，联邦学习能够保护隐私，并可应用于隐私敏感的网络管理场景 [137], [138]。此外，多智能体强化学习已经发展起来，用于以分布式方式实现强化学习，旨在处理网络智能体无法相互获取充分信息的场景。例如，多智能体强化学习技术可用于解决异构网络中的资源分配问题 [129], [130]。



图3：泛在网络智能四级人工智能架构示意图。

B. 动机与人工智能架构

在第六代移动通信技术中，人工智能有望渗透到网络的各个方面，包括终端用户、网络边缘和云，从而实现泛在网络智能。这一趋势得益于近几十年来机器学习、数据收集、边缘与云计算以及可编程网络控制领域的进步与创新。因此，人工智能将在许多方面从根本上改变现代网络，并催生众多令人振奋的应用。

人工智能应用可分为面向管理的应用和面向服务的应用，具体如下：

•面向管理的人工智能应用 - 在这些应用中，人工智能被用作网络管理工具，例如蜂窝网络中的发射功率分配 [139] 以及网络切片中的资源预留 [19]。人工智能技术（如强化学习）有望处理动态网络环境中的复杂决策问题。借助人工智能技术，面向管理的人工智能应用可以分析大量网络数据，

做出实时网络管理决策，然后根据新分析的数据更新网络管理策略。因此，对于此类应用，关键问题是如何利用先进的人工智能技术来管理和增强复杂网络，这属于面向网络的人工智能的范畴；

• 面向服务的人工智能应用 - 在这些应用中，人工智能以服务的形式提供给终端用户。在强大的计算服务器和精心整理的数据集的支持下，人工智能技术，尤其是深度学习，能够在广泛的应用中胜过传统技术，例如自动驾驶中的环境感知、智能医疗中的音频识别，以及移动虚拟现实中的物体检测 [140]–[142]。例如，

基于人工智能的YOLO算法能够以很高的准确率检测物体 [143], [144], ，而最先进的深度学习人脸识别算法的准确率可达99%或更高 [145]。在网络中支持面向服务的人工智能应用会消耗大量网络资源，包括用于模型训练/推理的存储和计算资源，以及用于数据收集和模型上传的通信资源。因此，对于此类应用，关键问题是如何设计和优化网络以支持新兴人工智能服务，这属于面向人工智能的网络的范畴。

需要注意的是，第六代移动通信技术中的人工智能范围包括面向网络的人工智能和面向人工智能的网络，其范围大于第五代移动通信技术；后者仅关注将人工智能应用于通信。

我们需要一种人工智能架构，用于描述人工智能在不同网络分段中的不同功能。文献中关于人工智能架构的研究并不多。边缘智能（或边缘人工智能）根据数据卸载的数量和路径长度分为六个级别 [131]。此外，边缘智能可分为两部分：面向边缘的人工智能（即利用人工智能技术增强和优化网络边缘）和边缘侧人工智能（即在网络边缘运行人工智能模型） [146], [147]。与这些关于边缘智能的研究不同，我们的研究关注范围更广的泛在网络智能，并根据人工智能在网络中的位置和功能将其划分为多个级别。

如图3所示，我们提出了一种四级人工智能架构，其中人工智能级别（AL）1和2侧重于面向服务的应用，而人工智能级别3和4则面向管理类应用。以下详细介绍各个级别。

人工智能级别 *1* - 终端用户托管的面向服务人工智能：利用终端用户的本地数据和计算资源，通过在本地处理人工智能任务，为终端用户提供终端用户托管的面向服务人工智能应用，例如服务。

例如移动键盘中的下一个词预测 [148],用户流量需求预测 [149], 以及车辆轨迹预测 [150]。当终端用户的计算资源不足以处理计算密集型人工智能任务时，可将部分计算工作负载卸载到附近的边缘服务器进行协同处理。

第*2*级人工智能 - 边缘托管的面向服务人工智能：面向服务的人工智能应用部署在靠近终端用户的网络边缘（例如，Wi-Fi接入点和基站），并作为低时延服务提供给终端用户，例如视频监控中的人脸识别 [151] 和虚拟现实中的目标检测 [152]。为支持边缘托管的面向服务人工智能应用，会收集、存储和分析终端用户的服务需求数据，然后利用分析结果提供服务。

第*3*级人工智能 - 边缘托管的面向管理人工智能：在此级别，人工智能托管于网络边缘的本地控制器上，用于实时执行网络管理，例如频谱分配、内容缓存 [153], 和计算卸载 [154]。具体而言，边缘托管的面向管理人工智能会向网络节点分配网络资源，以支持各项服务，包括第1级和第2级人工智能服务。例如，边缘托管的面向管理人工智能可用于在边缘网络之间执行服务迁移，从而确保高移动性用户（如车载用户）的服务连续性。

第*4*级人工智能 - 云托管的面向管理人工智能：云托管的面向管理人工智能部署在云端的集中式控制器上，用于网络管理，执行周期为数分钟或数小时，例如切片接纳控制 [155] 和虚拟网络功能部署 [156]。由于云拥有充足的计算与存储资源，因此可以训练和部署功能强大且复杂的人工智能模型，以管理大规模网络。

接下来，第 III-C 小节将详细阐述面向网络的人工智能，以说明人工智能在网络管理中的作用；第 III-D 小节将讨论面向人工智能的网络，以说明人工智能服务在6G网络中的提供方式。

C. 面向网络的人工智能

本小节讨论人工智能技术如何支持网络管理。我们首先回顾基于人工智能的网络切片相关研究，然后介绍我们提出的用于基于人工智能的网络切片的互联人工智能解答。

1）基于人工智能的网络切片：网络切片包括两个阶段：用于资源预留的网络规划阶段，以及用于资源调度的网络运行阶段 [11]。在网络规划阶段，需要在较大的时间尺度上为网络切片预留网络资源（例如，从几分钟到几小时）。在网络运行阶段，则在较小的时间尺度上将每个切片的预留资源分配给终端用户（例如，从几毫秒到几秒）。由于网络动态（例如网络流量随时空变化），基于模型的解决方案可能难以实现最优的网络切片策略。相比之下，人工智能技术能够通过分析收集到的网络数据来表征网络动态，并据此获得最优的网络切片策略。接下来，我们将综合考虑规划阶段与运行阶段之间的相互作用，回顾基于人工智能的网络切片研究。表V总结了基于人工智能的网络切片领域的代表性研究工作。

在较小的时间尺度上，本地控制器收集数据并提供资源调度策略，将为每个切片预留的资源分配给终端用户。具体而言，本地控制器根据两个因素确定资源调度策略：一是为每个切片预留的资源量，该数值由集中式控制器确定；二是来自与该切片相关的一级数字孪生的瞬时用户数据，例如服务类型、用户位置和用户移动性。确定最优资源调度策略的主要挑战有两个：终端用户数量庞大，以及服务需求动态变化。人工智能技术有望应对这两项挑战。首先，为大量终端用户调度资源时，可以采用无监督学习方法（例如k-均值 [167] 和基于深度神经网络的自编码器 [107],），根据终端用户的服务需求对其进行分类。对于服务需求相似的终端用户，可以采用相似的资源调度策略，从而实现可扩展的网络管理。例如，距离较近且移动模式相似的终端用户可能会经历相似的信道统计特性，因此可以对其采用相同的功率控制策略。其次，为应对网络动态变化，可以采用强化学习生成自适应资源调度策略 [168]。强化学习通过迭代分配资源来最大化长期奖励函数，并根据网络环境的反馈更新奖励函数。此外，强化学习还可以与深度神经网络相结合，例如循环神经网络 [27] 和传统神经网络 [123], ，以分析用户数据的时空模式，从而找到最优资源调度策略。

在较大的时间尺度上，本地控制器将收集到的用户级数据聚合为来自第二级数字孪生体的服务级信息，即切片数字孪生体。利用切片数字孪生体提供的信息，集中式控制器为每个切片预留网络资源。资源预留面临两方面的挑战。首先，在网络流量随时间变化的情况下，如何主动进行资源预留，避免资源过度配置或配置不足，是一项具有挑战性的任务。其次，资源预留策略与调度策略相互耦合，进一步增加了资源预留的复杂性。人工智能技术可以采取以下方式应对这些挑战。为应对主动资源预留的挑战，可以使用长短期记忆（LSTM）网络等监督学习方法，利用历史网络流量负载的特征来预测近期的流量负载 [149], [159]。随后，集中式控制器可以利用预测的流量负载进行主动资源预留。为处理资源预留与调度之间的相关性，可以采用强化学习，在预留资源的同时，将网络运行策略作为动态网络环境的一部分加以考虑 [19], [135], [164]。此外，基于选项的分层强化学习技术可以成为联合优化

表 V：基于人工智能的网络切片代表性研究工作

阶段	Work	研究重点	目标	人工智能方法
网络规划	[107]	网络容量预测	预测各个虚拟网络的容量	基于深度神经网络的自动编码器
	[86]	网络的虚拟表示 工作切片	捕捉切片之间的关系，并监测 动态网络环境中的端到端性能	图神经网络
	[157]	资源预留调整- ment	最大化从以下租户获得的总体奖励 切片	深度对偶神经网络
	[158]	带宽分配	联合最大化频谱效率和服务质量要求- 求满意度比率	生成对抗网络和 深度Q网络
	[159]	带宽分配	联合最大化频谱效率和整体服务 切片的级别协议满足率	长短期记忆和优势 演员-评论家
	[160]	流量 预测 and re- 资源配置	最小化切片服务级别协议的违约 违约	门控循环单元
网络操作	[161]	计算卸载	最小化服务的平均计算时间并最大 化用户计算体验	深度Q网络
	[162]	切片选择与信道分配 位置	降低无线传输的功耗 面向切片化雾无线接入网	强化学习
	[163]	内容缓存放置 和交付	管理缓存资源以最大化缓存命中率 同时满足资源预留约束	深度Q网络
	[164]	切片间协调	最大化竞争所带来的长期收益 服务提供商通过资源编排	深度Q网络
	[165]	切片间协调	通过调度最大化切片的QoS满意度 切片间的传输功率和资源共享	多智能体深度Q学习
两阶段相互作用	[19]	计算资源分配- 车联网中的	为切片分配频谱和计算资源 同时最小化计算服务时延	深度确定性策略梯度
	[166]	跨切片接纳 and 拥塞控制	通过资源预留最大化运营商收益，并 实时调整预留资源	状态-动作-奖励-状态-动作 （SARSA）

资源预留和网络运行策略并应对两个阶段网络动态的一种潜在解答。该技术通过根据决策时间尺度 [169]或决策代理 [170]对决策变量进行分组，然后确定决策变量，已被用于解决复杂的深度强化学习问题。通过这种新型强化学习技术，可以迭代地获得资源预留与资源调度策略之间的复杂相关性。要将该技术应用于网络切片，集中式控制器可以在较大的时间尺度上选择资源预留策略，而本地控制器则在较小的时间尺度上寻找最优资源调度策略，从而联合优化资源预留策略和调度策略。

2）面向网络管理的连接式人工智能解决方案：现有的网络管理人工智能应用大多侧重于单项控制功能。例如，基于学习的自编码器能够在信道状态信息有限的情况下，实现高速数据传输的可靠发射功率控制 [171]，；深度神经网络则能够以较低的通信和处理开销选择媒体接入控制协议参数 [172], [173]。尽管已有多种人工智能技术被用于网络管理，但各网络控制功能中的人工智能模型通常彼此隔离。这种隔离可能导致数据处理效率低下且重复冗余，因此亟需整合基于人工智能的网络控制功能中的人工智能模型。

集成人工智能模型 [6] 的解答有三种类型。第一种解答将整个网络视为一个黑盒，由单个人工智能模型描述整个网络并生成网络控制策略。这种结构简化了网络管理中的决策过程。然而，由于输入数据的维度很高，训练单个人工智能模型可能极其困难。第二种解答则在网络中针对不同的网络控制功能采用不同的人工智能模型，并且这些人工智能模型通常彼此独立，以降低训练复杂度。然而，这种方法忽略了网络功能之间的相关性与相互作用，因此无法获得全局最优的网络管理策略。此外，不同人工智能模型可能会对具有相似网络功能的网络数据进行重复处理，从而降低网络管理效率。例如，用于用户移动性管理和计算服务迁移的人工智能模型会重复分析最终用户的移动性。相比之下，第三种解答，即互联人工智能，利用网络控制功能之间的相关性，将其人工智能模型连接起来，并使它们能够协同制定网络控制决策。互联人工智能解决方案通过凸显模型之间的相互作用，并平衡训练复杂度与网络性能，在集成人工智能模型方面具有优势。因此，互联人工智能解决方案在促进基于人工智能的网络切片方面潜力巨大。然而，目前关于互联人工智能解决方案的研究仍然有限。如何将互联人工智能解决方案应用于网络管理，仍需进一步研究 [26]。

分布式学习技术的最新进展有助于开发用于网络管理的连接式人工智能解决方案。 [174], 研究的模型分割可将一个全局深度神经网络划分为多个子神经网络（子神经网络）。根据可用的计算和通信资源，这些子神经网络可部署在不同的网络实体上，并彼此通信 [175], [176]。此外，[177] 和 [178]提出并评估了嵌套式深度神经网络的构想：子神经网络既能拥有各自的功能，又能共同支持全局深度神经网络进行模型推理和训练。利用上述两种技术，每个子神经网络都能执行特定的网络控制功能。因此，多个子神经网络可以协作完成共同的控制功能，从而将连接式人工智能解决方案应用于网络管理。

基于 基于上述先进的深度神经网络技术，我们提 呈现

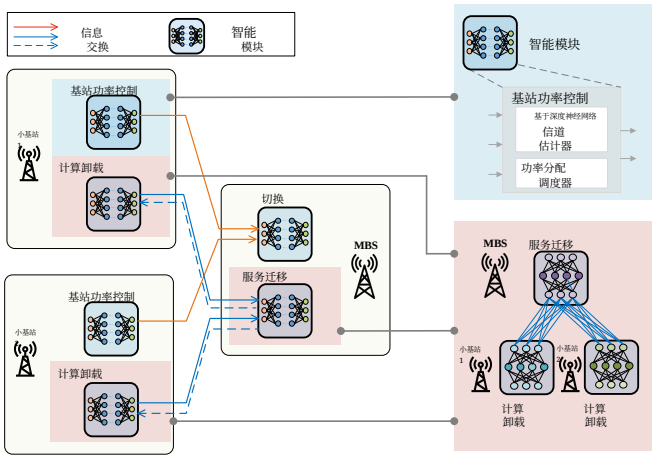


图4：用于网络管理的互联人工智能解决方案。

接下来，我们提出将连接式人工智能解决方案应用于网络管理。网络管理的控制功能被封装为智能模块。智能模块既可以仅通过深度神经网络实现，也可以由深度神经网络与传统的基于模型的技术协同实现。图 4 右上角展示了一个示例，其中用于功率控制的智能模块包括一个基于学习的信道估计器和一个基于模型的功率分配方案，例如注水功率分配 [179]。此外，每个智能模块中的深度神经网络都可以充当全局深度神经网络的子神经网络。智能模块彼此连接，以共享信息，例如其输出、模型训练中的梯度信息以及聚合用户数据。借助智能模块，控制功能可以采用分而治之的方式管理网络，从而避免无层人工智能所需的复杂模型训练。通过模型分割和嵌套式深度神经网络技术，多个网络控制功能可以协同制定控制决策，实现全局最优的网络管理。

图4展示了互联人工智能设计的另一个示例，即支持移动边缘计算。我们以车联网为例说明这一设计。需要注意的是，其他网络也可以采用相同或类似的设计。作为边缘服务器的小型基站（小型基站）可以处理车辆卸载的计算任务。小型基站中的智能模块根据网络状态和计算卸载请求，提供计算卸载决策，包括需要卸载的计算任务、用于卸载的传输功率、任务调度等。由于车辆具有较高的移动性，且小型基站的通信覆盖范围有限，计算任务通常会在小型基站之间迁移，这称为服务迁移，而迁移决策由宏基站确定。服务迁移决策与计算卸载决策高度耦合。例如，当车辆向某个小型基站卸载更多计算任务时，发生服务迁移的可能性会增加。此外，服务迁移需要多个小型基站协同完成。在我们的互联人工智能构想中，服务迁移决策和计算卸载决策由系统联合确定。具体而言，我们将深度神经网络通过深度神经网络分割和嵌套式深度神经网络技术分割为多个子神经网络。其中一些子神经网络部署在小型基站上，用于提供计算卸载决策。这些子神经网络还与部署在宏基站上的一个子神经网络相连接，后者可用于制定迁移决策。在此示例中，服务迁移模块的输入包括小型基站上智能模块的输出，例如计算卸载决策和子神经网络的参数；服务迁移模块的输出则是服务迁移策略。通过这种方式，智能模块可以协同为移动边缘计算制定决策。

D. 面向人工智能的网络

除了管理网络之外，人工智能还可以作为服务发挥作用，即人工智能服务，它们位于图3所示拟议人工智能架构中的 AL 1 和 AL 2 级。面向人工智能的网络 旨在设计和优化网络，以促进人工智能服务。本小节首先介绍面向人工智能的网络的动机。接下来，回顾现有研究工作，并提出研究挑战。最后，提出并阐述人工智能切片的构想。

1) 动机： 面向人工智能的网络正受到学术界和业界的广泛关注。在学术界，面向人工智能的网络需要网络研究人员与人工智能研究人员开展广泛的跨学科研究合作，以开发新的通信标准和技术，从而大规模支持人工智能服务 [141], [180]–[182]。在业界，国际电信联盟（ITU）正在讨论高层架构，以集成、编排和更新未来网络（包括 IMT-2020 网络）中的人工智能组件 [183], [184]。3 GPP 的一些工作组正在研究网络中的数据收集框架，以支持人工智能服务 [185], [186]。值得注意的是，面向人工智能的网络正成为促进网络中人工智能服务不可或缺的组成部分，并有望成为第六代移动通信技术的关键使能技术。

面向人工智能的网络应考虑以下因素：

- 分布式数据 - 随着各种物联网设备和小型基站的广泛部署，许多分布式网络节点（例如终端用户和网络边缘）会产生海量数据。在传统的基于云的人工智能范式中，云会收集海量分布式数据用于模型训练，并在云端部署训练良好的模型以进行模型推理。这种范式面临频谱资源短缺和用户隐私泄露等问题。[4] 为了解决这些问题，一种潜在的解答是以分布式方式在大量网络节点上提供人工智能服务 [148]，这需要新的网络协议来协调多个网络节点；

- 受限资源 - 网络节点（如终端用户）的资源有限，而最先进的人工智能模型（例如包含数十个神经网络层的深度神经网络模型）较为复杂。因此，运行复杂的

[4]谷歌的自动驾驶汽车每秒可生成超过 750 MB 的数据 [187]。

人工智能模型可能会耗尽单个网络节点的计算资源和能源。[5] 借助先进的模型划分技术（例如深度神经网络分区），可以将复杂的人工智能模型划分为多个子模型，并将其嵌入网络中，在子模型之间交换数据 [189]。执行子模型会消耗网络节点的计算资源，而在子模型之间交换数据也会消耗通信资源。因此，要以经济高效的方式在多个网络节点上运行人工智能模型，就需要创新的网络嵌入设计；

- 网络异构性与动态性- 第六代移动通信技术网络将具有高度异构性，其中网络节点拥有数量各异的通信、计算和存储资源。由于复杂的人工智能模型需要部署在多个网络节点上，执行人工智能任务需要审慎地分配这些网络节点的资源。此外，网络动态（例如网络节点之间随时间变化的信道条件以及时空服务需求）进一步加剧了资源分配决策问题的复杂性。因此，有必要设计定制化的资源管理算法，在适应网络动态的同时优化人工智能性能。

面向人工智能的网络的范围涵盖人工智能服务的整个生命周期，该生命周期包括三个阶段。第一阶段是通过通信链路进行数据收集，以开展模型训练。例如，需要收集终端用户的实时服务负载数据，以训练用于服务需求预测的人工智能模型。第二阶段是模型训练，即基于所收集的数据实现特定目标。例如，需要处理大量图像来训练基于深度神经网络的目标检测模块，直到满足目标精度要求。第三阶段是模型推理，即应用训练良好的模型来完成特定的计算任务。例如，基于人工智能的目标识别通过车载摄像头捕获的实时图像，检测并分类附近的车辆、行人和障碍物 [143]。

2) 最先进的方法：目前，面向人工智能的网络研究尚处于起步阶段，现有研究成果寥寥。本小节根据人工智能服务生命周期，将现有研究分为数据收集、模型训练和模型推理三类。表 VI 总结了具有代表性的相关研究。

数据收集 - 目标是高效地从终端用户处收集数据，以优化人工智能性能。由于数据分布在网络中的终端用户之间，因此需要为终端用户调度传输资源，以便上传其数据。例如，一级数字孪生需要定期与终端用户进行数据同步，而此类数据可用于人工智能服务。数据收集是无线传感器网络 [203]和无人机网络 [204],中广泛研究的经典研究问题，这些研究工作侧重于优化数据收集可靠性或所收集数据的数量。在人工智能服务中，收集的数据用于训练人工智能模型，并且数据样本对于模型训练可能具有不同的重要性级别。仅仅最大化收集数据的可靠性或数量并非最优。因此，需要将模型训练纳入考量的新型数据收集设计，以实现性能优化。

近年来，以人工智能为中心的数据收集主要沿以下两个研究方向展开：

- 资源分配- 为优化人工智能模型准确率，研究人员提出了数据重要性感知资源分配方案。其思路是在安排数据传输时，同时考虑终端用户的信道条件和数据重要性等级 [190]。数据重要性等级可通过数据不确定性来表征，即不确定性越高，重要性越高。数据不确定性可用熵来衡量 [205]。在多模型训练场景中，研究人员还探讨了数据收集的功率分配问题 [191]。由于收集的数据样本数量会影响模型准确率，以学习为中心的功率分配方案可以调整用户的传输功率，从而确定不同人工智能模型所收集的数据量，进而在传输功率预算约束下最大化整体模型准确率；

- 协议- 目前已有一些以人工智能为中心的数据收集协议。在信道条件较差的网络环境中，会采用数据重传来提高数据收集可靠性。现有的自动重传请求（ARQ）重传协议，例如长期演进（LTE）网络中的混合自动重传请求，会在满足最终用户的信噪比（SNR）阈值后，触发丢失数据包的重传。传输协议应纳入数据样本的重要性，以加快模型训练过程。文献 [192]提出了一种用于基于CNN的分类模型训练的重要性感知ARQ协议。该协议综合考虑数据重要性等级和信道条件，以确定数据重传阈值，从而提升模型训练性能。

模型训练 - 由于数据分布式存储以及用户隐私方面的顾虑，分布式训练适合在网络中训练人工智能模型 [206]。联邦学习（FL）是最有前景的分布式训练范式之一，可应用于智慧医疗和金融服务等多个领域 [138],[207], [208]。联邦学习的运行方式如下：每个最终用户使用自己的数据迭代训练一个本地模型，并将其上传至边缘服务器。随后，边缘服务器聚合这些本地模型，得到一个全局模型。模型训练会持续多个轮次，直至全局模型达到令人满意的准确率。

由于模型是在本地训练的，FL 具有通信效率高的特点，并且能够保护终端用户的数据隐私 [148],[209]。然而，随着最先进人工智能模型的数据规模不断增加，

[5]在定制的节能型 Eyeriss 芯片上使用 AlexNet 处理一幅图像，能耗最高可达 0.28 W [188]。

表 VI：面向人工智能的网络相关研究工作总结

主题	Work	贡献	亮点
Data 收集	[190]	根据用户数据重要性级别和信道条件调度数据传输	数据重要性感知频谱分配
	[191]	分配用户的传输功率，以调节所收集的数据样本数量，从而提升多个人工智能模型的整体模型准确率	数据量感知功率分配
	[192]	设计一种重要性感知ARQ协议，其中联合考虑用户数据重要性级别和信道条件，以触发数据重传	数据重要性感知重传协议
模型训练	[193]	提出一种边缘云辅助联邦学习框架，其中边缘服务器和云服务器交替聚合本地模型，以减少通信开销	双层联邦学习框架
	[194]	提出一种用于模型聚合的空中计算方法	空中模型聚合
	[195]	根据用户数据，选择对收敛贡献更大的用户参与模型聚合分发	数据分发感知型用户选择
	[196]	考虑用户间异质性，选择训练延迟较低的用户	训练时延感知型用户选择
	[197]	在资源预算范围内优化本地模型更新次数	本地更新频率优化
	[198]	根据终端用户的模型重要性等级和信道条件调度模型上传	模型重要性感知模型上传
模型推理	[144]	优化视频帧率和输入图像分辨率，以平衡服务时延和检测准确率，满足虚拟现实用户的需求	数据分辨率优化
	[199]	为实时视频分析选择最优深度神经网络模型	深度神经网络模型选择
	[200]	选择最优深度神经网络模型的切分层，以最小化用户-边缘深度神经网络的推理时延	用户-边缘深度神经网络模型划分协同
	[201]	在终端用户、网络边缘和云之间划分复杂的深度神经网络模型，以降低通信开销	用户-边缘-云端 DNN 模型分割
	[202]	设计一种协同深度神经网络模型推理方案，在物联网设备上使用轻量级模型并在网络边缘使用未压缩模型	协同深度神经网络模型推理 ence

上传本地人工智能模型仍然给频谱受限的无线网络带来了日益严峻的压力。此外，拥有强大计算服务器的终端用户可以以较低的时延进行本地模型训练。因此，由于无线电资源有限而产生的模型上传时延，可能成为整个 FL 时延的主要组成部分。因此，有必要在资源受限的无线网络中最大化 FL 性能。

近期研究从以下几个方面优化联邦学习的性能：

- 联邦学习框架设计 ——一类研究致力于设计创新的联邦学习框架，以降低通信开销。文献 [193], 提出了一种新颖的两层分层联邦学习框架，该框架协调终端用户、边缘服务器和云服务器执行联邦学习。每个联邦学习轮次中，各边缘服务器会聚合其覆盖范围内终端用户的本地模型；而云服务器则每隔若干个联邦学习轮次，聚合其覆盖范围内边缘服务器的模型。由于覆盖范围广，该两层框架能够支持大量终端用户进行模型训练；同时，由于模型聚合频率较低，还能减少云服务器与边缘服务器之间的回传数据流量。文献 [215]将此类框架应用于具有地理分布式数据的工业物联网网络；

- 模型聚合 ——另一类研究探讨频谱效率更高的模型聚合。文献 [194],[216], [217]研究了基于空中计算的方法。其基本思想是利用无线多址信道的叠加特性进行模型聚合，从而减少无线电资源消耗；

- 资源管理- 联邦学习的性能可以通过高效的资源管理得到优化。由于联邦学习的性能取决于多个因素，例如终端用户选择、本地模型更新的次数以及本地模型重要性等级，因此开发了以下不同的资源管理方案：(1) 用户选择——如何选择参与联邦学习过程的终端用户会影响模型收敛和训练时延，因此对联邦学习的性能至关重要。已有一些终端用户选择算法基于训练数据分布 [195] 和本地训练时延 [196]等原则提出；(2) 联邦学习参数——为降低通信开销，终端用户会在模型上传前进行若干次本地模型更新。在给定资源预算的情况下， [197],研究了本地模型更新次数的最优值，并为选择本地模型更新次数提供了理论指导；(3) 本地模型重要性等级——这一概念由数据重要性的理念扩展而来[7] 。 [198], 提出了一种重要性感知模型上传策略，其中，本地模型重要性等级较高且信道条件良好的终端用户会被优先安排，以加快联邦学习的收敛。

模型推理 - 对于网络中的许多人工智能服务，通常会将人工智能模型部署在终端用户和边缘服务器上，以实现较低的服务时延。模型推理需要大量计算，而终端用户和边缘服务器的计算能力和电池电量通常有限。执行模型推理任务通常会导致服务时延较长、能耗较高。因此，在执行模型推理时，应在节点能量约束下满足服务时延要求，这需要创新的模型推理方案。

[6]ResNet32 [210], Inception-v3 [211], AlexNet [212]和 VGG16 [213] 模型的数据规模分别为 50 MB、108MB、240MB 和 552 MB [214]。

[7]模型重要性可以通过逐层梯度范数来衡量。在联邦学习中，梯度范数较大的本地模型对全局模型收敛的贡献更大 [197]。

现有关于模型推理的研究可分为以下几类:

• 数据分辨率 -原始数据会被卸载到边缘或云服务器进行模型推理。输入数据分辨率会影响推理准确率。例如，目标检测的准确率与输入图像分辨率有关 [144],，而输入图像分辨率又会影响卸载数据量，因为高分辨率图像的数据量通常较大。考虑到推理准确率与卸载数据量之间的权衡，应优化输入图像分辨率，以满足目标人工智能服务需求。文献 [144]研究了如何优化视频帧率和输入图像分辨率，以平衡虚拟现实用户的服务时延和检测准确率;

• 模型选择 -选择适当的人工智能模型，以满足特定的人工智能服务需求。除了数据分辨率外，推理准确率还取决于人工智能模型的类型。与浅层深度神经网络模型相比，具有更多隐藏层的深度神经网络模型通常能实现更高的推理准确率。考虑到网络边缘部署了多种可用的深度神经网络模型，文献 [199]研究了如何为实时视频分析选择最优的深度神经网络模型;

• 模型分割 ——借助先进的模型分割技术，人工智能模型可以被划分为多个子模型，并嵌入不同的网络节点中以执行模型推理。例如，利用深度神经网络的分层结构，可以在适当的深度神经网络层（即切分层）将整个深度神经网络模型划分为终端用户侧模型和服务器侧模型。由此，终端用户和边缘服务器可以通过协同方式执行模型推理。深度神经网络模型可以根据不同目标进行分割。例如，文献 [200], 研究了如何通过模型分割最小化推理时延，其中在线学习算法可以自适应地确定最优切分层。为降低网络节点之间的通信开销，可以像文献 [201]那样，将复杂的深度神经网络模型划分为分别部署在终端用户、边缘服务器和云上的子模型;

•模型压缩- 使用轻量级模型，便于终端用户快速进行模型推理。通过各种模型压缩技术（如权重剪枝 [218], 知识蒸馏 [219]和快速退出 [220]），可以获得计算效率高的压缩模型。例如，权重剪枝技术会移除不太重要的模型权重，以降低模型推理的计算复杂度，同时使推理准确率接近未压缩模型。为提升服务性能，研究人员在工业物联网网络中提出了一种协同模型推理方案，在物联网设备上部署轻量级模型，在网络边缘部署未压缩模型 [202]。物联网设备根据时变的信道条件动态作出人工智能任务卸载决策，以尽可能降低服务时延，同时满足基于深度神经网络的故障诊断服务的准确率需求。

3）研究挑战： 尽管已开展上述研究工作，在网络中提供人工智能服务仍面临各种挑战，以下将讨论其中一些挑战。

复杂的实现方案选择 - 人工智能服务可以通过多种方案实现，这些方案具有不同的模型结构、训练流程和推理过程。例如，目标检测服务可以通过不同的神经网络实现，例如 AlexNet [212] 和 SqueezeNet [221]。即使模型结构相同，模型也可以采用不同的方式进行训练，例如集中式训练、分布式训练（如 FL [207]）以及半集中式训练（如分割训练 [126]）。此外，模型推理也可以采用多种方式进行，例如仅终端用户推理、仅边缘推理和协同推理。不同的实现方案会消耗不同数量的计算、存储和通信资源。因此，有必要为人工智能服务选择适合其服务特征和网络动态的实现解答。

多维服务质量要求 - 人工智能服务的服务质量要求是多维的。人工智能模型的准确率通常是一项关键性能指标。此外，在许多用例中，应以低时延向终端用户提供人工智能服务。例如，出于安全考虑 [143], , 自动驾驶中的目标检测服务时延应低于 100毫秒，而自动驾驶车辆要求三维目标检测具有超高准确率 [222]。此外，这些性能指标彼此相关。高准确率的目标检测通常需要以高分辨率图像作为输入，并使用先进的人工智能模型处理输入图像，这可能导致较长的服务时延。如何满足人工智能服务的多维服务质量要求仍需进一步研究。

4) 人工智能切片：为了更好地支持人工智能服务，我们扩展了网络切片的概念，并提出了由两个子切片组成的人工智能切片构想。其基本思路是构建用于模型训练的训练子切片和用于模型推理的推理子切片。这两个子切片在逻辑上相互隔离，并使用各自的网络资源。之所以要进行训练与推理分离，是因为这两个阶段可能具有不同的目标。

图 5 展示了人工智能切片的示意图。在人工智能切片中，训练子切片和推理子切片共享同一资源池，并相互协调以共同支持人工智能服务。首先，将人工智能切片的多维服务质量要求分解为两个子切片各自的服务质量要求。对于自动驾驶中的目标检测服务，需要同时满足较高的检测准确率（例如 99%）和较低的服务时延（例如 100毫秒）。训练子切片应满足检测准确率要求，而推理子切片应满足服务时延要求。其次，为满足两个子切片各自的服务质量要求，需要根据两个子切片的性能及其服务质量要求，合理地将为人工智能切片预留的资源分配给这两个子切片。随后，在已分配资源的基础上，对两个子切片进行配置，使其分别满足各自的服务质量要求，具体如下:

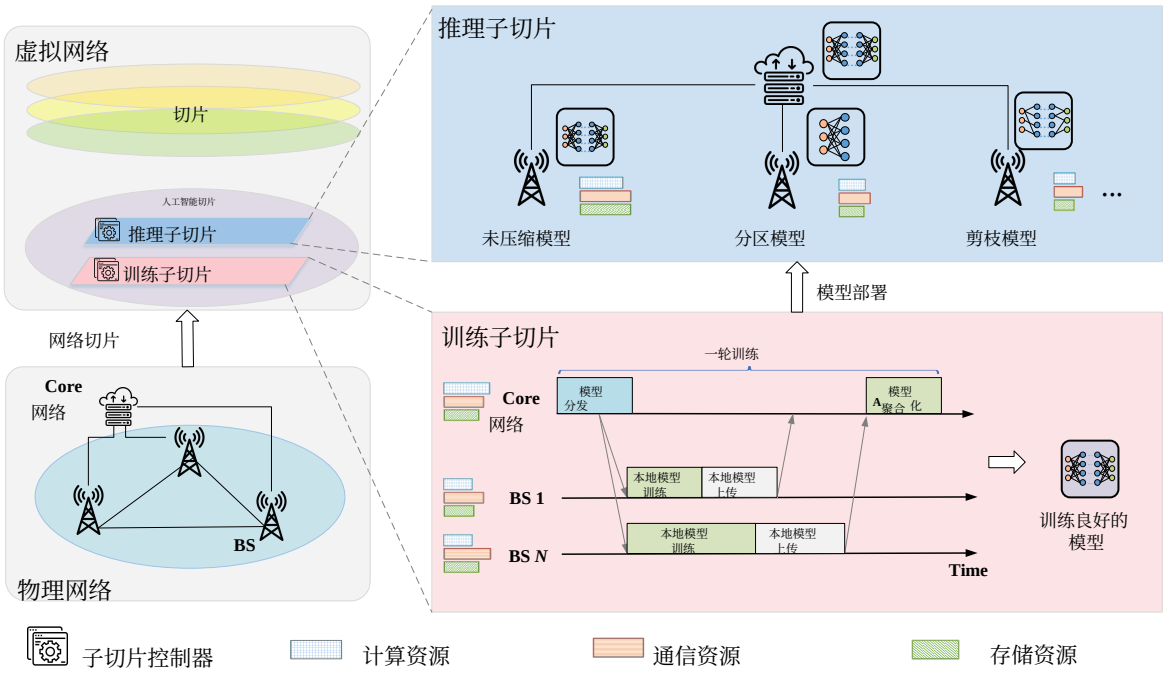


图 5：由训练子切片和推理子切片组成的人工智能切片概念图。

- 在训练子切片中，子切片控制器根据网络中的训练数据分布，确定训练配置（例如数据收集方案和模型训练方法），并为网络节点调度资源，以在给定目标准确率的情况下训练模型。此外，由于动态网络中的训练数据会随时间变化，人工智能模型可能需要不时重新训练。请注意，为训练子切片分配专用资源可以有效缓解大规模网络中分布式学习所面临的掉队者效应，从而加快模型训练过程；

- 在推理子切片中，子切片控制器会分析各个基站的服务需求模式，并确定推理配置（例如模型推理和输入数据压缩方案），以满足推理时延要求。例如，可在资源充足的基站上部署未压缩模型，在资源受限的基站上部署分区模型和剪枝模型。这样可以使不同基站的推理服务时延表现相近。

总体而言，这两个逻辑隔离的子切片侧重于满足不同的服务质量要求，并共同支持人工智能服务。

为阐释人工智能切片的概念，我们以车联网中的实时视频分析为例 [199]。在交叉路口部署智能摄像头，以提供车辆车牌识别等视频监控服务。在此类服务中，利用智能摄像头采集的视频流训练 CNN 模型，然后使用训练良好的模型执行视频分析任务。采用所提出的人工智能切片框架，CNN 模型训练在训练子切片中进行，而实时视频分析则在推理子切片中进行。具体而言，在训练子切片中，可通过联邦学习框架训练 CNN 模型，以保护数据隐私。为满足模型训练需求（如训练准确率），需分配智能摄像头的相应计算资源以及网络中的频谱资源。在推理子切片中，可根据随时间变化的服务需求以及车辆移动性导致的网络状况，配置不同的用户边缘编排方案（如深度神经网络模型分割）、输入数据压缩方案（如帧率降低）和网络资源管理策略，以满足视频分析服务的推理时延要求。借助用于视频分析的人工智能切片，即使在动态网络环境中，也能满足训练准确率和推理时延要求。

E. 总结

本节回顾了一些常见的人工智能技术，探讨了人工智能在6G网络中的作用，并提出了面向6G泛在智能的四层人工智能架构。我们讨论了人工智能在无线网络中的两个方面，即面向网络的人工智能和面向人工智能的网络，分别对应于将人工智能用作网络管理的强大工具，以及优化网络以支持人工智能应用。

机器学习算法的最新进展加速了人工智能在无线网络中的部署。在第五代移动通信技术中，人工智能技术用于解决特定的网络问题；而在第六代移动通信技术中，人工智能将渗透到无线网络的各个方面，从网络管理到网络服务。因此，我们需要一种面向人工智能的架构，以明确人工智能的作用，并描述其在网络中的各项功能。

在面向网络的人工智能领域，需要选择适当的人工智能技术，以应对具有不同特征、处于不同决策时间尺度上的网络问题。此外，智能模块之间的协作对于高效、灵活地实现人工智能驱动的网络至关重要。互联人工智能旨在实现智能模块之间的协同决策，从而进行网络控制。在面向人工智能的网络领域，人工智能算法的分布式架构将位于网络边缘的人工智能模型与网络资源连接起来。面向人工智能的网络研究仍处于起步阶段，但对于支持日益扩展的人工智能服务群体至关重要。网络切片仍将是提供人工智能服务的关键手段，但应根据人工智能算法的特征以及人工智能的训练与推理阶段定制切片策略。

IV. 面向6G的潜在网络架构

本节提出一种面向6G的概念性网络架构，该架构融合了整体网络虚拟化（包括数字孪生和网络切片）与泛在网络智能（包括互联人工智能和人工智能切片）。随后，我们阐述该架构所促成的三种相互作用，即数字孪生范式与网络切片、模型驱动方法与数据驱动方法，以及虚拟化与人工智能之间的相互作用。

*A. 6G*架构相关研究

已有多项研究针对6G网络提出了侧重点各异的架构，例如，实现全球覆盖的空天地一体化网络 [8], 用于缓解小区间干扰的无蜂窝大规模多输入多输出（MIMO）架构 [223], 以及用于提供泛在计算服务的多层计算架构 [224]。为实现先进网络管理这一目标，大多数所提出的架构都强调采用人工智能技术来优化网络架构、控制和管理 [12], [183]。例如，文献 [12]将基于人工智能的数据分析功能纳入网络架构，这些功能通过挖掘历史数据来排查网络运行故障、优化网络资源并预测网络流量。国际电信联盟（ITU）为未来网络制定了高层次的基于人工智能的架构框架，其中纳入了机器学习管理与编排功能等多个新型组件，以灵活部署基于人工智能的功能 [183]。除人工智能技术外，近期一些概念性网络架构也开始采用数字孪生技术 [75], [83]。例如，一种基于数字孪生的网络架构为每个最终用户构建一个数字孪生，作为其通信助手和数据资产管理器 [83]。另一种数字孪生赋能的网络架构采用三类数字孪生，即基于边缘、基于云以及混合数字孪生，以支持不同类型的服务 [75]。

不同于现有网络架构，我们所提出的网络架构具有新颖的整体网络虚拟化（融合了网络切片和数字孪生范式）以及泛在网络智能（融合了面向网络的人工智能和面向人工智能的网络）。此外，所提出的架构融入了第二节和第三节中的设计，使其关键要素能够以多种方式相互作用，从而赋能第六代移动通信技术。接下来的小节将详细介绍所提出的架构。

B. 架构概述

整体网络架构如图 6 所示，由物理空间和网络空间组成。物理空间包括边缘侧和核心网络中的终端用户与网络基础设施。从物理网络收集描述终端用户的数据，以创建第一级数字孪生，具体介绍见第 II-C 小节；同时，为各种服务创建网络切片。随后，将这些切片进一步抽象为第二级数字孪生，并补充从第一级数字孪生汇总而来的特定于服务的信息。图 2 所示的六层虚拟化架构应用于网络切片和数字孪生，两者均位于图 6 所示的网络空间中。

人工智能贯穿整个架构，该架构同时支持面向网络的人工智能和面向人工智能的网络。首先，如图 6 的逻辑网络控制部分所示，使用人工智能管理网络切片和数字孪生。在网络管理方面，应用第 III-C 小节讨论的连接式人工智能解决方案，以启用智能模块，进而管理网络切片和数字孪生。该连接式人工智能解决方案对应于图 3 中的 AL 3 和 4。其次，该架构支持专用人工智能切片，并通过训练与推理分离来提供人工智能服务，如第 III-D 小节所述。人工智能切片提供对应于图 3 中 AL 1 和 2 的服务，而人工智能切片的管理则由智能模块执行。

借助图 6 所示的整体网络架构，我们将整体网络虚拟化与普适网络智能融合于第六代移动通信技术中。虚拟化体现在网络和终端用户两个方面，而智能则通过面向网络的人工智能和面向人工智能的网络来体现。利用数字孪生范式、网络切片以及虚拟化和人工智能的优势，所提出的架构旨在实现更高的灵活性、可扩展性、适应性和智能化。

C. 组件与子系统

在物理空间中，所提出的架构包括无线接入网和核心网络。具体而言，涉及以下组件：

- 各类接入点：该组件包括宏基站、小基站、移动接入点（如无人机）、卫星以及其他非蜂窝接入点；
- 网络控制器：该组件包括部署在接入点或网络边缘服务器上的本地控制器，以及部署在核心网络或云端服务器上的集中式控制器。每个控制器可由计算服务器及其关联的网络存储服务器组成；
- 通用计算服务器：该组件包含用于实现网络功能（如路由和防火墙）以及托管虚拟网络功能的计算服务器；

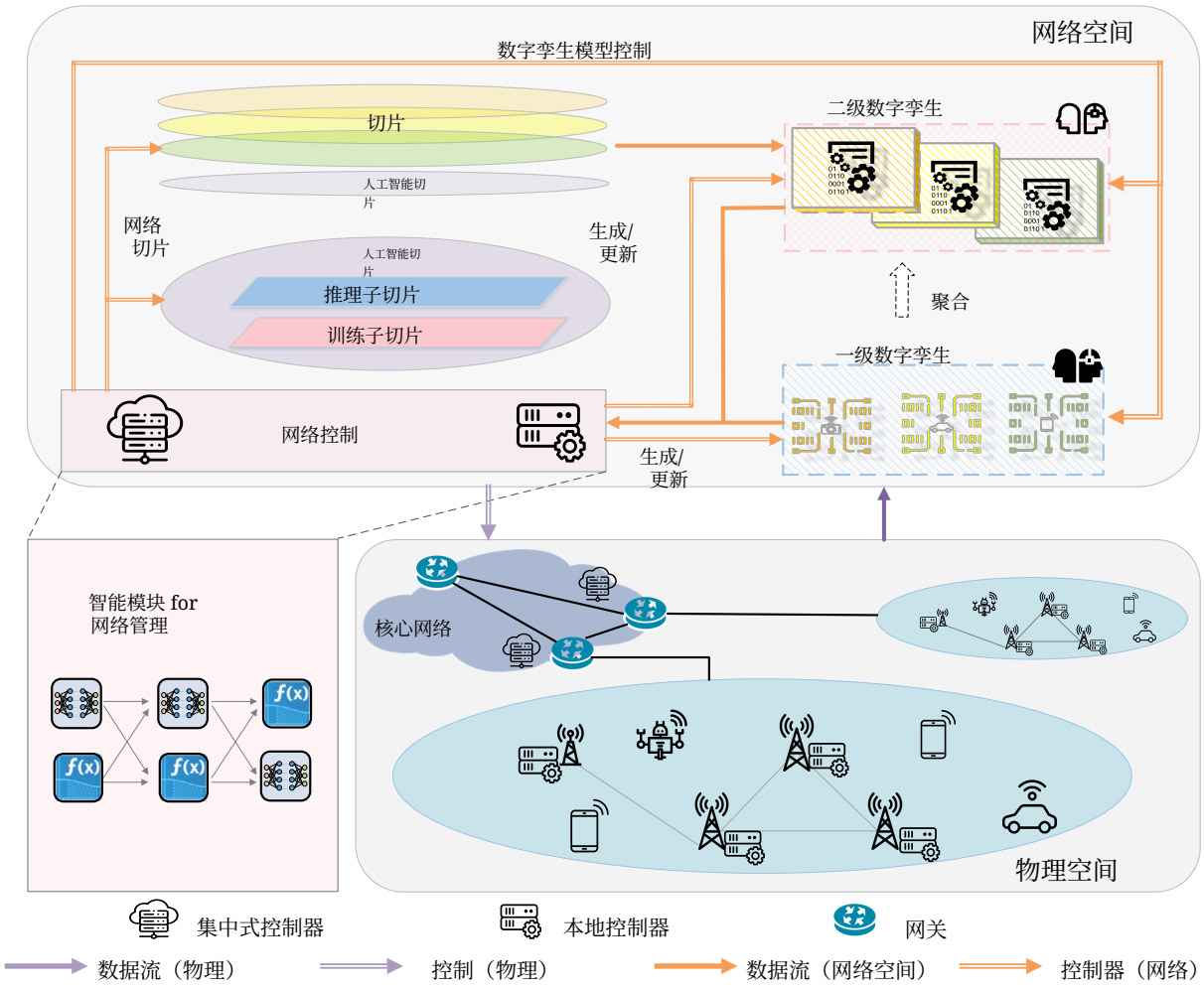


图 6：面向 6G网络的拟议网络架构。

- 应用服务器：该组件包含用于支持通用边缘计算和人工智能服务的计算服务器和网络存储服务器。这些服务器不用于网络管理或实现网络功能；
- 其他网络设备：此组件包括非通用计算服务器的专用网络硬件，例如基带处理单元和网络交换机；
- 终端用户：该组件包括人类移动用户、传感器、车辆以及各种物联网设备，例如仪表、执行器和机器人。

在网络空间中，所提出的架构包括三个子系统，即网络切片、数字孪生和互联人工智能，具体如下：

- 网络切片：该子系统包括网络切片中创建的所有虚拟网络，包括人工智能切片。一个网络切片可以涉及无线接入网、核心网络，或同时涉及两者。通用切片继承自现有网络，而人工智能切片将在第 III-D 小节中详细介绍；
- 数字孪生：该子系统包括一级和二级数字孪生。数字孪生子系统详见第 II-C 小节；
- 互联人工智能：该子系统包括部署在网络各处的智能模块，既部署于本地控制器，也部署于集中式控制器。互联人工智能子系统将在第 III-C 小节中详细介绍。

所提出架构的不同组件和子系统之间的互连将在第 IV-E 至 IV-G 小节中详细阐述，这些小节重点介绍所提出架构中数字孪生范式与网络切片、模型驱动方法与数据驱动方法，以及虚拟化与人工智能之间的相互作用。关于该架构的一些开放问题和挑战将在第五节中介绍。

请注意，尽管图 6 无法展示所有可能的网络场景，但所提出的概念架构可适用于各种类型的物理网络，例如车联网和天地一体化网络。在不同的物理网络中，整体网络虚拟化和泛在网络智能的实现方式可能有所不同，并且需要进行一定的定制化。例如，卫星网络段中智能模块的部署方式以及模块间的数据流，可能与地面网络段中的情况不同。此外，

数字孪生的迁移在车载网络中可能比在静态物联网中更为重要。相关讨论见第五节，我们将在其中介绍挑战与开放问题。尽管如此，所提出的概念架构中的基本理念，包括两级数字孪生、智能模块和人工智能切片，均适用于各种物理网络。

D. 实现

在本小节中，我们以车载网络为例，展示所提出网络架构的潜在实现方式。与边缘计算和缓存服务器共址的路侧基站，为道路上的车辆提供自动驾驶服务。为实现所提出的网络架构，需执行以下步骤。

- 网络切片建立：针对具有不同服务质量要求的自动驾驶服务建立多个网络切片，从而实现网络虚拟化。为非人工智能服务建立传统网络切片，例如高清地图下载；而对于人工智能服务，例如基于深度学习的协同感知，则建立由训练子切片和推理子切片组成的人工智能切片。这些网络切片由集中式控制器存储和管理。

- 数字孪生构建：通过从物理实体收集大量数据，为车辆用户、路侧基站和已建立的网络切片构建数字孪生，从而实现终端用户和切片的虚拟化。车辆用户和路侧基站的数字孪生位于边缘服务器，而网络切片的数字孪生位于云服务器。由于车辆移动性较高，车辆用户的数字孪生应在边缘服务器之间迁移，以确保服务连续性。除收集到的数据外，还可以通过挖掘历史数据，在数字孪生中加入生成的、针对特定用户和服务的数据，例如预测的车辆轨迹和时空服务需求。生成的车辆数据将用于网络管理和服务提供。

- 人工智能模块部署：可在集中式和本地网络控制器上部署具有不同功能的人工智能模块，以实现智能网络管理。集中式网络控制器上的人工智能模块负责网络规划。为满足不同切片的服务质量要求，这些人工智能模块可以根据路侧基站数字孪生预测的服务需求，以及从网络切片数字孪生收集的切片性能数据，做出资源预留决策。本地网络控制器上的人工智能模块负责网络运营。为提升车辆用户的感知性能，这些人工智能模块会根据从车辆用户数字孪生收集的数据（例如车辆用户的信道状况）以及生成的数据（例如预测的车辆轨迹），调度按需网络资源。

E. 数字孪生范式与网络切片之间的相互作用

作为整体网络虚拟化的两个组成部分，数字孪生范式与网络切片通过以下两个方面相互关联。

首先，面向终端用户虚拟化的数字孪生范式侧重于数据管理，而网络切片侧重于网络管理。在未来网络中，除通信、计算、缓存和感知资源之外，数据也可以被视为一种新型资源。同时，作为一种资源，数据具有其独特的特征。首先，数据可以被视为应用层资源，而非物理层资源。其次，与计算资源或通信资源不同，网络可用的数据资源量并非固定不变，而是逐步增长的。最后，数据资源的收集和处理是利用任何数据资源所必需的，同时也会消耗其他网络资源。一方面，有效利用数据资源将有益于网络管理，因此数字孪生范式能够增强网络切片。另一方面，网络管理应考虑到为利用数据资源而分配其他网络资源的需求和成本。因此，网络切片能够促进数字孪生。

其次，数字孪生将使未来网络能够实现以用户为中心的网络，而网络切片则支持以服务为中心的网络。为每项服务创建独立切片，并通过管理切片来提供服务，会使网络管理侧重于服务。与此同时，为每位终端用户创建数字副本，并管理用于描述该用户特征的数据，则能为网络管理提供以用户为中心的视角。集中式控制器通过数字孪生模型控制，选取一组信息来描述终端用户的各种特征，例如其位置、服务请求配置文件、资源利用率和信道信息，从而使网络运行阶段能够在每个切片内进行用户特定调度。例如，针对某位终端用户的接入控制和资源分配决策可能取决于其数字孪生中的数据配置文件，而不同的数据配置文件可能会导致采用不同的调度策略。因此，未来网络可能采用以服务为中心的网络规划和以用户为中心的网络运营，从而提高网络管理的粒度，以应对高度多样化的终端用户和动态网络环境。

F. 模型驱动方法与数据驱动方法的相互作用

所提出的架构促成的第二种相互作用，是网络运营和服务提供中模型驱动方法与数据驱动方法之间的相互作用。这种相互作用适用于图 6 所示的网络管理智能模块。

在第五代移动通信技术之前，网络管理主要依赖模型驱动或启发式方法。在人工智能普及之前，优化方法和博弈论等数学工具已广泛用于网络管理。优化方法以闭式形式构建目标和约束条件，并通过相应的网络

管理问题则使用优化算法求解 [225]–[227]。博弈论方法分析合作或非合作场景下网络实体之间的交互，以确定各实体的最优策略 [228]–[230]。机制设计是博弈论中的一种分析框架，也被用于协调掌握本地信息的网络实体，从而在网络效用最大化问题中实现理想的全网解决方案 [231]。

通过刻画若干关键变量之间的关系，模型驱动方法能够为网络管理问题推导出闭式解或算法。基于数学模型，模型驱动方法通常具有可解释性，并且能够很好地推广到不同的具体问题 [232].[8] 然而，当网络变得复杂（即变量数量众多和/或变量之间的相关性复杂）或高度动态（例如，网络环境变化过快，导致优化算法无法收敛或博弈无法达到均衡）时，模型驱动方法可能不再准确或适用。

自第五代移动通信技术以来，网络管理领域对数据驱动方法的研究日益深入。通过收集和利用真实世界数据，数据驱动方法隐式地刻画变量之间的关系，从而生成并微调网络管理策略。在数据充足且网络环境平稳的情况下，数据驱动方法能够为模型驱动方法难以处理的复杂问题提供近乎最优的解。然而，当网络环境非平稳，时常出现新的未知情况时，数据驱动方法的性能可能难以保证 [233]。此外，由于高度依赖从特定网络环境中收集的数据，数据驱动方法可能无法很好地泛化。

在第六代移动通信技术中，数据驱动方法与模型驱动方法应协同发挥作用。所提出的架构支持数据驱动方法与模型驱动方法相互作用，从而构建先进的混合数据-模型驱动方法。混合数据-模型驱动方法有多种选择，如图 7 所示，具体说明如下。前三种选择适用于面向网络的人工智能，最后一种则适用于面向人工智能的网络。

- 备份/切换 - 数据驱动方法和模型驱动方法可以互为备份。例如，当网络环境中出现未知情况，导致数据驱动方法的性能下降时，可以选择模型来备份数据驱动方法。同时，根据可用资源在数据驱动方法和模型驱动方法之间切换，有可能提高网络管理的适应性。

- 任务划分 - 数据驱动方法和模型驱动方法可以针对网络管理中的不同步骤，并解决不同的子问题。具体而言，数据驱动方法可以解决变量数量众多或变量间耦合关系复杂的子问题，而模型驱动方法可以解决相对独立且只有少数关键变量的子问题。这使数据驱动方法和模型驱动方法能够各展所长。

- 精炼 - 模型驱动方法可以基于通用数学模型提供粗略解答；随后，数据驱动方法以这些粗略解答为输入，并利用网络中的真实世界数据，针对特定网络场景精炼解答。通过模型生成初始解答，可能会减少数据驱动方法所需的数据量或时间。

- 混合 - 在面向人工智能的网络中，为人工智能服务部署服务功能链时，部分功能模块可以采用数据驱动方法，而同一服务功能链中的其他功能模块则可以采用模型驱动方法。例如，在基于人工智能的图像处理服务中，可以先使用模型驱动模块进行图像分辨率调整，再使用数据驱动

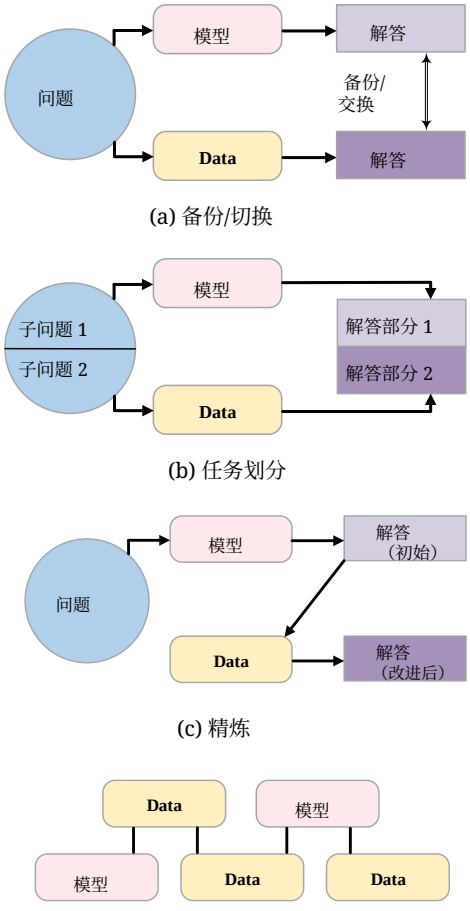


图 7：数据-模型混合驱动方法的选项。“数据”和“模型”模块分别表示“数据驱动方法”和“模型驱动方法”。

[8]例如，注水算法可应用于各种功率分配问题，而瑞利衰落模型可用于表征各种网络环境中的信道。

模块进行目标检测。这个想法与任务划分类似，只不过这里的应用场景是面向人工智能的网络，而不是面向网络的人工智能 [28]。

G. 虚拟化与人工智能的相互作用

所提出的架构促成的第三种相互作用，即虚拟化与人工智能之间的相互作用，如图8所示。

首先，虚拟化与人工智能通过数据紧密耦合。随着数字孪生的引入，关于终端用户（即一级数字孪生）和网络服务（即二级数字孪生）的大量有序数据得以提供。数字孪生中包含的数据可提供给智能模块、人工智能切片的训练或推理子切片，或同时提供给两者。例如，边缘托管人工智能可与终端用户托管人工智能协同，依据数字孪生中的数据执行针对特定用户的数据处理和预测。预测结果、资源调度方案或切片策略等结果可反馈至数字孪生，以记录终端用户的某些预测状态，例如位置和移动性。相应地，终端用户、网络基础设施和切片的数字孪生中的数据，既可作为人工智能模块的输入，也可作为其输出，从而实现虚拟化与人工智能之间的双向交互。[9]

虚拟化与人工智能之间的第二种联系体现在控制方面。基于数字孪生提供的数据，部署在边缘和核心网络中的人工智能功能可以做出网络管理和服务提供决策。这些决策可能包括网络切片控制，并将其反馈给物理网络和网络切片以供执行，同时反馈给二级数字孪生以更新数据。此外，这些决策还可能包括对一级和二级数字孪生的数字孪生模型控制。数字孪生模型控制可能包括确定数字孪生中纳入的数据类型和数量、数据收集的频率和方式，以及存储数据的格式和精度等。数字孪生模型会影响网络控制（尤其是人工智能驱动的网络控制）可用数据的可用性和质量，进而影响网络性能。因此，从网络控制的角度来看，虚拟化与人工智能之间的交互也是双向的。[10]

虚拟化与人工智能之间的第三种隐性联系
而人工智能则依赖资源。整体网络虚拟化需要大量资源，包括用于虚拟网络功能的计算资源、用于存储数字孪生的缓存资源，以及用于终端用户与其数字孪生之间同步的通信资源。同样，普适网络智能也需要大量计算资源，可能还需要其他资源，例如第三节提到的用于分布式训练的通信资源。因此，网络资源需要在

[9]有关人工智能与数据生命周期之间关系的讨论，请参阅 [234]，但其中的讨论未涉及虚拟化。[10] [235]讨论了数字孪生与人工智能在智能网络控制中的相互作用。请注意，其中对数字孪生的定义与我们的定义不同。

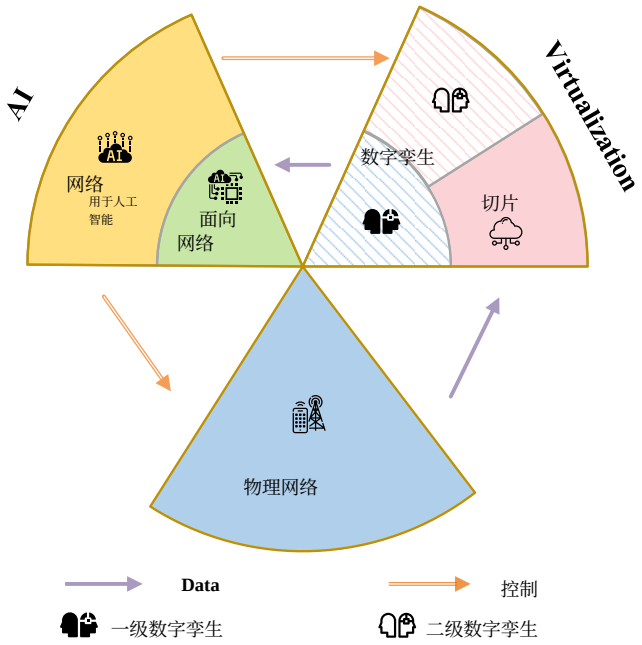


图 8：虚拟化与人工智能之间的相互作用。

虚拟化功能与人工智能功能之间共享并协调。然而，这并不意味着虚拟化功能与人工智能功能只会争夺资源。相反，它们可以相互协助，提高资源利用效率。数字孪生范式可以通过仅提供重要性较高的数据，降低人工智能功能的资源消耗。这可以通过前文提到的数字孪生模型控制来实现。同时，为每位终端用户创建数字孪生，在不久的将来可能会给网络带来过重的资源负担。利用人工智能挑选有代表性的终端用户来生成数字孪生，并优化数字孪生模型，可以减少维护和更新数字孪生所需的资源消耗。一种可能的实现方式是利用人工智能对终端用户进行分类，并从每个类别中选取部分用户来创建数字孪生。或者，由于为某些终端用户提供体验质量保证可能比为其他用户更具挑战性，利用人工智能挑选此类终端用户来创建数字孪生，有望降低6G网络在数字孪生方面的资源消耗。

H. 第六代移动通信技术的潜在网络架构：总结

本节提出了一种第六代移动通信技术的潜在网络架构，融合了两个关键要素，即泛在网络智能和整体网络虚拟化。在所提出的网络架构中，我们讨论了详细的网络组件、子系统及潜在实现。此外，本节还介绍了架构中的三种相互作用，以描述所提出的网络架构。

所提出的网络架构在实现先进的网络管理方案和支持6G网络中的人工智能服务方面具有巨大潜力。首先，融合数字孪生与网络切片有助于构建以用户为中心的网络，并提高网络管理的粒度。

其次，融合数据驱动方法与模型驱动方法，可形成新颖的数据-模型混合驱动方法，有望在适应性、粒度等方面优于现有的网络管理方法。第三，在人工智能服务中运用网络切片理念，有助于实现面向服务质量性能保证的人工智能服务。

V. 挑战与开放问题

第六代移动通信技术中的整体网络虚拟化和普遍网络智能仍有许多挑战与开放问题亟待解决。下面介绍一些关键的挑战与开放问题。

A. 数字孪生

第 II-C 小节中的六层架构为将数字孪生范式融入网络虚拟化提供了高层次设计。该架构在实际实现中仍需研究的开放问题包括数字孪生的定量性能表征、最优数字孪生模型、数字孪生迁移以及数据安全。

首先，有必要从服务质量/体验质量满意度或资源利用率的角度，对引入数字孪生所带来的网络性能提升进行定量表征。其次，集中式控制器配置的一级数字孪生模型可能因不同边缘网络而异，以适应网络异构性，因此如何确定有效的数字孪生模型是一项挑战。第三，车辆等终端用户的移动性使得在不同边缘网络之间更新和迁移数字孪生成为必要，这需要进一步研究。最后，在数字孪生范式中确保用户数据的安全也是一项挑战。由于本地和集中式网络控制器可以访问大量用户数据，因此必须制定适当的数据收集、聚合和迁移安全机制。有关上述部分挑战（例如数字孪生的异质性与迁移）以及与第六代移动通信技术中的数字孪生相关的更多开放问题，读者可参阅 [32], [236]–[238] 。

B. 面向网络管理的数据抽象与处理

虽然数字孪生能够提供数据以支持网络中的人工智能应用，包括自动化网络切片和人工智能赋能的网络控制，但高效的数据管理可能颇具挑战。首先，有必要开发数据抽象方法，以聚合具有不同粒度的数据，从而作出不同的网络管理决策。例如，在网络切片中，确定最优网络运行策略需要高粒度数据，而确定最优网络规划策略时，低粒度数据即可满足需求 [11], [239]。如何为不同的网络管理决策确定适当的数据粒度，是一个开放问题。一种潜在的解决方案是根据经验调整数据粒度和决策时间尺度 [240]。与此同时，随着变量数量以及网络管理中的数据类型可能非常庞大，因此需要更具可扩展性和更高效的解决方案。其次，在应用连接式人工智能解决方案进行网络管理时，应对智能模块的设置进行配置，例如算法选择、输入和输出属性以及智能模块之间的连接，以最大限度地利用数据，同时降低通信和处理开销；然而，确定最优设置颇具挑战。智能模块中的模型驱动方法与数据驱动方法相协作，可能是应对这一挑战的一种潜在途径，但如何支持不同类型智能模块之间的此类协作，仍需进一步研究。第三，由于数据可能在不同的网络利益相关方处生成、传输和处理，可配置且符合监管要求的数据管理同样是一项挑战。整合区块链和隐私增强技术可能是一种潜在解决方案，但隐私保护与处理效率之间的权衡仍需深入研究。关于上述挑战的讨论，例如隐私保护、人工智能模型选择、智能模块，以及有关数据抽象和处理的更多开放问题，读者可参阅 [4], [131], [182], [234]。

C. 模型与资源编排

第 III-D 小节中的面向人工智能的网络能够促进网络中的人工智能服务。一个关键问题是优化人工智能服务性能，这需要审慎配置网络，包括人工智能算法选择、数据收集和网络资源分配。主要挑战在于对人工智能性能与这些网络配置之间的关系进行建模。要建立准确的数学模型或经验模型，需要在真实网络中进行大量测量。即使建立模型可行，该模型也可能仅适用于选定的人工智能算法。此外，为适应网络动态（例如，服务需求快速波动），理想情况下应采用在线网络配置方案。由于强化学习算法能够在动态环境中进行在线决策，因此，针对高维网络配置问题开发高性价比的强化学习算法可能是一种有前景的方法。例如，有研究开发了一种强化学习算法，用于工业物联网中的人工智能模型选择与资源分配 [202]。关于上述挑战的更多讨论，感兴趣的读者可参阅 [175],[241], [242]。

D. 训练与推理协调

第 III-D 小节提出了人工智能切片的概念，以满足人工智能服务的特定服务质量要求。人工智能服务的训练与推理阶段会消耗多维网络资源 [131], [243]。在人工智能切片中，两个子切片共享虚拟化网络资源池，因此，两个子切片的资源预留决策密切相关。一方面，为训练子切片预留充足的资源可能有助于实现较高的训练准确率，但也可能导致推理子切片的资源不足，进而造成较长的服务

时延。另一方面，为训练子切片配置的资源不足可能产生低准确率的模型，从而为推理准确率造成瓶颈。为优化人工智能服务的性能，应协调训练子切片与推理子切片的资源预留。由于需要考虑大量系统因素，因此难以建立准确的数学模型来刻画训练与推理阶段之间的相互作用。因此，有必要研究高效的无模型方法来刻画这种相互作用。

E. 人工智能能效

神经网络层多达数百层、神经元多达数千个、参数多达数百万个，最先进的人工智能模型通常会消耗大量能源，并造成显著的环境成本。[11] 提高能源效率已成为广泛部署人工智能服务的一大难题。此外，近期研究表明，提高人工智能模型的准确率可能会导致计算量、环境成本和经济成本呈指数级增长 [245]。[12]因此，在网络中部署节能的人工智能服务，对于降低网络运营商的成本并满足环境标准至关重要。多种模型压缩技术，例如权重剪枝 [218], 、参数量化 [246], 和模型压缩 [247], ，都可用于缓解这一问题。此外，数据-模型混合驱动方法可以用更少的数据训练人工智能模型，从而降低能耗。

F. 数据-模型混合驱动方法

IV-F小节列出的四种选项为数据-模型混合驱动方法提供了初步构想。需要研究的相关开放问题包括以下几点。首先，有必要研究如何确定采用哪种选项，以及如何在不同选项之间切换。设计用于选择和切换选项的机制，将使网络能够灵活、自适应地集成数据驱动方法和模型驱动方法。其次，对于选定的选项，了解数据驱动和模型驱动组件对整体性能的影响程度，以及它们对彼此的影响程度十分重要。例如，在混合选项中，人工智能服务性能可能取决于数据驱动方法与模型驱动方法的组合选择，而找到合适的组合可能是一项挑战。第三，除了文中介绍的四种选项之外，数据-模型混合驱动方法还可能存在其他选项；识别其他有前景的选项是一项极其重要的开放问题。最后，由于现有数据驱动方法缺乏可解释性，应针对关键网络运营的管理开展审慎研究与分析。

[11]据估计，训练一个最先进的自然语言处理模型所产生的碳足迹，约为一辆普通汽车整个生命周期碳排放量的五倍 [244]。[12]据估计，要将ImageNet数据集上的分类错误概率从 11.5% 降至 5%，计算量需从10^{14}增加到 10^{19}十亿次浮点运算，碳排放需从 10^{6} 增加到 10^{10} 磅，经济成本需从 10^{6} 增加到 10^{11}美元 [245], 。

数据-模型混合驱动方法在增强系统鲁棒性方面所发挥的作用是一项值得进一步研究的开放问题。关于网络数据-模型混合驱动方法所面临挑战的更多讨论，感兴趣的读者可参阅 [232], [248], [249]。[13]

六、结论

为未来网络设计架构颇具挑战，尤其是在用例和定义性技术尚未浮现之时。然而，前几代网络的发展表明，网络必须支持日益异构的网络、多样化的服务以及严格的服务质量/体验质量要求。这推动了虚拟化的发展，也引发了人们对人工智能驱动的网络的浓厚兴趣。鉴于现有虚拟化和人工智能的范围与水平不足以满足未来第六代移动通信技术网络的需求，我们提出了一种融合整体网络虚拟化与泛在网络智能的概念架构设计。为完善并夯实整体网络架构，我们提出了若干具体设计，包括基于数字孪生的六层整体网络虚拟化、用于网络管理的连接式人工智能解决方案，以及人工智能切片和数据-模型混合驱动方法等构想。因此，所提出的网络架构有望凭借整体网络虚拟化实现前所未有的可扩展性和灵活性，并凭借泛在网络智能实现更强的适应性与智能化。最后，我们指出了与所提出架构相关的一些挑战与开放问题。我们希望这项研究能推动围绕第六代移动通信技术网络架构开展进一步讨论与发展。

致谢

作者谨此感谢刘东晓博士就数据隐私和安全性相关的未解决问题进行有益讨论。

参考文献

[1] X. You 等，“迈向第六代移动通信网络：愿景、使能技术与新范式变革”，《中国科学：信息科学》，第64卷第1期，第1–74页，2021年11月。[2] W. Saad、M. Bennis和M. Chen，“第六代移动通信系统展望：应用、趋势、技术与开放研究问题”，*IEEE Network*，第34卷第3期，第134–142页，2020年5/6月。[3] M. Giordani、M. Polese、M. Mezzavilla、S. Rangan和M. Zorzi，“迈向6G网络：用例与技术”，*IEEE Commun. Mag.*，第58卷第3期，第55–61页，2020年3月。[4] X. Shen、C. Huang、D. Liu、L. Xue、W. Zhuang、R. Sun和B. Ying，“未来无线网络的数据管理：架构、隐私保护与监管”，*IEEE Network*，第35卷第1期，第8–15页，2021年3/4月。[5] A. Sodhro、S. Pirbhulal、L. Zongwei、K. Muhammad和N. Zahid，“面向物联网赋能的智能自动化系统中高能效通信的第六代移动通信架构”，*IEEE*物联网期刊，第8卷第7期，第5141–5148页，2021年4月。[6] Z. Zhang、Y. Xiao、Z. Ma、M. Xiao、Z. Ding、X. Lei、G. Karagiannidis和P. Fan，“第六代移动通信网络：愿景、需求、架构与关键技术”，*IEEE Veh. Technol. Mag.*，第14卷第3期，第28–41页，2019年9月。

[13]有关在车载网络仿真中应用混合方法的案例，请参见 [250]。

[7] B. Zong、C. Fan、X. Wang、X. Duan、B. Wang和J. Wang，“第六代移动通信技术：关键驱动因素、核心需求、系统架构与使能技术”，IEEE Veh. Technol. Mag.，第14卷第3期，第18–27页，2019年9月。[8] N. Zhang、S. Zhang、P. Yang、O. Alhussein、W. Zhuang和X. Shen，“软件定义的空天地一体化车联网：挑战与解答”，IEEE Commun. Mag.，第55卷第7期，第101–109页，2017年7月。[9] S. Chen、S. Sun和S. Kang，“地面移动通信与卫星通信的系统集成——B5G和第六代移动通信技术中的趋势、挑战与关键技术”，《中国通信》，第17卷第12期，第156–171页，2020年12月。[10] E. Calvanese Strinati、S. Barbarossa、J. Gonzalez-Jimenez、D. Ktenas、N. Cassiau、L. Maret和C. Dehos，“第六代移动通信：下一前沿——从全息消息传递到利用亚太赫兹和可见光通信实现人工智能”，IEEE Veh. Technol. Mag.，第14卷第3期，第42–50页，2019年9月。[11] X. Shen、J. Gao、W. Wu、K. Lyu、M. Li、W. Zhuang、X. Li和J. Rao，“基于人工智能辅助网络切片的下一代无线网络”，IEEE Open J. Veh. Technol.，第1卷，第45–66页，2020年1月。[12] K. Letaief、W. Chen、Y. Shi、J. Zhang和Y. Zhang，“迈向第六代移动通信技术的路线图：人工智能赋能的无线网络”，IEEE Commun. Mag.，第57卷第8期，第84–90页，2019年8月。[13] H. Yang、A. Alphones、Z. Xiong、D. Niyato、J. Zhao和K. Wu，“人工智能赋能的智能第六代移动通信网络”，IEEE Network，第34卷第6期，第272–280页，2020年11月。[14] M. El-Sayed和J. Jaffe，“电信网络演进概述”，IEEE Commun. Mag.，第40卷第12期，第74–81页，2002年12月。[15] B. Bjerke，“LTE-Advanced与LTE部署的演进”，《IEEE无线通信》，第18卷第5期，第4–5页，2011年10月。[16] D. Kreutz、F. Ramos、P. Ver´ıssimo、C. Rothenberg、S. Azodolmolky和S. Uhlig，“软件定义网络：全面综述”，Proc. IEEE，第103卷第1期，第14–76页，2015年1月。[17] A. Checko、H. Christiansen、Y. Yan、L. Scolari、G. Kardaras、M. Berger和L. Dittmann，“面向移动网络的云无线接入网——技术概述”，IEEE Commun. Surveys Tuts.，第17卷，426，4 1，405 2015 no. pp. – th Quart. .[18] M. Bagaa、T. Taleb、A. Laghrissi、A. Ksentini和H. Flinck，“用于创建高效虚拟核心网络切片的联盟博弈”，IEEE J. Sel. Areas Commun.，第36卷第3期，第469–484页，2018年3月。[19] W. Wu、N. Chen、C. Zhou、M. Li、X. Shen、W. Zhuang和X. Li，“通过约束学习为面向服务的车联网实现动态无线接入网切片”，IEEE J. Sel. Areas Commun.，第39卷第7期，第2076–2089页，2021年7月。[20] K. Qu、W. Zhuang、Q. Ye、X. Shen、X. Li和J. Rao，“采用SDN-NFV集成的面向服务的5G网络流量工程”，IEEE Network，第34卷第4期，第234–241页，2020年7月。[21] I. Afolabi、T. Taleb、K. Samdanis、A. Ksentini和H. Flinck，“网络切片与软件化：原理、使能技术和解答综述”，IEEE Commun. Surveys Tuts.，第20卷第3期，第2429–2453页，2018年第3季度。[22] S. Ali等，“无线通信网络中机器学习的第六代移动通信白皮书”，arXiv:2004.*13875*，2020，[在线]。可访问：http://arxiv.org/abs/2004.13875。[23] H. Zhang，“未来无线网络：MyNET平台与端到端网络切片”，arXiv:1611.*07601*，2016，[在线]。可访问：https://arxiv.org/abs/1611.07601。[24] Z. Fadlullah、F. Tang、B. Mao、N. Kato、O. Akashi、T. Inoue和K. Mizutani，“最先进的深度学习：推动机器智能迈向未来的智能网络流量控制系统”，IEEE Commun. Surveys Tuts.，第19卷第4期，第2432–2455页，2017年第4季度。[25] A. Toma、A. Krayani、M. Farrukh、H. Qi、L. Marcenaro、Y. Gao和C. Regazzoni，“通过学习生成模型，在认知无线电物理层进行基于人工智能的异常检测”，IEEE认知通信与网络汇刊，第6卷第1期，第21–34页，2020年3月。[26] S. Han、T. Xie、C. I、L. Chai、Z. Liu、Y. Yuan和C. Cui，“人工智能赋能的第六代移动通信空中接口：解答、挑战与标准化影响”，IEEE Commun. Mag.，第58卷第10期，第73–79页，2020年10月。[27] M. Li、J. Gao、L. Zhao和X. Shen，“边缘辅助自动驾驶的自适应计算调度”，IEEE车辆技术汇刊，第70卷第6期，第5318–5331页，2021年6月。[28] M. Li、J. Gao、C. Zhou、W. Zhuang 和 X. Shen，《基于切片的网络边缘人工智能服务供给》，IEEE Veh. Technol. Mag.，即将发表，doi:10.1109/MVT.2021.3114655。[29] P. Chemouil、P. Hui、W. Kellerer、N. Limam、R. Stadler 和 Y. Wen，《客座编者按：网络人工智能与机器学习进展专刊》，IEEE J. Sel. Areas Commun.，第38卷，第10期，第2229–2233页，2020年10月。[30] 3GPP，《核心网络和终端技术规范组；5G系统；网络数据分析服务；第3阶段（Release 17）》，技术报告3GPP TS29.520 V0.0.0，2017年10月。[31] ——，《服务与系统方面技术规范组；Release 16说明；Rel-16工作项目摘要（Release 16）》，技术报告3GPP TR21.916 V1.0.0，2020年12月。[32] R. Minerva、G. M. Lee 和 N. Crespi，《物联网环境中的数字孪生：技术特征、场景和架构模型综述》，Proc. IEEE，第108卷，第10期，第1785–1824页，2020年10月。[33] R. Boutaba、M. A. Salahuddin、N. Limam、S. Ayoubi、N. Shahriar、F. Estrada-Solano 和 O. M. Caicedo，《网络机器学习全面综述：演进、应用与研究机会》，J. Internet Serv. Appl.，第9卷，第1期，第1–99页，2018年6月。[34] C. Zhang、P. Patras 和 H. Haddadi，《移动与无线网络中的深度学习：综述》，IEEE Commun. Surveys Tuts.，第21卷，第3期，第2224–2287页，2019年第3季度。[35] N. Chowdhury 和 R. Boutaba，《网络虚拟化：现状与研究挑战》，IEEE Commun. Mag.，第47卷，第7期，第20–26页，2009年7月。[36] G. Rossi 和 C. Garavaglia，《改进局域网网络层的提案》，ACM SIGCOMM Computer Communication Review，第16卷，第1期，第1–5页，1986年2月。[37] K. Sato、S. Ohta 和 I. Tokizawa，《基于虚拟路径的宽带ATM网络架构》，《通信汇刊》，第38卷，第8期，第1212–1222页，1990年8月。[38] K. Qu、W. Zhuang、Q. Ye、X. Shen、X. Li 和 J. Rao，《SDN/NFV赋能的5G核心网络中嵌入式服务的动态流迁移》，《通信汇刊》，第68卷，第4期，第2394–2408页，2020年4月。[39] M. Chiosi 等，《网络功能虚拟化：简介、优势、推动因素、挑战及行动倡议》，载于 SDN and OpenFlow World Congress 会议论文集，德国达姆施塔特，2012年10月。[40] A. Dalla-Costa、L. Bondan、J. Wickboldt、C. Both 和 L. Granville，《Orchestra：适用于动态云无线接入网的可定制、具备切分感知能力的网络功能虚拟化编排器》，IEEE J. Sel. Areas Commun.，第38卷，第6期，第1014–1024页，2020年6月。[41] G. Zhang、J. Shu、W. Xue 和 W. Zheng，《大型存储区域网络带外虚拟化系统的设计与实现》，IEEE Trans. Comput.，第56卷，第12期，第1654–1665页，2007年12月。[42] F. Xu、F. Liu、H. Jin 和 A. Vasilakos，《云计算中虚拟机性能开销管理：综述、研究现状与未来方向》，Proc. IEEE，第102卷，第1期，第11–31页，2014年1月。[43] P. Bellavista、A. Corradi、L. Foschini、S. Luciano 和 M. Solimando，《云数据中心网络中虚拟化资源的仿真框架》，IEEE J. Sel. Areas Commun.，第37卷，第8期，第1808–1819页，2019年8月。[44] X. Yuan、M. Sun 和 W. Lou，《移动环境下边缘云系统的动态深度学习虚拟边缘节点放置方案》，IEEE Trans. Cloud Comput.，即将发表，doi: 10.1109/TCC.2020.2974948。[45] S. Yang、P. Wieder、R. Yahyapour、S. Trajanovski 和 X. Fu，《云中的可靠虚拟机部署与路由》，IEEE Trans. Parallel Distrib. Syst.，第28卷，第10期，第2965–2978页，2017年10月。[46] M. Nagy、J. Tapolcai 和 G. R´etv´ari，《面向互联网协议层弹性的节点虚拟化》，IEEE/ACM Trans. Netw.，第26卷，第3期，第1250–1263页，2018年6月。[47] I. Khan、F. Belqasmi、R. Glitho、N. Crespi、M. Morrow 和 P. Polakos，《无线传感器网络虚拟化：早期架构与研究视角》，IEEE Network，第29卷，第3期，第104–112页，2015年5月。[48] S. Zaidi、O. Ben Smida、S. Affes、U. Vilaipornsawai、L. Zhang 和 P. Zhu，《面向未来5G网络的以用户为中心的基站无线接入虚拟化》，《通信汇刊》，第67卷，第7期，第5190–5202页，2019年7月。[49] H. Du、W. Wu、Q. Ye、D. Li、W. Lee 和 X. Xu，《无线传感器网络中具备路由成本保证的基于连通支配集的虚拟骨干网构建》，IEEE Trans. Parallel Distrib. Syst.，第24卷，第4期，第652–661

，2013年4月。

[50] F. Hosseini、A. James 和 M. Ghaderi，“需求不确定条件下的概率虚拟链路嵌入”，IEEE Trans. Netw. Service Manag.，第16卷，第4期，1552–1566页，2019年12月。[51] S. Tomovic 和 I. Radusinovic，“面向基于 SDN 的 ISP 网络的可扩展、稳健且满足服务质量要求的虚拟链路配置”，IEEE Trans. Netw. Service Manag.，第16卷，第3期，1032–1045页，2019年9月。[52] C. Papagianni、A. Leivadeas、S. Papavassiliou、V. Maglaris、C. Cervello-Pastor 和 A. Monje，“云计算网络中虚拟资源的最优分配”，IEEE Trans. Comput.，第62卷，第6期，1060–1071页，2013年6月。[53] T. Wood、K. Ramakrishnan、P. Shenoy、J. Van der Merwe、J. Hwang、G. Liu 和 L. Chaufournier，“Cloudnet：通过虚拟机 WAN 实时迁移动态汇集云资源”，IEEE/ACM Trans. Netw.，第23卷，第5期，1568–1583页，2015年10月。[54] M. Kalil、A. Moubayed、A. Shami 和 A. Al-Dweik，“无线资源虚拟化的高效低复杂度调度器”，《IEEE无线通信》快报，第5卷，第1期，56–59页，2016年2月。[55] X. Lu、Q. Ni、D. Zhao、W. Cheng 和 H. Zhang，“面向上行 VMIMO-SC-FDMA 系统定制化时延受限 QoS 保障的资源虚拟化”，《通信汇刊》，第67卷，第4期，2951–2967页，2019年4月。[56] S. Zhang、H. Luo、J. Li、W. Shi 和 X. Shen，“通过分层软切片满足缓存赋能车联网的多维 QoS 需求”，IEEE无线通信汇刊，第19卷，第3期，2150–2162页，2020年3月。[57] O. Alhussein、P. Do、Q. Ye、J. Li、W. Shi、W. Zhuang、X. Shen、X. Li 和 J. Rao，“面向 NFV 赋能的核心网络中组播服务的虚拟网络定制框架”，IEEE J. Sel. Areas Commun.，第38卷，第6期，1025–1039页，2020年6月。[58] N. Zhang、Y. Liu、H. Farmanbar、T. Chang、M. Hong 和 Z. Luo，“资源约束下的面向服务网络切片”，IEEE J. Sel. Areas Commun.，第35卷，第11期，2512–2521页，2017年11月。[59] J. Tang、B. Shim 和 T. Q. S. Quek，“融合超可靠低时延通信与组播 eMBB 的切片化云无线接入网中的业务复用与收益最大化”，IEEE J. Sel. Areas Commun.，第37卷，第4期，881–895页，2019年4月。[60] Q. Ye、W. Zhuang、S. Zhang、A. Jin、X. Shen 和 X. Li，“双层异构无线网络的动态无线资源切片”，IEEE车辆技术汇刊，第67卷，第10期，9896–9910页，2018年10月。[61] T. Guo 和 A. Su´arez，“通过 EDF 切片调度实现 5G RAN 切片”，IEEE车辆技术汇刊，第68卷，第3期，2865–2877页，2019年3月。[62] J. Ni、X. Lin 和 X. Shen，“支持 5G 赋能的物联网网络切片的高效安全面向服务认证”，IEEE J. Sel. Areas Commun.，第36卷，第3期，644–657页，2018年3月。[63] M. Kessler、A. Reifert、D. Lamp 和 T. Voith，“用于提供虚拟化网络的面向服务基础设施”，贝尔实验室技术期刊，第13卷，第3期，111–127页，2008年秋。[64] J. van de Belt、H. Ahmadi 和 L. Doyle，“无线链路虚拟化与无线网络虚拟化的定义和综述”，IEEE Commun. Surveys Tuts.，第19卷，第3期，1603–1627页，2017年第3季度。[65] N. McKeown、T. Anderson、H. Balakrishnan、G. Parulkar、L. Peterson、J. Rexford、S. Shenker 和 J. Turner，“OpenFlow：为校园网络创新赋能”，ACM SIGCOMM Comput. Commun. Review，第38卷，第2期，69–74页，2008年4月。[66] A. Gudipati、D. Perry、L. E. Li 和 S. Katti，“SoftRAN：软件定义无线接入网”，发表于 ACM SIGCOMM HotSDN 研讨会论文集，中国香港，2013年8月。[67] X. Foukas、N. Nikaein、M. M. Kassem、M. K. Marina 和 K. Konto-vasilis，“FlexRAN：面向软件定义无线接入网的灵活可编程平台”，发表于 ACM CoNEXT，尔湾，加利福尼亚州，美国，2016年12月，427–441页。[68] O-RAN 联盟， 2021， [在线]。网址：https://www.o-ran.org/software。[69] J. Breen 等，“POWDER：开放式无线数据驱动实验研究平台”，《计算机网络》，第197卷，文章编号108281，2021年10月。[70] D. Johnson、D. Maas 和 J. Van Der Merwe，“POWDER 上的开源 RAN 切片：自上而下的 O-RAN 用例”，发表于 ACM MobiSys，虚拟会议，2021年6月。[71] X. Foukas 和 B. Radunovic，“Concordia：教会 5G 虚拟化无线接入网共享计算资源”，发表于 ACM SIGCOMM，虚拟会议，2021年8月。[72] A. Conte、S. Kerboeuf 和 L. Thomas，“网络托管化身：网络中的用户终端虚拟化”，贝尔实验室技术期刊，第13卷，第2期，117–126页，2008年夏。

[73] M. Nitti、V. Pilloni、G. Colistra 和 L. Atzori，“作为物联网核心要素的虚拟对象：综述”，IEEE Commun. Surveys Tuts.，第18卷，第2期，第1228–1240页，2016年第二季度。[74] M. Grieves，“数字孪生：通过虚拟工厂复制实现卓越制造”，白皮书，第1卷，第1–7页，2015年3月。[75] L. U. Khan、W. Saad、D. Niyato、Z. Han 和 C. S. Hong，“数字孪生赋能的第六代移动通信技术：愿景、架构趋势与未来方向”，arXiv:2102.*12169*，2021，[在线]。可访问：https://arxiv.org/abs/2102.12169。[76] E. Glaessgen 和 D. Stargel，“面向未来美国国家航空航天局和美国空军车辆的数字孪生范式”，载于《AIAA结构、结构动力学与材料会议论文集》，美国夏威夷州檀香山，2012年4月。[77] Z. Gao、A. Paul 和 X. Wang，“客座编者按：数字孪生：整合人工智能与机器学习及大数据分析，实现虚拟表示”，IEEE工业信息学汇刊，第18卷，第2期，第1355–1358页，2022年2月。[78] A. M. Madni、C. C. Madni 和 S. D. Lucero，“在基于模型的系统工程中运用数字孪生技术”，Systems，第7卷，第1期，第7页，2019年1月。[79] P. Jia、X. Wang 和 X. Shen，“数字孪生赋能工业物联网系统中的智能分布式时钟同步”，IEEE物联网期刊，第8卷，第6期，第4548–4559页，2021年3月。[80] Y. Dai、K. Zhang、S. Maharjan 和 Y. Zhang，“数字孪生网络中的随机计算卸载深度强化学习”，IEEE工业信息学汇刊，第17卷，第7期，第4968–4977页，2021年7月。[81] L. Zhao、G. Han、Z. Li 和 L. Shu，“基于智能数字孪生的软件定义车联网”，IEEE Network，第34卷，第5期，第178–184页，2020年10月。[82] J. Taylor 和 H. Sharif，“利用数字孪生提升弱势网络中物联网的性能”，载于IEEE IWCMC论文集，塞浦路斯利马索尔，2020年6月。[83] Q. Yu、J. Ren、Y. Fu、Y. Li 和 W. Zhang，“网络孪生：下一代网络架构的起源”，《IEEE无线通信》，第26卷，第6期，第111–117页，2019年12月。[84] A. Barbie、N. Pech、W. Hasselbring、S. Flogel、F. Wenzhofer、M. Walter、E. Shchekinova、M. Busse、M. Turk、M. Hofbauer 和 S. Sommer，“利用数字孪生原型开发海洋观测系统水下网络——波罗的海实地报告”，IEEE Internet Comput.，即将发表，doi:10.1109/MIC.2021.3065245。[85] X. Xu、B. Shen、S. Ding、G. Srivastava、M. Bilal、M. Khosravi、V. Menon、M. Jan 和 W. Maoli，“边缘计算中，利用深度Q网络为数字孪生赋能的车联网进行服务卸载”，IEEE工业信息学汇刊，第18卷，第2期，第1414–1423页，2022年2月。[86] H. Wang、Y. Wu、G. Min 和 W. Miao，“用于网络切片管理的图神经网络数字孪生”，IEEE工业信息学汇刊，第18卷，第2期，第1367–1376页，2022年2月。[87] T. Wang、J. Cheng、Y. Yang、C. Esposito、H. Snoussi 和 F. Tao，“通过范围检测控制优化数字孪生输送机系统的自适应优化方法”，IEEE Trans. Autom. Sci. Eng.，即将发表，doi:10.1109/TASE.2020.3043393。[88] O. E. Marai、T. Taleb 和 J. Song，“ 道路基础设施数字孪生：迈向智慧城市实现的一步 ”，IEEE Network，第35卷，第2期，第136–143页，2021年3月/4月。[89] R. Minerva、F. Awan 和 N. Crespi，“利用数字孪生实现合成感知”，IEEE Internet Comput.，即将发表，doi:10.1109/MIC.2021.3051674。[90] A. Castellani、S. Schmitt 和 S. Squartini，“利用数字孪生系统和弱监督学习进行真实场景异常检测”，IEEE工业信息学汇刊，第17卷，第7期，第4733–4742页，2021年7月。[91] H. Elayan、M. Aloqaily 和 M. Guizani，“面向智能情境感知物联网医疗系统的数字孪生”，IEEE物联网期刊，第8卷，第23期，第16 749–16 757页，2021年12月。[92] M. Schluse、M. Priggemeyer、L. Atorf 和 J. Rossmann，“可实验数字孪生——为工业4.0精简基于仿真的系统工程”，IEEE工业信息学汇刊，第14卷，第4期，第1722–1731页，2018年4月。[93] R. Dong、C. She、W. Hardjawana、Y. Li 和 B. Vucetic，“面向移动边缘计算系统中混合5G服务的深度学习：从数字孪生中学习”，IEEE无线通信汇刊，第18卷，第10期，第4692–4707页，2019年10月。[94] C. Gehrmann 和 M. Gunnarsson，“基于数字孪生的工业自动化与控制系统安全架构”，IEEE工业信息学汇刊，第16卷，第1期，第669–680页，2020年1月。[95] K. Alam 和 A. El Saddik，“C2PS：基于云的信息物理系统数字孪生架构参考模型”，IEEE Access，第5卷

，第2050–2062页，2017年1月。

[96] L. Rivera、M. Jimenez、N. Villegas、G. Tamura 和 H. Muller，《自主协作数字孪生体的构建》，IEEE Internet Comput.，第 1–10 页，待出版，doi:10.1109/MIC.2021.3051902。[97] C. Zhang、G. Zhou、H. Li 和 Y. Cao，《面向数据与知识驱动的数字孪生制造单元配置的物联网区块链制造》，IEEE物联网期刊，第 7 卷，第 12 期，第 11 884–11 894 页，2020年12月。[98] Y. Fang、C. Peng、P. Lou、Z. Zhou、J. Hu 和 J. Yan，《面向智能制造的数字孪生车间作业调度》，IEEE工业信息学汇刊，第 15 卷，第 12 期，第 6425–6435 页，2019年12月。[99] Y. Xu、Y. Sun、X. Liu 和 Y. Zheng，《基于深度迁移学习的数字孪生辅助故障诊断》，IEEE Access，第 7 卷，第 19 990–19 999 页，2019年1月。[100] P. Jain、J. Poon、J. P. Singh、C. Spanos、S. R. Sanders 和 S. K. Panda，《分布式光伏系统故障诊断的数字孪生方法》，IEEE Trans. Power Electron.，第 35 卷，第 1 期，第 940–956 页，2019年1月。[101] W. Sun、P. Wang、N. Xu、G. Wang 和 Y. Zhang，《空中辅助车联网资源分配中的动态数字孪生体与分布式激励》，IEEE物联网期刊，第 1–14 页，2021年，待出版，doi:10.1109/JIOT.2021.3058213。[102] K. Zhang、J. Cao、S. Maharjan 和 Y. Zhang，《数字孪生赋能的社交感知车载边缘网络内容缓存》，IEEE Trans. Comput. Soc. Syst.，第 1–13 页，2021年，待出版，doi:10.1109/TCSS.2021.3068369。[103] J. A. Hartigan 和 M. A. Wong，《算法 AS 136：一种K均值聚类算法》，J. Roy. Stat. Soc. C (Appl. Stat.)，第 28 卷，第 1 期，第 100–108 页，1979年。[104] M. Parwez、D. Rawat 和 M. Garuba，《移动无线网络中的用户活动分析与用户异常检测的大数据分析》，IEEE工业信息学汇刊，第 13 卷，第 4 期，第 2058–2065 页，2017年8月。[105] A. Topchy、A. Jain 和 W. Punch，《用于聚类集成的混合模型》，载于 SIAM国际数据挖掘会议论文集，美国佛罗里达州布埃纳维斯塔湖，2004年。[106] M. Shih 和 A. Hero，《利用有限混合模型对网络链路时延分布进行基于单播的推断》，IEEE Trans. Signal Process.，第 51 卷，第 8 期，第 2219–2228 页，2003年8月。[107] D. Bega、M. Gramaglia、M. Fiore、A. Banchs 和 X. Costa-Perez，《DeepCog：利用基于人工智能的容量预测优化网络切片中的资源配置》，IEEE J. Sel. Areas Commun.，第 38 卷，第 2 期，第 361–376 页，2020年2月。[108] C. Ledig、L. Theis、F. Husz´ar、J. Caballero、A. Cunningham、A. Acosta、A. Aitken、A. Tejani、J. Totz、Z. Wang 等，《利用生成对抗网络实现照片级真实感单图像超分辨率》，载于 IEEE计算机视觉与模式识别会议论文集，美国夏威夷州檀香山，2017年。[109] T. Erpek、Y. Sagduyu 和 Y. Shi，《用于发起和缓解无线干扰攻击的深度学习》，IEEE认知通信与网络汇刊，第 5 卷，第 1 期，第 2–14 页，2019年3月。[110] S. Hu、Y. Yao 和 Z. Yang，《利用支持向量机识别认知无线电网络的MAC协议》，《IEEE无线通信》，第 21 卷，第 1 期，第 52–60 页，2014年2月。[111] C. You 和 R. Zhang，《莱斯衰落环境下无人机支持的数据采集三维轨迹优化》，IEEE无线通信汇刊，第 18 卷，第 6 期，第 3192–3207 页，2019年6月。[112] S. Peng、H. Jiang、H. Wang、H. Alwageed、Y. Zhou、M. Sebdani 和 Y. Yao，《基于信号星座图和深度学习的调制分类》，IEEE Trans. Neural Netw. Learn. Syst.，第 30 卷，第 3 期，第 718–727 页，2019年3月。[113] Z. Zhao、W. Chen、X. Wu、P. Chen 和 J. Liu，《长短期记忆网络：一种短期交通流预测的深度学习方法》，IET Intell. Transp. Syst.，第 11 卷，第 2 期，第 68–75 页，2017年2月。[114] S. Gu、T. Lillicrap、I. Sutskever 和 S. Levine，《基于模型加速的连续深度Q学习》，载于 ICML会议论文集，美国纽约州纽约市，2016年。[115] J. Zhu、Y. Song、D. Jiang 和 H. Song，《一种面向认知物联网的新型基于深度Q学习的传输调度机制》，IEEE物联网期刊，第 5 卷，第 4 期，第 2375–2385 页，2018年8月。[116] Y. Sun、M. Peng 和 S. Mao，《面向绿色雾无线接入网络的基于深度强化学习的模式选择与资源管理》，IEEE物联网期刊，第 6 卷，第 2 期，第 1960–1971 页，2019年4月。[117] D. Silver、G. Lever、N. Heess、T. Degris、D. Wierstra 和 M. Ried-miller，《确定性策略梯度算法》，载于 ICML会议论文集，北京，中国，2014年。

[118] S. O. Somuyiwa、A. Gyorgy 和 D. Gunduz，“无线网络中主动缓存的强化学习方法”，IEEE J. Sel. Areas Commun.，第 36 卷，第 6 期，第 1331–1344 页，2018 年 6 月。[119] D. T. Hoang、D. Niyato、P. Wang 和 D. I. Kim，“ 具备射频能量采集能力的协作式多用户认知无线电网络性能优化”，IEEE Trans. Wireless Commun.，第 14 卷，第 7 期，第 3614–3629 页，2015 年 7 月。[120] V. R. Konda 和 J. N. Tsitsiklis，“行动者-评论家算法”，载于 NIPS 论文集，美国科罗拉多州丹佛，2000年。[121] N. Cheng、F. Lyu、W. Quan、C. Zhou、H. He、W. Shi 和 X. Shen，“面向物联网应用的空间/空中辅助计算卸载：一种基于学习的方法”，IEEE J. Sel. Areas Commun.，第 37 卷，第 5 期，第 1117–1129 页，2019 年 5 月。[122] T. P. Lillicrap、J. J. Hunt、A. Pritzel、N. Heess、T. Erez、Y. Tassa、D. Silver 和 D. Wierstra，“基于深度强化学习的连续控制”，arXiv:1509。*02971*，2015，[在线]。可访问：https://arxiv.org/abs/1509.02971。[123] M. Li、J. Gao、L. Zhao 和 X. Shen，“车联网协作式边缘计算的深度强化学习”，IEEE Trans. Cogn. Commun. Netw.，第 6 卷，第 4 期，第 1122–1135 页，2020 年 12 月。[124] J. Kone˘cn`y、H. B. McMahan、F. X. Yu、P. Richt´arik、A. T. Suresh 和 D. Bacon，“联邦学习：提升通信效率的策略”，arXiv:1610。*05492*，2016，[在线]。可访问：https://arxiv.org/abs/1610.05492。[125] Z. Yang、M. Chen、W. Saad、C. S. Hong 和 M. Shikh-Bahaei，“无线通信网络中的节能联邦学习”，IEEE Trans. Wireless Commun.，第 20 卷，第 3 期，第 1935–1949 页，2021 年 3 月。[126] C. Thapa、M. Chamikara 和 S. Camtepe，“联邦学习在隐私保护方面的进展：从联邦学习到分割学习”，arXiv:2011。*14818*，2020，[在线]。可访问：https://arxiv.org/abs/2011.14818。[127] W. Shi、J. Li、H. Wu、C. Zhou、N. Cheng 和 X. Shen，“高移动性网络中的无人机蜂窝轨迹规划与资源分配：一种分层深度强化学习方法”，IEEE Internet Things J.，第 8 卷，第 12 期，第 9800–9813 页，2021 年 6 月。[128] M. Sana、A. De Domenico、W. Yu、Y. Lostanlen 和 E. Calvanese Strinati，“动态毫米波网络中自适应用户关联的多智能体强化学习”，IEEE Trans. Wireless Commun.，第 19 卷，第 10 期，第 6520–6534 页，2020 年 10 月。[129] R. Ding、Y. Xu、F. Gao 和 X. Shen，“基于多智能体深度强化学习的空地协同通信系统轨迹设计与接入控制”，IEEE Internet Things J.，第 1–14 页，2021 年，10 1109 2021 3062091 – 即将发表，doi: . /JIOT. . .。[130] Y. Xu、H. Zhou、T. Ma、J. Zhao、B. Qian 和 X. Shen，“利用多智能体学习实现无信号交叉路口自动驾驶车辆调度”，IEEE Internet Things J.，第 8 卷，第 14 期，第 11 427–11 439 页，2021 年 7 月。[131] Z. Zhou、X. Chen、E. Li、L. Zeng、K. Luo 和 J. Zhang，“边缘智能：借助边缘计算铺就人工智能的最后一英里”，Proc. IEEE，第 107 卷，第 8 期，第 1738–1762 页，2019 年 8 月。[132] S. Samarakoon、M. Bennis、W. Saad 和 M. Latva-aho，“无线小蜂窝网络的动态聚类与开关策略”，IEEE Trans. Wireless Commun.，第 15 卷，第 3 期，第 2164–2178 页，2016 年 8 月。[133] Y. Zhu、X. Dong 和 T. Lu，“用于无线信道预测的自适应无参数循环神经结构”，IEEE Trans. Com- mun.，第 67 卷，第 11 期，第 8086–8096 页，2019 年 11 月。[134] K. Yang、C. Shen 和 T. Liu，“基于深度强化学习的无线网络优化：比较研究”，载于 IEEE INFOCOM Workshops 论文集，虚拟会议，2020年。[135] M. Alsenwi、N. Tran、M. Bennis、S. Pandey、A. Bairagi 和 C. Hong，“面向增强型移动宽带与超可靠低时延通信在第五代移动通信技术及后续网络中共存的智能资源切片：一种基于深度强化学习的方法”，IEEE Trans. Wireless Commun.，第 20 卷，第 7 期，第 4585–4600 页，2021 年 7 月。[136] Y. Mao、C. You、J. Zhang、K. Huang 和 K. B. Letaief，“移动边缘计算综述：通信视角”，IEEE Commun. Surveys Tuts.，第 19 卷，第 4 期，第 2322–2358 页，2017 年第四季度。[137] S. Niknam、H. S. Dhillon 和 J. H. Reed，“无线通信中的联邦学习：动机、机遇与挑战”，IEEE Commun. Mag.，第 58 卷，第 6 期，第 46–51 页，2020 年 6 月。[138] W. Lim、N. Luong、D. Hoang、Y. Jiao、Y. Liang、Q. Yang、D. Niyato 和 C. Miao，“移动边缘网络中的联邦学习：全面综述”，IEEE Commun. Surveys Tuts.，第 22 卷，第 3 期，第 2031–2063 页，2020 年第三季度。[139] H. Sun、X. Chen、Q. Shi、M. Hong、X. Fu 和 N. Sidiropoulos，“学习优化：训练深度神经网络以管理干扰

”，IEEE Trans. Signal Process.，第 66 卷，第 20 期，第 5438–5453 页，2018 年 10 月。[140] H. Zhu、K. Yuen、L. Mihaylova 和 H. Leung，“智能车辆环境感知概述”，IEEE Trans. Intell. Transp. Syst.，第 18 卷，第 10 期，第 2584–2601 页，2017 年 10 月。[141] D. Gunduz 、 P. de Kerret、N. Sidiropoulos、D. Gesbert、C. Murthy 和 M. van der Schaar，“空中机器学习”，IEEE J. Sel. Areas in Commun.，第 37 卷，第 10 期，第 2184–2199 页，2019 年 10 月。[142] J. Chen 和 X. Ran，“边缘计算中的深度学习：综述”，Proc. IEEE，第 107 卷，第 8 期，第 1655–1674 页，2019 年 8 月。[143] S. Lin、Y. Zhang、C. Hsu、M. Skach、M. Haque、L. Tang 和 J. Mars，“自动驾驶的体系结构影响：约束与加速”，载于 ASPLOS 论文集，美国弗吉尼亚州威廉斯堡，2018 年 3 月。[144] Q. Liu、S. Huang、J. Opadere 和 T. Han，“面向移动增强现实的边缘网络编排器”，载于 IEEE INFOCOM 论文集，美国檀香山，2018 年 4 月。[145] G. Huang、M. Mattar、T. Berg 和 E. Learned-Miller，“野外标注人脸：用于研究非受限环境下人脸识别的数据库”，载于“现实生活”图像中的人脸研讨会：检测、对齐与识别，美国马萨诸塞州阿默斯特，2008 年 10 月。[146] S. Deng、H. Zhao、W. Fang、J. Yin、S. Dustdar 和 A. Zomaya，“边缘智能：边缘计算与人工智能的融合”，IEEE Internet Things J.，第 7 卷，第 8 期，第 7457–7469 页，2020 年 8 月。[147] E. Peltonen 等人，“第六代移动通信技术边缘智能白皮书”，arXiv:2004。*14850*， 2020， [在线]。可访问：http://arxiv.org/abs/2004.14850。[148] A. Hard、K. Rao、R. Mathews、S. Ramaswamy、F. Beaufays、S. Au- genstein、H. Eichner、C. Kiddon 和 D. Ramage，“用于移动键盘预测的联邦学习”，载于 IEEE IJCNN 论文集，匈牙利布达佩斯，2019 年 7 月。[149] W. Wang、C. Zhou、H. He、W. Wu、W. Zhuang 和 X. Shen，“基于长短期记忆网络和高斯过程回归的蜂窝网络流量负载预测”，载于 IEEE ICC 论文集，爱尔兰都柏林，2020 年 6 月。[150] N. Deo 和 M. Trivedi，“用于车辆轨迹预测的卷积社交池化”，载于 IEEE CVPR Workshops 论文集，美国犹他州盐湖城，2020 年 6 月。[151] M. Cos¸kun、A. Uc¸ar、 ¨O. Yildirim 和 Y. Demir，“基于卷积神经网络的人脸识别”，载于 IEEE MEES 论文集，乌克兰克列缅丘克，2017年。[152] D. Erhan、C. Szegedy、A. Toshev 和 D. Anguelov，“使用深度神经网络进行可扩展目标检测”，载于 IEEE CVPR 论文集，美国俄亥俄州哥伦布，2014 年 6 月。[153] S. M¨uller、O. Atan、M. van der Schaar 和 A. Klein，“无线网络中结合服务差异化的上下文感知主动内容缓存”，IEEE Trans. Wireless Commun.，第 16 卷，第 2 期，第 1024–1036 页，2017 年 2 月。[154] C. Zhou、W. Wu、H. He、P. Yang、F. Lyu、N. Cheng 和 X. Shen，“面向时延的空天地一体化网络物联网任务调度的深度强化学习”，IEEE Trans. Wireless Commun.，第 20 卷，第 2 期，第 911–925 页，2021 年 2 月。[155] V. Sciancalepore、X. Costa-Perez 和 A. Banchs，“RL-NSB：基于强化学习的第五代移动通信技术网络切片代理”，IEEE/ACM Trans. Netw.，第 27 卷，第 4 期，第 1543–1557 页，2019 年 8 月。[156] K. Qu、W. Zhuang、X. Shen、X. Li 和 J. Rao，“非平稳流量下虚拟网络功能的动态资源扩缩：一种学习方法”，IEEE Trans. Cogn. Commun. Netw.，第 7 卷，第 2 期，第 648–662 页，2021 年 6 月。[157] N. Van Huynh、D. Thai Hoang、D. Nguyen 和 E. Dutkiewicz，“基于深度对偶神经网络的最优快速实时资源切片”，IEEE J. Sel. Areas Commun.，第 37 卷，第 6 期，第 1455–1470 页，2019 年 6 月。[158] Y. Hua、R. Li、Z. Zhao、X. Chen 和 H. Zhang，“基于生成对抗网络的深度分布式强化学习，用于网络切片中的资源管理”，IEEE J. Sel. Areas Commun.，第 38 卷，第 2 期，第 334–349 页，2020 年 2 月。[159] R. Li、C. Wang、Z. Zhao、R. Guo 和 H. Zhang，“结合用户移动性的网络切片资源管理：基于长短期记忆网络的优势演员-评论家学习”，IEEE Commun. Lett.，第 24 卷，第 9 期，第 2005–2009 页，2020 年 9 月。[160] H. Chergui 和 C. Verikoukis，“面向第五代移动通信技术网络可靠动态端到端切片的离线服务级别协议约束深度学习”，IEEE J. Sel. Areas Commun.，第 38 卷，第 2 期，第 350–360 页，2020 年 2 月。[161] X. Chen、H. Zhang、C. Wu、S. Mao、Y. Ji 和 M. Bennis，“通过深度强化学习优化虚拟边缘计算系统中的计算卸载性能”，IEEE Internet Things J.，第 6 卷，第 3 期，第 4005–4018 页，2019 年 6 月。

[162] H. Xiang、M. Peng、Y. Sun 和 S. Yan，《切片雾无线接入网络中的模式选择与资源分配：一种强化学习方法》，IEEE车辆技术汇刊，第69卷，第4期，第4271–4284页，2020年4月。[163] H. Xiang、S. Yan 和 M. Peng，《通过深度强化学习实现雾无线接入网切片》，IEEE 无线通信汇刊，第19卷，第4期，第2515–2527页，2020年4月。[164] X. Chen、Z. Zhao、C. Wu、M. Bennis、H. Liu、Y. Ji 和 H. Zhang，《多租户跨切片资源编排：一种深度强化学习方法》，IEEE J. Sel. Areas Commun.，第37卷，第10期，第2377–2392页，2019年10月。[165] S. Messaoud、A. Bradai、O. Ben Ahmed、P. Quang、M. Atri 和 M. Hossain，《基于深度联邦Q学习的工业物联网网络切片》，IEEE工业信息学汇刊，第17卷，第8期，第5572–5582页，2021年8月。[166] G. Dandachi、A. De Domenico、D. Hoang 和 D. Niyato，《用于第五代移动通信技术网络切片部署与编排的人工智能框架》，IEEE认知通信与网络汇刊，第6卷，第2期，第858–871页，2020年6月。[167] H. Xiao、Y. Chen、Q. Zhang、A. Chronopoulos、Z. Zhang 和 S. Ouyang，《协作车联网交叉路口拥堵场景下的联合聚类与功率分配》，IEEE智能交通系统汇刊，第20卷，第6期，第2267–2277页，2019年6月。[168] Y. Yu、T. Wang 和 S. Liew，《异构无线网络中的深度强化学习多址接入》，IEEE J. Sel. Areas Commun.，第37卷，第6期，第1277–1290页，2019年6月。[169] Y. Zhang、Z. Mou、F. Gao、L. Xing、J. Jiang 和 Z. Han，《多无人机反向散射数据采集的分层深度强化学习》，IEEE物联网期刊，第8卷，第5期，第3786–3800页，2021年3月。[170] Y. Qian、R. Wang、J. Wu、B. Tan 和 H. Ren，《移动边缘网络中的强化学习最优计算与缓存》，IEEE J. Sel. Areas Commun.，第38卷，第10期，第2343–2355页，2020年10月。[171] S. D¨orner、S. Cammerer、J. Hoydis 和 S. Brink，《基于深度学习的空中通信》，IEEE信号处理选定主题期刊，第12卷，第1期，第132–143页，2018年2月。[172] J. Gao、W. Zhuang、M. Li、X. Shen 和 X. Li，《工业物联网中机器类型通信的媒体访问控制——第一部分：协议设计与分析》，IEEE物联网期刊，第8卷，第12期，第9945–9957页，2021年6月。[173] J. Gao、M. Li、W. Zhuang、X. Shen 和 X. Li，《工业物联网中机器类型通信的媒体访问控制——第二部分：调度与数值结果》，IEEE物联网期刊，第8卷，第12期，第9958–9969页，2021年6月。[174] Y. Kang、J. Hauswald、C. Gao、A. Rovinski、T. Mudge、J. Mars 和 L. Tang，《 Neurosurgeon：云与移动边缘之间的协作智能》，载于ACM ASPLOS 会议论文集，中国西安，2017年4月。[175] E. Li、L. Zeng、Z. Zhou 和 X. Chen，《边缘人工智能：通过边缘计算按需加速深度神经网络推理》，IEEE无线通信汇刊，第19卷，第1期，第447–457页，2020年1月。[176] W. He、S. Guo、S. Guo、X. Qiu 和 F. Qi，《面向物联网中时延敏感深度学习推理的联合深度神经网络分区部署与资源分配》，IEEE物联网期刊，第7卷，第10期，第9241–9254页，2020年10月。[177] Y. Cui、Z. Liu、W. Yao、Q. Li、A. B. Chan、T. Kuo 和 C. Xue，《用于自适应压缩与量化的全嵌套神经网络》，载于IJCAI会议论文集，日本横滨，2021年1月。[178] K. Lin、Y. Li、Q. Zhang 和 G. Fortino，《面向第六代移动通信技术赋能海量物联网任务执行的人工智能驱动协作资源分配》，IEEE物联网期刊，第8卷，第7期，第5264–5273页，2021年4月。[179] S. Haykin，《认知无线电：脑启发的无线通信》，IEEE J. Sel. Areas Commun.，第23卷，第2期，第201–220页，2005年2月。[180] P. Chemouil、P. Hui、W. Kellerer、Y. Li、R. Stadler、D. Tao、Y. Wen 和 Y. Zhang，《人工智能与机器学习在网络和通信中的应用专刊》，IEEE J. Sel. Areas Commun.，第37卷，第6期，第1185–1191页，2019年6月。[181] S. Sorour、U. Mohammad、A. Abutuleb 和 H. Hassanein，《投桃报李：无线组网能为人工智能和边缘学习带来什么》，arXiv:2006.*07453*，2020，[在线]。网址：https://arxiv.org/abs/2006.07453。[182]W. Wu、C. Zhou、M. Li、H. Wu、H. Zhou、N. Zhang、W. Zhuang 和 X. Shen，《面向第六代移动通信技术网络的人工智能原生网络切片》，arXiv:2105.*08576*， 2021， [在线]。网址：https://arxiv.org/abs/2105.08576。

[183]国际电信联盟电信标准化部门，《包括IMT-2020在内的未来网络机器学习架构框架》， 2019，[在线]。网址：https://https://www.itu.int/rec/T-REC-Y.3172/en。[184] 国际电信联盟，《第五代移动通信技术及未来网络机器学习统一架构》，2019年1月。[185] 3GPP，《第五代移动通信技术网络自动化使能器研究》，第3GPP TR 23.791 V16.2.0号，2019年6月。[186] F. Wilhelmi、S. Barrachina-Mu¯noz、B. Bellalta、C. Cano、A. Jonsson 和 V. Ram，《面向未来无线局域网的灵活机器学习感知架构》，IEEE Commun. Mag.，第58卷，第3期，第25–31页，2020年3月。[187] V. Va、T. Shimizu、G. Bansal 和 R. W. Heath Jr，《毫米波车载通信：综述》，Found. Trends Netw.，第10卷，第1期，第1–126页，2016年6月。[188] Y. Chen、T. Krishna、J. Emer 和 V. Sze，《Eyeriss：用于深度卷积神经网络的节能可重构加速器》，IEEE J. Solid-State Circuits，第52卷，第1期，第127–138页，2017年1月。[189] C. Hu、W. Bao、D. Wang 和 F. Liu，《用于边缘推理加速的动态自适应DNN剪枝》，载于IEEE INFOCOM会议论文集，法国巴黎，2019年4月。[190] D. Wen、X. Li、Q. Zeng、J. Ren 和 K. Huang，《边缘机器学习数据重要性感知无线资源管理概述》，J. Commun. Netw.，第4卷，第4期，第1–14页，2019年12月。[191] S. Wang、Y. C. Wu、M. Xia、R. Wang 和 V. Poor，《通过以学习为中心的功率分配实现边缘机器智能》，IEEE无线通信汇刊，第19卷，第11期，第7293–7308页，2020年11月。[192] D. Liu、G. Zhu、Q. Zeng、J. Zhang 和 K. Huang，《面向边缘学习的无线数据采集：数据重要性感知重传》，IEEE无线通信汇刊，第20卷，第1期，第406–420页，2021年1月。[193] L. Liu、J. Zhang、S. Song 和 K. Letaief，《客户端—边缘—云分层联邦学习》，载于IEEE ICC会议论文集，虚拟会议，2020年6月。[194] K. Yang、T. Jiang、Y. Shi 和 Z. Ding，《通过空中计算实现联邦学习》，IEEE无线通信汇刊，第19卷，第3期，第2022–2035页，2020年3月。[195] H. Wang、Z. Kaplan、D. Niu 和 B. Li，《利用强化学习优化非独立同分布数据上的联邦学习》，载于IEEE INFOCOM会议论文集，加拿大安大略省多伦多，2020年7月。[196] T. Nishio 和 R. Yonetani，《面向移动边缘异构资源联邦学习的客户端选择》，载于IEEE ICC会议论文集，中国上海，2019年5月。[197] S. Wang、T. Tuor、T. Salonidis、K. Leung、C. Makaya、T. He 和 K. Chan，《资源受限边缘计算系统中的自适应联邦学习》，IEEE J. Sel. Areas Commun.，第37卷，第6期，第1205–1221页，2019年6月。[198] J. Ren、Y. He、D. Wen、G. Yu、K. Huang 和 D. Guo，《兼顾重要性与信道感知的蜂窝联邦边缘学习调度》，IEEE无线通信汇刊，第19卷，第11期，第7690–7703页，2020年11月。[199] C. Wang、S. Zhang、Y. Chen、Z. Qian、J. Wu 和 M. Xiao，《面向边缘实时视频分析的联合配置自适应与带宽分配》，载于IEEE INFOCOM会议论文集，加拿大安大略省多伦多，2020年7月。[200] L. Zhang、L. Chen 和 J. Xu，《Autodidactic neurosurgeon：通过在线学习实现移动边缘智能的协同深度推理》，arXiv:2102.*02638*，2021，[在线]。网址：https://arxiv.org/abs/2102.02638。[201] S. Teerapittayanon、B. McDanel 和 H.-T. Kung，《云、边缘与终端设备上的分布式深度神经网络》，载于IEEE ICDCS会议论文集，美国佐治亚州亚特兰大，2017年6月。[202] W. Wu、P. Yang、W. Zhang、C. Zhou 和 X. Shen，《通过深度强化学习实现工业物联网中准确性有保障的协同深度神经网络推理》，IEEE工业信息学汇刊，第17卷，第7期，第4988–4998页，2020年7月。[203] K. Li、W. Ni、L. Duan、M. Abolhasan 和 J. Niu，《无线传感器网络中的无线能量传输与数据收集》，IEEE车辆技术汇刊，第67卷，第3期，第2686–2697页，2017年3月。[204] C. Zhan、Y. Zeng 和 R. Zhang，《无人机赋能无线传感器网络中的节能数据收集》，IEEE Commun. Lett.，第7卷，第3期，第328–331页，2017年6月。[205] A. Holub、P. Perona 和 M. Burl，《用于目标识别的基于熵的主动学习》，载于IEEE CVPR Workshops会议论文集，美国阿拉斯加州安克雷奇，2020年6月。[206] J. Verbraeken、M. Wolting、J. Katzy、J. Kloppenburg、T. Verbelen 和 J. Rellermeyer，《分布式机器学习综述》，ACM Computing Surveys，第53卷，第2期，第1–33页，2020年3月。[207] Q. Yang、Y. Liu、T. Chen 和 Y. Tong，《联邦机器学习：概念与应用》，ACM Trans. Intell. Syst. Technol.，第 10 卷，第 2 期，第 1–19 页，2019年1月。[208] K. Bonawitz、H. Eichner、W. Grieskamp、D. Huba、A. Ingerman、V. Ivanov、C. Kiddon、J. Kone˘cn`y、S. Mazzocchi、B. McMahan 等，《迈向大规模联邦学习：系统设计》，arXiv:1902.*01046*，2019，[在线]。获取地址：https://arxiv.org/abs/1902.01046。[209] K. He、X. Zhang、S. Ren 和 J. Sun，《用于图像识别的深度残差学习》，发表于 IEEE计算机视觉与模式识别会议论文集，美国内华达州拉斯维加斯，2016年6月。[211] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna，《重新思考用于计算机视觉的 Inception架构》，发表于 IEEE计算机视觉与模式识别会议论文集，美国内华达州拉斯维加斯，2016年6月。[212] A. Krizhevsky、I. Sutskever 和 G. Hinton，《使用深度卷积神经网络进行 ImageNet 分类》，发表于 IEEE神经信息处理系统大会论文集，美国内华达州太浩湖，2012年12月。[213] K. Simonyan 和 A. Zisserman，《用于大规模图像识别的超深卷积网络》，arXiv:1409.*1556*，2014，[在线]。获取地址：https://arxiv.org/abs/1409.1556。[214] C. Chen、W. Wang 和 B. Li，《轮询同步：缓解参数服务器中的通信瓶颈》，发表于 IEEE INFOCOM 会议论文集，法国巴黎，2019年4月。[215] W. Zhang、D. Yang、W. Wu、H. Peng、N. Zhang、H. Zhang 和 X. Shen，《优化分布式工业物联网中的联邦学习：一种多智能体方法》，IEEE J. Sel. Areas Commun.，第 39 卷，第 12 期，第 3688–3703 页，2021年12月。[216] M. Amiri 和 D. G¨und¨uz，《无线衰落信道上的联邦学习》，IEEE无线通信汇刊，第 19 卷，第 5 期，第 3546–3557 页，2020年5月。[217] J. Zhang、N. Li 和 M. Dedeoglu，《无线网络上的联邦学习：一种带宽受限协调下降方法》，arXiv 预印本 arXiv:2102.*07972*，2021，[在线]。获取地址：https://arxiv.org/abs/2102.07972。[218] S. Han、H. Mao 和 W. Dally，《深度压缩：通过剪枝、训练量化和霍夫曼编码压缩深度神经网络》，发表于 ICLR 论文集，波多黎各圣胡安，2016年5月。[219] G. Chen、W. Choi、X. Yu、T. Han 和 M. Chandraker，《利用知识蒸馏学习高效的目标检测模型》，发表于 IEEE神经信息处理系统大会论文集，美国加利福尼亚州长滩，2017年12月。[220] S. Teerapittayanon、B. McDanel 和 H. Kung，《Branchynet：通过深度神经网络提前退出实现快速推理》，发表于 IEEE国际模式识别会议论文集，墨西哥坎昆，2016年12月。[221] F. Iandola、S. Han、M. Moskewicz、K. Ashraf、W. Dally 和 K. Keutzer，《SqueezeNet：以少 50 倍的参数量和 <0.5 MB 的模型大小达到 AlexNet 级准确率》，arXiv:1602.*07360*，2016，[在线]。获取地址：http://arxiv.org/abs/1602.07360。[222] X. Chen、H. Ma、J. Wan、B. Li 和 T. Xia，《用于自动驾驶的多视角三维目标检测网络》，发表于 IEEE计算机视觉与模式识别会议论文集，美国夏威夷州檀香山，2017年7月。[223] H. Q. Ngo、A. Ashikhmin、H. Yang、E. G. Larsson 和 T. L. Marzetta，《无小区大规模多输入多输出：为每个人提供始终优质的服务》，发表于IEEE信号处理进展与无线通信研讨会论文集，瑞典斯德哥尔摩，2015年7月。[224] Y. Yang，《面向智能物联网的多层计算网络》，《自然·电子学》，第 2 卷，第 1 期，第 4–5 页，2019年1月。[225] M. Li、N. Cheng、J. Gao、Y. Wang、L. Zhao 和 X. Shen，《能源高效的无人机辅助移动边缘计算：资源分配与轨迹优化》，IEEE车辆技术汇刊，第 69 卷，第 3 期，第 3424–3438 页，2020年3月。[226] M. Leconte、G. S. Paschos、P. Mertikopoulos 和 U. C. Kozat，《网络切片的资源分配框架》，发表于 IEEE INFCOM 会议论文集，美国夏威夷州檀香山，2018年4月。[227] H. Halabian，《第五代移动通信技术虚拟化网络中的分布式资源分配优化》，IEEE J. Sel. Areas Commun.，第 37 卷，第 3 期，第 627–642 页，2019年3月。[228] Z. Xiong、S. Feng、W. Wang、D. Niyato、P. Wang 和 Z. Han，《区块链网络的云/雾计算资源管理与定价》，IEEE物联网期刊，第 6 卷，第 3 期，第 4585–4600 页，2019年6月。[229] J. Nie、J. Luo、Z. Xiong、D. Niyato 和 P. Wang，《面向移动群智感知的社会感知激励机制：一种斯塔克尔伯格博弈方法》，IEEE无线通信汇刊，第 18 卷，第 1 期，第 724–738 页

，2019年1月。

[230] P. Caballero、A. Banchs、G. De Veciana 和 X. Costa-P´erez，“网络切片博弈：实现多租户移动网络中的定制化”，*IEEE/ACMTrans.Netw.*，第27卷，第2期，第662–675页，2019年4月。[231] J. Gao、L. Zhao 和 X. Shen，“基于激励机制如实报告本地信息的网络效用最大化”，*IEEE*车辆技术汇刊，第67卷，第8期，第7523–7537页，2018年8月。[232] A. Zappone、M. Di Renzo 和 M. Debbah，“深度学习时代的无线网络设计：基于模型、基于人工智能，还是两者兼有？”，《通信汇刊》，第67卷，第10期，第7331–7376页，2019年10月。[233] Y. Yang、F. Gao、Z. Zhong、B. Ai 和 A. Alkhateeb，“基于深度迁移学习的频分双工大规模 MIMO 系统下行链路信道预测”，《通信汇刊》，第68卷，第12期，第7485–7497页，2020年12月。[234] D. C. Nguyen、P. Cheng、M. Ding、D. Lopez-Perez、P. N. Pathirana、J. Li、A. Seneviratne、Y. Li 和 H. V. Poor，“赋能未来无线网络中的人工智能：数据生命周期视角”，*IEEE Commun. Surveys Tuts.*，第23卷，第1期，第553–595页，2021年第1季度。[235] C. Zhou、H. Yang、X. Duan、D. Lopez、A. Pastor、Q. Wu、M. Boucadair 和 C. Jacquenet，“数字孪生网络：概念与参考架构”，互联网工程任务组，互联网草案，2021年7月。 [在线]。可访问：
h
tt
ps://datatracker.ietf.org/doc/html/draft-zhou-nmrg-digitaltwin-network-concepts-04 [236] Y. Lu、S. Maharjan 和 Y. Zhang，“第六代移动通信技术无线数字孪生网络的自适应边缘关联”，*IEEE*物联网期刊，第8卷，第22期，第16 219–16 230页，2021年7月。[237] W. Jiang、B. Han、M. A. Habibi 和 H. D. Schotten，“迈向第六代移动通信技术：全面综述”，*IEEE Open J. Commun. Society*，第2卷，第334–366页，2021年2月。[238] P. Bellavista、C. Giannelli、M. Mamei、M. Mendula 和 M. Picone，“面向应用的网络感知数字孪生管理：工业边缘环境中的实践”，*IEEE*工业信息学汇刊，第17卷，第11期，第7791–7801页，2021年11月。[239] J. Mei、X. Wang、K. Zheng、G. Boudreau、A. B. Sediq 和 H. Abou-zeid，“面向第六代移动通信技术服务提供的智能无线接入网切片：一种分层深度强化学习方法”，*IEEE*《通信汇刊》，第69卷，第9期，第6063–6078页，2021年9月。[240] C. Marquez、M. Gramaglia、M. Fiore、A. Banchs和X. Costa-Perez，“我该如何切分我的网络？资源共享效率的多业务实证评估”，发表于*MobiCom*论文集，印度新德里，2020年10月。[241] T. Mohammed、C. Joe-Wong、R. Babbar和M. Francesco，“通过自适应深度神经网络划分和卸载加速分布式推理”，发表于*IEEE INFOCOM*论文集，虚拟会议，2020年7月。[242] Y. He、J. Ren、G. Yu和Y. Cai，“利用卷积神经网络优化移动增强现实系统的学习性能”，发表于*IEEE*无线通信汇刊，第19卷，第8期，第5333–5344页，2020年8月。[243] W. Lin、G. Wu、X. Wang和K. Li，“一种用于构建云数据中心服务器功耗模型的人工神经网络方法”，发表于《*IEEE*可持续计算汇刊》，第5卷，第3期，第329–340页，2019年7月。[244] E. Strubell、A. Ganesh和A. McCallum，“自然语言处理深度学习的能源与政策考量”，*arXiv:1906.02243*, 2019, [在线]。链接：http://arxiv.org/abs/1906.02243。[245] N. Thompson、K. Greenewald、K. Lee和G. Manso，“深度学习的计算极限”，*arXiv:2007.05558*, 2020, [在线]。链接：http://arxiv.org/abs/2007.05558。[246] C. Zhu、S. Han、H. Mao和W. Dally，“训练式三值量化”，发表于*ICLR*论文集，法国土伦，2017年4月。[247] G. Hinton、O. Vinyals和J. Dean，“提炼神经网络中的知识”，发表于*IEEE NIPS*研讨会论文集，加拿大蒙特利尔，2014年12月。[248] C.-X. Wang、M. D. Renzo、S. Stanczak、S. Wang和E. G. Larsson，“人工智能赋能的第五代移动通信技术及后续无线组网：最新进展与未来挑战”，发表于《*IEEE*无线通信》，第27卷，第1期，第16–23页，2020年2月。[249] T. Wang、S. Wang和Z.-H. Zhou，“第五代移动通信技术及后续技术的机器学习：从基于模型到数据驱动的移动无线网络”，发表于《中国通信》，第16卷，第1期，第165–175页，2019年1月。[250] B. Sliwa、M. Patchou和C. Wietfeld，“兼得两全：数据驱动与基于模型相结合的车载网络仿真”，发表于*IEEE GLOBECOM*论文集，虚拟会议，2020年12月。