计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

智能物联网：概念、体系架构与关键技术

郭斌[1)] 刘思聪[1)] 刘琰[2)] 李志刚[1)] 於志文[1)] 周兴社[1)]

[1)](西北工业大学 计算机学院，西安 710072)

[2)](北京大学 计算机学院，北京 100871)

摘 要 智能物联网是当前人工智能与物联网技术相融合的产物，正成长为一个具有广泛发展前景的新兴前沿领域，实现从“万物互联”到“万物智联”的演进。在人工智能、边缘计算、物联网、移动嵌入式硬件等技术发展背景下，本文系统性地介绍智能物联网这一新兴方向。它对物联网感知、通信、计算和应用通过人工智能技术赋能，呈现泛在智能感知、云边端协同计算、分布式机器学习、人机物融合等新特征，具有更高灵活性、自组织性、自适应性。本文首先介绍了智能物联网的基本概念特质；其次阐述了智能物联网的体系架构；进一步详细介绍了智能物联网中的研究挑战与关键技术，包括泛在智能感知、群智感知计算、智能物联网通信、终端适配深度计算、物联网分布式学习、云边端协同计算、安全与隐私保护；最后，基于最新研究动态展望了极具潜力的未来研究方向，包括软硬协同终端智能、面向 AIoT 的智能演进、新一代智能物联网络、动态场景模型持续演化、人机物融合群智计算和通用 AIoT 系统平台。

关键词 智能物联网；群体智能；深度模型；边缘智能；人机物融合群智计算

中图法分类号 TP18

AIoT：The Concept, Architecture and Key Techniques

GUO Bin[1)] LIU Si-Cong[1)] LIU Yan[2)] LI Zhi-Gang[1)] YU Zhi-Wen[1)] ZHOU Xing-She[1)]

[1)](School of Computer Science, Northwestern Polytechnical University, Xi’an, 710072)

[2)](School of Computer Science, Peking University, Beijing, 100871)

Abstract Artificial Intelligence of Things (AIoT) is an emerging research field with broad development prospects, realizing the evolution from "Internet of Everything" to "Intelligent Connection of Everything." With the technological development such as artificial intelligence, edge computing, the Internet of Things, and mobile/embedded devices, AIoT aims to build a self-organizing, self-learning, self-adaptive, and continuous-evolving smart IoT system based on the deep fusion of these advanced technologies. AIoT is the combination of Artificial Intelligence (AI) technology and Internet of Things (IoT) infrastructure to achieve more intelligent IoT applications and provide more efficient services. Artificial intelligence models are good at analyzing and mining the potential patterns and strategies from massive amounts of data, while IoT has the ability to establish extensive connectivity for hundreds of millions of physical devices. AIoT empowers the perception, communication, computing, and application of the Internet of Things through various artificial intelligence technologies, and draws new features such as ubiquitous intelligent sensing, cloud-edge-end collaborative computing, distributed machine learning, and human-machine-things fusion. It has higher flexibility, self-organization, and self-adaptability. Specifically, AIoT enables the collection of multi-modal real-word data in real-time, and then utilizes machine learning approaches on the end devices, edge clusters or

收稿日期：2022-04-23；在线发布日期：2023-01-17. 本课题得到国家自然科学杰出青年基金(No. 62025205)、国家自然科学基金（62032020, 61725205, 62102317）资助. 郭斌(**通信作者**)，博士，教授，CCF会员,主要研究领域为智能物联网、人机物融合群智计算. E-mail: guob@nwpu.edu.cn. 刘思聪，博士，副教授，CCF会员,主要研究领域为智能物联网、移动嵌入式智能. 刘琰，博士，助理研究员，CCF会员，主要研究领域为普适计算、智能物联网、机器学习应用. 李志刚，博士，副教授，CCF会员，主要研究领域为物联网、移动计算. 於志文，博士，教授，CCF杰出会员，主要研究领域为普适计算、群智感知计算. 周兴社，博士，教授，CCF会士，主要研究领域为嵌入式计算和普适计算.

cloud servers for intelligent processing and decision making. This paper systematically introduces the direction of AIoT. Specifically, this paper firstly introduces the essential conceptual characteristics of AIoT, and then elaborates its architecture. Furthermore, we detail the research challenges and key technologies in AIoT, including ubiquitous intelligent sensing, mobile crowd sensing and computing, AIoT communication, terminal-adapted deep computing, AIoT distributed learning, cloud-edge-end collaborative computing, as well as security and privacy protection. Finally, based on the latest research developments, we present the future research directions, including collaborative soft and hard intelligence, AIoT-oriented intelligent evolution, new generation of intelligent IoT networks, continuous evolution of dynamic models, crowd intelligence with the deep fusion of Human, Machine, and IoT, and universal AIoT platforms.

Key words Artificial Intelligence of Things; Crowd Intelligence; Deep Model; Edge Intelligence; Crowd Intelligence with the Deep Fusion of Human, Machine, and Things

1 背景与趋势

物联网（Internet of Things，IoT）[1]，即“万物相连的互联网”，被认为是继计算机、互联网之后的又一次信息产业浪潮，是新一代信息技术的重要组成部分。它是在互联网基础上进一步延伸和扩展的网络，将各种信息传感设备与网络结合起来而形成的一个巨大网络，实现任何时间、任何地点，人、机、物的互联互通、信息交换与智能服务。万物互联是人类科技史上的又一次重大革命，对社会生产及生活产生了巨大而深远的影响。

自诞生以来，物联网技术的飞速发展不断引领产业升级，同时对其技术的演进提出了更高的要求。具体来讲，有五个重要的发展趋势。

一是物联网终端设备大规模普及，导致终端数据和连接出现井喷式增长。根据华为 GIV（全球产业展望）①和思科②预测，到 2025 年全球连接的设备数将达到 1000 亿台，而到 2030 年将有超过 5000 亿物联网设备接入互联网，届时全球每年产生的数据总量达 1YB，相比 2020 年，增长 23 倍。海量数据连接需要计算能力更高的物联网体系架构以实现数据的及时分析和处理。

二是数据处理的实时性、隐私性要求更为迫切。新的物联网业务不断衍生，万物感知、万物互联带来的数据洪流将与各产业深度融合，催生产业物联网的兴起。许多特殊的领域应用场景，如安防监测、自动驾驶、在线医疗等，一方面对数据的实时性要求较高，需要较低的数据传输时延，另一方面因为逐步与人们的日常生活深度融合，对隐私性保护的要求也极为迫切。

三是深度学习等人工智能技术的兴起。近年来，以深度学习为代表的新一代人工智能技术快速发展。相比传统机器学习模型，深度学习在很多领域任务上都取得了更好的性能结果。但同时，随着网络层数的增加，其模型参数规模不断变大，计算成本不断提高，为其在物联网环境的部署和执行带来了很大挑战。

四是物联网终端计算能力不断提升。传统物联网终端主要负责数据的采集与传输，而随着智能芯片、嵌入式处理器、感知设备等的不断发展和小型化，终端设备被不断赋予了智能数据处理能力，能在成本约束下完成部分数据处理和智能推理任务，可以为提升计算的实时性和保护数据隐私性提供支撑。

五是边缘计算和边缘智能的兴起。边缘计算是指在用户或数据源的物理位置或附近进行的计算，能就近提供边缘智能数据处理服务，这样可以降低延迟，节省带宽[2]-[3]。边缘计算的兴起进一步提升了本地数据处理能力。Gartner 将边缘计算列为 2020 年十大战略技术趋势之一③，其诞生解决了智能物联网发展的瓶颈问题。

综上，传统物联网架构的处理和计算能力已不足以支撑物联网络的深度覆盖、海量连接、实时处理、和智能计算等需求，在终端智能及边缘计算等

① https://www.huawei.com/cn/giv

②https://www.cisco.com/c/en/us/solutions/service-provider/a-network-to-support-iot.html

③https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2020

发展背景下，智能物联网（Artificial Intelligence of Things, AIoT，一般也表示为 AI+IoT 或人工智能物联网）[4]-[6]作为未来物联网发展的新趋势近年来得到广泛关注。

智能物联网是 2017 年兴起的概念①，是人工智能与物联网技术相融合的产物，正成长为一个具有广泛发展前景的新兴前沿领域，实现从“万物互联”到“万物智联”的演进。据 Gartner 预测，未来超过 75%的数据需要在网络边缘侧分析、处理与存储。AIoT 首先通过各种传感器联网实时采集各类数据（环境数据、运行数据、业务数据、监测数据等），进而在终端设备、边缘设备或云端通过数据挖掘和机器学习方法进行智能化处理和理解。近年来，智能物联网应用已逐步融入国家重大需求和民生的各个领域，例如智慧城市、智能制造、社会治理等。

智能物联网带来了泛在智能感知、情境自适应通信、分布式群体智能、云边端协同计算等新的挑战问题。来自麻省理工学院、斯坦福大学、耶鲁大学、加州大学伯克利分校、剑桥大学，以及国内的研究人员都对智能物联网这一前沿领域开展了系统性研究。例如，麻省理工学院研究人员对资源受限物联网终端上的深度模型压缩等技术进行了系统性研究[7]。耶鲁大学研究人员提出了边端协同高效深度推理模型[8]。斯坦福大学研究团队基于多智能体深度强化学习对智能体间的分布式协作学习能力进行了研究[9]。剑桥大学研究人员就资源受限环境下深度学习模型的轻量级自动搜索提出了新的方法[10]。香港理工大学研究人员则对车联网背景下边缘智能计算的应用进行了深入分析和探索[11]。

在 AIoT 快速发展趋势下，国内外著名 IT 企业都加紧布局，在边缘智能、智能芯片、智能物联网软件平台等方面取得了很多基础性成果。微软在 2015 年正式发布了 Azure 物联网套件——Azure IoT Suite②。2021 年，又进一步发布全新的边缘计算平台 Azure Edge Zone 以支持实时数据处理。亚马逊也于 2015 率先发布 AWS IoT③平台，并于 2017 年推出 FreeRTOS 操作系统，适用于小型低功耗的边缘设备进行编程、部署、连接与管理。2018 年，阿里巴巴推出 AliOS Things④物联网操作系统，提供 IoT 连接、智能处理、云边端协同计算等服务。同年，京东发布“城市计算平台”，结合深度学习等构建时空关联模型及学习算法解决交通规划、火力发电、环境保护等城市不同场景下的智能应用问题。2019 年，华为推出了面向物联网的华为鸿蒙操作系统 HarmonyOS⑤，是一种基于微内核、面向 5G 的全场景分布式操作系统，在传统的单设备系统能力基础上，提出了基于同一套系统能力、适配多种终端形态的分布式理念。

综上，无论在学术界和产业界，智能物联网均成为新的发展趋势。鉴于此，本文将面向泛在计算、人工智能与物联网交叉学术前沿，阐述其基本概念、体系架构、关键技术及典型应用，并在此基础上探索其未来科学挑战及机遇。

2 智能物联网概念与特征

在新一代人工智能发展背景下，与物联网在实际应用中不断落地融合，促进了智能物联网 AIoT 的诞生。物联网应用的不断普及为人工智能提供了海量的物理世界数据，而人工智能技术的应用则为物联网领域应用的效能提升提供赋能支撑。本节将介绍智能物联网的基本概念及其特质。

2.1 智能物联网概念

智能物联网（AIoT）属于比较新的名词，学术界和业界对其定义并未达成一致。下面先给出几种主要的智能物联网定义：

维基百科⑥：智能物联网是人工智能(AI)技术与物联网(IoT)基础设施的结合，以实现更高效的物联网运营，改善人机交互，提高数据管理与分析能力。

《2020 年中国智能物联网（AIoT）白皮书》[12]中指出：AIoT 是人工智能与物联网的协同应用，它通过 IoT 系统的传感器实现实时信息采集，而在终端、边缘或云进行数据智能分析，最终形成一个智能化生态体系。

悉尼大学的研究人员发表在《IEEE Internet of Things》[4]文中指出先进的通信技术（如 5G、WiFi 等）将促进万物广泛连接，产生海量数据并推动智能物联网的产生。其通过融合边缘计算、雾计算和云计算的新体系架构来提升物联网系统的智能性和数据处理的及时性与安全性。

① https://report.iresearch.cn/report/202002/3529.shtml

② https://azure.microsoft.com/en-us/overview/iot/

③ https://aws.amazon.com/cn/iot/

④ https://www.aliyun.com/product/aliosthings

⑤ https://www.harmonyos.com/

⑥ https://en.wikipedia.org/wiki/Artificial_intelligence_of_things/

弗吉尼亚理工大学研究人员[13]发表在《IEEE Computational Intelligence》杂志的论文从未来网络结构角度探索智能物联网的发展，认为随着 5G 和 6G 的发展，越来越多的设备将通过联结形成超级网络。人工智能将在促进 IoT 网络更有效连接方面发挥重要作用，包括 AI 增强的随机接入和频谱共享等技术。

美国加州大学研究人员发表在《Proceedings of the IEEE》的文章[14]指出智能手机、物联网传感器等终端设备正在生成需要利用深度学习进行实时分析或用于训练深度学习模型的数据。然而，深度学习推理和训练需要大量的计算资源才能快速运行。边缘计算将计算节点的细网格放置在靠近终端设备的位置，是满足深度学习对边缘设备高计算量和低延迟要求的一种可行方式，同时还提供了隐私、带宽效率和可扩展性方面的额外优势。其认为智能物联网通过跨终端设备、边缘服务器和云的组合进行深度学习推理，实现高效深度计算。

香港科技大学杨强教授和南洋理工大学 Dusit Niyato 教授等在其文章中[15]强调智能物联网中边缘计算和联邦学习发挥的重要作用，能解决传统的基于云的机器学习方法所产生的延迟和通信效率低下问题，在数据隐私保护的前提下实现机器学习模型的协同训练。

基于以上论述，我们将智能物联网定义为：**通过人工智能、边缘计算、物联网等技术的深度融合，赋能感知、通信、计算和应用等路径实现万物智联，呈现泛在智能感知、云边端协同计算、分布式机器学习、人机物融合等新特征，具有更高灵活性、自组织性、自适应性、持续演化的物联网系统**。

智能物联网是人工智能和物联网两种技术相互融合的产物：一方面，物联网终端设备大规模普及，终端设备生成的数据量呈爆炸式增长，人工智能技术能帮助物联网实现智能感知与智慧互联，提升感知与连接的广度、深度和有效性。同时能为物联网系统中数据的智能化分析和处理提供支撑，为物联网领域应用的效能提升和自主优化提供赋能，为用户提供更为个性化和智能化的体验，即 AI for IoT；另一方面，物联网应用的不断普及为人工智能提供了海量的物理世界数据，实现了人-机-物的智能互联，也为人工智能的应用落地提供了客观的需求和丰富的路径。随着智能芯片、处理器、感知设备等的不断发展和小型化，终端设备被不断赋予了智能数据处理能力，通过人类便携终端（如手机、可穿戴）、物联网嵌入式实体（如摄像头、智能小车）、互联网应用（如边缘、云服务器）等异构智能体的协作感知计算可赋予人工智能新特点，人工智能 2.0 中的“群体智能”是一种通过聚集群体的智慧解决问题的新模式，即 IoT for AI。总的来说，物联网是异构、海量数据的来源，而人工智能用于实施大数据分析，其最终目标是实现万物智联。

传统物联网是实现终端数据收集到云端数据处理的过程。海量的传感器和设备收集来自环境的数据，将它们传输到云中心，并通过互联网接收反馈，实现连接和感知。传统物联网数据的计算和存储均在云计算中心，而智能物联网是以数据处理为中心，通过物联网系统的传感器实现实时信息采集，进而在终端设备、边缘设备或云端通过数据挖掘和机器学习方法进行智能化处理和理解，最终形成一个智能化系统。相比传统的物联网云端数据处理，在智能物联网时代从云端计算集群、边缘网络节点到物联网智能终端都可参与到感知、学习和决策的过程中。

2.2 智能物联网特质

智能物联网特质与内涵如下：

人机物融合计算[16]-[17]：随着物联网、人工智能等技术的发展，计算系统正从信息空间拓展到包含人类社会、信息空间和物理世界的三元空间，人机物三元融合计算成为重要形态。它能有效协同与融合人、机、物异质要素，进而构建新型智能计算系统，是解决智能制造、智慧城市、社会治理等国家重大需求的有力支撑。

- **人（Human）**，主要体现为社会空间广大普通用户及其所携带的移动或可穿戴设备，其发挥的作用一方面为人类智慧（包括个体或群体智能），另一方面则涵盖基于移动设备的群智感知计算。
- **机（Machine）**，主要体现为信息空间丰富的互联网应用及云端和边缘服务，在传统互联网和移动互联网等发展背景下，信息空间集聚了海量多模态的数据和多样化的计算资源。
- **物（Things）**，主要体现为具有感知、计算、通信、决策和移动等能力的物理实体，在物联网发展背景下，各种各样的移动/嵌入式终端不断涌现，为感知和理解物理空间动态提供了重要支撑。

人、机、物三种要素在同一环境或应用场景下相互联结，和谐共生，但彼此能力差异、数据互补，

需要通过协作交互来实现能力增强，进而完成复杂的感知和计算任务。自组织性是指跨社会、信息、物理空间的异构群智能体，基于实时状态与动态环境交互，通过系统内部个体的分布式自主交互和内在共识，以形成时间、空间、逻辑或功能上的自组织协作。自适应性是指在动态变化的开放环境中，智能体根据感知数据的多样性、设备资源的动态性等因素，自适应调整优化感知计算模式、分布式学习策略等。

泛在智能感知：在智能物联网时代，利用无处不在的感知资源，包括摄像头、RFID、WiFi、红外、声波、毫米波等，产生丰富的多模态感知数据，进而通过机器学习和深度学习等方法实现对目标（人、环境、或事件等）行为的准确感知[18]-[21]。

情境自适应通信：针对不断变化的网络资源、连接拓扑和数据传输等情境，从实时获取的网络数据中提取情境信息，进而通过自适应机制实现情境适配的低成本、高效通信[22]-[23]。

物联网终端智能：智能物联网场景中，将深度学习模型（如实时视频数据处理）离线部署在资源受限且环境多变的物联网终端设备执行逐渐成为一种趋势[24]-[26]，其具有低计算延时、低传输成本、保护数据隐私等优势，然而硬件资源限制和环境动态变化对终端智能算法带来很大挑战。针对受限环境设计相适应的轻量级深度学习模型是智能物联网的一个关键问题。

分布式群体智能：针对单个终端智能体数据和经验有限、模型训练能力弱、应用场景和任务多变等问题，与现有集中式学习模型和框架相区别，在分布式环境下实现多智能终端协作增强学习是智能物联网发展的重要趋势[15][17][27]。

云边端协同计算：针对海量智能物联网数据及实时性、隐私性等数据处理需求，将边缘计算技术引入物联网，形成“端—边—云”协同计算的智能物联网体系架构，高效及时地处理业务数据[28]-[30]。

3 智能物联网体系架构

物联网的核心是物与物以及人与物之间的信息交互。传统的物联网体系架构分为 3 层：**感知层**如同人的各种感觉器官，由各种各样的传感器设备组成，用来感知外界环境的温/湿度、压强、光照、气压、受力情况等信息；**网络层**相当于人的神经系统，由各种异构网络组成，将来自感知层的各类信息通过网络传输到应用层；**应用层**是用户和物联网间的桥梁，通过云计算、大数据、中间件等技术，为不同行业提供应用方案。

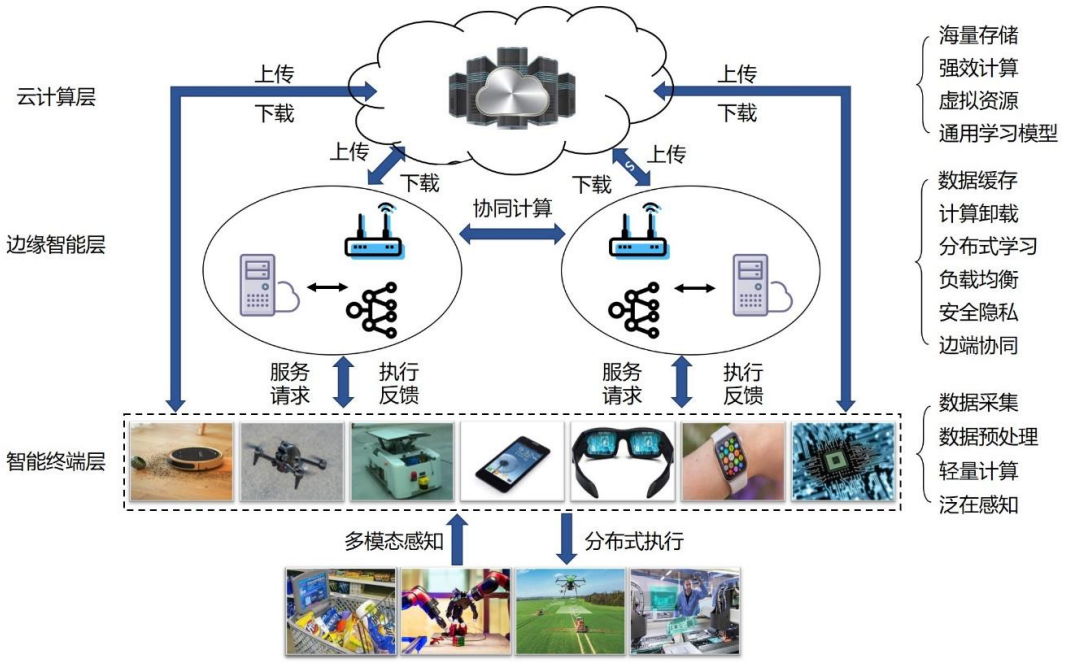

图 1　云边端协同的 AIoT 体系架构

智能物联网以数据处理为中心，面临新的机遇与挑战，将形成新的体系架构与系统软件平台，下面分别进行阐述。

3.1 云边端协同AIoT体系架构

智能物联网以智能信息的高效、实时处理为中心，随着边缘计算和边缘智能的引入，将形成云边端协同的 AIoT 体系架构。如图 1 所示，系统分为三层，包括智能终端层、边缘智能层、云计算层。

智能终端层由各种物联网终端（如机器人、无人机、智能车、移动/可穿戴设备等）组成，是 AIoT 的感知和执行单元。通过控制系统可以控制终端设备完成音/视频、位置、压力、温湿度等多模态感知数据采集，并执行运动、抓取、跟踪等行为。智能物联网的控制系统随着 AI 的引入，更强调终端设备的多样性、控制的智能化等方面。在复杂应用环境中，利用控制系统进行智能化的数据采集和分析，不仅可以让庞大复杂系统得到全面完善，还能让智能控制效果得到全面推动和促进，节省大量人力物力，降低工作成本。与传统物联网感知层不同，智能终端层将完成部分的数据处理任务，在终端部署传统机器学习或深度学习模型。由于终端资源受限，一般采用轻量级的模型设计方法，包括网络剪枝/压缩/量化等技术。

边缘智能层是指将计算和智能处理能力部署在靠近终端的边缘设备上，通过边端协同增强计算能力，可以减少计算延迟，支持实时服务。边缘智能层不仅具有传统物联网传输层和应用层的功能，还具有边端协同、负载均衡、分布式学习等功能。从计算层面来讲，可将终端计算任务部分卸载到边缘计算节点上；从智能处理层面来讲，由于终端资源和数据受限，通过边缘群智能体协同、边端协同深度模型分割可更好地执行模型训练和推理任务。

云计算层支持丰富的物联网应用服务，它使 AIoT 应用能够通过互联网虚拟地使用计算资源，具有灵活性、可伸缩、可靠性等特征。与传统物联网应用层中云计算平台类似，来自大规模分布式物联网终端和边缘设备的实时数据流通过网络传输到远程云中心，在那里它们被进一步集成、处理和存储。基于海量物联网数据和丰富的计算资源，在云上训练和部署具有较好泛化能力的机器学习模型成为可能。

不同于传统物联网体系架构中感知层、网络层和应用层之间的明确功能，在云边端协同的 AIoT 架构中，“智能终端层、边缘智能层和云计算层”动态分配计算量，有效缓解了云计算平台的数据处理负担，提高了数据处理效率。当实时响应和低时延是关键因素时，主要依靠更靠近用户的边缘计算架构；当计算决策的精确性是关键因素时，主要依靠云服务器完成。

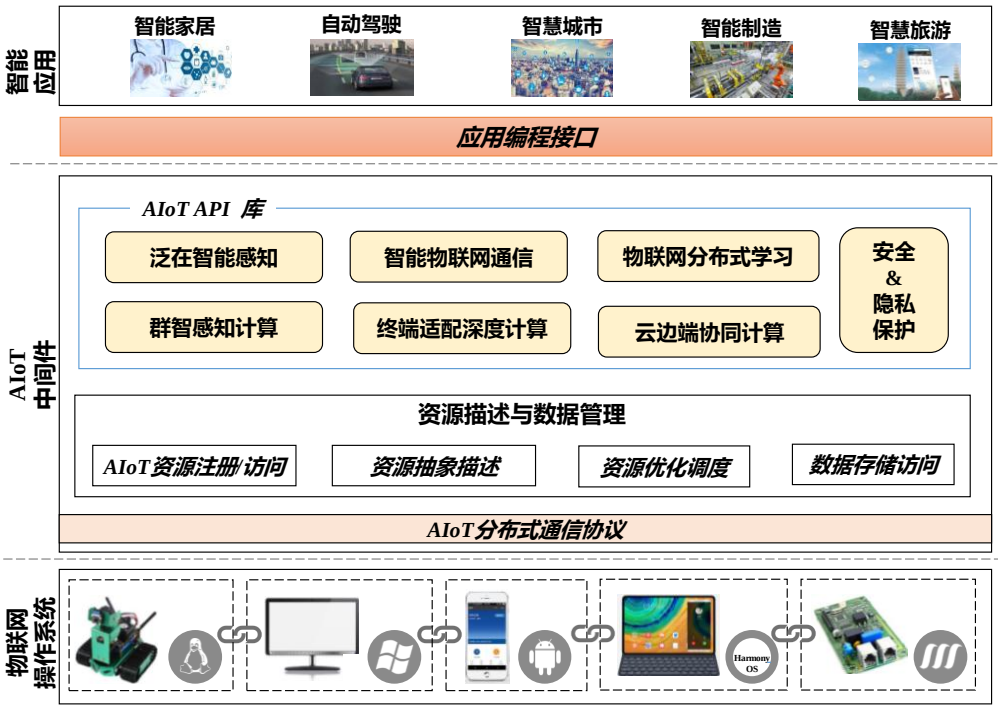

图 2 AIoT 系统软件平台

3.2 AIoT系统软件平台

智能物联网是“软硬协同”的智能系统，在云边端协同的智能物联网体系结构之上，软件平台也是智能物联网的核心组成要素。软件平台在设备和应用之间提供互操作能力，能够集成异构的计算和通信设备，简化应用的开发，并为运行在异构设备上的多种应用和服务之间提供互操作能力。一般来说，体现为中间件形式，如微服务框架。

传统物联网主要为云计算提供海量数据来源，产生了基于云计算的物联网中间件平台，如亚马逊 AWS IoT、微软 Azure IoT 等。AIoT 物联网中间件基于云边端协同的新型体系结构，且以数据和智能算法为中心要素。例如，在智能物联网中，提供 AI 服务的系统一般通过以下过程构建：数据收集与存储，数据分析与预处理 AI 模型训练，AI 模型部署与推理，以及监测并维护精度。经过上述过程训练好 AI 模型后，需要对 AI 模型进行封装部署，以提供 AI 推理服务，通常有云端 AI、边缘 AI 和端侧 AI 三种方式。物联网云-边-端 AI 功能的实现，使得可以在终端设备、边缘域或云中心通过 AI 对数据进行智能化分析，实现智能感知、智能连接与智能计算，提升物联网感知、连接与计算的广度与深度，从而有效支撑上层各种智能物联网应用。图 2 给出一个参考的 AIoT 系统软件平台架构。

物联网操作系统：物联网操作系统是运行在物联网设备上的提供物物相连能力的操作系统，其核心在于能够将各种物体连接到互联网，并提供数据通信能力。为适应物联网中异构硬件设备及操作系统的差异性，软件平台应充分考虑多样化的硬件需求，通过合理的架构设计，使软件平台本身具备充分的可伸缩性，能够便捷地应用于不同硬件设备上。例如，在智能物联网应用中，物联网操作系统需支持多种物联网协议，涵盖物联网主流通信协议，包括局域网连接能力（如 WiFi、BLE）、广域网连接能力（如 NB-IoT、LoRa）、网络应用协议（HTTP/s、MQTT、CoAP、WebSocket 等）。为了在设备端支持智能物联网应用开发，物联网操作系统需提供 AI 智能框架，提供常用 AI 算法集成的便捷框架，包括 Python/C++编程规范，隔离硬件差异，提供连云、控端、多媒体、机器学习等能力。

IoT 分布式通信协议：智能物联网软件平台中，存在感知、预测、决策和执行等多种模块，需要实时、频繁地交换数据，因此平台应具备灵活可靠的通信架构，以提供更好的互操作性，保障数据进行实时、高效、灵活地分发，可采用发布/订阅体系架构，以满足分布式通信需求。

资源描述与数据管理：物联网软件平台连接规模化异构物联设备，获取海量多模态感知数据，因此应具有设备资源描述与数据管理功能，包括资源注册与访问、资源抽象描述、资源优化调度、数据存储访问等模块。其中资源注册与访问面向动态连接拓扑环境，生成资源抽象描述，在此基础上提供资源优化调度服务。

AIoT API 库：为提供异构设备和多种应用间的互操作能力，软件平台以数据和智能算法为中心，如泛在智能感知、群智感知计算、智能物联网通信、终端适配深度计算、边端协同机器学习、云边端协同计算以及安全与隐私保护等模块。因此需要良好的接口设计以合理划分软件系统职责，降低各部分间的相互依赖，提高组成单元内聚性，以助力开发人员的高效开发。

应用编程接口：软件平台通过抽象和建模，对不同的底层硬件和功能部件进行抽象，为上层提供统一的应用编程接口。同时，编程接口为开发者屏蔽物联设备的硬件配置及资源状态，使得面向不同场景开发的智能应用可运行在多种异构的硬件平台，而只需硬件设备安装运行系统软件平台即可。

智能应用：智能物联网在智能家居、自动驾驶、城市计算、智能制造等领域均有重要应用前景。例如，智能家居利用室内大量的物联网设备（如温湿度传感器、家用电器、服务机器人、安防设施）收集多模态感知数据（可调用泛在智能感知 API），进而通过智能分析和处理进行室内状态和用户行为识别（可调用终端适配深度计算 API），并最终完成对家居环境的调控以及智能服务的提供，实现安全舒适、绿色健康、以人为本的智能家居体验（可调用安全&隐私保护 API）。在智能工厂中，智能物联网的应用主要分为两个层面，第一层次是通过工业互联网技术来实现连接并获取工厂各类主体的感知数据（可调用泛在智能感知 API），第二层次则是利用人工智能技术来对数据进行分析、识别和推理（可调用物联网分布式学习 API）。

4 研究挑战与关键技术

智能物联网的人机物融合、泛在计算、分布式智能、云边端协同等新特质，以及区别于传统物联网的体系及软件结构带来了很多新的挑战问题，下

面将简要阐述所面临的挑战及相关技术。本节从智能感知-网络通讯-协同计算-隐私保护四个层面分别介绍 AIoT 关键技术，如图 3 所示。

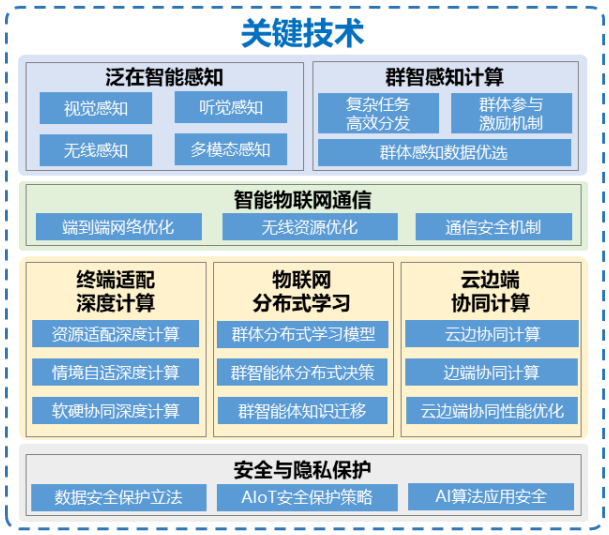

图 3 AIoT 关键技术

4.1 泛在智能感知

针对智能物联网感知的实时性、完整性等需求，如何对复杂场景中的目标进行全面且及时的感知是一大挑战。一方面，需要探索不同感知资源能力的差异性，并针对感知任务进行能力选择和聚合。另一方面，还需考虑感知对象行为的复杂性和个性化特征，以适应多样化的应用场景。“泛在智能感知”是普适计算、移动计算、物联网和人工智能等多个领域交叉的一个新兴研究方向。泛在智能感知主要通过内嵌在智能手机、手表、可穿戴设备、汽车、家电中的摄像头、加速度传感器、陀螺仪、WiFi、LTE、毫米波雷达、声波收发模块对人和环境进行多模态感知，并利用人工智能的算法、模型和技术对感知信息进行分析得到关于人和环境的情境状态；进而为人在合适的时间、地点提供智能的服务。泛在智能感知在智慧终端、智慧家居、智慧健康医疗、新型人机交互和自动驾驶等领域有着广泛应用，是智能物联网的重要组成部分。

智能视觉感知 视觉是人类从外界获得信息的主要途径，通过机器、计算机以及人工智能方法来模拟人类的视觉功能是人们多年的追求。较之于传统视觉技术，智能视觉感知采用机器学习与深度学习技术，具备更快更强的感知与运算能力，提升了边缘检测、语义分割、图像滤波等基础视觉处理能力，并在移动目标检测[31]、移动地图构建[32]、视频流目标跟踪[33]、视频动作识别[34]等关键视觉感知技术得到广泛应用，强化视觉感知能力的同时也拓宽了视觉感知的应用范畴。

智能听觉感知 听觉感知是感知主体检测、分析、识别和理解语音信号信息的过程，它允许我们与现实环境正确地互动，流畅地沟通。传统听觉感知在梅尔倒频谱系数等基础特征之上，通过混合高斯-隐马尔科夫等模型进行语音和语调的识别。而智能听觉感知通过深度神经网络增强语音模型的特征能力、感知精度和识别鲁棒性，并在更为广义的层面，利用语言理解、对话跟踪、语言生成等关键技术完成真实场景中的人机物互通[35]-[37]。

智能无线感知 智能无线感知是近年来新兴起的一个前沿研究热点，主要通过普适的无线信号如 WiFi、RFID、毫米波雷达、声波等对人和环境进行非接触式或与设备无关的（Device-Free）感知，从而为人类与物理设备、场景环境的融合奠定基础[38]。相较于视觉感知、听觉感知等技术，无线感知具有普适程度高、感知范围广、感知成本低、不侵扰用户、不泄露隐私等特点和优势，是实现泛在感知与普适计算的理想形式，在智能物联网中具有广阔的应用前景。无线感知的基本原理是环境中传播的无线信号，会由于感知目标（人或物）的存在而产生反射、衍射、散射等现象，通过检测和分析感知信号（如 RSSI、CSI 等）的变化特征，便可以推断感知目标的位置、状态等信息，达成感知之目的。通过多普勒效应模型[39]、菲涅尔区模型[40]等理论模型探索无线感知的一般机理，推导特定物理量（如位置、速度、角度等）与接收信号特征间的量化关系，进而基于物理量和信号特征识别人的行为。此外，通过三角度量、计算机指纹库、深度神经网络等智能计算方法对无线感知信号、无线感知数据、无线感知模式进行识别和处理，并通过室内定位[41]、目标跟踪[42]、行为识别[43]关键技术，在健康监护、人机交互、行为识别等领域得到了大量应用，因而受到学术界和产业界的广泛关注与重视。

多模态智能感知 多模态融合感知技术综合通过不同类型的感知设备收集具有不同特性的数据，避免了单个传感器的感知局限性和不确定性，形成了对环境或目标更全面的感知和识别，提高了系统的外部感知能力，是未来智能交互必不可少的研究课题。通过有效整合多模态数据，便可获得对感知目标的整体描述。例如，为了实现自动驾驶，智能汽车部署了激光雷达、毫米波雷达、超声波传感器、音频传感器、视频传感器、红外传感器等不同类型

的感知设备，以便获得更加全面的信息，进而增强系统的可靠性和容错性。传统多模态数据融合技术主要分为三种类型：数据级融合，即通过空间对齐直接融合不同模态的原始感知数据；特征级融合，即通过级联或元素相乘在特征空间中融合多模态感知数据；目标级融合，即通过融合各模态模型的预测结果完成最终决策。目前多模态融合技术的主要集中在视觉与音频之间的多模态学习，少量工作研究毫米波雷达、激光雷达和摄像头之间的多模态感知。多模态融合感知技术在行为识别[44]、机器人系统[45]、目标识别和跟踪[46]-[47]、定位与导航[48]、自动驾驶[49]等领域发挥着巨大的作用。

4.2 群智感知计算

群智感知由众包、参与感知等相关概念发展而来。2012 年，清华大学刘云浩教授首次提出“群智感知计算”概念[50]-[52]，它利用大量普通用户使用的移动设备作为基本感知单元，通过物联网/移动互联网进行协作，实现感知任务分发与感知数据收集利用，最终完成大规模、复杂的城市与社会感知任务[53]。群智感知计算利用群体智慧和泛在移动/可穿戴终端构建大规模移动感知网络，是一种新型智能感知模式，对传统静态传感网络互为补充[54]。

复杂任务高效分发 群智感知依赖参与用户的移动终端所具备的各种传感和计算能力等来进行感知。与传统感知网络相比，参与式感知节点具有规模大、分布广、能力互补等特点，而任务则具有需求多样、多点并发、动态变化等特征。需研究针对不同感知任务需求的参与者优选方法，根据任务的时空特征、技能需求及用户个人偏好、移动轨迹、移动距离、激励成本等设定优化目标和约束，设计任务分配模型，一般通过最优化理论（动态规划、博弈论、多目标优化等）和群智能优化算法（如遗传算法、粒子群算法、蚁群算法）等进行求解[55]-[56]。

群体参与激励机制 群智感知需要雇佣大量的参与者采集数据，很多任务还需要参与者前往特定的地点并有较高的数据传输和处理成本；此外，群体参与还存在数据质量难以保证的问题。针对以上问题，群智感知系统通过采用适当的激励方式（如报酬支付激励、虚拟积分激励、游戏娱乐激励、社会交互激励等），鼓励和刺激参与者参与到感知任务中，并提供优质可信的感知数据[57]。不同的激励方式在不同的场景下，对不同的参与者具有不同的激励效用，因此如何选择和设计合适的激励机制是群智感知计算的主要研究内容之一[58]。

群体感知数据优选 群智数据的质量直接影响数据分析的结果，进而影响群智服务的性能。由于不同用户在活动范围上有一定重叠，群智感知所采集到的数据中可能存在大量冗余。而大量未经训练的用户作为基本感知单元会带来感知数据多模态、不准确、不一致等质量问题。挑战在于如何实现优质数据选择和收集。在智能物联网中，一方面可以在终端进行数据预处理，剔除低质量数据；另一方面可以在边缘设备进行数据局部汇聚，及时发现来自不同终端的冗余数据[59]-[60]。从而在减少数据传输成本的同时为云数据挖掘与模型训练提供优质数据支持。

4.3 智能物联网通信

虽然完整物联网通信体系已经建立，但学术界和工业界近年来不断思考如何将 AI 融入到物联网通信系统中，实现物联网通信效能的大幅提升。已有研究集中于网络、资源管理和安全，主要思想是将机器学习、AI 的思想引入到相应算法和协议设计过程，实现通信与 AI 的结合。目前各项研究目前尚处于初步探索阶段，智能物联网通信的发展还需要一个长期的过程，机遇与挑战共存[23]。

端到端网络优化 在 MAC 协议中，机器学习为优化 IoT 网络的性能提供很好的解决方案。可以把物联网设备想象成一个能够借助机器学习访问信道资源的智能设备，通过机器学习，物联网设备能够观察和学习不同性能指标对网络性能的影响，然后利用这些学习到的经验来可靠地提升网络性能，同时生成后续的执行动作。强化学习、神经网络等 AI 方法的引入在物联网应用复杂多变的环境中提供了路由的自适应能力，在通信故障、拓扑变化和节点移动性等情况下提供了较好的性能[61]-[62]。例如通过深度神经网络学习网络拓扑、流量和路由之间的复杂关系以优化路由来降低网络负担[63]，通过强化学习的方法来动态选择合适的拥塞控制算法以提高数据的传输效率[64]。基于机器学习的拥塞控制方法可以更准确地估计网络流量，从而找到最佳路径，最小化节点与基站之间的端到端时延，并可根据网络的动态变化调整传输范围，更加灵活地控制传输层发生的拥塞，提高传输效率[65]-[66]。

无线资源优化 无线通信是 IoT 主要的通信方式，无线资源管理通过有限物理通信资源的合理利用，以满足各种 IoT 应用需求。现有无线资源管理方法通常是为静态网络设计，高度依赖于公式化的数学问题。而 IoT 网络的动态性，导致高复杂性的

算法频繁执行，带来了性能损失。因此，可将 AI 引入到无线资源管理，如强化学习可以仅基于环境反馈的回报/成本学习好的资源管理策略，可对动态网络做出快速决策[67]；深度学习模型优越的逼近能力，可以实现一些高复杂度的资源管理算法[68]；多智能体强化学习可赋予每个节点自主决定资源分配的能力[69]。因此，机器学习在功率控制、频谱管理、波束形成设计等具有较好的应用前景[70]。

通信安全机制 借助深度学习，通过对数据进行深入归纳、分析，从而获取新的、规律性的信息和知识，并利用这些知识建立用于支持决策的模型，进行网络风险分析或预测。如使用机器学习技术处理和分析收集的数据，可以更好的防范入侵检测，或利用人工智能对物联网系统中的恶意软件进行检测,挑战在于设计适合物联网设备的轻量级智能通信安全机制[71]-[72]。

4.4 终端适配深度计算

在智能物联网时代，在物联网终端执行深度计算实现智能推断逐渐成为一种趋势，其具有高可靠性、保护数据隐私的优势。然而，针对智能物联网终端平台资源受限、应用情境复杂多变，以及硬件优化能力不同等问题，亟需下列终端适配的深度计算方法。

资源适配深度计算 物联网的移动嵌入式终端资源（计算、存储和电量）通常较为受限，难以直接运行复杂的深度计算模型。因此，需基于深度计算模型的冗余性机理、平台资源约束以及物联网应用性能需求，探索不同深度模型压缩技术（如权重剪枝[73]、卷积分解[74]轻量化层结构替换[75]、量化[76]等）和超参数对不同深度模型精度、存储量、计算量、时延和能耗的影响，从而按需选择合适的压缩算法及超参数组合，以较少的精度损失并实现最低的终端资源消耗[77]-[78]。

情境自适深度计算 除了上述物联网平台资源约束以外，物联网终端运行深度模型还受综合情境因素影响，例如计算资源的动态性、输入数据的异质性，以及应用性能需求的差异性[7][79]。因此，需探索情境自适应的深度计算模型生成方法。近几年有一些相关研究进展，如自动化深度模型架构搜索（Neural Architecture Search, NAS）[80]，它采用合理的搜索空间、搜索策略和评价预估方法，可在不同情境需求下众多超参数和网络结构参数产生的爆炸性组合中完成自动搜索。

软硬协同深度计算 与深度模型算法层面的优化相结合，硬件优化通过合理利用不同设备的硬件性能和架构，可进一步实现深度模型加速。由于芯片内存带宽是十分受限的资源，因此将处理器性能与芯片内外存流量联系起来的模型可以指导软硬协同优化。例如，Roofline 模型[81]就是一个易于理解、可视化的性能模型。在资源极度受限的终端平台（如微控制器）上，软硬协同深度计算优化尤为重要。例如，Lin 等人[82]提出的面向微控制器的深度计算框架 MCUNet，通过联合设计一个高效神经网络架构和轻量级推理引擎，在微控制器上实现深度计算推理。

4.5 物联网分布式学习

在智能物联网时代，将会存在大量具有感知和计算能力的智能体，虽然单智能体数据和经验有限，但通过群体分布式协作可实现超越个体行为的集体智慧，构建具有自组织、自学习、自适应等能力的智能感知计算空间。

群体分布式学习模型 需基于生物群体交互式学习机理，探索融合协作、博弈、竞争、对抗等特征的群智能体分布式学习模型。此外，还要探索在单智能体数据有限且隐私要求高的情况下的可信群智学习方法。针对智能物联网分布式学习问题，近期有一些相关的研究进展，如联邦学习[83]、同伴学习[84]、多智能体深度强化学习[85]等。联邦学习的思想由谷歌最先提出[86]，它基于分布在多个设备上的数据集构建机器学习模型，在保障数据交换隐私安全的前提下，通过多设备协作开展高效率学习实现群体增强。作为一种分布式协同人工智能方法，联邦学习允许在分布式物联网设备上进行人工智能训练，已经在智慧健康、智能交通、无人机协作、智慧城市、智能制造等领域得到了应用[87][88][89]。

群智能体分布式决策 多智能体深度强化学习（Multi-Agent Deep Reinforcement Learning）利用智能体间的协作和博弈激发新的智能，产生智能行为决策，是机器学习领域的一个新兴的研究热点，并广泛应用于自动驾驶、路径规划、任务分配、集群编队、博弈对抗等现实领域，具有极高的研究价值和意义[90]-[91]。例如谷歌 DeepMind 在《科学》杂志上发表的论文[92]中通过让智能体在多玩家电子游戏中掌握策略、理解战术以及进行团队协作，展示了多智能体强化学习领域的最新进展。斯坦福大学的研究人员针对城市复杂交通网络的自适应交通信号控制问题，提出了一种新的基于多智能体深度强化学习的方法[91]。

群智能体知识迁移 由于云中心统一训练的模型与多样化边端部署环境之间的数据分布差异问题，所以会导致 AI 算法在实际部署中性能下降。域自适应（Domain Adaptation）方法[93]把分布不同的源域和目标域的数据，映射到一个特征空间中，使其在该空间中的距离尽可能近，可解决训练样本和测试样本概率分布不一致的学习问题[94]-[95]。例如，三星公司研究人员针对智能无线感知模型易受环境差异影响的问题，提出了一种基于域自适应的方法以适应不同的部署环境。元学习（Meta Learning）通过融合多个富经验智能体的训练模型来指导新的或缺少知识的智能体快速学习和成长，实现群智能体间的知识迁移和共享[96]。例如，西北工业大学和阿里巴巴的研究人员提出了一种基于多城市知识融合与元学习迁移的城市商业选址预测方法[97]。此外，加州大学伯克利分校的研究人员就机器人间的技能迁移提出了一种模块化的神经网络架构，具有更好的任务和情境泛化能力[98]。

4.6 云边端协同计算

云边端协同是智能物联网体系架构的重要特征。随着万物互联时代到来，海量数据和计算需求呈爆炸式增长，边缘设备大量部署，终端处理能力增强，因此将部分计算从云端下沉到边缘和终端可有效缓解云计算负载，产生更快的服务响应[30]。

云边协同计算 云边协同计算模式将大规模数据和复杂运算在云端集中处理，将小规模实时计算在边缘侧就近处理，从而提升数据传输性能，减少处理时延，保护数据隐私。云边协同的深度计算模式在视频实时处理、目标检测与追踪等复杂推理任务中应用较多，可分为边缘特征提取和云端深度识别两阶段[99]。此外，教师-学生模型的知识蒸馏[100]、深度计算模型的在线重训练[101]等人工智能学习任务的部署也都多采用云边协同计算模式[29]，可针对实际应用环境、数据分布等变化持续更新和提升边缘模型的知识与能力[102]。

边端协同计算 终端智能计算是智能物联网发展的一个重要趋势。针对单个智能终端计算资源不足的问题，可尝试由周边共存的多个移动、可穿戴或边缘设备等组成动态协作群。研究群智能体自组织协作高效计算模式，能根据性能需求（如时延、精度）和运行环境（如网络传输、能耗情况等），将原始任务进行自动“切分”并优选和调度合适的智能体协同完成感知计算任务。包括基于不同深度模型分割策略的串行[103]、并行[104]和混合协同计算模式[105]。此外，基于物联网中的分布式感知数据特点、边端通信及边缘服务器负载约束等实际因素[106]-[107]，需进一步研究综合性能更优的边端协调计算方法。

云边端协同性能优化 在上述技术基础上，需进一步结合复杂任务需求、部署环境和实时应用情境，探索云边端协同的高效计算任务分配、资源调度和负载平衡等方法，进一步提升和优化智能物联网系统云边端协同计算的整体效能[108]-[109]。此外，日益庞大而丰富的人工智能算法模型如何在智能物联网的云、边、端环境中进行有效部署和及时执行，且能够适应边端环境的复杂性、多样性和动态性也是一个关键的科学挑战问题[14][110]。

4.7 安全与隐私保护

智能物联网时代安全和隐私保护问题体现在多个方面。由于智能物联网终端在智能家庭、医院和城市中无处不在，在数据汇聚和处理过程中可收集大量的 AIoT 用户敏感数据（如面部图像、声音、动作、脉搏、图像数据等），存在数据窃取、误用和滥用的风险[111]。此外，在硬件层面，随着物联网的普及，少量未经严格认证、存在安全隐患的设备加入网络，也会威胁到其他联网设备的安全；其高度分散、随机加入退出的特性和分布式环境很难实施传统集中式信任认证[112]。最后，人工智能的应用使得在数据处理和算法训练/执行过程中也可能被攻击而泄露隐私或产生错误结果[113]。

数据安全保护立法 针对物联网数据安全问题，立法是一个重要途径，欧盟 2018 年起实施《通用数据保护条例》，该条例赋予个人对其个人数据的控制权。个人数据的控制器和处理器必须采取适当的措施来保护数据安全和隐私。我国于 2021 年出台《中华人民共和国数据安全法》，建立数据分类分级保护制度，禁止窃取或者以其他非法方式获取数据。

AIoT 安全保护策略 在技术层面，AIoT 的底层架构在泛在感知的物联网环境下，需研究分布式信任管理以提升 AIoT 交互的可靠性。与区块链技术结合来构建去中心信任管理是一个重要途径，通过数据和行为溯源，确保数据一致性和可靠性，保护数据隐私[114]。第二，由于涉及到云-边-端分层体系结构，需探索如何进行跨域和跨组织认证以提高云－边和多边协同的安全性。第三，由于不同用户或数据对安全和隐私保护需求的强度不同，可探索分级多粒度隐私保护策略[115]。

AI 算法应用安全 随着 AI 算法在智能物联网系统中的大量应用，也带来了新的安全性威胁问题。在数据收集、模型训练、模型测试以及系统部署等 AI 应用生命周期的不同阶段都可能引发安全与隐私泄露威胁[116]，如对抗攻击、数据投毒攻击和模型窃取攻击等。一方面应探索综合防御技术来应对实际应用场景中复杂的威胁，另一方面应从人工智能模型的可解释性等理论角度出发 **Error! Reference source not found.**，增强模型的泛化能力和鲁棒性，从根本上解决人工智能模型所面临的安全问题。

5 未来研究方向

作为物联网技术的最新演进方向，人工智能与物联网的融合已经在感知、网络、计算和推理等方面取得了许多进展。云边端协同的 AIoT 体系架构成为学术界和产业界的共识。多模态智能感知、群智感知等技术在自动驾驶、智能家居、智能工厂等领域得到有效应用。人工智能技术在 6G 等未来物联网通信技术发展中也成为主要驱动要素。物联网终端智能、边端协同计算、物联网联邦学习等成为当前研究热点并在 AIoT 应用中落地使用。然而，作为一个新兴领域，智能物联网发展还面临很多挑战，下面简要阐述智能物联网的未来研究方向。

5.1 软硬协同终端智能

尽管终端智能计算具有众多优势，但目前仍存在不足。（1）终端本地的智能计算具有隐私性高和稳定性强的优势，然而移动终端的计算、存储、电量资源十分受限，只能执行轻量级的智能计算模型。实现轻量化智能计算的方式包括深度模型压缩、量化和知识蒸馏等技术。然而，量化技术当前仍难以支持混合位宽[78][118]，深度模型压缩技术仍需依赖于模型重训练以修正模型精度。（2）移动终端设备的计算、存储等资源存在异构性和动态性[79]。因此，需要考虑终端设备的多级资源约束以及资源演化，并实现运行时自适应智能计算。

软硬协同的终端智能技术有助于打破硬件性能瓶颈，提升智能计算的综合性能[119]。近年来软硬协同的终端智能技术迅速发展，涌现出兼容多种移动嵌入式设备的深度计算框架、加速器，以及 FPGA 专用定制化芯片等[120]-[121]。然而，终端智能计算的研究起步较晚，仍然需要更多的探索和研究。（1）缺少成熟精准的硬件相关性能（如时延、能耗）评估方法[122]-[123]，在自动化深度模型架构搜索、自适应智能计算演化等框架中都需要评估智能计算方法的性能从而做出最优的搜索决策。然而，由于硬件相关性能受硬件资源影响难以推算、评估难度大，目前往往是基于回归、建表的方式构建时延、能耗预测模型，当更换终端平台或模型拓扑之后，便无法准确预估模型的硬件性能。（2）软硬协同的工程实现难度较高，普通研究人员往往需要较长时间才能胜任，缺少成熟的软件框架和硬件架构。

除此之外，基于物联网具体场景，设计定制化的智能芯片，能在大幅提升性能的同时，降低功耗和成本，满足异构物联网终端的需求[124]-[125]。传统的 CPU 和 GPU 芯片采用基于指令流的冯诺依曼式计算架构运行，未来的 AIoT 芯片将更多从脑科学或认知科学中汲取智慧，设计类脑 AI 芯片，实现成本、功耗、算力、推理能力等多样化需求之间的完美平衡[126]-[127]。

5.2 面向AIoT的智能演进

2017 年，国家发布《新一代人工智能规划》，对人工智能 2.0 时代进行规划和展望。AIoT 作为人工智能与物联网融合的产物，前沿人工智能理论和技术的应用具有重要意义和价值。

在 AIoT 应用场景下，集中式智能（如云计算）和分布式智能（如边端协同、多终端协同的智能计算）模式都发挥了重要的作用。集中式智能架构简单、云计算资源充足，边界清晰易于控制，然而存在网络链接丢失和泄露数据隐私性等缺点。分布式智能则充分利用物联网中的各级可用资源，靠近物联网终端及其边缘的智能计算提供多种选择性，但存在终端资源受限、分布式协同计算架构基础薄弱、设备资源异构性以及设备动态加入退出的不确定性等问题[11][15][128]。因此，综合二者优势的“分布&集中”混合式智能将成为未来有潜力的发展方向。

目前在 AIoT 中得到广泛应用的深度学习方法是基于大规模训练数据进行学习和获取知识的过程。除此之外，人类还具有根据已有知识进行推理的能力，而赋予 AIoT 推理能力对于做出准确和可解释的决策非常重要[129]-[130]。知识推理是指在已有知识的基础上推断出新知识的过程，基于知识图谱的推理方法近年来成为该领域研究热点[131]。知识图谱是一种在图中表示知识的结构化方法，可刻画实体之间的复杂语义关系。著名的知识图谱包括

WordNet①、DBpedia②、YAGO③等，已通过知识推理在许多应用程序中构建和使用。因果关系是一种特殊的知识，它描述的是系统中一个事件和第二个事件之间的作用关系，在许多 AIoT 应用中都很有用，智慧城市交通管理、自动驾驶安全策略等。最近基于深度神经网络的因果推理方法取得了很大进展，将在未来 AIoT 系统中发挥重要作用[132]-[134]。

通用人工智能算法的研究将为AIoT环境动态、情境复杂、场景多样等挑战问题提供解决途径[135]-[136]。深度学习模型的性能在很大程度上取决于大规模训练数据。然而，人类学习新概念不仅基于数据，还基于先验知识[137][138]。例如，人具有基于相关知识举一反三的联想学习能力，具有基于历史经验知识的自我纠错和提升能力，以及基于长期知识积累的思维演化能力。这些在现有的通用人工智能中都尚未提供支持。同样，先验知识对于以数据有效的方式训练深度学习模型非常有用。例如，基于数据在特征空间的条件分布知识引导训练过程，可以有助于解决不同训练阶段中数据异构性所带来的模型偏移和性能退化问题[139]-[140]。因此，数据和知识的融合对于提高 AIoT 的感知、学习、推理和行为非常重要。

5.3 新一代智能物联网络

海量的物联网设备量和巨大的系统规模，决定了未来智能物联网系统的终极形态是完全自主化的。传统物联网应用大多都是状态监测、远程控制等具有单一功能的形式，应用范围受限，智能程度低；而未来智能物联网的应用形式应是多功能集成，智能程度高。智能物联网可对整个系统进行实时监测，能够在开放的环境中持续学习、演化，从而不断满足用户个性化的需求，提升服务质量。

软件定义网络（SDN）和网络功能虚拟化（NFV）[141]，将现代通信网络转变为基于软件的虚拟网络。随着智能物联网的出现，网络变得越来越复杂，设备越来越异构，需要超越软件化的网络，实现智能化的体系结构。SDN 将控制平面与数据平面分离，逻辑集中的网络控制器能够从网络协议栈的不同层获取数据。应用 AI 技术，网络控制器可容易地做出最佳决策，使网络更易于控制和管理[142]。此外，为了支持智能物联网应用，网络实体不仅需要支持传统通信、内容缓存、无线传输等功能，还需要支持更先进的物联网功能，包括传感、数据收集、分析和存储。基于 AI 的方法可以实现快速学习和自适应，使网络变得智能、灵活，并能够根据不断变化的网络动态进行学习和调整。

在通信方面，物联网作为由大量无线设备组成系统，基于 AI 技术，可以解决无线通信随机接入、频谱接入和频谱感知等相关挑战，为从发射机到接收机的整个物理层链路端到端优化提供了可能[22][143]通过结合先进的传感和数据采集、人工智能技术和特定领域的信号处理方法，端到端系统能够自我学习和自我优化。智能网络协议的设计，需要运行复杂的机器学习模型，而物联网节点的资源约束给协议设计和实现带来了挑战。如何降低机器学习模型的复杂性或以分布方式实现，从而在资源受限设备上实现；如何实现大规模数据收集和传输，通过边缘分析和云计算增强整个系统的智能性，都是未来值得研究的方向[144]。

5.4 动态场景模型持续演化

与传统云端智能不同，智能物联网环境下学习模型的部署存在场景多变、领域差异、训练困难等问题，需要进一步探索动态场景的模型持续演化方法，包括动态场景下的云边协同模型演化和仿真-真实结合的模型训练与迁移。

随着智能物联网时代到来，智能计算也逐渐从云端下放至边端，在可穿戴设备、移动设备、嵌入式设备上为人类带来更安全、便捷的智能服务。由于现实世界是动态且复杂的，传统的“云端训练，边端推理”的静态部署方式无法适应变化，可能会导致性能损失。一方面，由于数据标注难和不可预见性，云端不可能涵盖众多边端的数据分布，不同边端处于动态复杂域之下——光强、遮挡、目标大小、拍摄清晰度等多因素随时间发生变化，静态的边端模型将导致性能低于云端训练的预期[7][145]；另一方面，边端的设备资源情境（如电量、内存）和应用需求情境（如能耗、时延、精度）同样处于动态变化的状态，静态部署也会导致边端资源浪费或资源过耗的现象[79][146]。因此，云-边协同的深度模型自演化方案亟待提出，这与一些著名的机器学习研究课题有关，包括少样本学习[147]、元学习[148]、持续学习[149]领域自适应[93]等。

智能物联网需要在物理环境中进行部署，而在真实环境中训练模型是困难和昂贵的，例如自动驾

① https://wordnet.princeton.edu/

② https://www.dbpedia.org/

③ https://yago-knowledge.org/

驶、机器人导航等。由于在真实环境样本采样的复杂性和安全性，在仿真环境中训练模型再应用到真实场景成为有效的替代方案。但是在仿真环境中学习到的策略在现实世界中往往表现不佳，这种现实差距是由于仿真的物理系统与真实物理系统之间的模型差异造成的[150]-[151]。为了缩小现实差距，可以从以下两方面进行改进：一方面，优化仿真环境中的机器人感知模型和机器人动力学模型是缩小现实差距的直接途径，可以通过专家知识和采集真实环境数据构建高精度仿真环境[152]-[153]；另一方面，即使仿真环境具有较强的系统辨识性，现实世界也有难以建模的物理效应，可以利用域随机化、教师学生网络、课程学习等方法训练鲁棒的控制器策略[154]-[155]。

5.5 人机物融合群智计算

在泛在计算、智能物联网、群体智能、移动群智感知等发展背景下，人机物融合群智计算（CrowdHMT, Crowd Intelligence with the Deep Fusion of Human, Machine, and Things）[17][156]正成长为一种新的智能感知计算模式。它通过人、机、物异构群智能体的有机融合，利用其感知能力的差异性、计算资源的互补性、节点间的协作性和竞争性，构建具有自组织、自学习、自适应、持续演化等能力的智能感知计算空间，实现智能体个体技能和群体认知能力的提升[17]。

在人机物融合群智计算背景下，人机物群智协同机理[156]、人机混合智能[157]、人在回路智能计算[158]-[159]、异构群智能体协作增强[160]、情境敏感的自适应协同[24][79]、群智能体分布式学习[27]、群智能体知识迁移与持续演进[161]-[162]等将成为新的研究问题，“以人为中心”、“人机协同”的计算理念也对智能物联网发展带来新的机遇和挑战。

5.6 通用AIoT系统平台

随着人工智能产业的快速发展和物联网技术的不断普及，越来越多兼具人工智能属性的物联网平台与应用应运而生，其在提供智能物联服务的基础上，也覆盖了医疗、交通、工业、家居等多个领域。国内外知名 IT 企业也推出了各具特色的智能物联网平台，当前主流的智能物联网平台通常依托于现有的工业级云平台，如 AWS IoT、Azure IoT、华为云 IoT、阿里云 IoT 等。然而，当前主流的智能物联网平台通常依托于现有的工业级云平台，并在云平台之上构建各具特色的物联网和人工智能服务，要构建具有不同领域适用性的通用 AIoT 系统平台，还需要进一步考虑以下方面内容。

现有物联网平台大多利用其云平台上自带的智能属性完成，如基于神经网络的深度学习模型、云边端算力分配的任务机制。智能物联网不仅具有大规模异构感知节点、多种类通信网络并存、AI 算法的分布式部署以及领域关联多样化应用等特点，而且要面对多模态感知、泛在互联、场景动态、资源受限、实时处理、普适服务等技术挑战。为更好应对以上挑战，未来的通用 AIoT 系统平台应该能够具备“自组织、可配置、抽象化”等特征。一方面提供软/硬件系统协同运行和优化调度的支撑环境，通过打造了一系列硬件产品为智能物联网提供底层支持，并提供相关接口以屏蔽底层细节；另一方面则支持应用系统的开发、部署与管理等功能，提供设备互联互通中心平台，以及帮助开发者快速开发应用程序的各类中间件。

在此背景下，“万物皆可互联、一切均可编程、软件定义一切”将成为智能物联网时代平台发展的主趋势，其涵盖的内容包含通用的物联网操作系统、物联网通信协议、智能物联网中间件、领域物联网应用开发支持软件等[16][54][163]。软件定义的本质是“资源虚拟化，功能可编程”[164]。就前者而言，物联网操作系统构成 AIoT 平台接触末端资源的触角，实现大规模异质异构泛在物联网终端的资源管理是其重要任务，例如建立开源实时操作系统提升其可靠性和易用性。就后者而言，智能物联网中间件需屏蔽异构物联网系统的技术细节，支撑 AI 算法和 AIoT 应用的快速开发和灵活部署[165]。

此外，智能物联网由于异构设备泛在连接与互操作、多模态数据融合汇聚、边端协同计算、群智能体分布式智能等特征还面临很多新的安全和隐私问题，成为通用 AIoT 系统平台需解决的一大关键挑战[114]。区块链技术的出现为克服上述挑战带来了机遇。区块链本质上是一个分布在整个分布式系统上的分布式账本。通过去中心化共识，区块链可以使交易在互不信任的分布式系统中发生并得到验证，而无需受信任的第三方的干预。此外，保存在区块链中的每笔交易本质上是不可变的，因为网络中的每个节点都将所有已提交的交易保存在区块链中。区块链本质上是对物联网的完美补充，具有改进的互操作性、隐私性、安全性、可靠性和可扩展性。

6 总结

智能物联网在物联网感知、网络、应用三层架构的基础上进行扩充，利用人工智能技术和物联网泛在设备平台的感知、存储、计算和学习能力，以智能信息的高效、实时、智能处理为目标，基于云边端协同的 AIoT 体系架构实现感知、通信、计算和应用的智能化提升。本文阐述了云边端协同 AIoT 体系架构和 AIoT 系统软件平台基本构想，介绍了泛在智能感知、群智感知计算、群智物联网通信、终端适配深度计算、物联网分布式学习、云边端协同计算、安全与隐私保护几个层面的关键技术及其前沿探索。

未来，智能物联网研究需要更多的研究者共同参与，深入物联网系统应用问题研究、关键技术瓶颈突破以及通用性平台的凝练与研发。一方面需要在软硬协同终端智能、面向 AIoT 的智能演进、新一代智能物联网网络、动态场景模型持续演化、人机物融合群智计算等关键技术方面实现不断突破。另一方面，面对多模态感知、泛在互联、场景动态、资源受限、实时处理、普适服务等技术挑战，亟需要研发具有“自组织、可配置、抽象化”等特征的通用 AIoT 操作系统、中间件等系统平台，推动生态发展。

参 考 文 献

[1] Liu Y. Introduction to Internet of Things. Third Edition. Beijing: Science Press, 2017(in Chinese)
(刘云浩. 物联网导论. 第三版. 北京: 科学出版社, 2017)

[2] Zhang K, Leng S, He Y, et al. Mobile edge computing and networking for green and low-latency Internet of Things. IEEE Communications Magazine, 2018, 56(5): 39-45

[3] Wang X, Han Y, Leung V C M, et al. Convergence of edge computing and deep learning: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2020, 22(2): 869-904

[4] Zhang J, Tao D. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. IEEE Internet of Things Journal, 2020, 8(10): 7789-7817.

[5] Samie F, Bauer L, Henkel J. From cloud down to things: An overview of machine learning in internet of things. IEEE Internet of Things Journal, 2019, 6(3): 4921-4934

[6] Guo Bin. Artificial Intelligence in Internet of Things and Future Manufacturing - Embracing the Era of Crowd Intelligence with the Deep Fusion of Human, Machine, and Things. Renming Luntan·Xueshu Qianyan, 2020(13): 32-42(in Chinese)
(郭斌. 论智能物联与未来制造--拥抱人机物融合群智计算时代. 人民论坛• 学术前沿, 2020 (13): 32-42)

[7] Cai H, Gan C, Wang T, et al. Once-for-all: Train one network and specialize it for efficient deployment. arXiv preprint arXiv:1908.09791, 2019

[8] Guo P, Hu B, Hu W. Mistify: Automating {DNN} Model Porting for {On-Device} Inference at the Edge// Proceedings of the 18th USENIX Symposium on Networked Systems Design and Implementation. Virtual Event, 2021: 705-719,

[9] Gupta J K, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning// Proceedings of the International conference on autonomous agents and multiagent systems. Sao Paulo, Brazil, 2017: 66-83,

[10] Abdelfattah M S, Mehrotra A, Dudziak Ł, et al. Zero-cost proxies for lightweight NAS. arXiv preprint arXiv:2101.08134, 2021

[11] Zhang J, Letaief K B. Mobile edge intelligence and computing for the internet of vehicles. Proceedings of the IEEE, 2019, 108(2): 246-261

[12] Li T, Lai Z, Chen L, et al. The White Paper of Chinese Artificial Intelligence of Things. Internet economy, 2020(03):90-97(in Chinese)
(李天慈, 赖贞, 陈立群, 等. 2020 年中国智能物联网（AIoT）白皮书. 互联网经济, 2020(03): 90-97)

[13] Song H, Bai J, Yi Y, et al. Artificial intelligence enabled Internet of Things: Network architecture and spectrum access. IEEE Computational Intelligence Magazine, 2020, 15(1): 44-51

[14] Chen J, Ran X. Deep learning with edge computing: A review. Proceedings of the IEEE, 2019, 107(8): 1655-1674

[15] Lim W Y B, Luong N C, Hoang D T, et al. Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2020, 22(3): 2031-2063

[16] Mei H, Cao D, Xie T. Ubiquitous Operating System: Toward the Blue Ocean of Human-cyber-physical Ternary Ubiquitous Computing. Bulletin of Chinese Academy of Sciences, 2022(1): 30-37(in Chinese)
(梅宏, 曹东刚, 谢涛. 泛在操作系统: 面向人机物融合泛在计算的新蓝海. 中国科学院院刊, 2022 (1): 30-37.)

[17] Guo B, Yu Z. Crowd Intelligence with the Deep Fusion of Human, Machine, and Things. Communications of the China Computer Federation, Vol. 17 No. 2, 2021, pp. 36-41(in Chinese)
(郭斌, 於志文. 人机物融合群智计算. 中国计算机学会通讯, 2021,17(2): 36-41)

[18] Zhou Z, Wu C, Yang Z, et al. Sensorless sensing with WiFi. Tsinghua Science and Technology, 2015, 20(1): 1-6

[19] Liu J, Liu H, Chen Y, et al. Wireless sensing for human activity: A survey. IEEE Communications Surveys & Tutorials, 2019, 22(3):

1629-1645

[20] Wu D, Zhang D, Xu C, et al. Device-free WiFi human sensing: From pattern-based to model-based approaches. IEEE Communications Magazine, 2017, 55(10): 91-97

[21] Wang P, Guo B, Wang Z, et al. ShopSense: Customer Localization in Multi-person Scenario with Passive RFID Tags. IEEE Transactions on Mobile Computing, 2020, 21(5): 1812-1828

[22] Zhang C, Patras P, Haddadi H. Deep learning in mobile and wireless networking: A survey. IEEE Communications surveys & tutorials, 2019, 21(3): 2224-2287

[23] Yang H, Alphones A, Xiong Z, et al. Artificial-intelligence-enabled intelligent 6G networks. IEEE Network, 2020, 34(6): 272-280

[24] Liu S, Du J, Nan K, et al. AdaDeep: a usage-driven, automated deep model compression framework for enabling ubiquitous intelligent mobiles. IEEE Transactions on Mobile Computing, 2020, 20(12): 3282-3297

[25] Yao S, Zhao Y, Zhang A, et al. Deepiot: Compressing deep neural network structures for sensing systems with a compressor-critic framework//Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems. Delft, Netherlands . 2017: 1-14

[26] Guo B, Wu Y, Wang H, et al. Context-aware adaptation of deep learning models for IoT devices. SCIENTIA SINICA Informationis,2020,50(11):1629-1644(in Chinese)
(郭斌, 仵允港, 王虹力, 等. 深度学习模型终端环境自适应方法研究.中国科学:信息科学,2020,50(11):1629-1644)

[27] Warnat-Herresthal S, Schultze H, Shastry K L, et al. Swarm learning for decentralized and confidential clinical machine learning. Nature, 2021, 594(7862): 265-270

[28] Deng S, Zhao H, Fang W, et al. Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 2020, 7(8): 7457-7469

[29] Ren J, Yu G, He Y, et al. Collaborative cloud and edge computing for latency minimization. IEEE Transactions on Vehicular Technology, 2019, 68(5): 5031-5044

[30] Kai C, Zhou H, Yi Y, et al. Collaborative cloud-edge-end task offloading in mobile-edge computing networks with limited communication capability. IEEE Transactions on Cognitive Communications and Networking, 2020, 7(2): 624-634

[31] Liu L, Li H, Gruteser M. Edge assisted real-time object detection for mobile augmented reality// Proceedings of the 25th Annual International Conference on Mobile Computing and Networking. ,Los Cabos, Mexico. 2019: 1-16

[32] Sucar E, Liu S, Ortiz J, et al. imap: Implicit mapping and positioning in real-time//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6229-6238

[33] Zhang W, He Z, Liu L, et al. Elf: accelerate high-resolution mobile deep vision with content-aware parallel offloading//Proceedings of the 27th Annual International Conference on Mobile Computing and Networking. New Orleans, USA, 2021: 201-214

[34] Zhang Q, Sun H, Wu X, et al. Edge video analytics for public safety: A review. Proceedings of the IEEE, 2019, 107(8): 1675-1696

[35] Lane N D, Georgiev P, Qendro L. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning//Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing. Osaka, Japan. 2015: 283-294

[36] Zhang Z, Geiger J, Pohjalainen J, et al. Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST), 2018, 9(5): 1-28

[37] Purwins H, Li B, Virtanen T, et al. Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(2): 206-219

[38] Wang Z, Guo B, Yu Z, et al. Wi-Fi CSI-based behavior recognition: From signals and actions to activities. IEEE Communications Magazine, 2018, 56(5): 109-115

[39] Niu, K, Wang, X, Zhang, F, et al. Rethinking Doppler effect for accurate velocity estimation with commodity WiFi devices. IEEE Journal on Selected Areas in Communications， 2022, 40(7): 2164-2178

[40] Dan W, Youwei Zeng, Fusang Z, and Daqing Z. WiFi CSI-based device-free sensing: from Fresnel zone model to CSI-ratio model. CCF Transactions on Pervasive Computing and Interaction (2021): 1-15.

[41] Li X, Li S, Zhang D, et al. Dynamic-music: accurate device-free indoor localization//Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing. Heidelberg ,Germany. 2016: 196-207

[42] Li X, Zhang D, Lv Q, et al. IndoTrack: Device-free indoor human tracking with commodity Wi-Fi. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017, 1(3): 1-22

[43] Wang W, Liu A X, Shahzad M, et al. Device-free human activity recognition using commercial WiFi devices. IEEE Journal on Selected Areas in Communications, 2017, 35(5): 1118-1131

[44] Bharti P, De D, Chellappan S, et al. HuMAn: Complex activity recognition with multi-modal multi-positional body sensing. IEEE Transactions on Mobile Computing, 2018, 18(4): 857-870

[45] Wellhausen L, Ranftl R, Hutter M. Safe robot navigation via multi-modal anomaly detection. IEEE Robotics and Automation Letters, 2020, 5(2): 1326-1333

[46] Huang Z, Lv C, Xing Y, et al. Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sensors Journal, 2020, 21(10): 11781-11790

[47] Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1341-1360

[48] Xu H, Yang Z, Zhou Z, et al. Indoor localization via multi-modal sensing on smartphones//Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg, Germany. 2016: 208-219

[49] Prakash A, Chitta K, Geiger A. Multi-modal fusion transformer for end-to-end autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Beijing ,China,2021: 7077-7087

[50] Liu Y. Crowd Sensing and Computing. Communications of the China Computer Federation, 2012, 8(10): 38-41(in Chinese)
(刘云浩. 群智感知计算. 中国计算机学会通讯, 2012, 8(10): 38-41)

[51] Yu Z, Guo B, Wang L. Crowd Sensing and Computing. Beijing: Tsinghua University Press, 2021(in Chinese)
(於志文, 郭斌, 王亮. 群智感知计算. 北京: 清华大学出版社, 2021)

[52] Ma H, Zhao D, Yuan P. Opportunities in mobile crowd sensing. IEEE Communications Magazine, 2014, 52(8): 29-35

[53] Guo B, Wang Z, Yu Z, et al. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. ACM computing surveys (CSUR), 2015, 48(1): 1-31

[54] Yu Z, Ma H, Guo B, et al. Crowdsensing 2.0. Communications of the ACM, 2021, 64(11): 76-80

[55] Liu Y, Guo B, Wang Y, et al. TaskMe: Multi-task allocation in mobile crowd sensing//Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing. Heidelberg, Germany.2016: 403-414

[56] Guo B, Liu Y, Wang L, et al. Task allocation in spatial crowdsourcing: Current state and future directions. IEEE Internet of Things Journal, 2018, 5(3): 1749-1764

[57] Zhao D, Ma H, Liu L. Frugal online incentive mechanisms for mobile crowd sensing. IEEE Transactions on Vehicular Technology, 2016, 66(4): 3319-3330

[58] Zhang X, Yang Z, Sun W, et al. Incentives for mobile crowd sensing: A survey. IEEE Communications Surveys & Tutorials, 2015, 18(1): 54-67

[59] Guo B, Chen H, Han Q, et al. Worker-contributed data utility measurement for visual crowdsensing systems. IEEE Transactions on Mobile Computing, 2016, 16(8): 2379-2391

[60] He Y, Ren J, Yu G, et al. Importance-aware data selection and resource allocation in federated edge learning system. IEEE Transactions on Vehicular Technology, 2020, 69(11): 13593-13605

[61] Zhao L, Zhao W, Hawbani A, et al. Novel online sequential learning-based adaptive routing for edge software-defined vehicular networks. IEEE Transactions on Wireless Communications, 2020, 20(5): 2991-3004

[62] Di Valerio V, Presti F L, Petrioli C, et al. CARMA: Channel-aware reinforcement learning-based multi-path adaptive routing for underwater wireless sensor networks. IEEE Journal on Selected Areas in Communications, 2019, 37(11): 2634-2647

[63] Xiao Xu, You L, and Qiming F. Applying Graph Neural Network in Deep Reinforcement Learning to Optimize Wireless Network Routing//2021 Ninth International Conference on Advanced Cloud and Big Data (CBD). Xi'an, China , 2022, 218-223.

[64] Xiaohui N, Youjian Z, Zhihan L, et al. Dynamic TCP initial windows and congestion control schemes through reinforcement learning. IEEE Journal on Selected Areas in Communications 2019,37(6): 1231-1247.

[65] Jay N, Rotman N, Godfrey B, et al. A deep reinforcement learning perspective on internet congestion control// Proceedings of the International Conference on Machine Learning, Lugano, Switzerland. 2019: 3050-3059

[66] Xu Z, Tang J, Yin C, et al. Experience-driven congestion control: When multi-path TCP meets deep reinforcement learning. IEEE Journal on Selected Areas in Communications, 2019, 37(6): 1325-1336

[67] Chen X, Wu C, Chen T, et al. Age of information aware radio resource management in vehicular networks: A proactive deep reinforcement learning perspective. IEEE Transactions on wireless communications, 2020, 19(4): 2268-2281

[68] Calabrese F D, Wang L, Ghadimi E, et al. Learning radio resource management in RANs: Framework, opportunities, and challenges. IEEE Communications Magazine, 2018, 56(9): 138-145

[69] Naderializadeh N, Sydir J J, Simsek M, et al. Resource management in wireless networks via multi-agent deep reinforcement learning. IEEE Transactions on Wireless Communications, 2021, 20(6): 3507-3523

[70] Zhang H, Zhang H, Long K, et al. Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation. IEEE Transactions on Network Science and Engineering, 2020, 7(4): 2406-2415

[71] Al-Garadi M A, Mohamed A, Al-Ali A K, et al. A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Communications Surveys & Tutorials, 2020, 22(3): 1646-1685

[72] Roopak M, Tian G Y, Chambers J. Deep learning models for cyber security in IoT networks// Proceedings of the 2019 IEEE 9th annual computing and communication workshop and conference (CCWC). Las Vegas, USA, 2019: 0452-0457

[73] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149, 2015

[74] Liu B, Wang M, Foroosh H, et al. Sparse convolutional neural networks//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, Massachusetts .2015: 806-814

[75] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016

[76] Zhou Y, Moosavi-Dezfooli S M, Cheung N M, et al. Adaptive

quantization for deep neural network//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA. 2018, 32(1): 4596-4604

[77] Liu S, Lin Y, Zhou Z, et al. On-demand deep model compression for mobile devices: A usage-driven model selection framework//Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services. Munich, Germany .2018: 389-400

[78] Wang K, Liu Z, Lin Y, et al. Haq: Hardware-aware automated quantization with mixed precision//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8612-8620

[79] Liu S, Guo B, Ma K, et al. AdaSpring: Context-adaptive and Runtime-evolutionary Deep Model Compression for Mobile Applications. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, online,2021, 5(1): 1-22

[80] Tan M, Chen B, Pang R, et al. Mnasnet: Platform-aware neural architecture search for mobile//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA. 2019: 2820-2828

[81] Williams S, Waterman A, Patterson D. Roofline: an insightful visual performance model for multicore architectures. Communications of the ACM, 2009, 52(4): 65-76

[82] Lin J, Chen W M, Lin Y, et al. Mcunet: Tiny deep learning on iot devices. Advances in Neural Information Processing Systems, 2020, 33: 11711-11722

[83] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19

[84] Zhang Y, Xiang T, Hospedales T M, et al. Deep mutual learning//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA. 2018: 4320-4328,

[85] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 2019, 575(7782): 350-354

[86] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 2019, 1: 374-388

[87] Nguyen D C, Ding M, Pathirana P N, et al. Federated learning for internet of things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1622-1658

[88] Zheqi Z, Shuo W, Pingyi F, and Khaled B. Letaief. Federated multiagent actor–critic learning for age sensitive mobile-edge computing. IEEE Internet of Things Journal 9, no. 2 (2021): 1053-1067.

[89] Jed M, Jia H, and Geyong Min. Multi-task federated learning for personalised deep neural networks in edge computing. IEEE Transactions on Parallel and Distributed Systems, 2021,33(3): 630-641.

[90] Wang L, Wang K, Pan C, et al. Multi-agent deep reinforcement learning-based trajectory planning for multi-UAV assisted mobile edge computing. IEEE Transactions on Cognitive Communications and Networking, 2020, 7(1): 73-84

[91] Chu T, Wang J, Codecà L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(3): 1086-1095

[92] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. nature, 2016, 529(7587): 484-489

[93] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, USA.2017: 7167-7176.

[94] Yang J, Zou H, Cao S, et al. MobileDA: Toward edge-domain adaptation. IEEE Internet of Things Journal, 2020, 7(8): 6909-6918

[95] Li H, Chen X, Wang J, et al. DAFI: WiFi-based Device-free Indoor Localization via Domain Adaptation. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(4): 1-21

[96] Vanschoren, J. Meta-learning: A survey. arXiv preprint arXiv:1810.03548, 2018.

[97] Liu Y, Guo B, Zhang D, et al. MetaStore: a task-adaptative meta-learning model for optimal store placement with multi-city knowledge transfer. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(3): 1-23

[98] Devin C, Gupta A, Darrell T, et al. Learning modular neural network policies for multi-task and multi-robot transfer// Proceedings of the 2017 IEEE international conference on robotics and automation (ICRA). Singapore, 2017: 2169-2176

[99] Ding C, Zhou A, Liu Y, et al. A cloud-edge collaboration framework for cognitive service. IEEE Transactions on Cloud Computing, 2020, 10(3): 1489-1499

[100] Mirzadeh S I, Farajtabar M, Li A, et al. Improved knowledge distillation via teacher assistant//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA. 2020, 34(04): 5191-5198

[101] Kolcun R, Popescu D A, Safronov V, et al. The case for retraining of ML models for IoT device identification at the edge. arXiv preprint arXiv:2011.08605, 2020

[102] Bhardwaj R, Xia Z, Ananthanarayanan G, et al. Ekya: Continuous learning of video analytics models on edge compute servers// Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22). Renton, USA. 2022: 119-135

[103] Zeng L, Li E, Zhou Z, et al. Boomerang: On-demand cooperative deep neural network inference for edge intelligence on the industrial Internet of Things. IEEE Network, 2019, 33(5): 96-103

[104] Mao J, Chen X, Nixon K W, et al. Modnn: Local distributed mobile computing system for deep neural network// Proceedings of the

Design, Automation & Test in Europe Conference & Exhibition. Lausanne, Switzerland. 2017: 1396-1401

[105] Hadidi R, Cao J, Woodward M, et al. Distributed perception by collaborative robots. IEEE Robotics and Automation Letters, 2018, 3(4): 3709-3716

[106] Li Y, Padmanabhan A, Zhao P, et al. Reducto: On-camera filtering for resource-efficient real-time video analytics//Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication. USA. 2020: 359-376

[107] Zeng X, Fang B, Shen H, et al. Distream: scaling live video analytics with workload-adaptive distributed edge intelligence//Proceedings of the 18th Conference on Embedded Networked Sensor Systems. Japan. 2020: 409-421

[108] Zhang S, Li Y, Liu X, et al. Towards real-time cooperative deep inference over the cloud and edge end devices. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4(2): 1-24

[109] Teerapittayanon S, McDanel B, Kung H T. Distributed deep neural networks over the cloud, the edge and end devices// Proceedings of the 2017 IEEE 37th international conference on distributed computing systems (ICDCS). Atlanta, USA. 2017: 328-339

[110] Wang X, Chen N, Zhang R. Intelligent adaptive edge systems：exploration and open issues. Chinese Journal on Internet of Things, 2021, 5(1): 1-10
王旭, 陈南希, 张柔佳. 智能自适应边缘系统: 探索与挑战. 物联网学报 5.1: 1-10

[111] Hu Y, Kuang W, Qin Z, et al. Artificial intelligence security: threats and countermeasures. ACM Computing Surveys (CSUR), 2021, 55(1): 1-36

[112] Xiao L, Wan X, Lu X, et al. IoT security techniques based on machine learning: How do IoT devices use AI to enhance security?. IEEE Signal Processing Magazine, 2018, 35(5): 41-49

[113] Chio C, Freeman D. Machine learning and security: Protecting systems with data and algorithms. USA:O'Reilly Media, Inc., 2018

[114] Dai H N, Zheng Z, Zhang Y. Blockchain for Internet of Things: A survey. IEEE Internet of Things Journal, 2019, 6(5): 8076-8094

[115] Zheng S, Apthorpe N, Chetty M, et al. User perceptions of smart home IoT privacy. Proceedings of the ACM on human-computer interaction, 2018, 2: 1-20

[116] Finlayson S G, Bowers J D, Ito J, et al. Adversarial attacks on medical machine learning. Science, 2019, 363(6433): 1287-1289

[117] Samek W, Montavon G, Vedaldi A, et al. Explainable AI: interpreting, explaining and visualizing deep learning. Lecture Notes in Computer Science. 2019.

[118] Wu B, Wang Y, Zhang P, et al. Mixed precision quantization of convnets via differentiable neural architecture search. arXiv preprint arXiv:1812.00090, 2018

[119] Deng L, Li G, Han S, et al. Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE, 2020, 108(4): 485-532

[120] Imani M, Razlighi M S, Kim Y, et al. Deep learning acceleration with neuron-to-memory transformation// Proceedings of the 2020 IEEE international symposium on high performance computer architecture. San Diego, USA. 2020: 1-14

[121] Daghaghi S, Meisburger N, Zhao M, et al. Accelerating slide deep learning on modern cpus: Vectorization, quantizations, memory optimizations, and more. Proceedings of Machine Learning and Systems, 2021, 3: 156-166

[122] Zhang L L, Han S, Wei J, et al. Nn-Meter: Towards accurate latency prediction of deep-learning model inference on diverse edge devices//Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. Virtual Event. 2021: 81-93

[123] Xu Z, Zhao L, Liang W, et al. Energy-aware inference offloading for DNN-driven applications in mobile edge clouds. IEEE Transactions on Parallel and Distributed Systems, 2020, 32(4): 799-814

[124] Whatmough P N, Lee S K, Brooks D, et al. DNN engine: A 28-nm timing-error tolerant sparse deep neural network processor for IoT applications. IEEE Journal of Solid-State Circuits, 2018, 53(9): 2722-2731

[125] Wang J, Wang X, Eckert C, et al. A 28-nm compute SRAM with bit-serial logic/arithmetic operations for programmable in-memory vector computing. IEEE Journal of Solid-State Circuits, 2019, 55(1): 76-86

[126] Pei J, Deng L, Song S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. Nature, 2019, 572(7767): 106-111

[127] Zhang W, Gao B, Tang J, et al. Neuro-inspired computing chips. Nature electronics, 2020, 3(7): 371-382

[128] Bloembergen D, Tuyls K, Hennes D, et al. Evolutionary dynamics of multi-agent learning: A survey. Journal of Artificial Intelligence Research, 2015, 53: 659-697

[129] Tran T, Le V, Le H, et al. From deep learning to deep reasoning//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Singapore. 2021: 4076-4077

[130] Li O, Liu H, Chen C, et al. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions//Proceedings of the AAAI Conference on Artificial Intelligence. California, USA. 2018, 32(1): 3530-3537

[131] Lin B Y, Chen X, Chen J, et al. Kagnet: Knowledge-aware graph networks for commonsense reasoning. arXiv preprint arXiv:1909.02151, 2019

[132] Luo Y, Peng J, Ma J. When causal inference meets deep learning. Nature Machine Intelligence, 2020, 2(8): 426-427

[133] Cui P, Athey S. Stable learning establishes some common ground between causal inference and machine learning. Nature Machine Intelligence, 2022, 4(2): 110-115

[134] Schölkopf B, Locatello F, Bauer S, et al. Toward causal representation learning. Proceedings of the IEEE, 2021, 109(5): 612-634

[135] Adams S, Arel I, Bach J, et al. Mapping the landscape of human-level artificial general intelligence. AI magazine, 2012, 33(1): 25-42

[136] Hu B, Guan Z H, Chen G, et al. Neuroscience and network dynamics toward brain-inspired intelligence. IEEE Transactions on Cybernetics, 2021, 52(10): 10214-10227

[137] Brod, G. Toward an understanding of when prior knowledge helps or hinders learning. npj Science of Learning 6, no. 1, 2021: 1-3.

[138] Raina R, Alexis B, Honglak L, et al. Self-taught learning: transfer learning from unlabeled data. Proceedings of the 24th international conference on Machine learning. Corvalis, USA, 2007, 759-766

[139] Li L, Hoyer S, Pederson R, et al. Kohn-Sham equations as regularizer: Building prior knowledge into machine-learned physics. Physical review letters, 2021, 126(3): 036401

[140] Chen D, Bai Y, Ament S, et al. Automating crystal-structure phase mapping by combining deep learning with constraint reasoning. Nature Machine Intelligence, 2021, 3(9): 812-822

[141] Nguyen V G, Brunstrom A, Grinnemo K J, et al. SDN/NFV-based mobile packet core network architectures: A survey. IEEE Communications Surveys & Tutorials, 2017, 19(3): 1567-1602

[142] Mao Q, Hu F, Hao Q. Deep learning for intelligent wireless networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2018, 20(4): 2595-2621

[143] Letaief K B, Chen W, Shi Y, et al. The roadmap to 6G: AI empowered wireless networks. IEEE communications magazine, 2019, 57(8): 84-90

[144] Xiao Y, Shi G, Li Y, et al. Toward self-learning edge intelligence in 6G. IEEE Communications Magazine, 2020, 58(12): 34-40

[145] Zheng Z, Wang Y, Dai Q, et al. Metadata-driven Task Relation Discovery for Multi-task Learning// Proceedings of the IJCAI. Macao ,China. 2019: 4426-4432

[146] Chen Q, Zheng Z, Hu C, et al. On-edge multi-task transfer learning: Model and practice with data-driven task allocation. IEEE Transactions on Parallel and Distributed Systems, 2019, 31(6): 1357-1371

[147] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 2020, 53(3): 1-34

[148] Hospedales T, Antoniou A, Micaelli P, et al. Meta-learning in neural networks: A survey. arXiv preprint arXiv:2004.05439, 2020

[149] Mitchell T, Cohen W, Hruschka E, et al. Never-ending learning. Communications of the ACM, 2018, 61(5): 103-115

[150] Peng X B, Andrychowicz M, Zaremba W, et al. Sim-to-real transfer of robotic control with dynamics randomization// Proceedings of the 2018 IEEE international conference on robotics and automation, Brisbane, QLD, Australia. 2018: 3803-3810

[151] Chebotar Y, Handa A, Makoviychuk V, et al. Closing the sim-to-real loop: Adapting simulation randomization with real world experience// Proceedings of the 2019 International Conference on Robotics and Automation. Montreal, Canada. IEEE, 2019: 8973-8979

[152] Nygaard T F, Martin C P, Torresen J, et al. Real-world embodied AI through a morphologically adaptive quadruped robot. Nature Machine Intelligence, 2021, 3(5): 410-419

[153] Chang P, Padif T. Sim2real2sim: Bridging the gap between simulation and real-world in flexible object manipulation// Proceedings of the 2020 Fourth IEEE International Conference on Robotic Computing. Shanghai, China. 2020: 56-62

[154] Lee J, Hwangbo J, Wellhausen L, et al. Learning quadrupedal locomotion over challenging terrain. Science robotics, 2020, 5(47): eabc5986

[155] Miki T, Lee J, Hwangbo J, et al. Learning robust perceptive locomotion for quadrupedal robots in the wild. Science Robotics, 2022, 7(62): eabk2822

[156] Guo B, Liu J, Liu S, et al. CrowdIM: Crowd-Inspired Intelligent Manufacturing Space Design. IEEE Internet of Things Journal, 2022, 9(19): 19387-19397

[157] Yu Z, Guo B. Human-Machine Intelligence. Communications of the China Computer Federation, 2017,13(12): 64-68 (in Chinese)
(於志文，郭斌．人机共融智能．中国计算机学会通讯，2017, 13(12): 64-68)

[158] Monarch R M. Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. USA:Simon and Schuster, 2021

[159] Nunes D S, Zhang P, Silva J S. A survey on human-in-the-loop applications towards an internet of all. IEEE Communications Surveys & Tutorials, 2015, 17(2): 944-965

[160] Cheraghi A R, Shahzad S, Graffi K. Past, present, and future of swarm robotics//Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2021: 190-233

[161] Da Silva F L, Warnell G, Costa A H R, et al. Agents teaching agents: a survey on inter-agent transfer learning. Autonomous Agents and Multi-Agent Systems, 2020, 34(1): 1-17

[162] Gupta A, Savarese S, Ganguli S, et al. Embodied intelligence via learning and evolution. Nature communications, 2021, 12(1): 1-12

[163] Liu Y, Yu Z, Guo B, et al. CrowdOS: A ubiquitous operating system for crowdsourcing and mobile crowd sensing. IEEE Transactions on Mobile Computing, 2020, 21(3): 878-894

[164] Mei H, Guo Y. Toward ubiquitous operating systems: A software-defined perspective. Computer, 2018, 51(1): 50-56

[165] Cao D, Xue D, Ma Z, et al. XiUOS: an open-source ubiquitous operating system for industrial Internet of Things. Science China. 2022: 1-2

Guo Bin, Ph.D. ， professor His research interests include ubiquitous computing, mobile crowd sensing, and HCI.

LIU Si-Cong, Ph.D., associate professor. Her research interests include ubiquitous computing, AIoT.

LIU Yan, Ph.D., assistant researchers. Her research interest is ubiquitous computing.

LI Zhi-Gang, Ph.D., associate professor. His research interests include IoT, mobile computing.

YU Zhi-Wen, Ph.D., professor. His research interests include ubiquitous computing, mobile crowd sensing.

ZHOU Xing-She, Ph.D., professor. His research interest is ubiquitous computing.

Background

The rapid development and fusion of Intelligence of Things (IoT), big data and artificial intelligence technologies has given rise to a promising emerging frontier field of Artificial Intelligence of Things (AIoT). AIoT is the combination of Artificial Intelligence (AI) technology and Internet of Things (IoT) infrastructure to achieve more intelligent IoT applications and provide more efficient services. Artificial intelligence models are good at analyzing and mining the potential patterns and strategies from massive amounts of data, while IoT has the ability to establish extensive connectivity for hundreds of millions of physical devices. Therefore, the deep fusion of AI and IoT will bring more powerful sensing, computing potential to boost the quality of intelligent service management, including smart cities, intelligent manufacturing, etc.

Based on the deep fusion of artificial intelligence, edge computing, Internet of Things and other technologies, AIoT aims to build a self-organizing, self-learning, self-adaptive, and continuous-evolving smart IoT system, which is achieved by empowering the sensing, communication, computing and application. As the latest research direction of IoT technology, the deep fusion of AI and IoT has made a lot of progress in many areas. For example, multi-modal intelligent sensing, mobile crowd sensing and other technologies have been effectively applied in smart home, smart factory, and so on. Intelligent IoT terminals, cloud-edge-end collaborative computing, and federal learning have become current research hotspots in many AIoT applications. Different from existing technologies, AIoT has some new characteristics:

- **Ubiquitous Intelligent Sensing**: AIoT utilizes ubiquitous sensing resources, including cameras, RFID, WiFi, infrared, acoustic waves, millimeter waves, etc., to collect rich multi-modal sensing data, and then enables accurate sensing of target (e.g., person, environment, or event, etc.) behavior via machine learning and deep learning.
- **Context-Adaptive Communication**: Considering of the constantly changing context of network resources, connection topology and data transmission, etc., AIoT extracts contextual information from the real-time network data, and then achieves low-cost and efficient communication by adaptive mechanisms for contextual adaptation.
- **On-Device Deep Computing**: With the dramatic increase in hardware capability of terminal devices in IoT, processing data (e.g., feature extraction, model training and inference) on IoT devices has become a new trend in many scenarios, which have

the advantages of low computational latency, low transmission cost, and protecting the data privacy.

- **Cloud-Edge-End Collaborative Architecture:** For the requests of real-time and data-privacy data processing for massive amount of IoT data, the edge computing technologies are introduced into the IoT system to form the collaborative “cloud-edge-end” system architecture, which could process a large amount of data in an efficient and timely manner.

This paper aims to systematically introduce the concept, architecture and key techniques of AIoT, to give a more comprehensive and in-depth elaboration for this emerging field. We first introduce the essential conceptual characteristics of AIoT, and then elaborates its architecture. Furthermore, we detail the research challenges and key technologies in AIoT, Finally, based on the latest research developments, we present the future research directions.