Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications

Khaled B. Letaief, *Fellow, IEEE*, Yuanming Shi, *Senior Member, IEEE*,
Jianmin Lu, and Jianhua Lu, *Fellow, IEEE*

(Invited Paper)

***Abstract*— The thriving of artificial intelligence (AI) applications is driving the further evolution of wireless networks. It has been envisioned that 6G will be transformative and will revolutionize the evolution of wireless from "connected things" to "connected intelligence". However, state-of-the-art deep learning and big data analytics based AI systems require tremendous computation and communication resources, causing significant latency, energy consumption, network congestion, and privacy leakage in both of the training and inference processes. By embedding model training and inference capabilities into the network edge, edge AI stands out as a disruptive technology for 6G to seamlessly integrate sensing, communication, computation, and intelligence, thereby improving the efficiency, effectiveness, privacy, and security of 6G networks. In this paper, we shall provide our vision for scalable and trustworthy edge AI systems with integrated design of wireless communication strategies and decentralized machine learning models. New design principles of wireless networks, service-driven resource allocation optimization methods, as well as a holistic end-to-end system architecture to support edge AI will be described. Standardization, software and hardware platforms, and application scenarios are also discussed to facilitate the industrialization and commercialization of edge AI systems.**

***Index Terms*— 6G, edge AI, edge training, edge inference, federated learning, over-the-air computation, task-oriented communication, service-driven resource allocation, large-scale optimization, end-to-end architecture.**

I. Introduction

A. Roadmap to 6G: Vision and Technologies

WITH the standardization and worldwide deployment of 5G networks, researchers, companies, and governments have initiated the vision, usage scenarios, and disruptive technologies for future 6G. In particular, the United States [1], European Union [2], and China [3] have recently funded 6G projects with a common goal of enabling connected intelligence. Besides, the International Telecommunication Union (ITU) has published the system requirements and driving characteristics for Network 2030 [4]. To improve a real-time immersive experience and interaction, as well as accelerate intelligence upgrades for industrial internet-of-things (IoT) and digital twins, multiple companies are now considering new usage scenarios. For example, based on typical use cases in 5G [5], [6] (i.e., enhanced mobile broadband (eMBB), ultra-reliable and low-latency communications (URLLC), and massive machine type communications (mMTC)), Huawei has recently proposed three additional application scenarios in the vision of 5.5G. These include uplink centric broadband communication (UCBC), real-time broadband communication (RTBC), and harmonized communication and sensing (HCS) [7]. It is expected that 6G will go beyond the mobile internet to support ubiquitous artificial intelligence (AI) services and Internet of Everything (IoE) applications [1]–[4], [8], including sustainable cities, connected autonomous systems, brain-computer interfaces, digital twins, tactile and haptic internet, high-fidelity holographic society, extended reality (XR) and metaverse [9], e-health, etc. Researchers in industry and academia have published many visionary 6G proposals [10]–[12] to provide a better understanding, sensing, controlling, and interacting for a physical world. In particular, three new application services were envisioned for 6G, including computation oriented communications (COC), contextually agile eMBB communications (CAeC), and event defined uRLLC (EDuRLLC) [12]. Based on these quoted usage scenarios, we present the evolution of visionary use cases for 6G in Fig. 1 by integrating intelligence, coordination, sensing, and computing for a connected cyber-physical world.

To shape the future of 6G use cases in 2030, multidisciplinary research and various disruptive technologies are required, including spectrum exploration technologies, devices and circuit technologies, as well as networking, computing, sensing, and learning functionalities. In particular, AI, especially deep learning (DL), provides a revolutionary approach to design and optimize 6G wireless networks across the physical, medium-access, and application layers [12], [13]. Specifically, DL provides a novel way to design 6G air interface by optimizing the radio environment [14], communication algorithms [15], hardware, and applications in a

Manuscript received June 22, 2021; revised September 8, 2021; accepted October 11, 2021. Date of publication November 8, 2021; date of current version December 17, 2021. This work was supported in part by Project No. 20210400L016 under RDC Corporation Ltd. and by the Natural Science Foundation of Shanghai under Grant 21ZR1442700. *(Corresponding author: Yuanming Shi.)*

Khaled B. Letaief is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: eekhaled@ust.hk).

Yuanming Shi is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: shiym@shanghaitech.edu.cn).

Jianmin Lu is with Huawei Technologies Company Ltd., Shenzhen 518066, China (e-mail: lujianmin@huawei.com).

Jianhua Lu is with the Department of Electronic Engineering and the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: lhh-dee@mail.tsinghua.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2021.3126076.

Digital Object Identifier 10.1109/JSAC.2021.3126076

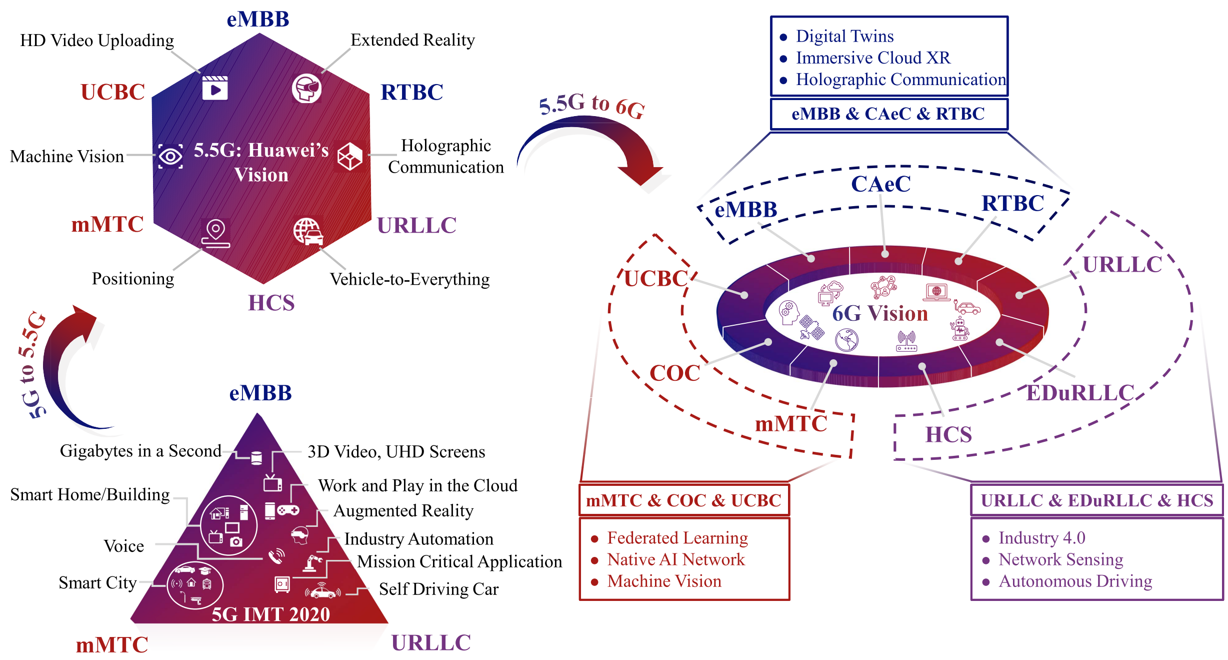


Fig. 1. Towards 6G: the evolution of use cases from 5G to 6G.

unified way [16], [17]. This has inspired the recent success applications for joint source-channel coding (JSCC) [18], task-oriented communication [19], [20], semantic communication [21]. Besides, machine learning (ML) also provides a paradigm shift for automatically learning high performance and fast optimization algorithms to solve the resource allocation problems in wireless networks [22]–[25]. The domain knowledge (e.g., optimization models and theoretical tools) was further incorporated into the DL framework for optimizing ultra-reliable and low-latency communication networks [26]. An ML approach was also developed for addressing the communication, networking, and security challenges for vehicular applications [27]. With the development of wireless data collection, learning models and algorithms, as well as software and hardware platforms, we envision that AI will become a native tool to design disruptive wireless technologies for accelerating the design, standardization, and commercialization of 6G. On the other hand, the evolution of 6G wireless communication technologies and communication theory will also inspire the progress and development of AI techniques in terms of novel learning theory, new deep neural network (DNN) architectures, customized software and hardware platforms.

Given the requirements of emerging 6G, connected intelligence is expected to be the central focus and an indispensable component in 6G [28]. This shall revolutionize the evolution of wireless from “connected things” to “connected intelligence”, thereby enabling the interconnections between humans, things, and intelligence within a hyper-connected cyber-physical world [12]. *Edge AI* provides a promising solution for connected intelligence by enabling data collection, processing, transmission, and consumption at the network edge [29], [30]. Specifically, by embedding the training capabilities across the network nodes, edge training is able to preserve privacy and confidentiality, achieve high security and fault-tolerance, as well as reduce network traffic congestion and energy consumption. For instance, over-the-air federated learning (FL) provides a collaborative ML framework to train a global statistical model over wireless networks without accessing edge devices’ private raw data [31]. By directly executing the AI models at the network edge, edge inference can provide low-latency and high-reliability AI services by requiring less computation, communication, storage, and engineering resources. For example, edge device-server co-inference is able to remove the communication and computation bottlenecks by splitting a large DNN model between edge devices and edge servers [32]. However, edge AI will cause task-oriented data traffic flows over wireless networks, for which disruptive wireless techniques, efficient resource allocation methods and holistic system architectures need to be developed. To embrace the era of edge AI, wireless communication systems and edge AI algorithms need to be co-designed for seamlessly integrating communication, computation, and learning.

B. Edge AI: Challenges and Solutions

Creating a trustworthy and scalable edge AI system will be of utmost importance for imbuing connected intelligence in 6G. The challenges of trustworthiness and scalability are multidisciplinary spanning ML, wireless networking, and operation research. Specifically, *trustworthiness* in terms of privacy and security is one of the key requirements for 6G intelligent services and applications, for which the general data protection regulation (GDPR) needs to be satisfied, and directly transmitting or collecting data from users are forbidden. To tame privacy leakages and adversarial attacks, various edge learning models and architectures have been proposed, including FL (i.e., server-client network architecture with data partition among edge devices) [33], [34], swarm learning (i.e., decentralized device-to-device (D2D) communication architecture without central authority) [35], and split learning (i.e., model parameters partitioned among edge devices and edge servers) [36], [37]. Distributed reinforcement learning (RL) [38], [39] and trustworthy learning techniques [40], [41] were further proposed to address the dynamic and adversarial learning environments, respectively. In particular, differential privacy [42], lagrange coded computing [43], security multi-party computation, quantum computing, blockchain, and distributed ledger technologies can be further leveraged to build trustworthy edge AI architectures. However, with limited storage, computation, and communication resources in the wireless edge networks, deploying an edge AI system causes a significant *scalability* issue in terms of latency, energy and accuracy. To address this challenge, a paradigm shift for wireless system design is required from data-oriented communication (i.e., maximizing communication rate or reliability based on Shannon theory) to *task-oriented communication* (i.e., achieving fast and accurate intelligence distillation at the network edge).

In this paper, we shall provide a comprehensive picture for the design of scalable and trustworthy edge AI systems by matching the principles and architectures of wireless networks with the task structures of edge AI models and algorithms. The system performance metrics for edge AI are further characterized to facilitate efficient resource allocations based on operation research and ML. Specifically, to design a communication-efficient edge AI training system, we will provide novel multiple access schemes (e.g., over-the-air computation (AirComp) for model aggregation [31], [44], [45]) to support massive access for edge devices, new multiple antenna techniques (e.g., cell-free massive MIMO [46], [47] and reconfigurable intelligent surface (RIS) [48], [49]) to support fast exchange for high-dimensional model updates, and next-generation network architectures (e.g., space-air-ground integrated network (SAGIN) [50], [51]) to support diverse edge learning models and topologies. To design a communication-efficient edge inference system with low-latency and reliability guarantees, interference management, cooperative transmission, and task-oriented communication will be introduced to support edge device distributed inference [52], edge server cooperative inference [53], [54], and edge device-server co-inference [32], respectively. We then provide a holistic view for mathematically modeling the resource allocation problems in edge training and inference systems, which are categorized as mixed combinatorial optimization, nonconvex optimization and stochastic optimization models. A "learning to optimize" framework is further introduced to facilitate scalable, real-time, robust, parallel, distributed, and automatic optimization algorithms design for service-driven resource allocation in edge AI systems [22], [23], [25], [55]. We also provide a holistic end-to-end architecture for edge AI systems. Moreover, standardizations, resource allocation optimization solvers, software and hardware platforms, and application scenarios are discussed. The roadmap to edge AI ecosystem is demonstrated in Fig. 2 to encourage multidisciplinary collaborations among information science, computer science, operation research, and integrated circuits.

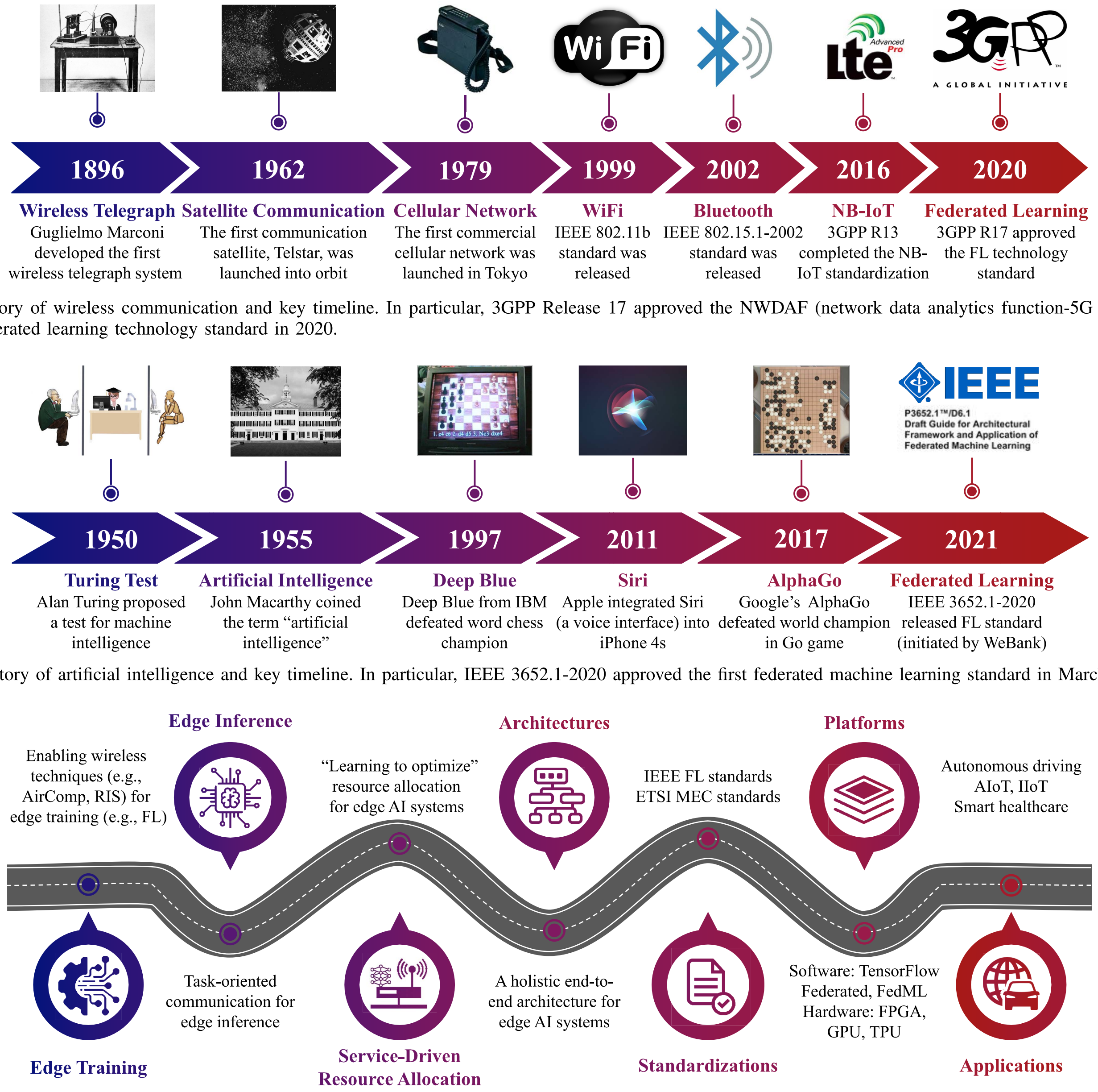
C. Edge AI Empowered 6G Networks

The developed edge AI technology will serve as a distributed neural network to accelerate the evolution of sensing capabilities, communication strategies, network optimizations, and application scenarios in 6G networks. Specifically, edge AI paves the way for network sensing and cooperative perception to understand the network environments and services for an agile and intelligent decision making. For example, edge simultaneous localization and mapping (SLAM) [56], [57] has recently been developed to deploy DL based visual SLAM algorithms on vehicles by edge inference. Edge AI can also help design AI-native communication strategies for the physical layer (e.g., task-oriented semantic communication [58]) and medium access control layer (e.g., random access protocol [59]). For instance, edge DL approach has been developed in [58] to deliver low-latency semantic tasks (e.g., text messages) by learning the communication strategies in an end-to-end fashion based on JSCC. Furthermore, edge AI provides a new paradigm for optimization algorithms design to enable service-driven resource allocation in 6G networks [60]. For instances, distributed RL [55], decentralized graph neural networks [23], and distributed DNN [61], are able to automatically learn the distributed resource allocation optimization algorithms. By seamlessly integrating sensing, communication, computation, and intelligence, edge AI shall empower 6G networks to support diversified intelligent applications, including autonomous driving, industrial IoT, smart healthcare, etc.

To further imbue native intelligence, native trustworthiness, and native sensing in 6G, mimicking nature for innovating edge AI empowered future networks can be envisioned. Inspired by the dynamic spiking neurons in the human brain, the energy consumption and latency of edge AI can be significantly reduced by processing the learning tasks in an event-driven manner [62], [63]. The brain-inspired stigmergy-based federated collective intelligence mechanism was proposed in [64] to accomplish multi-agent tasks (e.g., autonomous driving) through simple indirect communications. By leveraging the prior knowledge of the immune system and brain neurotransmission, a brand-new network security architecture and fully-decoupled radio access network have recently been proposed in [65] and [66], respectively. These results on nature-inspired edge AI models and network architectures provide a strong evidence that one can establish an integrated data-driven and knowledge-guided framework to design and optimize 6G networks. Further details and description of the edge AI empowered 6G network are provided in Fig. 3, which highlight the integration of sensing, communication, computation and intelligence in a closed-loop ecosystem.

D. Key Contributions

We provide extensive discussions, visions, and summaries of wireless techniques, resource allocations, standardizations, platforms, and application scenarios to embrace the era of edge AI for 6G. The major contributions are summarized as follows:


1896
Wireless Telegraph
Guglielmo Marconi developed the first wireless telegraph system
1962
Satellite Communication
The first communication satellite, Telstar, was launched into orbit
1979
Cellular Network
The first commercial cellular network was launched in Tokyo
1999
WiFi
IEEE 802.11b standard was released
2002
Bluetooth
IEEE 802.15.1-2002 standard was released
2016
NB-IoT
3GPP R13 completed the NB-IoT standardization
2020
Federated Learning
3GPP R17 approved the FL technology standard
1950
Turing Test
Alan Turing proposed a test for machine intelligence
1955
Artificial Intelligence
John Macarthy coined the term "artificial intelligence"
1997
Deep Blue
Deep Blue from IBM defeated word chess champion
2011
Siri
Apple integrated Siri (a voice interface) into iPhone 4s
2017
AlphaGo
Google's AlphaGo defeated world champion in Go game
2021
Federated Learning
IEEE 3652.1-2020 released FL standard (initiated by WeBank)
IEEE
P3652.1™/D6.1
Draft Guide for Architectural Framework and Application of Federated Machine Learning
Edge Inference
Enabling wireless techniques (e.g., AirComp, RIS) for edge training (e.g., FL)
"Learning to optimize" resource allocation for edge AI systems
Architectures
IEEE FL standards
ETSI MEC standards
Platforms
Autonomous driving
AIoT, IIoT
Smart healthcare
Edge Training
Task-oriented communication for edge inference
Service-Driven Resource Allocation
A holistic end-to-end architecture for edge AI systems
Standardizations
Software: TensorFlow Federated, FedML
Hardware: FPGA, GPU, TPU
Applications

(a) History of wireless communication and key timeline. In particular, 3GPP Release 17 approved the NWDAF (network data analytics function-5G network AI) federated learning technology standard in 2020.

(b) History of artificial intelligence and key timeline. In particular, IEEE 3652.1-2020 approved the first federated machine learning standard in March 2021.

(c) Paper organization and structure: 1) technologies: communication-efficient edge training in Section II, and communication-efficient edge inference in Section III; 2) systems: resource allocation for scalable and trustworthy edge AI systems via theory-driven and data-driven optimizations in Section IV; 3) architectures: a holistic end-to-end architecture for edge AI systems in Section V; 4) commercializations: standardizations for edge learning and computing, software and hardware platforms, as well as potential applications including autonomous driving, IoT, and smart healthcare in Section VI.

Fig. 2. Roadmap to edge AI.

- The vision (i.e., connected intelligence for 6G), challenges (i.e., trustworthiness and scalability) and solutions (i.e., wireless techniques, resource allocations and system architectures) for edge AI, as well as edge AI empowered 6G network, are introduced and summarized in Section I.
- The communication-efficient edge training system is presented in Section II, including the edge learning models and algorithms, followed by the promising wireless techniques and architectures to support their deployment.
- The communication-efficient edge inference system is introduced in Section III. Here, we introduce horizontal edge inference and vertical edge inference by cooperative transmission and task-oriented communication, respectively.
- A unified framework for resource allocation in edge AI systems is provided in Section IV. Here, we present operation research based theory-driven and machine learning based data-driven approaches for designing efficient resource allocation optimization algorithms.
- A holistic end-to-end architecture for edge AI systems is proposed in Section V, including network infrastructure, data governance, edge network function, edge AI management and orchestration.
- The standardizations, software and hardware platforms, and application scenarios are discussed in Section VI.

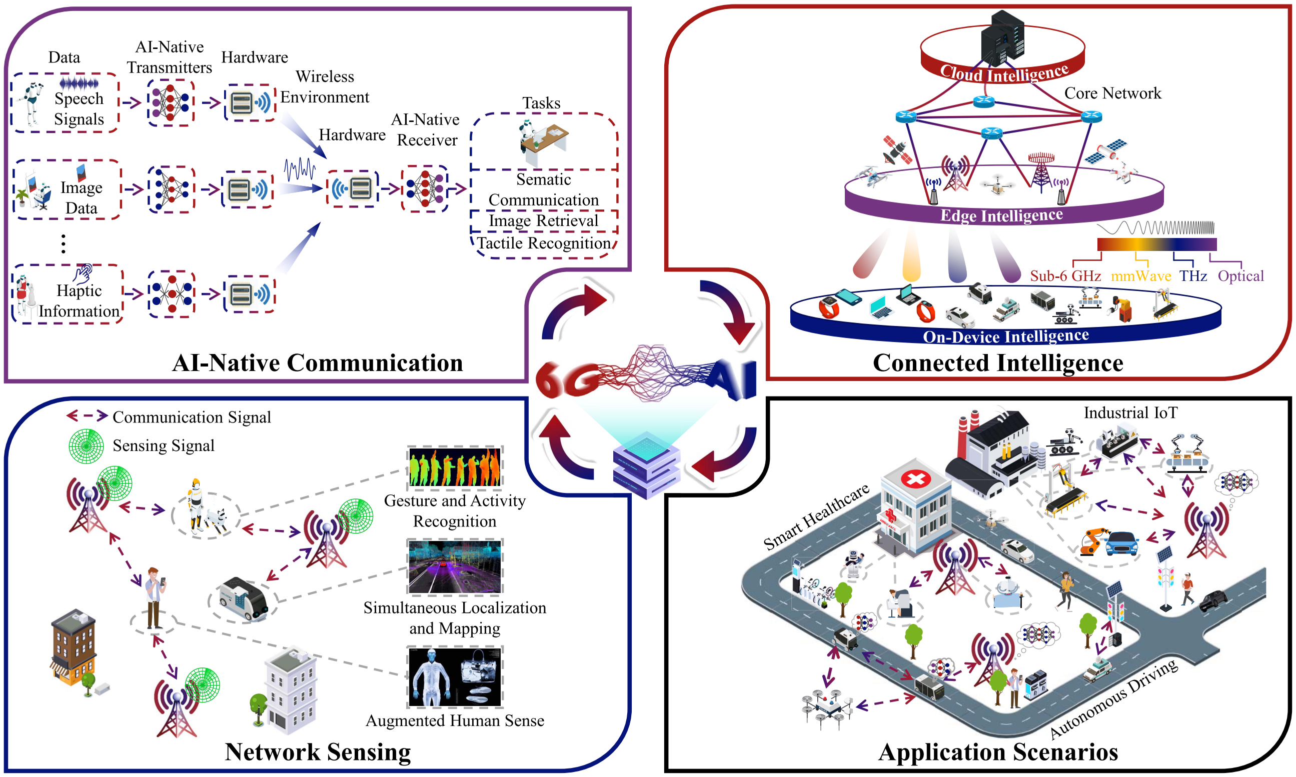


Fig. 3. Edge AI empowered 6G networks: integrated sensing, communication, computation, and intelligence.

This will help facilitate the booming market of edge AI in the 6G era.

We summarize the main topics and relevant technologies as well as highlight the representative results in Table I.

II. Communication-Efficient Edge Training

In this section, we shall present various communication-efficient distributed optimization algorithms for edge training, followed by promising enabling wireless techniques to support the deployment of edge learning models and algorithms.

A. Edge Learning Models and Algorithms

The training process of edge AI models typically involves minimizing a loss or empirical risk function to fit a global model from decentralized data generated by a massive number of intelligent devices. The goal of the distributed optimization for edge training is to minimize the global loss function $\mathcal{L}$, namely,

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{\theta}) := \sum_{k\in\mathcal{S}} w_k \mathcal{L}_k(\boldsymbol{\theta};\mathcal{D}_k), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ are the model parameters, $\mathcal{L}_k$ is the local loss function of device k over local dataset $\mathcal{D}_k$, $\mathcal{S}$ denotes the set of participating edge nodes, and $w_k \geq 0$ with $\sum w_k = 1$ denotes the weight for each local loss function. Considering the network topology for edge training, the heterogeneous local dataset $\mathcal{D}_k$, varying device participation $\mathcal{S}$, dynamic communication and computation environments, as well as privacy concerns and adversarial attacks, highly-efficient and trustworthy distribution optimization algorithms need to be developed. As shown in Fig. 4, based on the data partition and model partition principles [29], we will first introduce various edge training architectures, including FL, decentralized learning, and model split learning. We then present distributed RL and trustworthy learning techniques to accommodate dynamic and adversarial environments, respectively, as shown in Fig. 5.

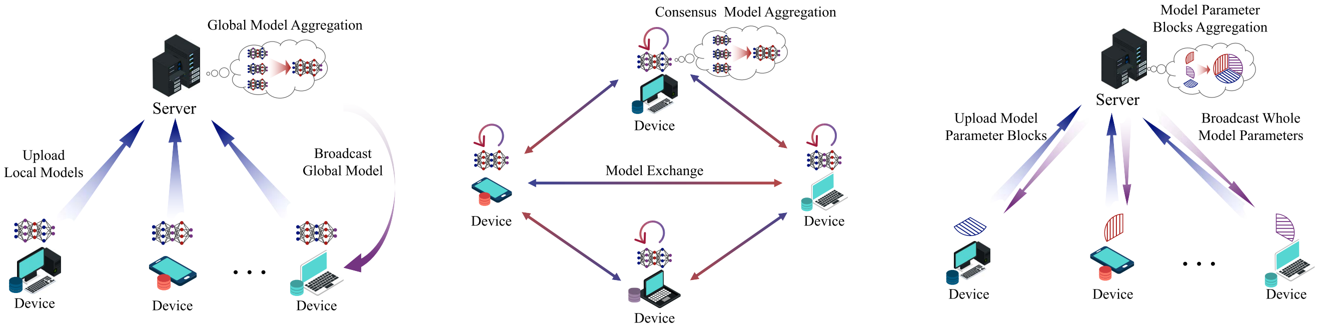
1) Federated Learning: FL is a collaborative ML framework to train a global statistical model without accessing edge devices' private raw data, wherein a dedicated edge server is responsible for aggregating local learning model updates and disseminating global learning model updates [34], as shown in Fig. 4 (a). FL is being adopted by many industrial practitioners, including Google's Gboard mobile keyboard for next word prediction and emoji suggestion, Apple's QuickType keyboard for vocal classifier, NVIDIA for COVID-19 patients oxygen needs prediction, and WeBank for money laundering detection [68]. Compared with the cloud data center based distributed learning, cross-device FL raises unique challenges for solving the distributed training optimization problems, including high communication costs with a large model frequently exchanged over wireless networks, statistical heterogeneity with non-identical local data distributions and sizes, system heterogeneity with varied storage, computation and communication capabilities, as well as dynamic devices participation [122]. A growing body of recent works have developed effective methods to address these unique challenges in FL.

To address the challenge of expensive communication overheads for intermediate local updates with a central server, federated averaging [67] turns out to be effective to reduce the number of communication rounds by performing multiple local updates, e.g., running multiple stochastic gradient

TABLE I
AN OVERVIEW OF THE MAIN TOPICS AND REPRESENTATIVE RESULTS

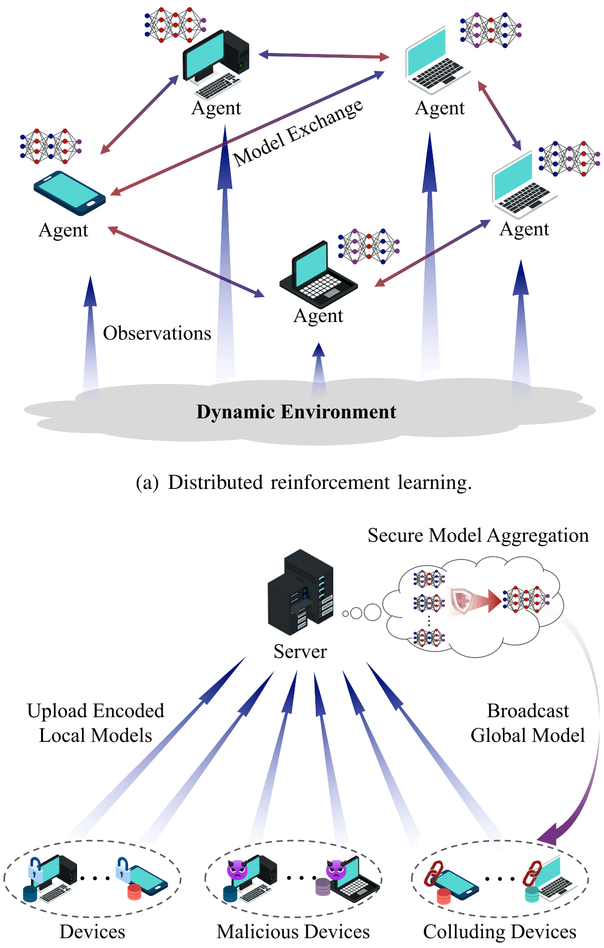
Sections	Topics	Methods	Representative Results
Section II: Communication-Efficient Edge Training	Section II-A: Edge Learning Models and Algorithms	Federated Learning	Horizontal and vertical federated learning [67], [34], [33]; federated optimization [68]
		Decentralized Learning	Swarm learning [35]; consensus-based methods [69]; diffusion strategies [70]; decentralized training [71]
		Model Split Learning	Model parameter partitioned edge learning [72]; split learning [36]
		Distributed Reinforcement Learning	Multi-agent reinforcement learning [73]
		Trustworthy Learning	Differential privacy [42]; secure model aggregation [74]; blockchain smart contract [35]
	Section II-B: Wireless Techniques for Edge Training	Over-the-Air Computation	Low-latency analog model aggregation [31], [44], [45]
		Massive Access Techniques	Grant-free random access [75], [76]; NOMA [77], [78]; blind demixing [79]
		Ultra-Massive MIMO	Cloud-RAN [80]; cell-free massive MIMO [46]
		Reconfigurable Intelligent Surfaces	RIS-empowered edge training [48] /edge inference [54]
		Space-Air-Ground Integrated Networks	UAV-aided model aggregation [81]; fog learning [50]
Section III: Communication-Efficient Edge Inference	Section III-A: Horizontal Edge Inference	Edge Device Distributed Inference	Wireless MapReduce [52]
		Edge Server Cooperative Inference	Energy-efficient edge cooperative inference [53]
	Section III-B: Vertical Edge Inference	Edge Device-Server Co-Inference	DNN inference via edge computing [82]
		Ultra-Reliable and Low-Latency Communication	Short packet communication [83]; ultra-reliable low-latency edge computing [84], [85]
		Task-Oriented Communication	Single-device/multi-devices edge inference [20], [86]
Section IV: Resource Allocation for Edge AI Systems	Section IV-A: Engineering Requirements and Methodologies	Accuracy	Convergence analysis for edge training algorithms [87], [45], [88]
		Latency	Delay analysis and optimization of wireless FL [89], [90]; low-latency edge inference [52], [32]
		Energy	Wireless-powered over-the-air FL [91]; energy-efficient FL [92]; green edge inference [53], [54]
		Trustworthiness	Privacy [41], security [93], optimality [94], interpretability [95]
		Service-Driven Resource Orchestration	Heterogeneous demands for edge AI services
	Section IV-B: Optimization Models and Algorithms	Mixed-Combinatorial Optimization	Sparse optimization models and algorithms [96]; learning to branch-and-bound [22]; algorithm unrolling [97]
		Nonconvex Optimization	Large-scale convex approximation [98]; DC programming [99], [31]; manifold optimization [100], [101]; GNN [23]
		Stochastic Optimization	Robust optimization [102]; chance constrained programming [53]; deep RL [103]; transfer learning [104]
		End-to-End Optimization	DNN [105]; GNN [25]
Section V: Architecture for Edge AI Systems	Section V-A: End-to-End Architecture for Edge AI Systems		Deep integration of sensing, communication, computation and intelligence [28]
	Section V-B: Data Governance		Independent data plane [106]; multi-player roles [107]
	Section V-C: Deeply Converged Communication and Computing at the Edge		AI for networks; networks for AI [12]
	Section V-D: Edge AI Management and Orchestration		Planning, deploying, maintaining, and optimizing edge AI models and edge network infrastructures [28]
Section VI: Standardizations, Platforms, and Applications	Section VI-A: Standardizations	Learning	IEEE 3652.1-2020 [108]
		Computing	ETSI ISG MEC [109]
	Section VI-B: Platforms	Software	FedML [110], FATE [111], HarmonyOS [112]
		Solver	CVX [113]; SCS [114]; Open-L2O [115]
		Hardware	Edge computing hardware [116]; RF hardware [117]
	Section VI-C: Applications	Autonomous Driving	Perception; HD mapping; edge SLAM [118]
		Internet of Things	AIoT [119], [58]; IIoT[120], [121]
		Smart Healthcare	Clinical disease detection [35]; haptic communication [122]

descent (SGD) iterations on each edge device. The local updating approach is able to learn a global model within much fewer communication rounds compared with the vanilla distributed SGD method, i.e., only running one mini-batch with SGD



(a) Federated learning. (b) Decentralized learning. (c) Model split learning.

Fig. 4. Edge learning models and architectures.



(a) Distributed reinforcement learning.

(b) Trustworthy learning.

Fig. 5. Edge learning modes in dynamic and adversarial environments.

at each edge device. Model compression, such as quantization and sparsification, is another notable way to address the communication bottleneck by reducing the size of the exchanged messages during each model update round. Scalar quantization is a typical way to implement lossy compression for the high-dimensional gradient vectors by quantizing each of their entries to a finite-bit low precision value [123]–[125], which was further improved by the recent proposal of vector quantization [126], [127]. Sparsification, on the other hand, proposes to only communicate the informative elements of the gradient or model vectors among nodes [128], [129]. A set of algorithms combining the local updates method and model compression have shown the capability of achieving high communication efficiency [130], [131]. In particular, a lazily aggregated quantized gradient method was further proposed in [132] to reduce both the amount of exchanged data and communication rounds by reusing the outdated gradients for the less informative quantized gradients.

Although the above periodical compressed update methods have shown empirical or theoretical success for tackling the communication challenge, the heterogeneity in systems and local datasets may slow down or even diverge the convergence [133], [134], for which various algorithms and models have been proposed to address the statistical and system heterogeneity challenges. To learn the AI models from statistically heterogeneous local datasets, various effective and personalized models have been proposed to rectify the original model (1), including regularizing local loss functions at each device [134]–[136], distributionally robust modeling [137], [138], multi-task learning [139], as well as the meta-learning approaches [140]. Running a local update at the devices with heterogeneous computation capabilities may yield objective inconsistency or client drift, i.e., the learned model can be far from the desired true model. To address this problem, an operator splitting method was proposed to avoid the local models drifting apart from the global model [141]. A normalized model aggregation method was also developed to ensure that the global model converges to the desired true model [142]. A novel federated aggregation scheme was further developed in [143] to address the system heterogeneity issue concerning the dynamic, sporadic and partial device participation. To leverage the computation capabilities across the device-edge-cloud heterogenous network, a hierarchical model aggregation approach was proposed in [130] to reduce the latency by controlling the two aggregation intervals.

2) Decentralized Learning: Decentralized ML learns a global model from inherently decentralized data structures via peer-to-peer communications over the underlying communication network topology without a central authority [144], as shown in Fig. 4 (b). It has great potentials for applications in the autonomous industrial systems, including cooperative automated driving, cooperative simultaneous localization and mapping, and collaborative robotics in advanced manufacturing environments [145]. The decentralized learning architecture harnesses the benefits of communication efficiency, computation scalability and data locality. In particular, swarm learning [35] provides a completely decentralized AI solution based on decentralized ML by keeping local datasets at each edge device. This can achieve high privacy, security, resilience and scalability. Compared with the sever-client learning architecture in FL, decentralized learning can accommodate the decentralized D2D communication network architectures and protocols with arbitrary connectivity graphs (e.g., cooperative driving and robotics networks). It can also overcome the straggler dilemma with heterogeneous hardware, as well as improve the robustness to data poisoning attacks and master node fails [35], [145]. The convergence behavior of decentralized learning highly depends on the decentralized averaging mechanism and the network topology for data exchange [71]. Typical decentralized aggregation approaches include the consensus-based methods [69] and diffusion strategies [70].

To improve the communication efficiency for exchanging the locally updated models at edge devices within their neighbors, one may reduce either the number of communication rounds (i.e., improve convergence rate) or the volume of exchanged data per round. Specifically, the variance reduction with the gradient tracking method was investigated in [146] to achieve a fast convergence rate. Periodic-averaging via running multiple local updates before decentralized averaging is an effective way to reduce the number of communication rounds among devices [147], [148]. Besides, quantizing or sparsifying the locally updated models can reduce the volume of the exchanged messages to address the communication bottleneck [149]. A consensus distance controlling framework was further developed in [150] to achieve the trade-off between the learning performance and the exactness of decentralized averaging for decentralized DL. Moreover, a communication network topology design is also critical to improve the communication efficiency [151], for which a group alternating direction method of multipliers [152] was proposed to form a connectivity chain by dividing the workers into head and tail workers. To address the heterogeneity issue of local datasets, the momentum-based method [153] has recently been developed to achieve good generalization performance.

3) Model Split Learning: Model split learning enables a collaborative learning process across the edge devices and edge servers by partitioning the model parameters across the edge nodes, as shown in Fig. 4 (c). That is, each edge node k, including edge devices and edge servers, is only responsible for updating $\boldsymbol{\theta}_k$ with $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{\mathcal{S}}]$ in (1). This model splitting architecture can achieve higher privacy levels and better trade-offs between communication and computation. It is thus particularly applicable for DL with a large model parameters size, whereas the data partition based training method, e.g., FL, normally requires the local update of a whole copied global model at each involved edge device. The model parameter partitioned edge learning approach [72] proposed to train only a block of model parameters based on the coordinate decent method for the decomposable ML models [154] or the alternating minimization approach for the general DL models [155]. However, this approach is prone to data privacy leakage as the datasets need to be shared across edge devices. Vertical FL, on the other hand, can directly learn the global model from the partitioned data features among different edge devices without sharing them [156]. Therefore, the data features and the associated model parametric blocks are split among edge devices, for which the asynchronous SGD method can be applied for vertical FL [157]. Consensus algorithms were also developed in [158] to jointly learn a model under the decentralized network while keeping the distributed data features locally.

Split DL further provides a flexible way to train a DNN by dividing it into lower and upper segments located at the edge device-side and edge server-side, respectively [36]. It can be typically applied to the medical diagnosis and millimeter wave channel prediction [37]. Split DL is able to preserve privacy without sharing raw data and enjoys computation scalability by allowing that only edge devices perform simple computation for the lower segments. Compared with FL, split DL can significantly improve computation efficiency, reduce communication costs, as well as achieve higher learning accuracy, data security and system scalability. Specifically, edge devices and edge server collaboratively train the whole neural network, which involves routing the activations of the edge device-side subnetwork to the edge sever via forward propagation, and downloading the gradients of the edge server-side subnetwork to update the lower segment via back propagation. However, exchanging the instantaneous intermediate values between edge devices and edge server becomes the communication bottleneck, especially in the case with multiple edge devices. Therefore, a joint communication strategy and neural network architecture design is required [37] for split training of various DNNs with heterogeneous edge devices. Considering the large-scale privacy-sensitive and delay-sensitive IoT applications, Lyu *et al.* [159] proposed a hybrid fog-based privacy-preserving DL framework, where a fog-level DNN is partitioned between the edge device and the fog server side.

4) Distributed Reinforcement Learning: RL provides a flexible framework for sequential decision making in dynamic settings by interacting with a dynamic environment, as shown in Fig. 5 (a). This can be frequently modeled as decision making and learning in a Markov decision process (MDP) [160]. Typical RL algorithms include the model based algorithm, policy-based algorithm (e.g., natural policy gradient), value based algorithm (e.g., Q-learning), and actor-critic method. In particular, an asynchronous method, by leveraging parallel computing, was developed in [161] to solve the large-scale nonconvex RL problem. However, in modern intelligent applications, e.g., autonomous driving and robotics, it is critical to consider multi-agent reinforcement learning (MARL), in which multiple agents collaboratively interact with a common environment to complete a common goal and maximize a shared team award with different local action spaces [73]. Due to the enormous state-action space, delayed rewards and feedback, as well as the non-stationary and unknown environments with heterogeneous agents' behaviors, efficient communication strategy among multiple agents shall play a key role to achieve good and stable performance for MARL.

For the server-client architecture based MARL, the edge server coordinates the learning process for all the edge agents. Lowe *et al.* [162] proposed a multi-agent actor-critic method involving decentralized actors at each agent and a centralized critic for parameter sharing among the agents. To improve the communication efficiency of the distributed policy gradient for MARL, a lazily aggregated policy gradient was developed in [38] to reduce the communication rounds by only communicating informative gradients of partial agents while reusing the outdated gradients for the remaining agents. For applications without central coordinators, e.g., autonomous driving, decentralized MARL is essential wherein the agents only allow the exchange of messages with their neighbors over a communication connectivity graph [163]. Zhang *et al.* [39] proposed decentralized actor-critic algorithms with function approximation, where each agent makes individual decisions based on both the information observed locally and the messages shared through a consensus step over the network. A decentralized entropy-regularized policy gradient method by only sharing information with neighbor agents was developed in [164] to learn a single policy for multi-task RL with multiple agents operating different environments.

5) Trustworthy Learning: To learn and deploy AI models for high-stake applications (e.g., autonomous driving) at the network edge, it is critical to ensure privacy, security, interpretability, responsibility, robustness, and fairness for the edge learning processes, as shown in Fig. 5 (b). However, the heterogeneity of massive scale edge systems and decentralized datasets raises unique challenges to design trustworthy edge AI techniques. Although FL addresses the local confidentiality issue by keeping datasets locally, the shared model updates still cause extreme privacy leakage (e.g., model inversion attack), the learned global model can be colluded by malicious attackers [40], [165], and the edge devices may be adversarial attackers (e.g., data or model poisoning). This calls for rigorous privacy-preserving mechanisms and secure aggregation rules [43]. Differential privacy provides a promising lightweight privacy-preserving mechanism to guarantee a level of privacy disclosure for local datasets by adding random perturbations [42]. The additive noise and signal superposition properties in the wireless channel can be naturally harnessed as the privacy-preserving mechanism [41]. The resulting inherent noisy model aggregation scheme can limit the privacy disclosure of local datasets at the edge server for free while keeping the learning performance unchanged [41], [166]. To improve the communication efficiency for private distributed learning, Chen *et al.* [167] developed efficient encoding and decoding mechanisms to simultaneously achieve optimal communication efficiency and differential privacy under typical statistical learning settings.

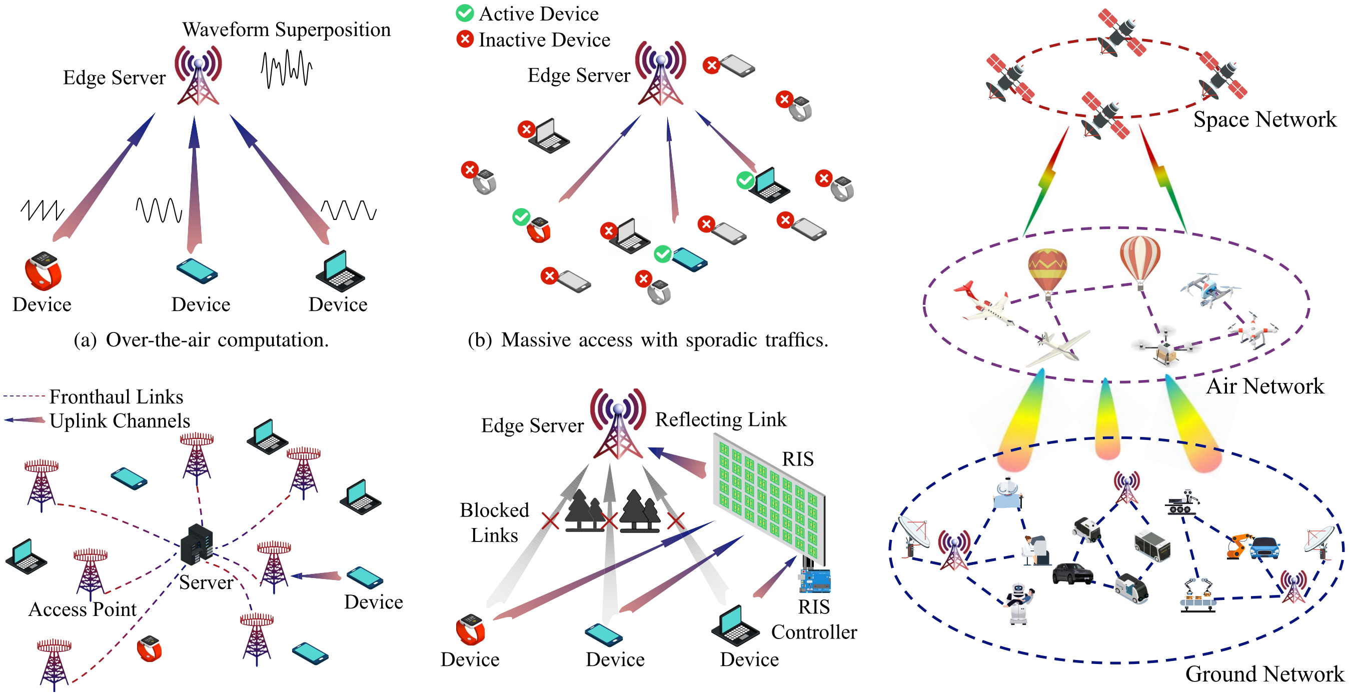
Apart from preserving privacy for individual users, edge AI also needs to be robust to errors and adversarial attackers, as the decentralized nature makes it easy to be unreliable in the learning process or even completely controlled by external attackers [74]. To address Byzantine attacks (i.e., the faulty edge device can behave arbitrarily badly by modifying its local updates) in FL with a server-client architecture, various robust and secure model aggregation schemes (e.g., geometric median [168], trimmed mean [169], and Krum [170]) were proposed to tolerate the Byzantine corrupted edge devices. To simultaneously preserve privacy for individual users while tolerating Byzantine adversaries, a Byzantine-resilient secure aggregation framework was developed in [171] to detect adversarial models without the knowledge of individual local models, as they are masked for privacy guarantees. To further avoid malicious edge servers, blockchain technology was utilized to provide a decentralized consensus environment to guarantee the validity of global models in every learning iteration. This is achieved by packing the local models and global model into blocks, which are confirmed under a consensus mechanism, followed by linking them into the blockchain [172]. To protect decentralized learning from attacks, a blockchain based peer-to-peer network was developed in [35] to support swarm learning without a central server. This high security level in decentralized learning is achieved by securely enrolling new nodes via blockchain smart contract to perform local model training.

To summarize, the presented edge learning models and algorithms provide a strong evidence that to deploy the edge training process in wireless networks, we need to develop new wireless communication techniques and strategies to support massive and flexible edge devices participation, as well as support efficient function computation for model aggregation (e.g., weighted sum global model aggregation in FL, consensus model aggregation in decentralized learning, and robust model aggregation in secure learning). Various edge training architectures (e.g., server-client, decentralized, and hierarchical network topologies), as well as high-dimensional model updates exchange motivate us to develop new wireless network principles and architectures to support edge AI training systems, which will be discussed in the following subsection.

B. Wireless Techniques for Edge Training

As the communication target for edge AI becomes the learning performance instead of the conventional data rates, we shall exploit the task structures of edge AI models and algorithms to match the principles and architectures of wireless networks. This helps demystify the efficiency of edge training in wireless networks, which yields a learning-communication co-design principle for future 6G wireless networks to enable AI functionalities sitting natively within 6G. As shown in Fig. 6, we will introduce next generation multiple access schemes (e.g., AirComp and massive random access) to accommodate a massive number of edge devices dynamically involved in the training process, new multiple antenna techniques (e.g., RIS and cell-free massive MIMO) to support high-dimensional model updates exchange, as well as new network architectures (e.g., SAGIN and unmanned aerial vehicle (UAV) network) to support diversified edge training models and topologies.

1) Over-the-Air Computation: Edge training tasks typically involve computing aggregation functions of multiple local



(a) Over-the-air computation.

(b) Massive access with sporadic traffics.

(c) Cell-free massive MIMO.

(d) Reconfigurable intelligent surface.

(e) Space-air-ground integrated network.

Fig. 6. Enabling wireless techniques for edge training.

model updates to update a global model. To accomplish weighted averaging aggregation in FL, consensus aggregation in decentralized learning, and robust aggregation (e.g., geometric median) in trustworthy learning, the local updates need to be transmitted from the edge devices, followed by computing the relevant aggregation function at the edge sever. However, the limited bandwidth and resource in wireless networks becomes one of the key bottlenecks to enable a massive number of edge devices that upload the local model updates for global aggregation. AirComp provides a new multiple access scheme for low-latency model aggregation. By concurrently transmitting the locally updated models, AirComp can harness interference to reduce communication bandwidth consumptions. The key idea is that the waveform superposition of a wireless multiple access can be exploited for computing the nomographic functions (e.g., the model aggregation function weighted average) over the same channel [118], as shown in Fig. 6 (a). Specifically, the transmitted signals at edge devices are first multiplied by the fading channels and then superposed over-the-air with additive channel noise, resulting in a noisy weighted sum of transmitted signals [31]. This perfectly matches the structure of model aggregation computation. Note that the robust aggregation function (i.e., geometric median) does not hold the additive structure. But we can still approximate it by computing a few number of weighted averaging functions via AirComp [93]. The communication latency and bandwidth requirement of AirComp will not increase with the number of edge devices, thus relieving the communication bottleneck in the edge training process.

Channel fading and noise perturbation in the model aggregation raise unique challenges for the edge training algorithm design and analysis. To tackle the channel fading perturbation, a channel inversion method was proposed in [44], [45], [173] by multiplying the inverse of channel gain for the transmit signal, which may however not satisfy the power constraint at edge devices. To address these issues, a transceiver design was provided in [31] to minimize the distortion for the perturbed model aggregation, whereas the perturbed model updates are directly incorporated in the FL algorithm design [166]. Although the analog transmission in AirComp is prone to channel noise, the additive noise in the model aggregation turns out to be controllable or even beneficial in the edge training process. Specifically, the channel noise in the model aggregation yields a new class of noisy FL algorithms. The convergence behavior demonstrates that the noisy iterates typically introduce non-negligible optimality gap in various FL algorithms, e.g., vanilla gradient method [174], quantized gradient method [175], sparsified gradient method [173], and operator splitting method [87]. The optimality gap can be further controlled by transmit power allocation [41], [173], [176], model aggregation receiver beamforming design [31], [177], [178], and device scheduling [31], [178], [179]. Besides, channel perturbation in algorithm iterates can also serve as the mechanism to design saddle points escaping algorithms [94], thereby establishing global optimality for training the non-convex over-parameterized neural networks in high-dimensional statistical settings [180]. The additive channel noise in model aggregation can also serve as an inherent privacy-preserving mechanism to guarantee differential-privacy levels for each edge device without sacrificing learning performance [41].

2) Massive Access Techniques: Deploying cross-devices FL in IoT networks raises practical challenges, i.e., the IoT devices have sporadic access to the wireless network [181]. It is thus critical to design practical FL systems to accommodate flexible device participation with sporadic access to the wireless network [143], as shown in Fig. 6 (b). The grant-free random access protocol provides a low-latency and low signaling overhead way to detect the active devices, followed by decoding their corresponding information data [75], [182], [183]. In this protocol, active devices can transmit the data signals directly without waiting for any permission. Sparse signal processing provides a promising modeling framework to simultaneously detect the active devices and estimate their channels [75], [76], which is supported by various efficient algorithms, including the approximate message passing algorithm [184], [185] and DNN algorithm unrolling approach [97], [186]. To further reduce the latency for data decoding in random access, a sparse blind demixing framework was developed in [187] by simultaneously performing active device detection, channel estimation and their data decoding. The key observation is that blind demixing is able to perform low-latency data decoding for multiple users from the sum of bilinear measurements without channel estimation at both the transmitters and receivers [188], [189]. To enhance the performance, the common sparsity pattern in pilot and user data has been exploited via joint activity detection and data decoding [190], [191].

Random access protocols are promising to support flexible and massive device participation in the edge training process by identifying active devices with sporadic traffic. It is still critical to develop massive access techniques to improve the learning performance by enrolling more active devices to perform local model update and exchange under digital transmission. Nonorthogonal multiple access (NOMA) [77], [78] is a key enabling candidate technology to simultaneously serve massive devices for model aggregation in the same radio resource block via superposition coding. Typical NOMA schemes include the power-domain NOMA with different transmit powers as weight factors and the code-domain NOMA (e.g., sparse code multiple access [192] and pattern division multiple access [78]) with different codes assigned to users. Therefore, the user's data can be decoded from the simultaneously transmitted signals via successive interference cancellation. In particular, DL provides a powerful method to design and optimize NOMA systems [193]–[195]. Under analog uncoded transmission, interference can be harnessed via the new massive access techniques AirComp, for which Dong *et al.* further proposed a blind AirComp for low-latency model aggregation without channel state information (CSI) access [79]. It is thus particularly interesting to integrate a massive random access protocol (e.g., grant-free random access) and massive access technique (e.g., AirComp based access technique) with analog uncoded transmission to simultaneously perform active device detection, channel estimation and model aggregation, thereby supporting flexible and low-latency edge devices enrolling for collaboratively training the models.

3) Ultra-Massive MIMO: Leveraging massive antenna arrays is a key enabling wireless technology to achieve high spectral and energy efficiency, which is envisioned to be further scaled up by an order-of-magnitude in 6G [8]. The recent advances in digital beamforming, analog beamforming, as well as hybrid beamforming have helped the roll-out of massive MIMO into practice by operating over a wider frequency band. It has been demonstrated that massive MIMO is able to bring enormous benefits for edge training systems, including high-accuracy and high-rate for model aggregation, as well as high-reliability for massive device connectivity. Specifically, massive MIMO can achieve a high computation accuracy for model aggregation via exploiting spatial diversity [196], and enable ultra-fast model aggregation with simultaneous multi-functions computation by spatial multiplexing [197]. Furthermore, for FL with edge devices sporadically enrolling, the device activity detection error goes to zero as the number of antenna elements in the BS goes to infinity, thereby achieving high-reliable devices participation for model updates.

To scale edge training to huge physical areas with massive geographically distributed edge devices, ultra-dense wireless network is a promising way to achieve low-latency, high-reliability and high-performance. This is achieved by simultaneously uploading massive local model updates with multiple distributed edge servers with abundant communication, computation, and storage resources, thereby mitigating the stragglers issues (i.e., devices with low communication and computation capabilities may prolong the training time) and unfavorable channel dynamics. Besides, compared with the single edge server architecture, distributed edge servers are robust to server failure issues for reliable edge training. In particular, cloud radio access network (Cloud-RAN) [198], [199] provides a cost-effective way to implement distributed antenna aided edge training systems, for which reliable model aggregation via AirComp can be achieved by centralized signal processing and shortening the communication distances between edge devices and edge servers [200]. The recent proposal of cell-free massive MIMO [201] serves a promising way to realize the wireless distributed FL systems by exploiting the channel hardening characterization (i.e., the effective channel gain is approximated by its expectation value) and avoiding sharing instantaneous CSI among edge servers [46], as shown in Fig. 6 (c).

4) Reconfigurable Intelligent Surfaces: To obtain the desired average function of local model updates for model aggregation via AirComp, magnitude alignment by scaling the transmit signals (e.g., channel inversion) is normally required to reduce the channel perturbation [202]. However, due to the resource-limited edge devices and the non-uniform fading channels, the unfavorable signal propagation environment inevitably leads to magnitude reduction and misalignment with perturbed model aggregation, which in turn degrades the learning performance of the edge training process. Besides, the massive edge devices with sporadic access to the edge servers can be located at a service dead zone, which makes device activity detection challenging for weak channel links [185]. To enroll multiple edge devices via simultaneously transmission with NOMA, sufficient diversified channel gains are normally required for successive interference cancellation, which however may not always hold in practical scenarios [203]. Heterogeneity in terms of computation, communication, and storage across edge devices is one of major challenges to deploy edge AI systems. Waiting for the straggler edge devices with slow computation and communication speeds for model aggregation causes significant delays, which can be tackled by computation offloading and task scheduling by mobile edge computing (MEC) technique [204]. However, fully unleashing the benefits of MEC for straggler mitigation is limited by the hostile wireless links [205].

To address the above challenges in terms of propagation impairments, RIS has been shown to be a cost-effective technology to support fast yet reliable model aggregation with massive edge devices participation by programming the propagation environment of electromagnetic waves [118], [206], as shown in Fig. 6 (d). Specifically, RIS is typically realized by planar or conformal artificial metamaterials or metasurfaces equipped with a large number of low-cost passive reflecting elements, which are capable of adjusting the phase shifts and amplitudes of the incident signals for directional signal enhancement or nulling, and thus altering the propagation of the reflected signals [49], [207], [208]. To design an RIS-empowered edge training system, RIS can be leveraged to align the magnitudes of the transmit signals by establishing favorable propagation links in waveform superposition for AirComp, resulting in boosted received signal power and accurate aggregated function at the edge server [209]. The boosted model aggregation via RIS can support efficient edge devices scheduling in over-the-air FL, thereby adapting to the time-varying local model updates and channel dynamics [48], [178]. The reliable sporadic access in edge training can be developed by establishing abundant propagation scatters using RIS for accurate activity detection [185]. The latency for local model updates of the active devices can be further reduced by establishing favorable propagation links via RIS, thereby mitigating stragglers [205].

5) Space-Air-Ground Integrated Networks: The typical SAGIN [51], [210] provides an integrated space information platform across the satellite networks (e.g., miniaturized satellites [211]), aerial networks (e.g., UAV communications [212]), and terrestrial communications (e.g., vehicular communications [213]) to provide ubiquitous connectivity for various edge training architectures, as shown in Fig. 6 (e). Edge learning over a vehicle-to-everything network is critical to enable autonomous driving with delay-sensitive applications [145]. In this scenario, the local model updates need to be fast and reliably aggregated within neighbors via vehicle-to-vehicle communications [181], or to the roadside units via vehicle-to-infrastructure communication. In particular, radar sensing provides a promising way to predict the vehicular links [214] and holds the potential to provide real-time model aggregation via predictive beamforming in the model aggregation procedure. In the scenario with sparsely deployed edge servers and moving edge devices (e.g., ground vehicles), UAV, serving as the flying edge servers, can provide a promising solution to aggregate local model updates in the whole procedure of edge training by joint UAV trajectory and transceivers design over dynamic wireless edge networks [81].

To build a scalable edge training system with massive devices participation for training extremely deep AI models [215], it is critical to access abundant computation resources across the continuum of nodes from edge devices, edge servers, to cloud servers [50]. It was shown in [130] that the client-server-cloud multi-layer architecture is able to significantly reduce the training time and energy consumption. In the scenario without abundant edge and cloud computing infrastructures, SAGIN provides an ubiquitous computing platform for the multi-layer hierarchical edge learning system, where the flying UAVs serve as the proximal edge computing, and the low earth orbit satellites serve as the relays to the cloud computing [216]. To realize SAGIN empowered edge training system, tier-adaptive aggregation interval management becomes critical to control the local and global model aggregation intervals [130] to achieve high communication efficiency. Besides, the client-edge-satellite association with dynamic scheduling and offloading is fundamental to tackle the heterogeneity challenges in terms of system resources and network topologies.

In summary, this section presented multiple access technologies (e.g., AirComp, grant-free random access, NOMA), multiple antenna techniques (e.g., Cloud-RAN, cell-free massive MIMO, RIS), and multiple layer networks (e.g., UAV, SAGIN) that are needed to support low-latency model aggregation and diversified learning architectures and environments. We hope this can inspire more advanced 6G wireless and information techniques (e.g., millimeter-wave and terahertz (THz) communications [217], [218], age of information [219]) to support edge AI systems for establishing integrated communication, computation and learning ecosystems.

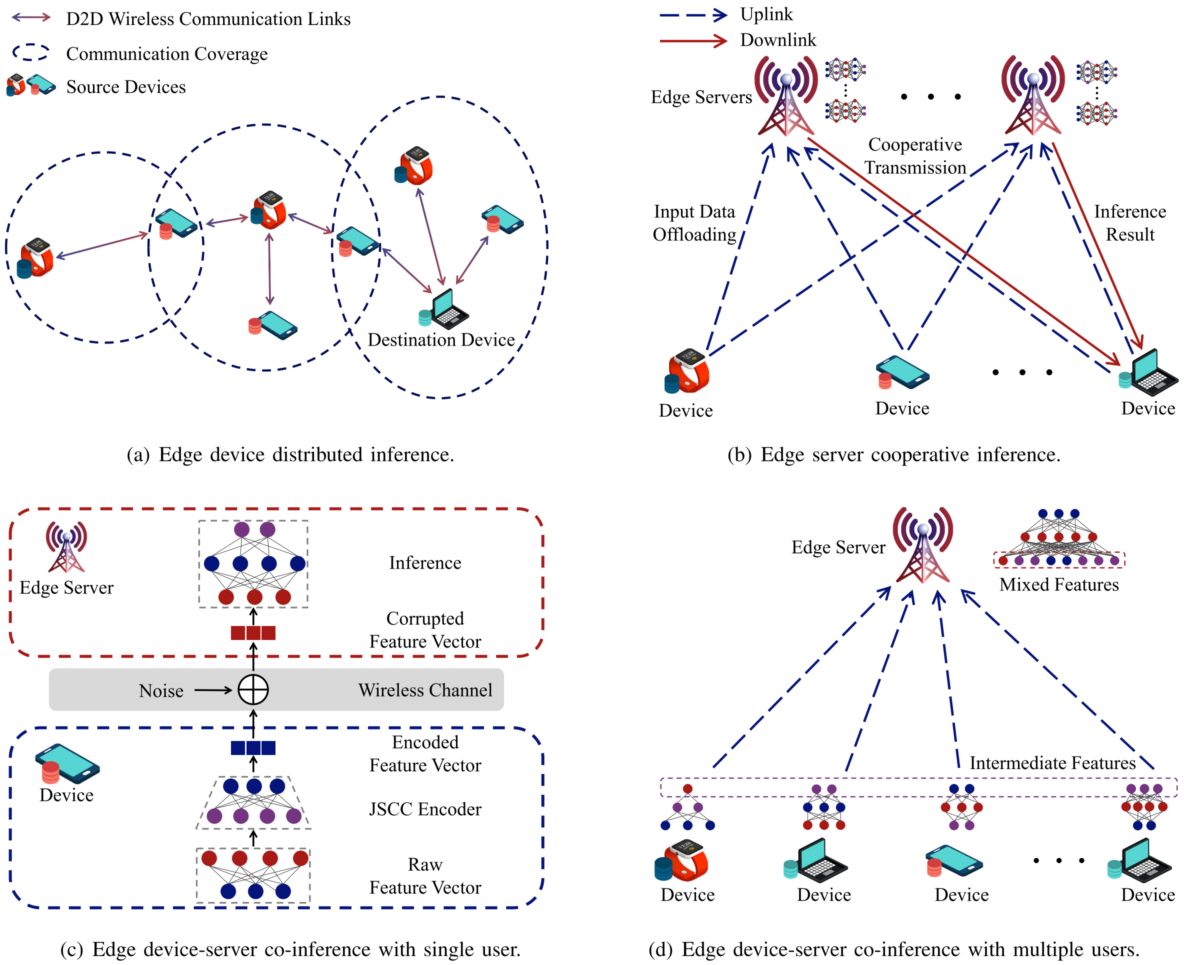
III. Communication-Efficient Edge Inference

In this section, we present communication-efficient techniques for edge inference tasks with latency and reliability guarantees. Based on the dataset distribution characteristics, Yang *et al.* [33] proposed to categorize FL as horizontal FL (i.e., datasets share the same feature space but different sample space) and vertical FL (i.e., datasets share the same sample space but differ in the feature space). Hosseinalipour *et al.* [50] further proposed a fog learning framework by allowing both vertical communications (i.e., model updates are only exchanged across different network layers) and horizontal communications (model updates can be exchanged between devices in the same network layer). In a similar way, based on different computing collaboration schemes, we shall propose to categorize edge inference as horizontal edge inference (i.e., computation resources can only be harvested among edge devices, or only be pooled among edge servers), and vertical edge inference (i.e., computation resources can be harnessed between edge devices and edge servers), which are discussed in the following two subsections, respectively.

A. Horizontal Edge Inference

We consider two different types of horizontal edge inference, as shown in Fig. 7 (a) and Fig. 7 (b).

1) Edge Device Distributed Inference: Enormous efforts on TinyML with DL model compression and neural network architecture search have been conducted to enable low-latency



(a) Edge device distributed inference.

(b) Edge server cooperative inference.

(c) Edge device-server co-inference with single user.

(d) Edge device-server co-inference with multiple users.

Fig. 7. Communication-efficient edge inference systems.

and energy-efficient model inference on a single device with limited storage and computation resources [220]. However, due to limited storage capability at edge devices, it becomes extremely difficult to accomplish inference computation tasks at a single device, for applications such as mobile navigation with a huge map information dataset [221]. Edge device distributed inference based on wireless MapReduce enjoys the advantages of providing low-latency, high-accurate, scalable, and resilient services for edge devices without accessing the cloud data center [12], [29]. Specifically, edge device distributed inference involves computing the intermediate values based on the local input datasets using the map function, followed by sharing the intermediate values via horizontal communication among edge devices, thereby constructing the desired computation or inference results using the reduce function [52], [222].

To tackle the communication bottleneck for shuffling intermediate values in the edge device distributed inference process, a coded distributed computing approach [223] was adopted in [221] to improve the scalability of wireless MapReduce by inducing the coded multicasting transmission opportunities. This, however, sacrifices computation efficiency as computation replication of the local dataset is needed. To further improve the spectral efficiency, instead of reducing the volume of communication bits [221], a joint uplink and downlink design approach based on the interference alignment principle was developed in [52] to improve the communication data rates for local intermediate values shuffling. In particular, to compute the nomographic function [224] for edge device distributed inference based on the MapReduce decomposition, a multi-layer hierarchical AirComp approach was proposed in [225] to improve the spectral efficiency over the multi-hop D2D communication network, as shown in Fig. 7 (a).

2) Edge Server Cooperative Inference: DL with high-dimensional model parameters is able to provide high accurate intelligent services. However, it is challenging to directly deploy such large AI models on IoT devices due to very limited onboard computation, storage and energy resources. Deploying and executing DL models on edge servers turns out to be a promising solution. However, the limited wireless bandwidth between edge devices and edge servers becomes the key bottleneck [53], [54] for edge server cooperative inference. Compressing and encoding the input source data at edge devices are essential to reduce the uplink communication overheads, for which various data dimensionality reduction approaches have been proposed by exploiting the specific computation tasks and communication environments [29]. Besides, for the applications with high-dimension output inference results (e.g., the output of the NVIDIA's AI system GauGAN is a large-sized photorealistic landscape image), it is equally important to design highly efficient downlink communication solutions for delivering the output inference results for the edge devices [53], [54].

Computation replication has been shown to be effective for reducing the communication latency in computation offloading when the output size is large [226]. This is achieved by executing each inference task at multiple edge servers, followed by delivering the inference results for multiple edge devices via downlink cooperative transmission [227]. Although edge server cooperative inference via downlink transmission cooperation is able to significantly improve communication efficiency by mitigating interference and alleviating channel uncertainties, it causes extra energy consumption to execute the same inference tasks at multiple edge servers. To design a green edge server cooperative inference system, joint inference task selection and downlink coordinated beamforming framework was proposed in [53] to minimize the overall computation and communication energy consumption, as shown in Fig. 7 (b). RIS was further leveraged in [54] to design green edge server cooperative inference systems by considering both uplink and downlink transmit power consumption. The rate splitting method is also anticipated to be able to further improve the energy-efficiency for edge server cooperative inference by partially decoding the inference result and partially treating it as noise in a flexible way [228].

B. Vertical Edge Inference

We consider two different cases of vertical edge inference, as shown in Fig. 7 (c) and Fig. 7 (d), with a single edge device and multiple edge devices, respectively. In the following, we shall first present effective techniques for communication-efficient vertical edge inference for these two cases, and then present a new general design principle for resource-constrained vertical edge inference, named *task-oriented communication*.

1) Edge Device-Server Co-Inference: Edge device distributed inference enjoys low-latency whereas it has limited accuracy due to limited processing capabilities and limited bandwidth. Although edge server cooperative inference is able to achieve high accuracy with DL models, it may raise data leakage issue and excessive communication delay. It thus becomes inapplicable for privacy-sensitive and delay-sensitive applications. To provide ubiquitous AI services across diversified application scenarios, edge device-server co-inference, as a complementary solution to horizontal edge inference, is promising to alleviate the communication overheads while achieving high accuracy and privacy for inferring the DNN models. This is achieved by dividing the DNN model into a computational friendly segment at the edge device, and the remaining computational heavily segment at the edge server [32], as shown in Fig. 7 (c). By adaptively partitioning the computation burdens between the edge devices and edge server, model split selection for the neural network is essential to achieve optimal computation-communication trade-off in the vertical edge inference system via edge device-server synergy and collaboration [82]. To further reduce the communication overheads, a communication-aware model compression approach was proposed in [32] to limit the number of the activated neurons at the last layer of neural network deployed at the edge device. However, the short message transmission [37] and data amplification effect [229] of the output features extracted by the on-device split model raise unique challenges to realize real-time vertical edge inference.

2) Ultra-Reliable and Low-Latency Communication: The packet length of the extracted output features transmitted from the edge devices can be very short [83], [145], for which the achievable data rate in such a finite block length regime is penalized by a non-vanishing decoding error probability [83]. Besides, the output inference results from the edge server should be delivered to the edge devices with latency and reliability guarantees for mission-critical applications. Considering the system dynamics, including task arrival dynamics in the network layer and the wireless channel dynamics in the physical layer, cross-layer optimization is needed to minimize the end-to-end delay for edge device-server co-inference [84], [85]. In particular, MDP supported by linear programming was adopted in [85] to jointly schedule the transmission at edge devices and computation at the edge server for achieving the optimal power-latency tradeoff for edge device-server co-inference via MEC. The random delay characteristics were also investigated in [230] by modeling the coupled transmission and computation process as a discrete-time two-stage tandem queueing system. To support multiple edge devices for uploading intermediate features using short packet transmission, massive MIMO can be adopted to combat channel fast fading and provide a nearly deterministic communication environment due to channel hardening [231]. The received multiple intermediate features can be further aggregated via the mixup augmentation technique [232] to enable scalable and cooperative inference at the edge server, as shown in Fig. 7 (d).

3) Task-Oriented Communication: As revealed in [32], there exists an intrinsic communication-computation trade-off in resource-constrained vertical edge inference. This is mainly caused by the data amplification issue in DL based inference, namely the dimension of the intermediate feature may be larger than the input data size. Thus, if only a few layers of the neural network were deployed on the edge device, the output feature would have a size larger than the input data, yielding too much communication overhead. To reduce the intermediate feature size, more layers have to be deployed on the edge device, which however will lead to high local computation burden. To resolve this tension between local computation and communication overhead, it is of critical importance to effectively compress and transmit the intermediate feature. Such a communication task is fundamentally different from *data-oriented communication* in current wireless networks, i.e., to transmit a binary sequence at the highest data rate for reliable reconstruction at the receiver. In vertical edge inference, the feature transmission is for the inference task, not for reconstructing the feature vector with high fidelity. Thus, as advocated in [32], we should rather design the communication scheme for feature transmission in a task-oriented manner, i.e., only transmitting the informative messages for the downstream inference task at the edge server. Instead of decoding the intermediate features, the received signal corrupted by channel fading and noise is directly processed at the edge server to obtain the inference results.

This *task-oriented communication* principle constitutes a paradigm shift for the communication system design from data recovery to task accomplishment. It was first tested in vertical edge inference via end-to-end training with joint source-channel coding in [233], which helps to reduce both the communication overhead and on-device computation cost. Such design principle has also been applied in other tasks. For example, the DL based end-to-end semantic communication system was developed in [21] via joint semantic source and wireless channel coding for recovering the meaning of sentences instead of the original transmitted data samples. The analog JSCC approach was presented in [234] to compress and then code the feature vectors, followed by leveraging the received perturbed signal directly for wireless image retrieval at the edge server via a fully-connected neural network. Recently, a novel and generic design framework for task-oriented communication was developed in [20], which is based on the information bottleneck formulation [235]. This framework provides a principled way to extract informative and concise representation from the intermediate feature, which is made mathematically tractable via variational approximation. Furthermore, it has been extended to the cooperative inference scenario with multiple edge devices in [86] based on distributed information bottleneck [236] and distributed source coding theory.

In summary, this section presented interference coordination techniques and task-oriented low-latency communication principles for horizontal edge inference and vertical edge inference, respectively. We hope this can motivate the co-design of wireless communication networks and deep learning models to deliver low-latency, energy-efficient and trustworthy edge AI inference services.

IV. Resource Allocation for Edge AI Systems

In this section, we shall characterize the engineering requirements for designing communication-efficient edge AI systems, including accuracy, latency, energy, privacy and security. Effective service-driven resource allocation methods based on mathematical programming and ML are then provided to achieve scalability and trustworthiness for edge AI systems.

A. Engineering Requirements and Methodologies

We identify the engineering requirements for designing scalable and trustworthy edge AI systems. Resource allocation strategies must cater to the needs of edge AI systems for achieving accurate intelligence distillation into the edge network at an ultra-low power and low-latency cost.

1) Accuracy: The edge training process involves designing the global iterates $\boldsymbol{\theta}^{[t]}$ with t as the iteration index to minimize the empirical loss function while achieving fast convergence rates with negligible optimality gap for problem (1). To design efficient resource allocation schemes in edge training systems, it is particularly important to characterize the convergence behaviors for the global iterates $\boldsymbol{\theta}^{[t]}$, which typically depend on the scheduled devices, local updates, aggregation behaviors, network topologies, propagation environments, function landscapes, and underlying algorithms. Specifically, for edge training systems via AirComp, the global model aggregation errors due to the wireless channel fading and noise will cause learning performance degradation [45], [87]. The optimality gap (i.e., the distance between the current iterate and the desired solution), characterized by the convergence behavior of the global iterate, can be further controlled by various resource allocation schemes, including edge devices transmit power control [41], [237], edge server receive beamforming [31], [179], passive beamforming at RIS [48], [178], as well as device scheduling policy [31], [48]. For digital design of the edge training system, the optimality basically depends on the edge devices selection, packet errors in the uplink transmission, and model parameter partition, for which user scheduling [238], power control [88], batchsize selection [239], aggregation frequency control [240], and bandwidth allocation [241] were provided to improve the accuracy in the edge training process.

For edge inference, the accuracy indicates the quality of the inference results for a given task. It is typically measured by the number of correct predictions from inference, e.g., the classification tasks. For computer vision applications in autonomous driving, ultra-high accuracy for the DNN model inference is demanded. For applications in radio resource allocation via distributed ML, the accuracy of inferring a DNN model can be moderate. The accuracy of edge inference depends on the difficulty of the tasks and datasets, the quality of the trained model, the dynamics of wireless communication and edge computation environments, as well as the methods for processing the models, datasets and features. In particular, for horizontal edge inference via AirComp aided wireless MapReduce, the accuracy for computing a nomographic function is fundamentally limited by the channel fading and noise, for which various transceivers were designed to minimize the mean square error for inference computation tasks [225]. The accuracy of vertical edge inference depends on the informativeness and reliability of the intermediate features transmitted from edge devices, as well as the dynamic wireless environments, for which an ultra-reliable communication and adaptive JSCC approach need to be developed to improve the inference performance. In particular, information bottleneck was adopted in [20] to characterize the relationship between the accuracy of the vertical edge inference and the communication overhead of the intermediate features.

2) Latency: For edge training, the latency consists of computation latency and communication latency. The computation latency highly depends on the computation capability of the edge devices and servers, as well as the size of the models and datasets. The communication latency is the sum of the transmission latency of one round with respective to the total learning rounds until convergence for training the global model. In one typical training round, the communication delay in the uplink and downlink transmissions for model updates, is mainly affected by the wireless communication techniques, bandwidth and power budgets, wireless channel conditions, as well as the scheduled edge devices. Li *et al.* characterized the delay distribution for FL over arbitrary fading channels via the saddle point approximation method and large deviation theory [89]. The trade-off between the convergence speed and the per-round latency was revealed in [242] based on the key observation that more scheduled devices yield faster convergence rate while prolonging the time of uploading the local updates at each iteration due to limited radio resources. A probabilistic device scheduling policy was further proposed in [90], [243] to minimize the overall training time in wireless FL. Besides, the trade-off between the local computation rounds for local model updates and the global communication rounds for global model updates is characterized to guide the resource allocation for minimizing the total learning time and energy consumption [244]. The convergence speeds of FL algorithms were characterized in [245] by considering non-identical dataset distributions, partial edge devices participation, and quantized model updates in both uplink and downlink communications.

In the case of edge inference, the latency measures the time between the data arrival to the generation of the inference results through the edge AI system. It consists of the data pre-processing, data transmission, model inference, and result post-processing, which highly depend on the computation hardware, communication schemes, DL models and tasks. For the real-time mobile computer vision application of AR/VR, stringent latency requirements are required, e.g., 100ms. For scalable radio resource allocation application via DL, the inference latency must be within the channel coherence time (e.g., 10ms) to yield a meaningful resource allocation decision [23]. A low-rank matrix optimization based transceiver design approach was proposed in [52] for fast shuffling intermediate values in wireless distributed computing, thereby reducing the latency for horizontal edge inference via edge devices collaboration. For vertical edge inference, the dynamic computation partition and early existing scheme was proposed in [82] to accelerate the inference speed via edge device-server synergy. The cross-layer design approach was adopted in [85] to reduce the communication and computation latency for the time-sensitive edge inference computing applications. In particular, the DL enabled task-oriented communication framework was developed to achieve low-latency edge device-server co-inference by merging feature compression, source coding and channel coding for the specific inference tasks [20], [234].

3) Energy: For edge training, the energy consumption consists of the computation and communication process. For AlphaGo, it may cost 280 GPUs and a $3000 electric bill per game [246]. It is therefore critical to design energy-efficient edge training systems to minimize carbon dioxide footprint for contributing the carbon neutrality target. Such a design is mainly dictated by the size of training models, model training algorithms, and wireless transmission strategies and hardware (e.g., the scaled SiGe bipolar technology [247]), and edge computing architectures and hardware. Both computation energy consumption for local model updates and communication energy consumption for uploading local updates are simultaneously minimized in [92] by considering the learning latency and accuracy constraints for wireless FL. The wireless power transfer approach was further adopted in [248] to power the edge devices for local model computation and communication, for which the active devices with enough harvested energy will contribute to accelerate the learning procedure. To deploy AirComp-assisted FL across massive IoT devices with a limited battery capability, microwave based wireless power transfer supported by RIS was adopted in [91] to recharge the IoT devices via energy beamforming at edge server and passive beamforming at RIS.

In the case of edge inference, it becomes particularly important to achieve high energy efficiency for processing the DNN models at the network edge with battery-limited devices. The energy consumption of executing a DNN model is highly dictated by the computation architecture and methods (e.g., ultra-low power compute-in-memory AI accelerator) at the edge computation nodes [249], the architecture of DNN models [250], and the wireless transmission for data exchange during the model inference procedure. For horizontal edge inference via wireless cooperative transmission at multiple edge servers, the sum of the computation and transmission power consumption for generating and delivering the inference results were minimized via downlink coordinated beamforming [53]. Energy consumption at the edge devices can be minimized in the cross-layer design for delay-sensitive edge device-server co-inference by computation offloading [85]. Besides, energy harvesting becomes a promising technology for the edge computing based vertical edge inference by providing renewable energy resources for edge devices [251].

4) Trustworthiness: Trustworthiness is one of the main drivers for developing the next generation AI technologies. Specifically, the developed AI models and algorithms must be privacy-preserving, adversarial-resilient, robust, fair, optimal and interpretable [95]. For edge training, privacy mainly depends on the offloading or coding of the raw data and intermediate features. Keeping datasets at devices is a direct and effective way to preserve user's privacy in FL. Besides, the wireless channel noise yields a noisy model aggregation procedure via AirComp, which provides an inherent privacy-preserving mechanism to enhance differential-privacy for each edge device. An adaptive power control method was further developed in [41] to control the differential-privacy levels in this over-the-air FL system, while avoiding the learning performance degradation. To address the adversarial attacks, the blockchain based decentralized learning was proposed in [252] to enable secure global model aggregation by using a consensus mechanism of blockchain. The block generation rate was optimized by considering the communication, computation and consensus delays in the blockchain enabled secure edge learning systems [252], [253]. For edge inference, privacy and security are mainly dictated by the way of processing the input data, of transmitting the inference results, as well as the computation methods for model inference (e.g., secure multi-party computation).

Establishing optimality for ML algorithms is important to deliver reliable and responsible AI services. However, empirical risk minimization for training the models is usually nonconvex, which poses significant challenges to guarantee global optimality for the learning algorithms and models [180]. Fortunately, under the high-dimensional statistical setting, the local strong convexity and smoothness of the nonconvex loss functions can be exploited to tame the nonconvexity for various learning models, e.g., blind demixing [189], phase retrieval [254], and shallow neural networks [255]. Besides, with high-dimensional datasets, the nonconvex loss functions of certain statistical learning models, including over-parameterized neural networks [180] and dictionary learning [256], can enjoy benign global geometric landscape such that all the local minima are global minima, and all the saddle points can be escaped efficiently using the algorithms including trust region method and perturbed gradient descent method [257]. In particular, for edge training, the channel noise yields a perturbed stochastic gradient descent method to escape saddle points for distributed principal component analysis via AirComp [94]. Therefore, channel noise can provide a mechanism for both preserving differential privacy [41] and achieving global optimality [94]. These evidences indicate that we should embrace channel fading and noise for achieving trustworthy edge AI.

5) Service-Driven Resource Orchestration: Edge AI systems need to incorporate various wireless network architectures and communication strategies by integrating communication and computation. This will result in a highly complex and dynamic network, which requires innovative technologies and solutions. Various use cases (e.g., autonomous driving, industrial IoT, and smart healthcare) and heterogeneous requirements in terms of accuracy, latency, energy and trustworthiness, would further aggravate the complexity for resource allocation in edge AI systems. Besides, the complex edge servers and base stations will be quite energy-consuming, which brings formidable challenges for achieving high energy efficiency. To enable efficient resource allocation, it is thus critical to precisely model the heterogeneous demands for edge AI services, and reversely matching them with proper network resource orchestration. This, however, relies on the quantitative relationship between network resources and user requirements for edge AI tasks. To pave the way for this paradigm shift for service-driven resource allocation in edge AI systems, in the next subsection, we shall provide various intelligent optimization models and algorithms to adapt to diversified network environments and services.

B. Optimization Models and Algorithms

The service-driven network resource management problems for edge AI systems can be classified as a parametric family of mathematical optimization problems:

$$\begin{aligned} &\underset{\boldsymbol{z}}{\mathsf{minimize}} \;\; f_0(\boldsymbol{z};\boldsymbol{\alpha}) \\ &\mathsf{subject\ to}\; g_i(\boldsymbol{z};\boldsymbol{\alpha}) \le 0, \quad i=1,\ldots,m, \\ &\qquad\qquad h_i(\boldsymbol{z};\boldsymbol{\alpha}) = 0, \quad i=1,\ldots,p, \end{aligned} \tag{2}$$

where $\boldsymbol{z} \in \mathbb{R}^n$ is the optimization variable vector consisting of both discrete and continuous variables, $\boldsymbol{\alpha} \in \mathcal{A}$ is the problem parameter vector with $\mathcal{A}$ denoted as the parameter space (e.g., CSI). For each fixed $\boldsymbol{\alpha} \in \mathcal{A}$, $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective function (e.g., optimality gap in edge training), $g_i : \mathbb{R}^n \to \mathbb{R}, i=1,\ldots,m$ are the inequality constraint functions (e.g., latency requirements in edge inference), and $h_i : \mathbb{R}^n \to \mathbb{R}, i=1,\ldots,p$ are the equality constraint functions. The resource allocation optimization problems are typically categorized as mixed-combinatorial optimization, nonconvex continuous optimization, stochastic optimization, and end-to-end optimization. To provide scalable, real-time, parallel, distributed and automatic resource allocation schemes, we shall propose to exploit the landscape of the underling optimization problems (2) by the theory-driven method based on mathematical programming, followed by developing the novel data-driven approach based on machine learning to achieve real-time and distributed implementations, as well as improved and robust performance, as shown in Fig. 8. Here, $\psi(\boldsymbol{\alpha})$ is a mapping function to map the problem parameter $\boldsymbol{\alpha}$ to the optimal solution of problem (2).

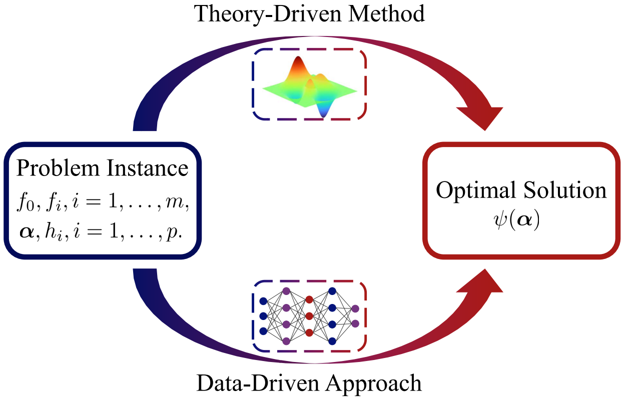

Theory-Driven Method
Problem Instance
$f_0, f_i, i=1,\ldots,m,$
$\boldsymbol{\alpha}, h_i, i=1,\ldots,p.$
Optimal Solution
$\psi(\boldsymbol{\alpha})$
Data-Driven Approach

Fig. 8. Resource allocation optimization methods for edge AI systems.

1) Mixed-Combinatorial Optimization: The resource allocation problems in edge AI systems involve optimizing across learning, computation and communication. Specifically, for edge training systems, we need to jointly optimize the subcarrier and bandwidth allocation [88], [90], [241], transmit power and receive beamforming [31], [48], [178], passive beamforming at RIS [48], [178], device selection [31], [242] and activity detection [76], local updates computation [92], and global aggregation frequency control [130], thereby reducing the optimality gap and energy consumption in the distributed learning procedure. For edge inference via collaboration among edge servers, task selection, coordinated downlink beamforming among edge servers, as well as passive beamforming at RISs were jointly optimized to achieve green edge inference [53], [54]. All of these resource allocation schemes can be formulated as a mixed combinatorial optimization problem, which needs to jointly optimize continuous-valued variables (e.g., beamforming and power control) and discrete-valued variables (e.g., device selection and subcarrier allocation). In particular, sparse optimization provides a powerful modeling approach to solve the mixed combinatorial resource allocation problems by exploiting the sparsity structures in the optimal solutions [96]. For instance, the group sparsity can represent the combinatorial variables for edge devices selection in FL [31], edge devices activity detection [76], and inference tasks selection [53]. The algorithmic advantages of the sparse optimization modeling approach are supported by various convex relaxation algorithms [198], [258], e.g., mixed ℓ_1/ℓ_2-norm minimization [80]. A typical sparse and low-rank optimization modeling and algorithmic framework was developed in [31] to support the joint device selection and transceiver design for improving learning performance in over-the-air FL systems.

Although operation research provides a theory-driven approach for solving the mixed combinatorial optimization problem or its equivalent sparse optimization problem, the existing algorithms are either heuristic with noticeable performance loss or optimal with intolerably high computation complexity. To address these challenges, "learning to optimize" provides a data-driven design paradigm to improve the computation efficiency and system performance for resource allocation [114], [259]. This is achieved by developing computationally efficient optimization methods by learning from the sampled problem instances using training models and methods. The learned algorithms can be furthered executed online and distributed for real-time resource allocations in edge AI systems. To solve the large-scale mixed combinatorial optimization problem efficiently, imitation learning was adopted in [22] to learn an aggressive pruning policy in the globally optimal-achieving branch-and-bound algorithm. This learning based brand-and-bound method can significantly save the time for pruning the nodes in the search tree, achieve near-optimal performance with few training samples, as well as guarantee feasibility of constraints without performance degradation. To further speed up the sparse optimization method for the mixed combinatorial optimization problem in edge device activity detection, the DNN based algorithm unrolling framework was developed in [97] to achieve theoretical guarantees, performance improvements, interpretability and robustness for the learned sparse optimization algorithms [186]. This is achieved by mapping the theory-driven iterate operations, i.e., iterative shrinkage thresholding algorithm, into an unrolled recurrent neural network, followed by training the model parameters based on supervised learning. Besides, a multi-agent RL approach was developed in [260] to solve the distributed mixed combinatorial optimization problem for task offloading and resource allocation in multi-layer edge inference systems.

2) Nonconvex Optimization: Most of the resource allocation problems in edge AI need to solve a series of nonconvex optimization problems, e.g., nonconvex sparse optimization for device selection in wireless FL, nonconvex quadratic programming for transceiver design in over-the-air FL [31], low-rank matrix optimization for interference management in edge device distributed inference [52], and unit modulus constrained phase shifts optimization [261] in RIS-empowered edge AI systems [48], [54]. Convex approximation provides a natural way to design polynomial time complexity algorithms for nonconvex programs based on the principle of majorization-minimization [262] or successive convex approximation [263]. A two-stage framework was provided in [98] for solving general large-scale convex programs with infeasibility detection and scalable computation. This is achieved by matrix stuffing technique for fast conic program modeling in the first stage, and operator splitting method for scalable conic program solving in the second stage [98]. Although the semidefinite relaxation approach [264] is able to convexify the general quadratic programs by matrix lifting and dropping the resulting rank constraints, it fails to return high quality solutions in the high-dimensional settings. This issue was addressed by a difference-of-convex-functions (DC) programme [31], [52], [99] by representing the rank function via an equivalent DC function. This DC optimization modeling and algorithmic framework was typically applied to solve the nonconvex passive beamforming problem in the RIS-empowered FL systems [48] and edge inference systems [54]. To solve the large-scale rank constrained matrix optimization problems, Riemannian manifold optimization was proposed to optimize such nonconvex programs directly by exploiting the manifold geometric structures of fixed-rank matrices [100], [101].

To further enable real-time, automatic and distributed design of nonconvex optimization algorithms for resource allocation in edge AI systems, DL was shown to have great potentials for achieving this goal. A multi-layer perceptron was adopted in [265] to directly learn the mapping from the problem instance to the output solution generated by the weighted minimum mean square error (WMMSE) algorithm for nonconvex precoding design [266]. Instead of running the iterates, the learned algorithm via deep learning can be executed in real-time, as neural networks only involve computationally cheap operations, e.g., matrix-vector multiplication. To reduce the model and sample complexity, as well as improve the performance and interpretability, unfolded neural networks were developed in [267]–[269] to parameterize the iterative policy via unfolding one iteration of the existing structured algorithm into one layer of a neural network. Graph neural network (GNN) has recently been shown to be able to harness the benefits of generalizability, interpretability, robustness, scalability, superior performance, real-time and distributed implementation for learning to optimize nonconvex problems, including power control [270], beamforming [23], and phase shift design [25]. This is achieved by modeling wireless network as a graph, followed by using a GNN to parameterize the mapping function $\psi(\boldsymbol{\alpha})$ for the optimal solution.

3) Stochastic Optimization: In large-scale edge AI systems, the estimated CSI will be inevitably imperfect or partially available [177], [271]. It is thus critical to design practical resource allocation schemes by considering the CSI uncertainty, for which robust optimization and stochastic optimization are two typical approaches. Specifically, robust optimization approach aims at guaranteeing the worst-case but conservative performance over the uncertainty set. The robust optimization method can usually yield computationally tractable optimization models [102]. The stochastic optimization approach, e.g., chance constrained programming, only relies on the probabilistic description of the uncertainty of the problem parameter $\boldsymbol{\alpha}$ in problem (2) and is able to provide a trade-off between conservativeness and probabilistic guarantees for the achievable performance [272]. In particular, a statistical learning approach was presented in [53] to learn a tractable uncertainty set to approximate the chance constrained programming for achieving high computation efficiency and system performance in the energy-efficient edge inference systems. However, due to the limited historical samples, it is difficult to characterize the true probability distribution for the CSI uncertainty. Distributionally robust optimization [273] provides a promising way to achieve worst-case probabilistic performance by incorporating all sample-generating distributions into an ambiguity set. However, finding the globally optimal solution for this method is often computationally intractable.

DL provides an alternative way to address the uncertainty and dynamics of environment parameters to achieve modeling flexibility and computational efficiency for resource allocation in the complicated edge AI systems. Specifically, DL can provide acceptable performance for resource allocation based only on geographic locations information of the transmitters and receivers [274]. By considering CSI variations [55], [275], [276] and stochastic task arrivals [84], [85], the dynamic communication and computation resource allocation problem can be formulated as a MDP, for which deep RL, a model-free approach, can provide efficient and robust solutions [73]. Besides, the learned algorithms can be distributively executed in the multi-agent edge AI systems. However, due to the distribution shift for system parameters in episodically dynamic environment, the trained model may suffer from performance deterioration when the dataset follows a different distribution in the inference stage [22]. Transfer learning [103] and continual learning [277] have recently been adopted to address such task mismatch issue in the "learning to optimize" framework considering the system distribution dynamics.

4) End-to-End Optimization: Channel estimation plays a pivotal role to support effective resource allocation in large-scale edge AI systems [198], [206]. In particular, exploiting the low-dimensional structures of wireless channels becomes a promising way to address the curse of dimensionality for CSI acquisition in various networks. Specifically, in ultra-dense Cloud-RAN, a high-dimensional structured channel estimation framework was proposed in [278] by inducing the spatial sparsity and temporal correlation prior information using a convex regularizer. Sparsity structures of a massive MIMO channel was exploited in [279] to reduce the training overheads for CSI acquisition. The signal superposition property of a wireless multiple access channel was exploited to directly obtain the weighted sum of channels for receive beamformer design, thereby avoiding global CSI estimation [280]. The sparsity in the activity pattern was leveraged to develop the sparse signal processing framework for joint activity detection and channel estimation in grant-free massive access [75]. Due to the passive nature of RIS, it becomes infeasible to directly perform signal processing for channel estimation at RIS and the cascaded channel can only be estimated either at the edge servers or edge devices [206]. To address this unique challenge, the common reflective channels among all edge devices [281], quasi-static property between RIS and edge server channel links [282], [283], spatial features of noisy channels and additive nature of noises [284], as well as channel sparsity [285] and device activity sparsity [185], were exploited to reduce the training overhead.

However, all of the above works follow the "estimate-then-optimize" framework by first performing pilots-based channel estimation, followed by allocating resources based on the estimated CSI. However, this two-stage approach fails to achieve a low signaling overhead and superior system performance. Although the low-dimensional structures have been exploited for designing efficient channel estimation methods, the additional information (e.g., user location and mobility), are difficult to be modelled and incorporated into a unified mathematical model for CSI acquisition overhead reduction, which may exceed latency. Besides, the artificially defined criterion (e.g., mean square error) for channel estimation may not be aligned with the ultimate goal for resource allocation in edge AI systems. To address this challenge, a DL approach has recently been proposed to merge the two stages into an "end-to-end optimization" framework for resource allocation [104]. This is achieved by directly mapping the received pilots (i.e., the problem parameters $\boldsymbol{\alpha}$ in (2) can be the received pilots) into the resource allocation policy without explicit channel estimation. This mapping function is further parameterized by a DNN to capture the inherent structures of the resource allocation problems. For instance, the GNN was adopted to model the permutation invariant and equivalent properties of the mapping function for resource allocation in the RIS empowered TDD wireless networks [25]. The neural calibration approach [286] was developed in FDD massive MIMO systems to map the received pilots at edge devices into feedback bits, followed by directly mapping the feedback bits into the downlink beamformers [104].

In summary, this section presented the operation research based theory-driven and ML based data-driven methods for designing effective, real-time, distributed and robust resource allocation strategies in edge AI systems. We hope these results can stimulate more service-driven resource allocation methods (e.g., network slicing [287]) and optimization approaches (e.g., multi-objective optimization [288]). The presented "learning to optimize" framework is also promising for resource allocation in various future wireless networks.

V. Architecture for Edge AI Systems

In this section, we present a new mobile network architecture for edge AI systems, supported by the wireless network infrastructures in Section II and Section III, as well as the service-driven resource allocations in Section IV. We will provide an end-to-end (E2E) architecture design across the network infrastructure, data governance, network function, network management, as well as operations and applications.

A. End-to-End Architecture for Edge AI Systems

For each new generation of mobile networks, new services and capabilities have been introduced at the architecture level in order to meet more and typically more stringent demands. The mobile network was originally designed to deliver voice services. Since then, both the architecture and deployment of mobile networks have followed a centralized and hierarchical

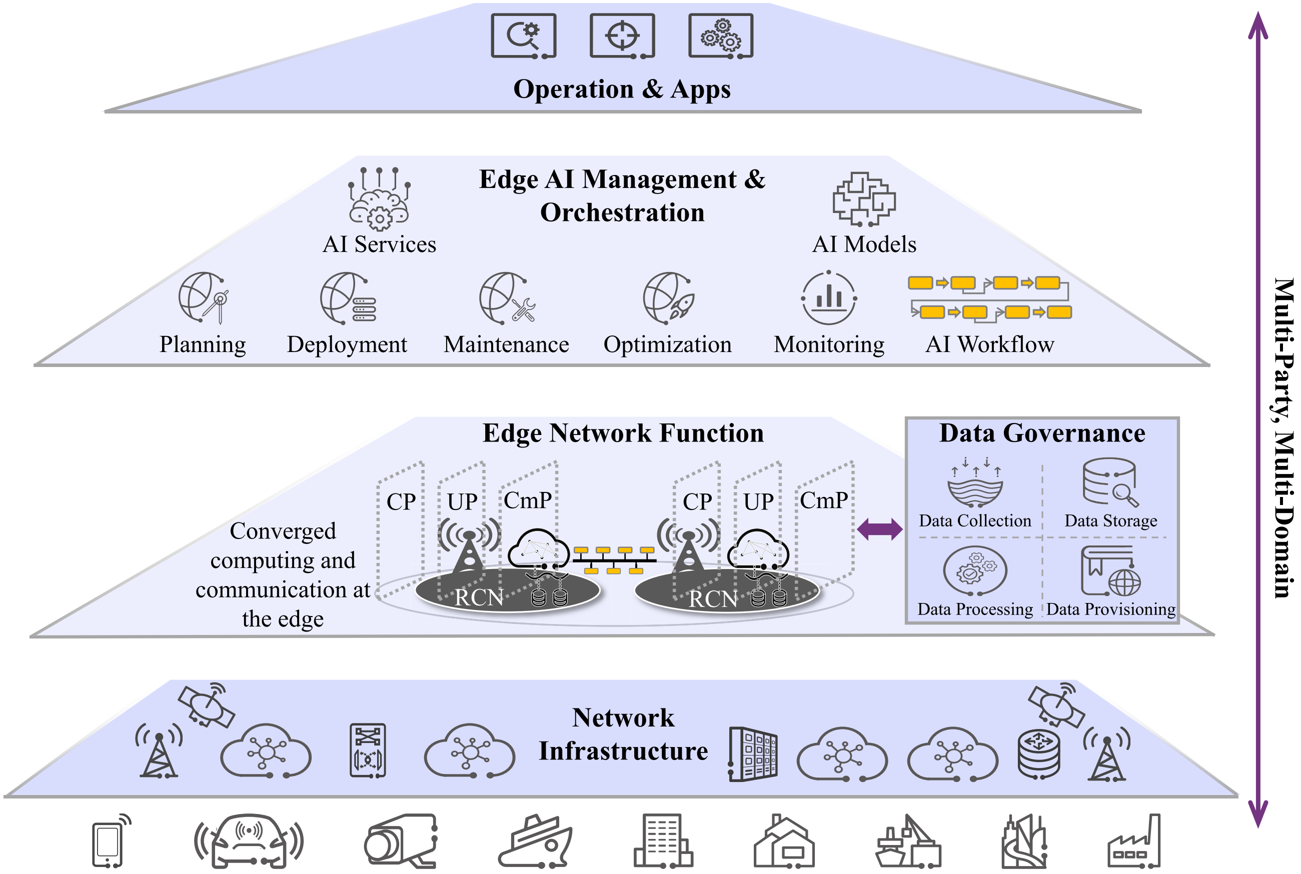


Fig. 9. E2E architecture for edge AI systems with radio computing nodes (RCN) to allow seamless integration of communication and computing capabilities. New independent computing planes (CmP) in RCN will also be used to host AI tasks and collaborate with communication functions in the control & user planes (CP and UP).

paradigm that reflects the nature of voice traffic and packet traffic of the mobile internet. To realize the vision of "connected intelligence", 6G will break and shift these traditional paradigms towards a novel architecture and design that meet new requirements for the deep integration of communication, AI, computing, and sensing at the network edge with new integrated capabilities empowered by evolutionary, as well as, revolutionary enabling technologies.

Under this new design philosophy, we introduce a holistic E2E architecture for scalable and trustworthy 6G edge AI systems, as illustrated in Fig. 9. By providing new wireless network infrastructures, enabling efficient data governance, integrating communication and computation at the network edge, as well as performing automated and scalable edge AI management and orchestration, the proposed E2E architecture will provide a scalable and flexible platform to support diversified edge AI applications with heterogeneous service requirements.

B. Data Governance

Due to the expected huge energy consumption, as well as, security and privacy concerns, we envisioned that data in future 6G networks need to be collected, processed, stored and consumed at the network edge. Since data and AI applications in 6G are expected to be much more diverse than ever before, it is incentive that there will be a provision for a unified and efficient data governance framework at the architecture level. Data governance goes far beyond conventional data collection and storage, which will also consider the data availability and quality, data sovereignty, knowledge management and legal implication. Data governance also must consider the mechanism to comply with the regional or national data protection policies and regulations of the data source in terms of usage rights and obligations such as GDPR.

1) Independent Data Plane: 5G has introduced a new network data analytics function (NWDAF) in the core network to implement AI-based network automation, optimize the related network functions (e.g., AI-based mobility management [105]), and improve user service experience, etc. One of its main goals is to collect and analyze data from other 5G network elements to train AI models and implement AI inference for automated and scalable network optimization. Meanwhile, similar mechanisms such as collecting and analyzing data based on the existing SON/MDT (self-organizing networks and minimization of drive tests), was adopted for 5G radio access networks (RAN). In 6G, such a separated data collection and analytics mechanism needs to evolve to a unified and more efficient paradigm. An independent data plane in 6G could contribute to organizing and managing data efficiently while also considering privacy protection [28]. This paves the way for natively embedding edge AI into the 6G networks by leveraging multi-domain data.

2) Multi-Player Roles: The data governance ecosystem includes different roles: data customer, data provider, data owner and data steward, etc. These could be taken by the same or different business entities, including individual users. Hence, data governance is a typical scenario that involves multiple players. It thus becomes essential to establish a multi-party data trading platform to negotiate data rights and prices among different business entities while achieving trustworthiness, fairness and efficiency. This can be achieved using decentralized technologies such as blockchain with smart contracts design [28], [106]. This will improve data efficiency and business ecosystem for the deployment of edge AI.

C. Deeply Converged Communication and Computing at the Edge

In 5G, the superior performance has been achieved by leveraging the AI capability into RAN [289], [290]. For instance, we can optimize radio resource scheduling and mitigate interference using machine learning methods [23]. Such utilization of AI in 5G can be referred to as *AI for networks*. The targets of edge AI are not only AI for networks, but also *networks for AI* [12], as presented in Section II and Section III. This will depend on the new functional capabilities of future networks, including how to make computing as a foundational capability of future 6G networks. A new type of radio equipment may emerge, which we refer to as a radio computing node (RCN), which allows the computing resources to be seamlessly converged with the communication capability. This will require the introduction of a new independent computing plane (CmP) in RCN to host AI tasks and collaborate with the communication functions in the control plane (CP) and user plane (UP) [28]. This will also enable the flexible integration of computation, communication and intelligence for edge AI.

D. Edge AI Management and Orchestration

Edge AI involves a diverse set of learning models and algorithms, network infrastructures, as well as complicated collaboration for communication, computation and intelligence. Developing a framework for edge AI management and orchestration thus becomes an essential aspect for the design of the native AI support at the architecture level. This framework needs to be designed so as to facilitate the seamless integration and deployment of AI services, especially from third-parties. This can be achieved by planning, deploying, maintaining, and optimizing the decentralized machine learning models and algorithms, as well as the edge network infrastructures and functions. The edge AI management and orchestration shall also include AI workflow, distributed and streaming data, along with heterogeneous network resources, etc. Scale and cross-domain issues will be huge challenges for such a framework and this may involve complicated standardization efforts. Hence, building such new framework which will fully rely on standardization may not be feasible. We may instead leverage the open-source approach [28] to commercialize some of the components in this framework.

In summary, this section presented the edge AI system architecture from an E2E perspective detailing its network infrastructure, data governance, edge network function, as well as edge AI management and orchestration. The standardization efforts, hardware and software platform, and application scenarios will be further discussed in the next section. We hope this novel E2E architecture can stimulate more innovative and out-of-the-box ideas for the evolution of edge AI system architectures.

VI. Standardizations, Platforms, and Applications

In this section, we will first discuss the standardization for edge learning models and algorithms, as well as integrated computing functionalities at the network edge. The research-oriented and production-oriented platforms are then provided, including distributed optimization based FL software, large-scale optimization based resource allocation solvers, as well as edge AI computing and communicating hardware. To accelerate commercialization for edge AI, the application scenarios are also investigated, including autonomous driving, industrial IoT, and smart healthcare.

A. Standardizations

The standardization of 6G will not be limited to the communications part, but also to the deep integration of communications, intelligence, and computing. The 3rd Generation Partnership Project (3GPP) may start an overall study into 6G systems around the end of 2025 (3GPP Release 20), while starting research into technical specifications around the end of 2027 [28]. In this subsection, we will introduce the standardizations on trustworthy edge learning models and algorithms, as well as wireless computing functionalities implemented in digital or analog communication systems.

1) Learning: The first technical standard for FL was approved on March 2021 as IEEE 3652.1-2020 [107], *IEEE Guide for Architectural Framework and Application of Federated Machine Learning*. This IEEE standard for FL is developed by the Learning Technology Standards Committee of the IEEE Computer Society with participants from the shared machine learning working group, including 4Paradigm, AI Singapore, Alipay, Huawei, JD iCity, Tencent, WeBank, and Xiaomi, etc. Specifically, the IEEE 3652.1-2020 standard provides the guidelines for architectures and categories of FL from the perspectives of data, user and system, followed by identifying the associated application scenarios, performance evaluations, and regulatory requirements. Standardization plays a vital role in creating a private and secure FL ecosystems at large-scale to provide consumer products and services in the market. Besides, various standards for data privacy and security have been developed by the information security, cybersecurity and privacy protection technical committee from the International Organization for Standardization (IOS) and the International Electrotechnical Commission (IEC). For example, ISO/IEC TS 27570 [291], *Privacy Protection - Privacy Guidelines for Smart Cities*, provides guidelines and recommendations for the management of privacy and the usage of standards. ISO/IEC DIS 27400 [292], *Guidelines for Security and Privacy in Internet of Things (IoT)*, provides guidance for principles and controls to provide private and secure IoT systems, services and solutions. The technical committee on cybersecurity of the European Telecommunications Standards Institute (ETSI) has recently unveiled ETSI EN 303 645 [293], *Cyber Security for Consumer Internet of Things: Baseline Requirements*, to provide cybersecurity standard and baseline for IoT consumer products and certification schemes. All these standards are applicable for developing private and secure edge AI models and algorithms to provide trustworthy products and services.

2) Computing: The computing functionality can be implemented in wireless networks by either digital modulation or analog modulation. Specifically, MEC provides a promising solution for deploying edge AI systems in current wireless systems with digital modulation [115]. The standardization activities on MEC thus pave a way to integrate edge AI into mobile networks at a maturity level. Specifically, the ETSI ISG MEC (Industry Specification Group for Multi-access Edge Computing) has established a standardized and open ecosystem for both edge-aware and edge-unaware applications at the network edge. It has published a set of white papers and specifications covering across user equipment application, service application, as well as management, mobility, and orchestration related application programming interfaces (APIs). Besides, 3GPP 5G specifications define the key enablers and architectures for edge computing to allow traffic routing, policy control, and network management for collaboration in a MEC system and a 5G system [294]. The collaboration between two independent systems of MEC and 5G can be further optimized in 6G, where communication and computing can be converged into one system by adding a computing plane [28]. In particular, ETSI ISG MEC has recently developed a synergized mobile edge cloud architecture by leveraging and harmonizing the existing and ongoing standards (including 3GPP, ETSI ISG MEC, GSMA, and 5GAA) [108]. Although 5G is rolled out globally, the modern mobile systems are widely deployed based on digital modulation instead of analog modulation [295]. To support analog communication based AirComp for edge training in current wireless networks [6], one may either directly leverage the existing digital modulator with quantized analog signals or introduce an additional analog modulator with a matched filter for decoding the received signals [37]. It is obvious that more efforts are needed to incorporate AirComp functionalities into the future 6G standards to mature edge AI systems.

B. Platforms

We present the software and hardware platforms for deploying edge AI models and algorithms, as well as the optimization solvers for resource allocation in edge AI systems.

1) Software: There is a rapidly growing body of software platforms for simulations and productization of edge AI algorithms and models. FL library, TensorFlow Federated, Leaf, and PySyft have provided excellent open software frameworks for FL simulations and evaluations. To further accelerate research progress and facilitate algorithmic innovation and performance comparison in realistic FL environments, FedML [109], a research-oriented open FL library, has recently been established to support diverse FL computing environments and topological architectures with standardized FL algorithm implementations and benchmarks. As a production-oriented software project, FATE [110] has been developed in the Webank's AI Department for financial industry by supporting various secure computing protocols and FL architectures. Besides, existing edge computing frameworks (e.g., Baidu "Baetyl" and Huawei "KubeEdge") provide promising solutions to deliver edge AI services. For edge AI empowered IoT applications, Microsoft "Azure IoT Edge", Google "Cloud IoT", Amazon "Web Services (AWS) IoT" and NVIDIA "EGX" provide edge AI platform to bring real-time AI services across a wide range of applications, including smart retail, home, manufacturing, and healthcare. Huawei has recently released a next-generation operating system, HarmonyOS [111], to enable seamless collaboration and interconnection among smart edge devices across diverse platforms. This empowers connected intelligence by deploying edge AI in the operating systems.

2) Solver: Resource allocations for edge AI systems and wireless networks are booming through the development of various large-scale optimization models and algorithms. General-purpose large-scale optimization software solvers are important to enable rapid prototyping and deploying resource allocation optimization algorithms for edge AI systems. Specifically, CVX [112] provides a two-stage software framework for modeling and solving general large-scale convex optimization problems. This is achieved by automatically transforming the original problem instances into standard conic programming forms, followed by calling the advanced off-the-shelf conic solvers, e.g., MOSEK [296] and SCS [113]. To further speed up the modeling phase and avoid repeatedly parsing and re-generating conic forms, a matrix stuffing technique was presented in [98] to generate the mapping function between the original problem and the conic form in a symbolic way instead of the time-consuming numerical way using CVX. It is thus particularly interesting to develop a solver to automatically generate the mapping functions for conic transformation in a symbolic forms. Besides, Gurobi [297] and MOSEK [296] are among the fastest solvers for solving the general mixed-integer second-order conic programs. Chen *et al.* recently released the software package "Open-L2O" [114] to implement the "learning to optimize" framework for benchmarking performance fairly and designing algorithm automatically.

3) Hardware: The achievable performance and benefits of edge AI systems are conditioned upon the availability of edge AI computing hardware and radio frequency (RF) hardware technologies. Specifically, edge AI computing hardware can be categorized as graphic processing unit (GPU)-based hardware (e.g., NVIDIA's GPUs), field programmable gate array (FPGA)-based hardware (e.g., Xilinx's SDSoC), and application specific integrated circuit (ASIC)-based hardware (e.g., Google's TPU). The detailed comparisons for various edge AI computing hardware can be found in [115]. In particular, the chip design procedure for edge AI hardware can be significantly accelerated by the recent proposal of deep RL assisted fast chip floorplanning [298]. Besides, the massive broadband connectivity requirements for edge AI systems motivate the innovations in RF hardware technologies. The benefits of RIS-empowered FL systems highly depend on the capabilities of manipulating electromagnetic waves at the metasurfaces [49], whose reconfigurability is typically enabled by switches, tunable material, topological metasurfaces, and hybrid metasurfaces [299]. THz communication with frequency band 0.1-10 THz, is envisioned as a promising enabler for achieving sensing, communication, and learning in an integrated edge AI system. To approach this THz region, RF hardware technologies and solutions were thoroughly investigated in [116], including semiconductor circuits, antenna forms, packaging and testing of transceivers.

C. Applications

We discuss edge AI enabled application scenarios by inspiring new communication algorithms, resource allocation optimization algorithms, as well as data processing methods.

1) Autonomous Driving: Autonomous driving basically refers to self-driving vehicles that move without the intervention of human drivers. Self-driving vehicle integrates various innovative technologies, including advanced sensor technologies, new energy automobiles, next generation AI technologies, as well as future vehicular networks. Autonomous driving can significantly improve the safety, passenger comfort, travel and logistics efficiency, collision avoidance, and energy efficiency. Edge AI shall provide a pivotal role for achieving ultra-low latency communication, intelligent networking, real-time data analytics, as well as high security for intelligent vehicles [27], [117]. A general DL framework was proposed in [26] to enable ultra-reliable and low-latency vehicular communication, by incorporating the domain knowledge including information theoretical tools and cross-layer optimization design. To minimize the vehicles' queuing latency, a FL approach was developed in [300] to learn the tail distribution of the queue lengths. To cope with the high mobility and heterogeneous structures in vehicular networks, DL becomes powerful for dynamic resource allocation [24] and network traffic control [27]. In particular, edge AI techniques, including distributed RL [55], [301], decentralized GNN [23], as well as distributed DNN with binarized output layer [61], are able to learn and execute the distributed resource allocation polices in an automatic and real-time manner.

The data processing tasks for autonomous driving mainly include perception, high-definition (HD) mapping, as well as SLAM [117], [302]. Specifically, to understand the environments for intelligent decision making, various sensory data from onboard sensors (e.g., light detection and ranging (LiDAR), cameras, radar and sonar) need to be processed for the perception tasks, including localization, object detection and tracking. The perception capability can be enhanced by edge AI systems, e.g., edge device-server co-inference of DNN models for vision based perception tasks [303]. HD mapping aims at constructing a representation of the vehicles operating environments, e.g., obstacles, landmarks position, curvature and slope. This is imperative to achieve high accurate localization for autonomous driving. The edge server cooperative inference method in Section III-A.2 can be adopted to reduce the storage and communication overheads for updating the HD map by collecting fresh data from the vehicles in the dynamic environments [117]. SLAM comprises simultaneously estimating the state of a vehicle and constructing a map of the environment [304], which paves the way for achieving full autonomy in autonomous driving [305]. Edge SLAM [56], [57] has recently been developed to execute DL based visual SLAM algorithms on edge vehicles. This is achieved by deploying the tracking computation parts on the edge vehicles while offloading the remaining parts (e.g., local mapping and loop closure) to the roadside edge server via vertical edge inference in Section III-B.

2) Internet of Things: Artificial Intelligence of Things (AIoT) leverages AI technologies and IoT infrastructures to improve the human-machine interactions and enable multi-agent communications and collaborations. AIoT goes beyond the conventional communication paradigm for audio, video and data delivery. It will enable semantic communication [58] to exchange semantic information among agents. Shannon and Weaver categorize communication into three levels, including transmission level (i.e., transmit symbols accurately), semantic level (i.e., convey the desired meanings precisely), and effectiveness level (i.e., produce the desired actions effectively) [306]. Sematic communication is able to significantly improve the communication efficiency by only transmitting the extracted relevant information for sematic information delivery tasks with the semation error as the performance metric. A distributed edge DL approach has been recently developed in [58] to enable low-latency semantic communication over IoT networks. This is achieved by jointly optimizing the compressed DNN based transmitters at the edge IoT devices and the quantized DNN based receivers at the edge server over the wireless fading channels.

Industrial IoT (IIoT) is a production-oriented industrial network for connecting industrial devices and equipments, processing and exchanging generated data, as well as optimizing the production system [307]. Besides, digital twin is becoming a key technology for smart manufacturing in industrial 4.0 by connecting physical machines and digital representations in a cyber-physical system [308], [309]. This is achieved by providing a virtual representation of the industrial entities and products' life-cycle to predict and optimize the behaviors of the manufacturing process. Edge AI provides a promising way to model and deploy digital twins for IIoT networks to process the high volume of industrial streaming data with low-latency and high-security guarantees. Specifically, edge computing provides a general platform for inferring DNN models via computation offloading to reduce network latency and operation cost in IIoT [119]. FL becomes a key enabling technology to support intelligent IIoT applications (e.g., smart grid and smart manufacturing) and provide IIoT services (e.g., data offloading and mobile crowdsensing) [120]. In particular, blockchain empowered FL was proposed in [310] to provide secure communication and private data sharing schemes for constructing digital twin IIoT networks, followed by reducing communication overheads via asynchronous model aggregation.

3) Smart Healthcare: Smart healthcare aims to realize a common platform for efficient and personalized healthcare, intelligent health monitoring, and precision medicine development via collaboration among multiple participants (e.g., doctors, patients, hospitals, and research institutions). This is achieved by emerging advanced technologies, including DL [311], [312], Tactile Internet, IoT, edge AI, and wireless communications. In particular, edge AI with distributed and secure DL has been demonstrated to be able to significantly improve the reliability, accuracy, scalability, privacy and security for precision medicine and Internet of Medical Things [313], including medical imaging, drug development, and chronic disease management [314]. Specifically, Kaissis *et al.* [315] presented a FL approach for medical imaging to preserve privacy and avoid potential attacks against the datasets or learning algorithms. Besides, swarm learning has recently been developed in [35] to provide a decentralized and confidential clinical disease detection solution for diseases (e.g., COVID-19, tuberculosis, and leukaemia). This is achieved by leveraging the blockchain and edge computing techniques to develop a secure and private decentralized learning architecture while keeping the medical data locally. MIT Media Lab established a split learning project to allow health entities collaboration for training patient diagnostic models without sharing sensitive raw data [316]. An RL approach for decisions making in patient treatment was introduced in [317] to realize safe and risk-conscious healthcare practice.

Haptic communication [121] aims at delivering the skill set (e.g., the manipulation skills representation learned from the multisensory tactile and visual data [318], and the signatures of the human grasp learned using a tactile glove [319]) over the Tactile Internet in an ultra-reliable and low-latency manner. It has potentials in healthcare applications including telediagnosis, tele-rehabilitation, and tele-surgery, which turns out to be essential during the ongoing COVID-19 pandemic. Edge AI becomes a key enabling technique for the Tactile Internet with human-in-the loop to facilitate ultra-responsive and truly immersive tactile actuation in the tele-operation systems [320]. This is achieved by enabling the network edge with intelligent prediction capability for haptic information (e.g., tactile feedback and control traffics) [321], as well as the intelligent resource allocations across the whole network layers [322]. Specifically, a distributed optimization framework was developed in [323] to design an edge computing assisted Tactile Internet for achieving both the ultra-low latency and high energy efficiency. Such a distributed optimization algorithm can be further learned via the distributed DL techniques [61]. Besides, a variational optimization framework was proposed in [324] to enjoy low-latency and high-reliability for massive access in the Tactile Internet. The variational decision function can be further parameterized via DNNs with the capability of distributed training and inference for practical deployments [23], [324].

In summary, this section presented standardizations, platforms, and applications for practical deployment of edge AI systems. Combining the presentations of edge training in Section II, edge inference in Section III, resource allocation in Section IV, and system architecture in Section V, we complete the roadmap for edge AI ecosystem, as shown in Fig. 2. We hope these results can encourage more communities and stakeholders to engage in industrializing and commercializing edge AI in the era of 6G.

VII. Conclusion

Embedding low-power, low-latency, reliable, and trustworthy intelligence into the network edge is an inevitable trend and disruptive shift in both academia and industry. Edge AI serves as a distributed neural network to imbue connected intelligence in 6G, thereby enabling intelligent and seamless interactions among the human world, physical world, and digital world. The challenges for building edge AI ecosystems are multidisciplinary spanning wireless communications, machine learning, operation research, domain applications, regulations and ethics. In this paper, we have investigated the key wireless communication techniques, effective resource management approaches and holistic network architectures to design scalable and trustworthy edge AI systems. The standardizations, platforms, and applications were also discussed for productization and commercialization of edge AI. We hope that this article will serve as a valuable reference and guideline for further considering edge AI opportunities across theoretical, algorithmic, systematic, and entrepreneurial considerations to embrace the exciting era of edge AI.

References

[1] *Resilient and Intelligent NextG Systems (RINGS)*. [Online]. Available: https://www.nsf.gov/pubs/2021/nsf21581/nsf21581.pdf

[2] *Expanded 6G Vision, Use Cases and Societal Values*. [Online]. Available: https://hexa-x.eu/wp-content/uploads/2021/05/Hexa-X_D1.2.pdf

[3] *IMT-2030 (6G) Promotion Group, 6G Vision and Candidate Technologies*. [Online]. Available: http://www.caict.ac.cn/english/news/202106/P020210608349616163475.pdf

[4] *Network 2030: A Blueprint of Technology, Applications and Market Drivers Towards the Year 2030 and Beyond*, document FG-NET-2030, ITU Focus Group on Technologies for Network, May 2019.

[5] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[6] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[7] *Huawei 5.5G*. [Online]. Available: https://www.huawei.com/en/news/2020/11/mbbf-shanghai-huawei-david-wang-5dot5g

[8] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1–34, Jul. 2021.

[9] M. Maier, A. Ebrahimzadeh, S. Rostami, and A. Beniiche, "The internet of no things: Making the internet disappear and 'see the invisible,'" *IEEE Commun. Mag.*, vol. 58, no. 11, pp. 76–82, Nov. 2020.

[10] X.-H. You *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–74, Jan. 2021.

[11] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[12] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J.-A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[13] S. Ali, W. Saad, D. Steinbach, I. Ahmad, and J. Huusko, "White paper on machine learning in wireless communication networks," *6G Res. Vis.*, no. 7, pp. 1–36, Jun. 2020.

[14] J. Wang *et al.*, "Interplay between RIS and AI in wireless communications: Fundamentals, architectures, applications, and open research problems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2271–2288, Aug. 2021.

[15] H. Kim, Y. Jiang, R. B. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2018, pp. 1–19.

[16] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[17] J. Hoydis, F. A. Aoudia, A. Valcarce, and H. Viswanathan, "Toward a 6G AI-native air interface," *IEEE Commun. Mag.*, vol. 59, no. 5, pp. 76–81, May 2021.

[18] Y. M. Saidutta, A. Abdi, and F. Fekri, "Joint source-channel coding over additive noise analog channels using mixture of variational autoencoders," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2000–2013, Jul. 2021.

[19] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, May 2021, Art. no. 107930.

[20] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," 2021, *arXiv:2102.04170*.

[21] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[22] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "LORM: Learning to optimize for resource management in wireless networks with few training samples," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 665–679, Jan. 2020.

[23] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "Graph neural networks for scalable radio resource management: Architecture design and theoretical analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 101–115, Jan. 2021.

[24] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, Feb. 2020.

[25] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1931–1945, Jul. 2021.

[26] C. She *et al.*, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proc. IEEE*, vol. 109, no. 3, pp. 204–246, Mar. 2021.

[27] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6G: Machine-learning approaches," *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.

[28] W. Tong and P. Zhu, *6G: The Next Horizon: From Connected People and Things to Connected Intelligence*. Cambridge, U.K.: Cambridge Univ. Press, 2021.

[29] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2167–2191, 4th Quart., 2020.

[30] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[31] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.

[32] J. Shao and J. Zhang, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 20–26, Dec. 2020.

[33] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, Feb. 2019.

[34] E. B. P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1, pp. 1–210, 2021.

[35] S. Warnat-Herresthal *et al.*, "Swarm learning for decentralized and confidential clinical machine learning," *Nature*, vol. 594, pp. 265–270, Jun. 2021.

[36] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *J. Netw. Comput. Appl.*, vol. 116, pp. 1–8, Aug. 2018.

[37] J. Park *et al.*, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proc. IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[38] T. Chen, K. Zhang, G. B. Giannakis, and T. Basar, "Communication-efficient policy gradient methods for distributed reinforcement learning," *IEEE Trans. Control Netw. Syst.*, early access, May 6, 2021, doi: 10.1109/TCNS.2021.3078100.

[39] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 5872–5881.

[40] Y. Dong, J. Cheng, M. J. Hossain, and V. C. M. Leung, "Secure distributed on-device learning networks with byzantine adversaries," *IEEE Netw.*, vol. 33, no. 6, pp. 180–187, Nov. 2019.

[41] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2021.

[42] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[43] Q. Yu, S. Li, N. Raviv, S. M. M. Kalan, M. Soltanolkotabi, and S. A. Avestimehr, "Lagrange coded computing: Optimal design for resiliency, security, and privacy," *Proc. Mach. Learn. Res.*, vol. 89, pp. 1215–1225, Apr. 2019.

[44] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[45] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[46] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

[47] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[48] Z. Wang *et al.*, "Federated learning via intelligent reflecting surface," *IEEE Trans. Wireless Commun.*, early access, Jul. 30, 2021, doi: 10.1109/TWC.2021.3099505.

[49] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[50] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 41–47, Dec. 2020.

[51] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart., 2018.

[52] K. Yang, Y. Shi, and Z. Ding, "Data shuffling in wireless distributed computing via low-rank optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3087–3099, Jun. 2019.

[53] K. Yang, Y. Shi, W. Yu, and Z. Ding, "Energy-efficient processing and robust wireless cooperative transmission for edge inference," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9456–9470, Oct. 2020.

[54] S. Hua, Y. Zhou, K. Yang, Y. Shi, and K. Wang, "Reconfigurable intelligent surface for green edge inference," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 2, pp. 964–979, Jun. 2021.

[55] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[56] A. J. B. Ali, Z. S. Hashemifar, and K. Dantu, "Edge-SLAM: Edge-assisted visual simultaneous localization and mapping," in *Proc. Annu. Int. Conf. Mobile Syst. Appl. Service (MobiSys)*, Jun. 2020, pp. 325–337.

[57] J. Xu *et al.*, "Edge assisted mobile semantic visual SLAM," in *Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM)*, Jul. 2020, pp. 1828–1837.

[58] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, Jan. 2021.

[59] J. Liu, Z. Shi, S. Zhang, and N. Kato, "Distributed Q-learning aided uplink grant-free NOMA for massive machine-type communications," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2029–2041, Jul. 2021.

[60] Y. Shen, J. Zhang, S. Song, and K. B. Letaief, "AI empowered resource management for future wireless networks," 2021, *arXiv:2106.06178*.

[61] H. Lee, S. H. Lee, and T. Q. S. Quek, "Deep learning for distributed optimization: Applications to wireless resource management," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2251–2266, Oct. 2019.

[62] H. Jang, O. Simeone, B. Gardner, and A. Gruning, "An introduction to probabilistic spiking neural networks: Probabilistic models, learning rules, and applications," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 64–77, Nov. 2019.

[63] N. Skatchkovsky, H. Jang, and O. Simeone, "Spiking neural networks—Part III: Neuromorphic communications," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 1746–1750, Jun. 2021.

[64] R. Li, Z. Zhao, X. Xu, F. Ni, and H. Zhang, "The collective advantage for advancing communications and intelligence," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 96–102, Aug. 2020.

[65] Q. Yu *et al.*, "An immunology-inspired network security architecture," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 168–173, Oct. 2020.

[66] Q. Yu *et al.*, "A fully-decoupled RAN architecture for 6G inspired by neurotransmission," *J. Commun. Inf. Netw.*, vol. 4, no. 4, pp. 15–23, Dec. 2019.

[67] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 54, 2017, pp. 1273–1282.

[68] J. Wang *et al.*, "A field guide to federated optimization," 2021, *arXiv:2107.06917*.

[69] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.

[70] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks: An examination of distributed strategies and network behavior," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.

[71] Y. Lu and C. De Sa, "Optimal complexity in decentralized training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 7111–7123.

[72] M. Li, D. G. Andersen, A. J. Smola, and K. Yu, "Communication efficient distributed machine learning with the parameter server," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 19–27.

[73] D. Lee, N. He, P. Kamalaruban, and V. Cevher, "Optimization for reinforcement learning: From a single agent to cooperative agents," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 123–135, May 2020.

[74] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, pp. 1–25, Dec. 2017.

[75] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[76] T. Jiang, Y. Shi, J. Zhang, and K. B. Letaief, "Joint activity detection and channel estimation for IoT networks: Phase transition and computation-estimation tradeoff," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6212–6225, Aug. 2019.

[77] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[78] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[79] J. Dong, Y. Shi, and Z. Ding, "Blind over-the-air computation and data fusion via provable Wirtinger flow," *IEEE Trans. Signal Process.*, vol. 68, pp. 1136–1151, 2020.

[80] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[81] M. Fu, Y. Zhou, Y. Shi, W. Chen, and R. Zhang, "UAV aided over-the-air computation," 2021, *arXiv:2106.00254*.

[82] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.

[83] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[84] C.-F. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4132–4150, Jun. 2019.

[85] D. Han, W. Chen, and Y. Fang, "Joint channel and queue aware scheduling for latency sensitive mobile edge computing with power constraints," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3938–3951, Jun. 2020.

[86] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multi-device cooperative edge inference," 2021, *arXiv:2109.00172*.

[87] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, "Fast convergence algorithm for analog federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[88] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[89] L. Li *et al.*, "Delay analysis of wireless federated learning based on saddle point approximation and large deviation theory," 2021, *arXiv:2103.16994*.

[90] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, Apr. 2021.

[91] Z. Wang, Y. Shi, Y. Zhou, H. Zhou, and N. Zhang, "Wireless-powered over-the-air computation in intelligent reflecting surface-aided IoT networks," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1585–1598, Feb. 2021.

[92] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.

[93] S. Huang, Y. Zhou, T. Wang, and Y. Shi, "Byzantine-resilient federated machine learning via over-the-air computation," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.

[94] Z. Zhang, G. Zhu, R. Wang, V. K. N. Lau, and K. Huang, "Turning channel noise into an accelerator for over-the-air principal component analysis," 2021, *arXiv:2104.10095*.

[95] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021.

[96] Y. Shi, J. Zhang, W. Chen, and K. B. Letaief, "Generalized sparse and low-rank optimization for ultra-dense networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 42–48, Jun. 2018.

[97] Y. Shi, H. Choi, Y. Shi, and Y. Zhou, "Algorithm unrolling for massive access via deep neural network with theoretical guarantee," *IEEE Trans. Wireless Commun.*, early access, Aug. 6, 2021, doi: 10.1109/TWC.2021.3100500.

[98] Y. Shi, J. Zhang, B. O'Donoghue, and K. B. Letaief, "Large-scale convex optimization for dense wireless cooperative networks," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4729–4743, Sep. 2015.

[99] J. Y. Gotoh, A. Takeda, and K. Tono, "DC formulations and algorithms for sparse optimization problems," *Math. Program.*, vol. 169, no. 1, pp. 141–176, 2018.

[100] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a MATLAB toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1455–1459, 2014.

[101] Y. Shi, J. Zhang, and K. B. Letaief, "Low-rank matrix completion for topological interference management by Riemannian pursuit," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4703–4717, Jul. 2016.

[102] Y. Shi, J. Zhang, and K. Letaief, "Robust group sparse beamforming for multicast green cloud-RAN with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4647–4659, Sep. 2015.

[103] Y. Yuan, G. Zheng, K.-K. Wong, B. Ottersten, and Z.-Q. Luo, "Transfer learning and meta learning-based fast downlink beamforming adaptation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1742–1755, Mar. 2021.

[104] F. Sohrabi, K. M. Attiah, and W. Yu, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4044–4057, Jul. 2021.

[105] C. Shen, C. Tekin, and M. van der Schaar, "A non-stochastic learning approach to energy efficient mobility management," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3854–3868, Dec. 2016.

[106] W. Dai, C. Dai, K.-K.-R. Choo, C. Cui, D. Zou, and H. Jin, "SDTE: A secure blockchain-based data trading ecosystem," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 725–737, 2020.

[107] *IEEE Guide for Architectural Framework and Application of Federated Machine Learning*, IEEE Standard 3652.1-2020, 2021, pp. 1–69.

[108] *Harmonizing Standards for Edge Computing*, ETSI ISG MEC, Sophia Antipolis, France, Jul. 2020.

[109] C. He *et al.*, "FedML: A research library and benchmark for federated machine learning," 2020, *arXiv:2007.13518*.

[110] *Fate*. [Online]. Available: https://fate.fedai.org/

[111] *HarmonyOS*. [Online]. Available: https://www.harmonyos.com/en/

[112] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

[113] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd, "Conic optimization via operator splitting and homogeneous self-dual embedding," *J. Optim. Theory Appl.*, vol. 169, no. 3, pp. 1042–1068, Jun. 2016.

[114] T. Chen *et al.*, "Learning to optimize: A primer and a benchmark," 2021, *arXiv:2103.12828*.

[115] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869–904, 2nd Quart., 2020.

[116] S. Amakawa *et al.*, "White paper on RF enabling 6G—Opportunities and challenges from technology to spectrum," *6G Res. Vis.*, no. 13, pp. 1–68, Apr. 2021.

[117] J. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the internet of vehicles," *Proc. IEEE*, vol. 108, no. 2, pp. 246–261, Feb. 2020.

[118] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Netw.*, vol. 34, no. 5, pp. 16–22, Sep. 2020.

[119] T. Qiu, J. Chi, X. Zhou, Z. Ning, M. Atiquzzaman, and D. O. Wu, "Edge computing in industrial Internet of Things: Architecture, advances and challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2462–2488, 4th Quart., 2020.

[120] D. C. Nguyen *et al.*, "Federated learning for industrial Internet of Things in future industries," *IEEE Wireless Commun.*, early access, Aug. 6, 2021, doi: 10.1109/MWC.001.2100102.

[121] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Toward haptic communications over the 5G tactile internet," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3034–3059, 4th Quart., 2018.

[122] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[123] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2017, pp. 1707–1718.

[124] W. Wen *et al.*, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Dec. 2017, pp. 1508–1518.

[125] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 80, Jul. 2018, pp. 560–569.

[126] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE Trans. Signal Process.*, vol. 68, pp. 2128–2142, 2020.

[127] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, 2021.

[128] H. Wang, S. Sievert, S. Liu, Z. B. Charles, D. S. Papailiopoulos, and S. J. Wright, "ATOMO: Communication-efficient learning via atomic sparsification," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 9872–9883.

[129] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Sep. 2017, pp. 440–445.

[130] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Hierarchical quantized federated learning: Convergence analysis and system design," 2021, *arXiv:2103.14272*.

[131] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, Mar. 2021, pp. 2350–2358.

[132] J. Sun, T. Chen, G. B. Giannakis, Q. Yang, and Z. Yang, "Lazily aggregated quantized gradient innovation for communication-efficient federated learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 23, 2020, doi: 10.1109/TPAMI.2020.3033286.

[133] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–26.

[134] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst. (MLSys)*, vol. 2, 2020, pp. 429–450.

[135] D. A. E. Acar, Y. Zhao, R. Matas, M. Mattina, P. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–36.

[136] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 21394–21405.

[137] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 15111–15122.

[138] R. Chen and I. C. Paschalidis, "Distributionally robust learning," *Found. Trends Optim.*, vol. 4, nos. 1–2, pp. 1–243, 2020.

[139] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4427–4437.

[140] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 3557–3568.

[141] R. Pathak and M. J. Wainwright, "FedSplit: An algorithmic framework for fast federated optimization," 2020, *arXiv:2005.05238*.

[142] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 1–13.

[143] Y. Ruan, X. Zhang, S.-C. Liang, and C. Joe-Wong, "Towards flexible device participation in federated learning," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 130, Apr. 2021, pp. 3403–3411.

[144] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5330–5340.

[145] S. Savazzi, M. Nicoli, M. Bennis, S. Kianoush, and L. Barbieri, "Opportunities of federated learning in connected, cooperative, and automated industrial systems," *IEEE Commun. Mag.*, vol. 59, no. 2, pp. 16–21, Feb. 2021.

[146] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 102–113, May 2020.

[147] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," 2018, *arXiv:1808.07576*.

[148] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Nov. 2020, pp. 5381–5393.

[149] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *Proc. Int. Conf. Mach. Learn. (ICML)*, May 2019, pp. 3478–3487.

[150] L. Kong, T. Lin, A. Koloskova, M. Jaggi, and S. U. Stich, "Consensus control for decentralized deep learning," 2021, *arXiv:2102.04828*.

[151] G. Neglia, C. Xu, D. Towsley, and G. Calbi, "Decentralized gradient methods: Does topology matter?" in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 108, Aug. 2020, pp. 2348–2358.

[152] A. Elgabli, J. Park, A. S. Bedi, M. Bennis, and V. Aggarwal, "GADMM: Fast and communication efficient framework for distributed machine learning," *J. Mach. Learn. Res.*, vol. 21, no. 76, pp. 1–39, 2020.

[153] T. Lin, S. P. Karimireddy, S. U. Stich, and M. Jaggi, "Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data," 2021, *arXiv:2102.04761*.

[154] S. J. Wright, "Coordinate descent algorithms," *Math. Program.*, vol. 151, no. 1, pp. 3–34, 2015.

[155] A. Choromanska *et al.*, "Beyond backprop: Online alternating minimization with auxiliary variables," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 97, Jun. 2019, pp. 1193–1202.

[156] B. Gu, Z. Dang, X. Li, and H. Huang, "Federated doubly stochastic kernel learning for vertically partitioned data," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*, Aug. 2020, pp. 2483–2493.

[157] Y. Hu, D. Niu, J. Yang, and S. Zhou, "FDML: A collaborative machine learning framework for distributed features," in *Proc. Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*, Jul. 2019, pp. 2232–2240.

[158] B. Ying, K. Yuan, and A. H. Sayed, "Supervised learning under distributed features," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 977–992, Feb. 2019.

[159] L. Lyu, J. C. Bezdek, J. Jin, and Y. Yang, "FORESEEN: Towards differentially private deep inference for intelligent Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 10, pp. 2418–2429, Oct. 2020.

[160] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[161] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1928–1937.

[162] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multiagent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6382–6393.

[163] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," 2019, *arXiv:1911.10635*.

[164] S. Zeng, A. Anwar, T. Doan, A. Raychowdhury, and J. Romberg, "A decentralized policy gradient approach to multi-task reinforcement learning," 2020, *arXiv:2006.04338*.

[165] G. Damaskinos, E. M. El Mhamdi, R. Guerraoui, A. H. A. Guirguis, and S. L. A. Rouault, "AGGREGATHOR: Byzantine machine learning via robust gradient aggregation," in *Proc. Conf. Syst. Mach. Learn. (SysML)*, 2019, pp. 1–19.

[166] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5194–5208, Aug. 2021.

[167] W.-N. Chen, P. Kairouz, and A. Ozgur, "Breaking the communication-privacy-accuracy trilemma," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3312–3324.

[168] Z. Wu, Q. Ling, T. Chen, and G. B. Giannakis, "Federated variance-reduced stochastic gradient descent with robustness to Byzantine attacks," *IEEE Trans. Signal Process.*, vol. 68, pp. 4583–4596, 2020.

[169] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 5650–5659.

[170] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 118–128.

[171] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2168–2181, Jul. 2021.

[172] H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Blockchained on-device federated learning," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1279–1283, Jun. 2020.

[173] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

[174] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, 2020.

[175] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, Mar. 2021.

[176] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," 2021, *arXiv:2101.02198*.

[177] M. M. Amiria, T. M. Dumanb, D. Gündüzc, S. R. Kulkarni, and H. V. Poor, "Collaborative machine learning at the wireless edge with blind transmitters," *IEEE Trans. Wireless Commun.*, 2021.

[178] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach," *IEEE Trans. Wireless Commun.*, early access, Jun. 10, 2021, doi: 10.1109/TWC.2021.3086116.

[179] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," 2021, *arXiv:2104.03490*.

[180] C. Fang, H. Dong, and T. Zhang, "Mathematical models of overparameterized neural networks," *Proc. IEEE*, vol. 109, no. 5, pp. 683–703, May 2021.

[181] S. Savazzi, M. Nicoli, and V. Rampa, "Federated learning with cooperating devices: A consensus approach for massive IoT networks," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4641–4654, May 2020.

[182] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[183] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148–156, Aug. 2020.

[184] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.

[185] S. Xia, Y. Shi, Y. Zhou, and X. Yuan, "Reconfigurable intelligent surface for massive connectivity," 2021, *arXiv:2101.10322*.

[186] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[187] Y. Shi, J. Dong, and J. Zhang, *Low-Overhead Communications in IoT Networks*. Singapore: Springer, 2020.

[188] J. Dong, K. Yang, and Y. Shi, "Blind demixing for low-latency communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 897–911, Feb. 2019.

[189] J. Dong and Y. Shi, "Nonconvex demixing from bilinear measurements," *IEEE Trans. Signal Process.*, vol. 66, no. 19, pp. 5152–5166, Oct. 2018.

[190] X. Bian, Y. Mao, and J. Zhang, "Supporting more active users for massive access via data-assisted activity detection," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

[191] Y. M. X. Bian and J. Zhang, "Joint activity detection and data decoding in massive random access via a turbo receiver," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 1–5.

[192] H. Cheng, Y. Xia, Y. Huang, Z. Lu, and L. Yang, "Deep neural network aided low-complexity MPA receivers for uplink SCMA systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9050–9062, Sep. 2021.

[193] N. Ye, X. Li, H. Yu, L. Zhao, W. Liu, and X. Hou, "DeepNOMA: A unified framework for NOMA using deep multi-task learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2208–2225, Apr. 2020.

[194] C. Huang, G. Chen, Y. Gong, P. Xu, Z. Han, and J. A. Chambers, "Buffer-aided relay selection for cooperative hybrid NOMA/OMA networks with asynchronous deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2514–2525, Aug. 2021.

[195] Y. Lu, P. Cheng, Z. Chen, W. H. Mow, Y. Li, and B. Vucetic, "Deep multi-task learning for cooperative NOMA: System design and principles," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 61–78, Jan. 2021.

[196] X. Zhai, X. Chen, J. Xu, and D. W. K. Ng, "Hybrid beamforming for massive MIMO over-the-air computation," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2737–2751, Apr. 2021.

[197] G. Zhu and K. Huang, "MIMO over-the-air computation for high-mobility multimodal sensing," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6089–6103, Aug. 2019.

[198] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 84–91, Jun. 2015.

[199] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, 3rd Quart., 2016.

[200] L. Xing, Y. Zhou, and Y. Shi, "Over-the-air computation via cloud radio access networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.

[201] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[202] G. Zhu, J. Xu, K. Huang, and S. Cui, "Over-the-air computing for wireless data aggregation in massive IoT," 2020, *arXiv:2009.02181*.

[203] M. Fu, Y. Zhou, Y. Shi, and K. B. Letaief, "Reconfigurable intelligent surface empowered downlink non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3802–3817, Jun. 2021.

[204] M. M. Amiri and D. Gündüz, "Computation scheduling for distributed machine learning with straggling workers," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6270–6284, Dec. 2019.

[205] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2666–2682, Nov. 2020.

[206] X. Yuan, Y.-J. A. Zhang, Y. Shi, W. Yan, and H. Liu, "Reconfigurable-intelligent-surface empowered wireless communications: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 136–143, Apr. 2021.

[207] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[208] C. Huang *et al.*, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[209] W. Fang, Y. Jiang, Y. Shi, Y. Zhou, W. Chen, and K. B. Letaief, "Over-the-air computation via reconfigurable intelligent surface," 2021, *arXiv:2105.05113*.

[210] X. Zhu, C. Jiang, L. Yin, L. Kuang, N. Ge, and J. Lu, "Cooperative multigroup multicast transmission in integrated terrestrial-satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 981–992, May 2018.

[211] N. Saeed, A. Elzanaty, H. Almorad, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "CubeSat communications: Recent advances and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1839–1862, 3rd Quart., 2020.

[212] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.

[213] H. Zhou, W. Xu, J. Chen, and W. Wang, "Evolutionary V2X technologies toward the internet of vehicles: Challenges and opportunities," *Proc. IEEE*, vol. 108, no. 2, pp. 308–323, Feb. 2020.

[214] F. Liu, W. Yuan, C. Masouros, and J. Yuan, "Radar-assisted predictive beamforming for vehicular links: Communication served by sensing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7704–7719, Nov. 2020.

[215] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," 2019, *arXiv:1902.01046*.

[216] N. Cheng *et al.*, "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, May 2019.

[217] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[218] T. S. Rappaport *et al.*, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[219] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[220] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 126–136, Jan. 2018.

[221] S. Li, Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "A scalable framework for wireless distributed computing," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2643–2654, Oct. 2017.

[222] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[223] S. Li, M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr, "A fundamental tradeoff between computation and communication in distributed computing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 109–128, Jan. 2018.

[224] M. Goldenbaum, H. Boche, and S. Stańczak, "Nomographic functions: Efficient computation in clustered Gaussian sensor networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2093–2105, Apr. 2015.

[225] F. Wang and V. K. N. Lau, "Multi-level over-the-air aggregation of mobile edge computing over D2D wireless networks," 2021, *arXiv:2105.00471*.

[226] K. Li, M. Tao, and Z. Chen, "Exploiting computation replication for mobile edge computing: A fundamental computation-communication tradeoff study," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4563–4578, Jul. 2020.

[227] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[228] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May 2016.

[229] J. Shao, H. Zhang, Y. Mao, and J. Zhang, "Branchy-GNN: A device-edge co-inference framework for efficient point cloud processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8488–8492.

[230] J. Cao, W. Feng, N. Ge, and J. Lu, "Delay characterization of mobile-edge computing for 6G time-sensitive services," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3758–3773, Mar. 2021.

[231] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint pilot and payload power allocation for massive-MIMO-enabled URLLC IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816–830, May 2020.

[232] V. Verma *et al.*, "Manifold mixup: Better representations by interpolating hidden states," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6438–6447.

[233] J. Shao and J. Zhang, "BottleNet++: An end-to-end approach for feature compression in device-edge co-inference systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[234] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 89–100, Jan. 2021.

[235] N. Tishby, F. C. N. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, 2000, pp. 368–377.

[236] I. E. Aguerri and A. Zaidi, "Distributed variational representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 120–138, Jan. 2021.

[237] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," 2021, *arXiv:2106.09316*.

[238] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, Jan. 2020.

[239] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 219–232, Jan. 2021.

[240] S. Wang *et al.*, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 1205–1221, Jun. 2019.

[241] D. Wen, M. Bennis, and K. Huang, "Joint parameter-and-bandwidth allocation for improving the efficiency of partitioned edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8272–8286, Dec. 2020.

[242] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 453–467, Jan. 2021.

[243] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 17, 2021, Art. no. e2024789118.

[244] C. T. Dinh *et al.*, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 398–409, Feb. 2021.

[245] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2150–2167, Jul. 2021.

[246] T. F. de Lima *et al.*, "Machine learning with neuromorphic photonics," *J. Lightw. Technol.*, vol. 37, no. 5, pp. 1515–1534, Mar. 1, 2019.

[247] M. Alioto, V. De, and A. Marongiu, "Energy-quality scalable integrated circuits and systems: Continuing energy scaling in the twilight of Moore's law," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 4, pp. 653–678, Dec. 2018.

[248] Q. Zeng, Y. Du, and K. Huang, "Wirelessly powered federated edge learning: Optimal tradeoffs between convergence and power transfer," 2021, *arXiv:2102.12357*.

[249] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "How to evaluate deep neural network processors: TOPS/W (alone) considered harmful," *IEEE Solid State Circuits Mag.*, vol. 12, no. 3, pp. 28–41, Aug. 2020.

[250] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[251] Y. Mao, J. Zhang, Z. Chen, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[252] Y. Qu *et al.*, "Decentralized privacy using blockchain-enabled federated learning in fog computing," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5171–5183, Jun. 2020.

[253] S. R. Pokhrel and J. Choi, "Federated learning with blockchain for autonomous vehicles: Analysis and design challenges," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4734–4746, Aug. 2020.

[254] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5239–5269, Oct. 2019.

[255] S. Xia and Y. Shi, "Learning shallow neural networks via provable gradient descent with random initialization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5616–5620.

[256] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: Overview and the geometric picture," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 853–884, Feb. 2017.

[257] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1724–1732.

[258] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed L_p-minimization for green cloud-RAN with user admission control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1022–1036, Apr. 2016.

[259] Y. Bengio, A. Lodi, and A. Prouvost, "Machine learning for combinatorial optimization: A methodological tour d'horizon," *Eur. J. Oper. Res.*, vol. 290, no. 2, pp. 405–421, Apr. 2021.

[260] Y. Zhang, B. Di, Z. Zheng, J. Lin, and L. Song, "Distributed multi-cloud multi-access edge computing by multi-agent reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2565–2578, Apr. 2021.

[261] B. Feng, J. Gao, Y. Wu, W. Zhang, X.-G. Xia, and C. Xiao, "Optimization techniques in reconfigurable intelligent surface aided networks," 2021, *arXiv:2106.15458*.

[262] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[263] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 26, no. 4, pp. 681–683, 1978.

[264] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[265] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.

[266] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[267] A. Chowdhury, G. Verma, C. Rao, A. Swami, and S. Segarra, "Unfolding WMMSE using graph neural networks for efficient power allocation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6004–6017, Sep. 2021.

[268] Q. Hu, Y. Liu, Y. Cai, G. Yu, and Z. Ding, "Joint deep reinforcement learning and unfolding: Beam selection and precoding for mmWave multiuser MIMO with lens arrays," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2289–2304, Aug. 2021.

[269] Q. Hu, Y. Cai, Q. Shi, K. Xu, G. Yu, and Z. Ding, "Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1394–1410, Feb. 2021.

[270] M. Eisen and A. Ribeiro, "Optimal wireless resource allocation with random edge graph neural networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 2977–2991, 2020.

[271] M. M. Wadu, S. Samarakoon, and M. Bennis, "Joint client scheduling and resource allocation under channel uncertainty in federated learning," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5962–5974, Sep. 2021.

[272] Y. Shi, J. Zhang, and K. B. Letaief, "Optimal stochastic coordinated beamforming for wireless cooperative networks with CSI uncertainty," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 960–973, Feb. 2015.

[273] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Math. Program.*, vol. 171, no. 1, pp. 115–166, 2018.

[274] W. Cui, K. Shen, and W. Yu, "Spatial deep learning for wireless scheduling," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1248–1261, Jun. 2019.

[275] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

[276] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[277] H. Sun, W. Pu, M. Zhu, X. Fu, T.-H. Chang, and M. Hong, "Learning to continuously optimize wireless resource in episodically dynamic environment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4945–4949.

[278] X. Liu, Y. Shi, J. Zhang, and K. B. Letaief, "Massive CSI acquisition for dense cloud-RANs with spatial-temporal dynamics," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2557–2570, Apr. 2018.

[279] J.-C. Shen, J. Zhang, E. Alsusa, and K. B. Letaief, "Compressed CSI acquisition in FDD massive MIMO: How much training is needed?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4145–4156, Jun. 2016.

[280] L. Chen, N. Zhao, Y. Chen, X. Qin, and F. R. Yu, "Computation over MAC: Achievable function rate maximization in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5446–5459, Sep. 2020.

[281] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.

[282] H. Liu, X. Yuan, and Y.-J.-A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2621–2636, Nov. 2020.

[283] C. Hu, L. Dai, S. Han, and X. Wang, "Two-timescale channel estimation for reconfigurable intelligent surface aided wireless communications," *IEEE Trans. Commun.*, early access, Apr. 12, 2021, doi: 10.1109/TCOMM.2021.3072729.

[284] C. Liu, X. Liu, D. W. K. Ng, and J. Yuan, "Deep residual learning for channel estimation in intelligent reflecting surface-assisted multi-user communications," *IEEE Trans. Wireless Commun.*, early access, Aug. 3, 2021, doi: 10.1109/TWC.2021.3100148.

[285] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[286] Y. Ma, Y. Shen, X. Yu, J. Zhang, S. Song, and K. B. Letaief, "Neural calibration for scalable beamforming in FDD massive MIMO with implicit channel estimation," 2021. *arXiv:2108.01529*.

[287] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018.

[288] E. Björnson, E. A. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective signal processing optimization: The way to balance conflicting metrics in 5G systems," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 14–23, Nov. 2014.

[289] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2595–2621, 4th Quart., 2018.

[290] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[291] *Privacy Protection—Privacy Guidelines for Smart Cities*, Standard ISO/IEC TS 27570:2021, 2021. [Online]. Available: https://www.iso27001security.com/html/27570.html

[292] *Guidelines for Security and Privacy in Internet of Things*, Standard ISO/IEC 27030, 2021. [Online]. Available: https://www.iso27001security.com/html/27030.html

[293] *Cyber Security for Consumer Internet of Things: Baseline Requirements*, Standard ETSI EN 303 645, 2020. [Online]. Available: https://www.etsi.org/newsroom/press-releases/1789-2020-06-etsi-releases-world-leading-consumer-iot-security-standard

[294] *Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 15)*, document 3GPP TS 23.501, Version 15.1.0, 3rd Generation Partnership Project, Mar. 2018.

[295] S. Haykin, *An Introduction to Analog and Digital Communication*. Hoboken, NJ, USA: Wiley, 1994.

[296] *Mosek*. [Online]. Available: https://www.mosek.com/

[297] *Gurobi*. [Online]. Available: https://www.gurobi.com/

[298] A. Mirhoseini *et al.*, "A graph placement methodology for fast chip design," *Nature*, vol. 594, no. 7862, pp. 207–212, 2021.

[299] C.-W. Qiu, T. Zhang, G. Hu, and Y. Kivshar, "Quo vadis, metasurfaces?" *Nano Lett.*, vol. 21, no. 13, pp. 5461–5474, Jun. 2021.

[300] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146–1159, Feb. 2020.

[301] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3177–3192, Oct. 2021.

[302] S. Liu, L. Liu, J. Tang, B. Yu, Y. Wang, and W. Shi, "Edge computing for autonomous driving: Opportunities and challenges," *Proc. IEEE*, vol. 107, no. 8, pp. 1697–1716, Aug. 2019.

[303] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, Dec. 2011.

[304] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[305] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 2, no. 3, pp. 194–220, Sep. 2017.

[306] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, vol. 96. Urbana, IL, USA: Univ. Illinois Press, 1949.

[307] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial Internet of Things: Challenges, opportunities, and directions," *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4724–4734, Nov. 2018.

[308] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019.

[309] G. N. Schroeder, C. Steinmetz, R. N. Rodrigues, R. V. B. Henriques, A. Rettberg, and C. E. Pereira, "A methodology for digital twin modeling and deployment for industry 4.0," *Proc. IEEE*, vol. 109, no. 4, pp. 556–567, Apr. 2021.

[310] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4177–4186, Jun. 2020.

[311] Z. Obermeyer and E. J. Emanuel, "Predicting the future—Big data, machine learning, and clinical medicine," *New England J. Med.*, vol. 375, no. 13, p. 1216, 2016.

[312] S. K. Zhou *et al.*, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.

[313] Y. Sun, F. Lo, and B. Lo, "Security and privacy for the internet of medical things enabled healthcare systems: A survey," *IEEE Access*, vol. 7, pp. 183339–183355, 2019.

[314] M. Subramanian *et al.*, "Precision medicine in the era of artificial intelligence: Implications in chronic disease management," *J. Transl. Med.*, vol. 18, no. 1, pp. 1–12, Dec. 2020.

[315] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020.

[316] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," 2018, *arXiv:1812.00564*.

[317] O. Gottesman *et al.*, "Guidelines for reinforcement learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 16–18, Jan. 2019.

[318] N. Fazeli, M. Oller, J. Wu, Z. Wu, J. B. Tenenbaum, and A. Rodriguez, "See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion," *Sci. Robot.*, vol. 4, no. 26, pp. 1–22, Jan. 2019.

[319] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, pp. 698–702, May 2019.

[320] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[321] L. Ruan, M. P. I. Dias, and E. Wong, "Achieving low-latency human-to-machine (H2M) applications: An understanding of H2M traffic for AI-facilitated bandwidth allocation," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 626–635, Jan. 2021.

[322] N. Promwongsa *et al.*, "A comprehensive survey of the tactile internet: State-of-the-art and research directions," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 472–523, 1st Quart., 2021.

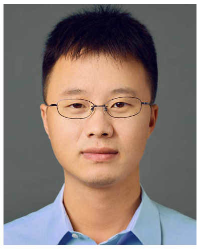
[323] Y. Xiao and M. Krunz, "Distributed optimization for energy-efficient fog computing in the tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2390–2400, Nov. 2018.

[324] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free NOMA toward reliable low-latency access in tactile Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 2995–3005, May 2019.

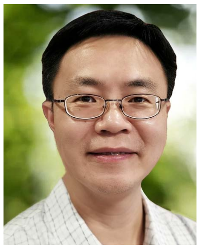
Khaled B. Letaief (Fellow, IEEE) received the B.S. (Hons.), M.S., and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in December 1984, August 1986, and May 1990, respectively.

From 1990 to 1993, he was a Faculty Member with the University of Melbourne, Australia. Since 1993, he has been with The Hong Kong University of Science and Technology (HKUST), where he is currently a New Bright Professor of engineering. While at HKUST, he has held many administrative positions, including the Dean of Engineering, the Head of the Electronic and Computer Engineering Department, the Director of the Wireless IC Design Center and the Hong Kong Telecom Institute of Information Technology, and the Founding Director of Huawei Innovation Laboratory. He served as a Consultant for different organizations, including Huawei, ASTRI, ZTE, Nortel, PricewaterhouseCoopers, and Motorola. He is currently an Internationally Recognized Leader in wireless communications and networks with research interest in artificial intelligence, big data analytics systems, mobile cloud and edge computing, tactile internet, and 5G systems and beyond. In these areas, he has over 720 papers along with 15 patents, including 11 U.S. inventions.

Dr. Letaief is a member of the United States National Academy of Engineering and the Hong Kong Academy of Engineering Sciences, and a fellow of the Hong Kong Institution of Engineers. He is well recognized for his dedicated service to professional societies and IEEE, where he has served in many leadership positions. These include the IEEE Communications Society Vice President for conferences, an Elected Member of the IEEE Product Services and Publications Board, and the IEEE Communications Society Vice President for technical activities. He also served as the President for the IEEE Communications Society for the period 2018–19, the world's leading organization for communications professionals with headquarter in New York City and members in 162 countries. From 2022 to 2023, he will serve as member for the IEEE Board of Directors. He is also recognized by Thomson Reuters as an ISI Highly Cited Researcher and was listed among the 2020 top 30 of AI 2000 Internet of Things Most Influential Scholars. He was a recipient of many distinguished awards and honors, including the 2021 IEEE Communications Society Best Survey Paper Award, the 2019 Distinguished Research Excellence Award by the HKUST School of Engineering (Highest Research Award and only one recipient/three years is honored for his/her contributions), the 2019 IEEE Communications Society and Information Theory Society Joint Paper Award, the 2018 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 IEEE Cognitive Networks Technical Committee Publication Award, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2016 IEEE Marconi Prize Paper Award in Wireless Communications, the 2011 IEEE Wireless Communications Technical Committee Recognition Award, the 2011 IEEE Communications Society Harold Sobol Award, the 2010 Purdue University Outstanding Electrical and Computer Engineer Award, the 2009 IEEE Marconi Prize Award in Wireless Communications, the 2007 IEEE Communications Society Joseph LoCicero Publications Exemplary Award, and 19 IEEE best paper awards. He is the Founding Editor-in-Chief of the prestigious IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and been involved in organizing many flagship international conferences.



Yuanming Shi (Senior Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology (HKUST) in 2015. Since September 2015, he has been with the School of Information Science and Technology, ShanghaiTech University, where he is currently a Tenured Associate Professor. He visited the University of California, Berkeley, CA, USA, from October 2016 to February 2017. His research areas include optimization, statistics, machine learning, wireless communications, and their applications to 6G, the IoT, and edge AI. He was a recipient of the 2016 IEEE Marconi Prize Paper Award in Wireless Communications, the 2016 Young Author Best Paper Award by the IEEE Signal Processing Society, and the 2021 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He is also an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.



Jianmin Lu joined Huawei Technologies in 1999. He is currently the Executive Director of Huawei Wireless Technology Laboratory. During the last two decades, he conducted various researches on wireless communications especially on physic layer and MAC layer and developed 3G, 4G, and 5G products. He received more than 50 patents during the research. He was deeply involved in 3GPP2 (EVDO/UMB), WiMAX/802.16m, and 3GPP(LTE/NR) standardization and contributed several key technologies, such as flexible radio frame structure, radio resource management, and MIMO. His current research interest is in the area of signal processing, protocol, and networking for the next generation wireless communication.

Jianhua Lu (Fellow, IEEE) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1986 and 1989, respectively, and the Ph.D. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, China, in 1998.

Since 1989, he has been with the Department of Electronic Engineering, Tsinghua University, where he currently serves as a Professor. He has authored/coauthored over 300 refereed technical papers published in international renowned journals and conferences and over 80 Chinese invention patents. His research interests include broadband wireless communications, multimedia signal processing, and satellite communications.

Prof. Lu is a member of the Chinese Academy of Sciences. He is the Vice President of the National Natural Science Foundation of China. He was a recipient of the Best Paper Awards at the IEEE ICCCS 2002, *China Communications* 2006, IEEE Embedded-Com 2012, IEEE WCSP 2015, IEEE IWCMC 2017, and IEEE ICNC 2019. He served as the program committee co-chair and a TPC member for many international conferences, and an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2008 to 2011. He is the Editor-in-Chief of *China Communications*.