SCIENCE CHINA
Information Sciences

March 2023, Vol. 66 130301:1–130301:19
https://doi.org/10.1007/s11432-022-3652-2

• REVIEW •
Special Topic: Spectrum, Coverage, and Enabling Technologies for Intelligent 6G

Pushing AI to wireless network edge: an overview on integrated sensing, communication, and computation towards 6G

Guangxu ZHU[1], Zhonghao LYU[2,3,1], Xiang JIAO[4,1], Peixi LIU[4,1], Mingzhe CHEN[5], Jie XU[3,2*], Shuguang CUI[3,2,1,7*] & Ping ZHANG[6,7]

[1]*Shenzhen Research Institute of Big Data, Shenzhen* 518172*, China;*
[2]*Future Network of Intelligence Institute (FNii), The Chinese University of Hong Kong (Shenzhen), Shenzhen* 518172*, China;*
[3]*School of Science and Engineering (SSE), The Chinese University of Hong Kong (Shenzhen), Shenzhen* 518172*, China;*
[4]*State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics, Peking University, Beijing* 100871*, China;*
[5]*Department of Electrical and Computer Engineering and Institute for Data Science and Computing, University of Miami, Coral Gables* FL 33146*, USA;*
[6]*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing* 100876*, China;*
[7]*Peng Cheng Laboratory, Shenzhen* 518066*, China*

Received 31 August 2022/Revised 2 November 2022/Accepted 13 December 2022/Published online 14 February 2023

Abstract Pushing artificial intelligence (AI) from central cloud to network edge has reached board consensus in both industry and academia for materializing the vision of artificial intelligence of things (AIoT) in the sixth-generation (6G) era. This gives rise to an emerging research area known as edge intelligence, which concerns the distillation of human-like intelligence from the vast amount of data scattered at the wireless network edge. Typically, realizing edge intelligence corresponds to the processes of sensing, communication, and computation, which are coupled ingredients for data generation, exchanging, and processing, respectively. However, conventional wireless networks design the three mentioned ingredients separately in a task-agnostic manner, which leads to difficulties in accommodating the stringent demands of ultra-low latency, ultra-high reliability, and high capacity in emerging AI applications like auto-driving and metaverse. This thus prompts a new design paradigm of seamlessly integrated sensing, communication, and computation (ISCC) in a task-oriented manner, which comprehensively accounts for the use of the data in downstream AI tasks. In view of its growing interest, this study provides a timely overview of ISCC for edge intelligence by introducing its basic concept, design challenges, and enabling techniques, surveying the state-of-the-art advancements, and shedding light on the road ahead.

Keywords sixth-generation (6G), edge intelligence, artificial intelligence of things (AIoT), integrated sensing, communication, and computation (ISCC)

Citation Zhu G X, Lyu Z H, Jiao X, et al. Pushing AI to wireless network edge: an overview on integrated sensing, communication, and computation towards 6G. Sci China Inf Sci, 2023, 66(3): 130301, https://doi.org/10.1007/s11432-022-3652-2

1 Introduction

With the commercialization of fifth-generation (5G) wireless networks, we are moving toward a new era where everything is connected. The convergence of modern information and communications technologies, such as the internet of things (IoT), cloud computing, mobile edge computing (MEC), and big data analytics, is prompting a huge leap in social productivity and management efficiency. Moreover, artificial intelligence (AI) is achieving great success in various applications and continues its explosive growth and penetration into all walks of life, driving the ongoing convergence of communication networks with AI technology. Specifically, in the current 5G network, AI has been used as an add-on module to boost

* Corresponding author (email: xujie@cuhk.edu.cn, shuguangcui@cuhk.edu.cn)

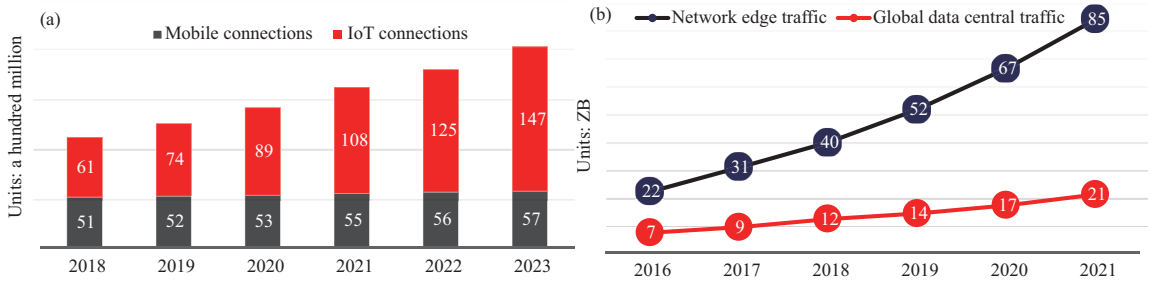

Figure 1 (Color online) Edge AI connections and data growth. (a) Number of mobile and IoT connections [7]; (b) global data traffic growth [9].

network performance, while in the future 6G era, it will be deeply integrated into the network design to achieve the so-called AI-native network. As foreseen, the future 6G wireless communication networks will go beyond a pure data delivery pipeline and become a comprehensive platform integrating sensing, communication, computation, and intelligence to deliver pervasive AI services [1–3].

1.1 Edge AI

The interplay between wireless communication networks and AI technology can be generally classified into two categories: AI-assisted communication [4] and communication-assisted AI [5]. AI-assisted communication, which refers to the use of AI to help existing communication systems (e.g., channel estimation and signal detection [6]), boosts the end-to-end performance of communication links to achieve higher rates, lower latency, and better connectivity. Conversely, communication-assisted AI, which refers to the use of communication networks to help AI acquisition, allows for distributed AI training and inference across the entire network, as well as the delivery of pervasive AI services.

The current study focuses on communication-assisted AI, which has received increasing attention from academia and industry in recent years, as various AI applications such as industrial Internet, smart cities, virtual reality (VR), augmented reality (AR), smart health, metaverse, and auto-driving, continue to mature and gain popularity. Traditional communication-assisted AI based on cloud data processing, which requires the delivery of a large amount of data collected by the edge devices to the cloud for AI distillation, cannot support the emerging mission-critical AI applications mentioned above due to the following challenges.

• **Massive data and network connectivity.** According to Cisco, global business volumes increased at a rate of nearly 42% per year between 2018 and 2020. As shown in Figure 1(a), the total number of connected devices worldwide is expected to reach 29.3 billion by 2023, with 5.7 billion mobile connections and 14.7 billion IoT connections [7]. Furthermore, according to Huawei [8], the total number of global network connections will reach 200 billion by 2030, with wireless and passive connections constituting approximately half of the total. Moreover, besides the massive number of temperature, humidity, pressure, and photoelectric sensors in the industrial sector, the network will also include a large number of smart vehicles, robots, and drones. The ever-growing amounts of data and network connections increase the demand for communication capacity and computing power.

• **Data sinking.** Previously, big data, such as online shopping records, social media content, and business information, was primarily generated and stored in hyperscale data centers. However, with the proliferation of mobile and IoT devices, this trend is now being reversed. Till 2021, all people, machines, and things have generated nearly 85 zettabytes of usable data at the network edge. In contrast, as shown in Figure 1(b), global data center traffic had only reached 21 zettabytes by 2021 [9]. In the traditional cloud computing design, the massive data sinking to the network edge need to be transferred to a central cloud server far away from the edge for analysis and processing, which will undoubtedly result in an unacceptable communication cost and delay.

• **Ultra-low latency requirements.** Most new AI services require high-demand network connections with ultra-low latency and ultra-high reliability. For example, smart industrial Internet requires real-time state feedback, data analysis, and highly accurate control [1]. Similarly, VR and AR applications request real-time aggregation, analysis, and reconstruction on three-dimensional images for complex control feedback. In such cases, the required closed-loop sensing-communication-computation latency must

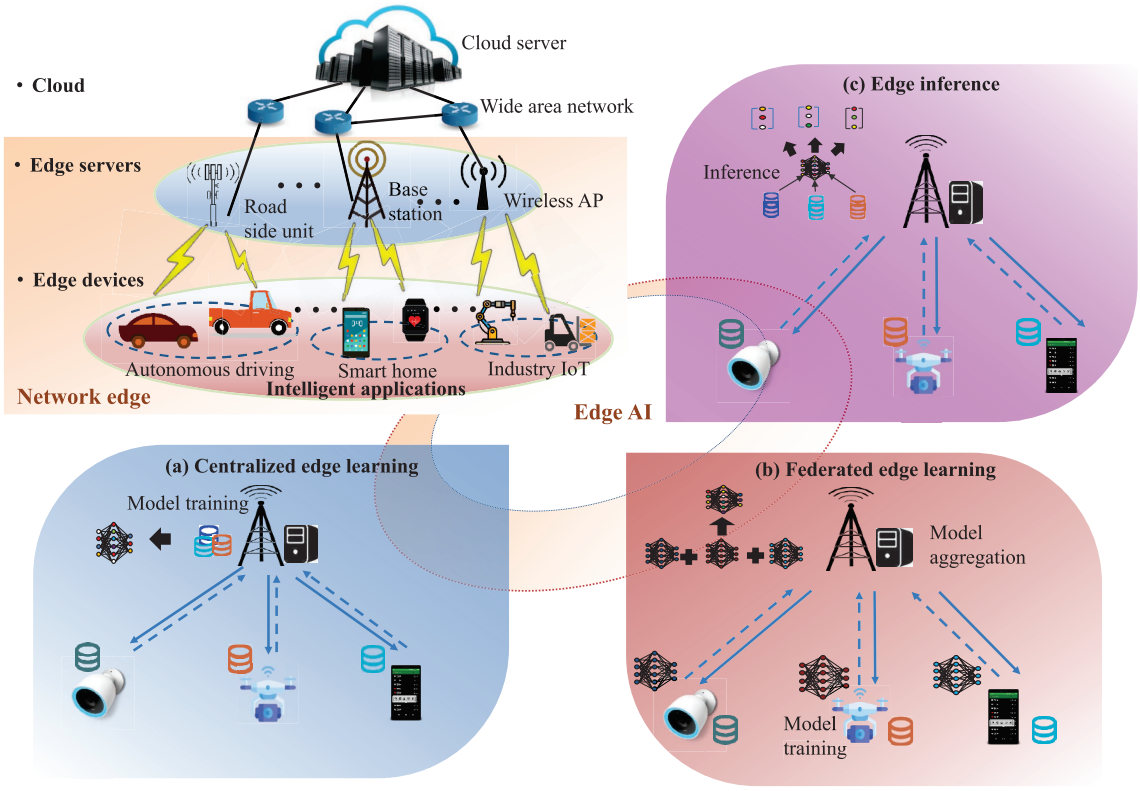

Figure 2 (Color online) The concept of edge AI and three related scenarios. (a) Centralized edge learning; (b) federated edge learning; (c) edge inference.

be within approximately one millisecond, which poses formidable challenges to the current communication networks.

To address these issues, industry and academia have agreed on bringing computing power and AI functionality close to data, leading to a new technical breakthrough called "edge computing" or "fog computing" [10]. As its name implies, edge computing aims to relocate some of the data processing and storage for specific services from the central cloud to the distributed edge network nodes, which are physically and logically closer to the data provider, so as to achieve the desired low latency. Moreover, AI seeks to emulate human intelligence in a machine by learning from the data. Naturally, the convergence of edge computing and AI gives rise to a new domain called edge AI [9], which aims to provide mobile terminals with low-latency AI services by exploiting both the computing resources and data scattered at the network edge. Due to its promising performance gain, edge AI has received increasing attention from both academia and industry and has become a popular area in the field of communication-assisted AI [2, 11, 12].

As shown in Figure 2, edge AI can include two types of learning, depending on where the data is processed: centralized edge learning and distributed edge learning. After the AI model is well-trained, it can be deployed for edge inference. In the early stage of edge AI, the AI model is attained via centralized edge learning in many large AI companies (e.g., Google, Facebook, and Microsoft) [11]. However, centralized edge learning requires uploading the private raw data (e.g., personal photos in smart phones) from edge devices to an edge server, which may pose a huge challenge to user data privacy. However, owing to the Moore's law, the power of computing chips, such as central processing units (CPUs) and graphics processing units (GPUs), has continuously been upgraded with decreasing costs and size. Particularly, with the emergence of dedicated AI chipsets, the computing power of edge devices has increased tremendously and can now support the running of machine learning (ML) tasks, driving the rapid development of distributed edge learning, such as federated edge learning (FEEL), to exploit the rich distributed computing resources at the network edge. Moreover, in FEEL, data privacy is preserved because the need to upload raw data is waived in favor of sharing the less privacy-sensitive gradient or model updates.

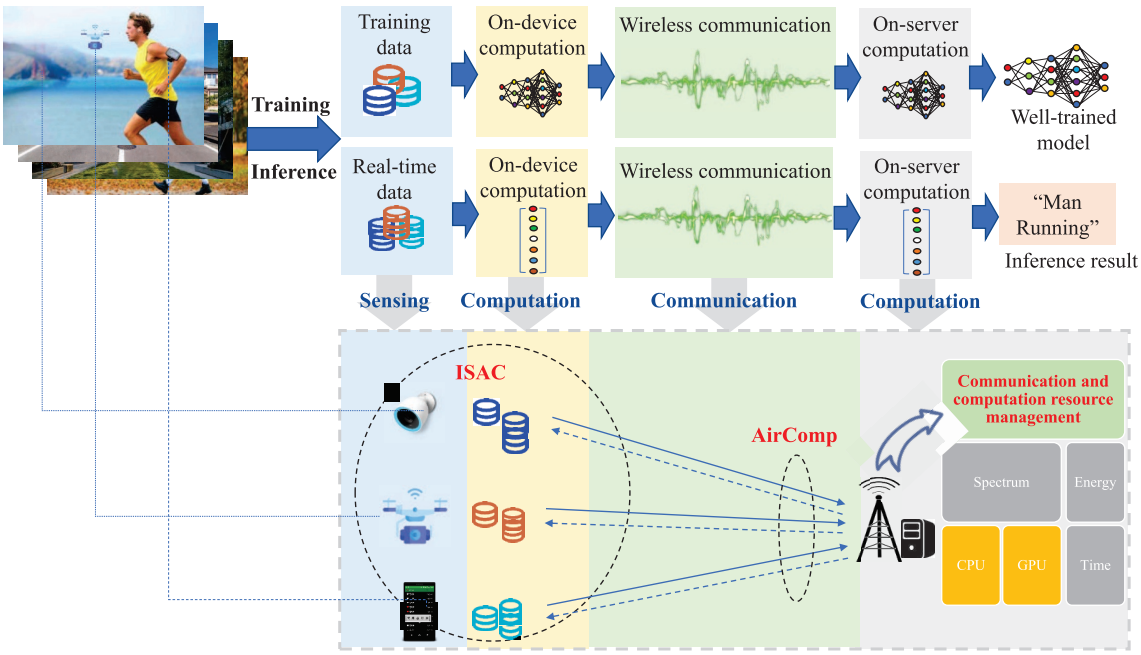

Figure 3 (Color online) Sensing, communication, and computation in edge AI.

1.2 Integrated sensing, communication, and computation (ISCC)

In practice, a complex system (e.g., the edge AI system) generally consists of three coupled processes, namely sensing, communication, and computation, as shown in Figure 3. However, in traditional wireless networks, these three processes are designed separately for different goals: sensing for obtaining high-quality environmental data, communication for data delivery, and computation for executing the downstream task within a certain deadline. Such a separate design principle encounters difficulty in accommodating the stringent demands of ultra-low latency, ultra-high reliability, and high capacity in emerging 6G applications such as auto-driving and metaverse. This thus prompts a new wireless design paradigm of ISCC [13] in a task-oriented manner, which comprehensively accounts for the use of the data in the downstream tasks (e.g., AI applications) in 6G. In the literature, there have been some prior studies on the integration of two of the above three entities. Some examples include joint communication and computation resource management, over-the-air computation (AirComp), and integrated sensing and communication (ISAC), which are elaborated in the following.

1.2.1 *Joint communication and computation resource management*

Data acquisition and model computation are typically separate processes in traditional complex communication systems. With the rapid increase in data volumes in MEC, the communication and computation capabilities at the network edge become the bottleneck. Particularly, the limited wireless resources make it challenging for the edge server to receive significant amounts of data from edge devices swiftly through wireless links. Hence, many researches have focused on joint communication and computation resource management to tackle this issue in MEC. For example, in [14], in order to minimize the energy and delay cost of the multi-user multi-task MEC system, the authors used a separable semidefinite relaxation method to jointly optimize the offloading decision and communication resource allocation. In [15], in order to solve the resource allocation problem for MEC, the authors proposed an effective solution to maximize the quality of service (QoS) of all mobile devices, by transforming the problem into a linear programming model. Besides, in [16], by using an online algorithm based on Lyapunov optimization, the energy-latency trade-off problem of multi-user MEC systems is studied, where the computation tasks arrive at the mobile devices in a stochastic manner. In addition, Ref. [17] solved the problem of reducing the total energy consumption at the edge server in wireless-powered multi-user MEC systems, by jointly optimizing the AP's energy transmit beamforming, the user's CPU frequency, the number of offloaded bits, and the time allocation among users to improve MEC performance. Furthermore, joint communication and computation cooperation was exploited in MEC in [18], where a neighboring user node serves

as not only a relay node for helping task offloading of a user, but also a computation helper to help remotely execute some of tasks of that user. Joint communication and computation resource management is essential to ensure the information security, mobile energy saving, and so on.

1.2.2 *AirComp*

AirComp has emerged as a promising technology in recent years. As opposed to "communication then computing", AirComp integrates computing into communication, resulting in a new technique featuring "communication while computing". In contrast to traditional wireless communication over a multi-access channel (MAC), which requires separate transmission and decoding of information, AirComp allows edge devices to simultaneously transmit their respective signals on the same frequency band with proper processing, such that the functional computation of the distributed data is accomplished directly over the air. This thus significantly improves the communication and computing efficiency, and considerably reduces the latency required for multiple access and data fusion.

Ref. [19] provided a comprehensive overview of AirComp by introducing the basic principles, discussing the advanced techniques and promising applications, and identifying promising research opportunities. In order to achieve reliable AirComp in practice, Ref. [20] focused on the power control problem in AirComp, and the optimal power allocation under both deterministic and fading channels was derived by minimizing the mean-squared error (MSE) of the aggregated signals. Similarly, Ref. [21] minimized the computation MSE at the receiver by optimizing the transmitting and receiving policy under the maximum power constraint of each sensor. While only a single cell was considered in [20, 21], the power control problem of AirComp in the multi-cell scenario was considered in [22]. To quantify the fundamental AirComp performance trade-off among different cells, in [22], the Pareto boundary of the multi-cell MSE region was characterized by minimizing the MSE subject to a set of constraints on individual MSE. Note that the work in [20–22] only considered the scenario with a single antenna. Ref. [23] generalized AirComp to the multiple-input-multiple-output (MIMO) setup to support multi-modal sensing with high mobility, where MIMO-AirComp equalization and efficient channel feedback were designed for spatially multiplexing multi-function computation. Subsequently, AirComp in more complex systems has been also considered in the literatures. For example, Ref. [24] considered to use reconfigurable intelligent surface (RIS) to assist AirComp. Besides, under imperfect channel state information, Ref. [25] investigated the joint optimization of transceiver and RIS phase design for AirComp systems. In [26], when the ground receiver is unavailable, unmanned aerial vehicles (UAVs) are utilized to establish line-of-sight (LoS) connections by tracking mobile sensors, and thus improving the performance of AirComp.

1.2.3 *ISAC*

ISAC generally refers to the integration of sensing and communication into a unified design in wireless networks to enhance the efficiency of spectrum use allowing a mutual benefit via sensing-assisted communication and communication-assisted sensing [27]. In comparison to traditional wireless networks, ISAC can use the wireless infrastructure as well as limited spectrum and power resources for both communication and sensing, which can potentially improve the system performance at a lower cost.

ISAC is one of the potential key technologies in 6G networks that have received a lot of attention in the literature. Many studies have focused on joint sensing and communication in [28]. For example, Ref. [29] proposed a dual-functional MIMO radar communication system that consists of a transmitter with multiple antennas that can communicate with downlink cellular users and detect radar targets at the same time. In [30], a joint transmit beamforming model for a dual-function MIMO radar and multiuser MIMO communication transmitter was proposed and the weighting coefficients of the radar beamforming were designed. Ref. [31] also considered the beamforming optimization problem in ISAC system, and the radar sensing performance is maximized subject to the communication users' minimum signal-to-interference-plus-noise ratio (SINR) requirements and the transmit power constraint of the base station (BS). Ref. [32] employed the Cramér-Rao bound (CRB) as a performance metric of target estimation, and the CRB of radar sensing is minimized while guaranteeing a pre-defined level of SINR for each communication user. Ref. [33] considered a UAV-enabled ISAC system, where UAV trajectory/deployment and beamforming design are jointly considered to balance the sensing-communication performance trade-off under quasi-stationary and mobile UAV scenarios, respectively. Furthermore, using RIS to facilitate radar sensing and ISAC has attracted growing research interests (see [34–38]). For instance, Ref. [34] derived the fundamental CRB for RIS-enabled non-line-of-sight (NLoS) sensing. Ref. [35] jointly designed the transmit

beamforming and the RIS reflective beamforming for ensuring both sensing and communication performance. Ref. [36] used RIS for a joint design of constant-modulus waveform and discrete phase shift to mitigate multi-user interference in ISAC. In addition, Ref. [37] elaborated the benefits of RIS in wireless communication, sensing, and security, and envisioned that the RIS-assisted communication and sensing will mutually benefit each other. Ref. [38] considered the combination of ISAC and AirComp to improve the spectral efficiency and sensing performance, and the beamformers for sensing, communication, and computation were jointly optimized. Ref. [39] used ISAC in smart homes to provide inconspicuous sensing and ubiquitous connectivity. Besides joint sensing and communication, there is also a line of research on sensing-assisted communication. For example, a radar-assisted predictive beamforming design for vehicle-to-infrastructure communication was investigated in [40], and it was found that the communication beam tracking overheads can be drastically reduced by exploiting the radar functionality of the road side unit.

1.2.4 *Task-oriented ISCC towards edge AI*

As previously stated, sensing, computation, and communication have a symbiotic relationship, especially in the context of edge AI. Specifically, the ultimate performance of edge AI depends on the input feature vector's distortion level arising from three processes, i.e., data acquisition (sensing), feature extraction (computation), and feature uploading to the edge server (communication). Particularly, sensing and communication compete for radio resources, and the allowed communication resource further determines the required quantization (distortion) level such that the quantized features can be transmitted reliably to the edge server under a delay constraint. Thereby the three processes are highly coupled and need to be jointly considered. Furthermore, the implementation of ISCC should be designed under a new task-oriented principle that concerns the successful completion of the subsequent AI task. Different from conventional communication system design aiming at maximizing the data-rate throughput, the ultimate performance metrics of interest for the system become the inference/training accuracy, latency, energy efficiency, etc. For instance, an edge AI task-oriented ISCC scheme can be designed to maximize the inference/training accuracy under constraints on low latency and on-device resources. This is in sharp contrast to the classic separation-based design approach that considers the sensing, communication, and computation processes in isolation.

1.3 Structure of the survey

Different from prior studies considering the integration of sensing and communication or computation and communication in generic wireless networks, we focus on their integration towards edge AI applications. As a result, we classify them based on three application scenarios, i.e., centralized edge learning, FEEL, and edge inference. The remaining of the survey is organized as shown in Figure 4. In Section 2, we first review the joint communication and computation resource management, mixup data augmentation with AirComp, and ISAC, respectively, in centralized edge learning, and then discuss the related research opportunities. In Section 3, we discuss the joint communication and computation resource management in federated learning, the application of AirComp into FEEL, and the combination of ISAC and FEEL, as well as the related research opportunities. In Section 4, we first discuss the joint source and channel coding (JSCC), then present joint communication and computation resource management in edge inference, over-the-air edge inference, and co-inference with ISAC, respectively, and highlight several key research opportunities. Finally, Section 5 concludes the article.

2 Centralized edge learning

With the continuing development of deep learning (DL) in recent years, the increasing DL model complexity has posed a grand challenge in training due to the demand for computation power and storage capacity. There are two basic strategies to accommodate the increasing demands for the resources required by DL at the centralized cloud: scaling-up, which involves adding extra processing and storage resources to a single central server, and scaling-out [41], which involves forming a server cluster by networking multiple servers each with certain computing and storage capacity. The recent rapid development of MEC makes it possible to deploy the mentioned two strategies at the network edge so as to "bring the computation power close to the data", leading to an emerging research area known as centralized edge learning.

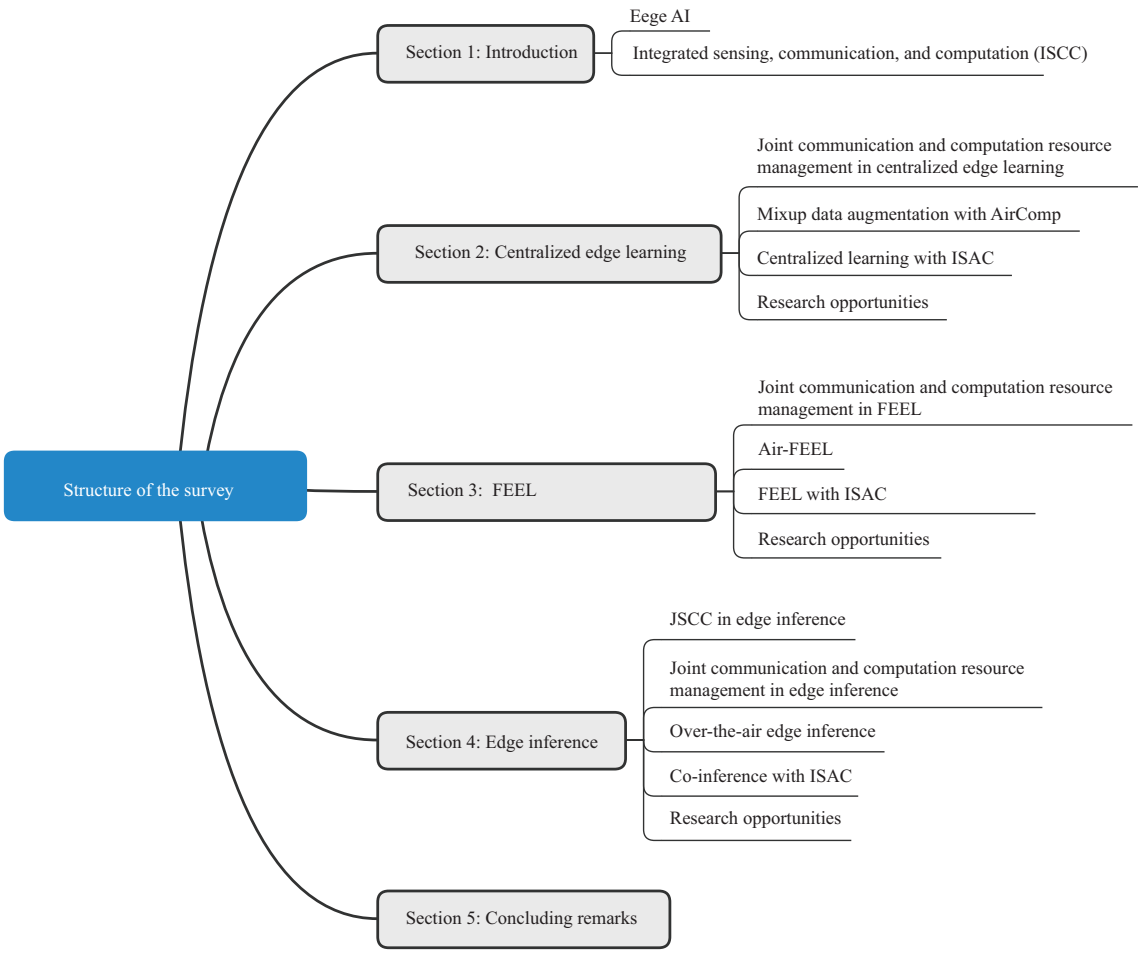

Figure 4 (Color online) The structure of this article.

In the centralized edge learning, the data are first collected on the client side and then transferred through the wireless channel to a central edge server. The central server then stores and processes the data, and finally returns the learned model back to the client. This architecture is simple to deploy and manage, particularly in circumstances where the data are scattered across geographically dispersed nodes. However, due to the need for centralized data processing, long delays and high transmission costs accompany when the communication channel between the devices and the central server is poor. Furthermore, due to the central edge server's limited computational power and storage resources, it is difficult to enable the construction of complicated models based on massive datasets using centralized edge learning. Therefore, in order to improve the efficiency of centralized edge learning, the sensing, communication, and computation processes need to be jointly designed and the associated resources should be judiciously managed as described in the sequel.

2.1 Joint communication and computation resource management in centralized edge learning

As previously stated, centralized edge learning may introduce a significant delay, which will be catastrophic in delay-sensitive applications such as autonomous driving and VR games. Furthermore, the massive data communication places a significant strain on the backbone network, resulting in significant computation overhead for the central server. Thus, centralized edge learning exists at the crossroads of two domains: communication and computing. This thus brings many interdisciplinary research opportunities, and joint design is required to manage the resource in two domains in order to overcome the challenges to accomplish fast and efficient intelligence acquisition. Different from the goal of traditional communication systems in throughput maximization, centralized edge learning systems aim to maximize learning performance. With that said, classic resource management strategies in wireless communication

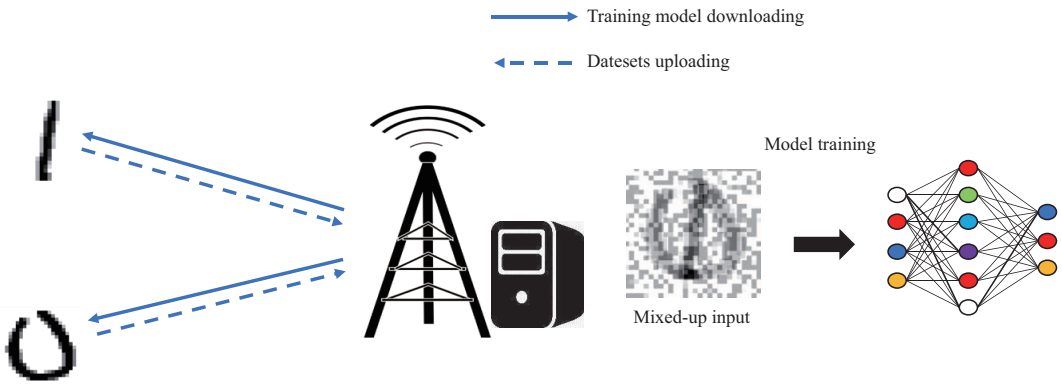

Figure 5 (Color online) Transmitting procedure in AirMixML.

literatures may not lead to the optimal learning performance as the key learning factors such as model and data complexity are not taken into account. For example, considering a support vector machine (SVM) and a convolutional neural network (CNN), they can achieve different learning accuracies with the same size of training data set. Furthermore, the communication cost of transferring a data sample in various ML tasks can vary significantly. As a result, new resource allocation algorithms are required for centralized edge learning.

There have been several prior studies proposing optimized resource allocation schemes for centralized edge learning in recent years. One efficient design is to schedule the joint communication and computation processes based on the data importance. Ref. [42] proposed a data-importance-aware user scheduling scheme for edge ML systems targeting SVM models, where the data are classified into different importance levels according to a derived data importance measure, so that more resources can be allocated to the data with higher importance. Furthermore, the design based on SVM in [42] was then extended to CNN, in which the authors considered the retransmission decision problem featuring quantity-quality tradeoff in training data in the presence of transmission data errors. Unlike the traditional automatic-repeat-request which focuses solely on reliability, the proposed scheme in [42] selectively retransmitted data samples based on their importance in order to accelerate the convergence speed of the training process. The other line of research focuses on the learning-centric wireless resource allocation. Ref. [43] proposed a nonlinear classification error model via data fitting for ML tasks. Based on this, a learning-centric power allocation scheme was proposed. Furthermore, Ref. [44] employed statistical learning to forecast the relationship between the learning accuracy of various tasks and the amount of training data. This thus yields a tractable learning performance to be maximized by using differential convex programming (DCP) and the derived optimal solution shows that the optimal transmission time is inversely proportional to the generalization error.

2.2 Mixup data augmentation with AirComp

AirComp is another integrated communication-and-computation design that can be exploited to enhance the efficiency of centralized edge learning, and particularly combined with the mixup data augmentation. Mixup is a well-known data augmentation technology [45], which in essence trains neural networks using convex combinations of data and labels, thus regularizing neural networks to make it easier to learn simple linear behaviors between training samples.

As shown in Figure 5, Ref. [46] combined mixup and AirComp, in which an ML framework called over-the-air mixup ML (AirMixML) is proposed to take advantage of the natural distortions and superpositions properties of wireless channels. Multiple users in AirMixML send analog modulated signals of their private data samples to a central server, which uses the received aggregated noisy samples to train ML models, while protecting the user's privacy. AirMixML was shown to achieve the same accuracy as the learning using raw data samples, but with much stronger privacy protection. Specifically, AirMixML adjusts the privacy disclosure level of the transmitted signal by controlling the user's transmit power, and the Dirichlet dispersion ratio controls the local power contribution of each worker to the superimposed signal. Despite the pioneering contribution of AirMixML as the first privacy-preserving centralized edge ML framework exploiting over-the-air signal superposition and additional channel noise, there is still much room for improvement before practical application.

2.3 Centralized edge learning with ISAC

Sensing and communication are carried out sequentially in traditional communication systems. Nevertheless, in ISAC systems, the sensing and communication are implemented with a shared signal waveform. This thus motivates many prior studies that proposed to accelerate the centralized edge learning by ISAC. In particular, ISAC allows a highly efficient use of wireless signals for simultaneous dataset generation and transmission. Nevertheless, additional interference between sensing and communication is introduced by ISAC, which needs to be dealt with before practical use.

As a pioneering work in this direction, Ref. [47] proposed a classification error minimization method for joint beamforming design and time allocation. After that, several follow-up studies appear. For example, in [48], the user conducted sensing and communication on the same spectrum using an MIMO array. In this work, the authors proposed a multi-objective optimization problem for jointly optimizing transmit precoding for sensing, communication, and allocating computation resources, by considering both beampattern design and energy consumption in multi-user MIMO radar processing. As another example, Ref. [49] studied the throughput maximization in a multi-user MEC system using a sense-then-offload protocol. Furthermore, Ref. [50] investigated a traffic-aware task offloading scheme in a vehicular network and proposed an offloading mechanism based on the sensed environment data. Ref. [51] proposed a brandnew sensory system with ISAC based on analog spike signal processing.

2.4 Research opportunities

Despite the research efforts discussed above for efficient centralized edge learning, there are still many unexplored territories yet, which are discussed as follows.

• **Secure data uploading.** The centralized edge learning architecture may suffer from the single point of failure issue, that is, the central server is prone to attack by malicious users who may upload forged or poisoned data to mislead the entire training process. Therefore, how to build a trustworthy mechanism to guard against the potential attack in the data uploading process is a critical issue that warrants further study.

• **Data-importance aware systems.** Due to the limited communication resources in a real communication system, it is often not possible to transmit all of the datasets, so it is necessary to selectively transmit the data in order to train the network more efficiently. Furthermore, the importance of data changes during the transmission process (for example, a certain large category of data is important until it is transmitted, but after a certain amount of data have been transmitted, the same large category of data has little impact on the training), and this aspect has not been considered in previous studies.

• **Task-oriented ISCC in centralized edge learning.** The various learning tasks in centralized edge learning frequently require the simultaneous support of sensing, communication, and computation functions. As a result, joint resource management for ISCC is required to improve the performance of centralized edge learning. There is still an unexplored territory for task-oriented ISCC targeting edge inference, warranting further investigation.

3 FEEL

FEEL is a machine learning paradigm where multiple edge devices collaborate in training a shared ML model, and each device's raw data is stored locally and not exchanged or transferred. From the networking perspective, FEEL can be divided into two classes, including centralized FEEL and decentralized FELL. Centralized FEEL is the most popular FEEL architecture, in which an edge server coordinates the ML model training among the edge devices. Unlike centralized FEEL, decentralized FELL is a network topology without any central server to coordinate the training process, in which all edge devices are connected in a peer-to-peer manner to perform the model training. In FEEL, the sensing, communication, and computation are three coupled processes for training an ML model. To begin with, the edge devices must sense the environment (e.g., by using the equipped radio sensors) in order to obtain data for training ML models. Second, the edge devices compute model updates using local computing power. Finally, the edge devices upload local model updates to the edge server via the uplink channel, and the edge server broadcasts the global model to each edge device via the downlink channel. Wireless communication, sensing, and computation all have different effects on FEEL. This thus prompts the necessity for a joint

design of the three processes especially under the stringent constraints on the on-device resources (e.g., bandwidth and energy), as surveyed in the sequel.

3.1 Joint communication and computation resource management in FEEL

In FEEL, the heterogeneity of edge devices in terms of radio sources, channel status, and computational capabilities is a common issue that has a direct impact on the ultimate learning performance such as accuracy and latency. This thus calls for joint communication and computation resource allocation for FEEL performance optimization. Specifically, the mentioned device heterogeneity essentially leads to distinct uploading time between different devices and edge servers. As a result, if the edge server uses synchronized aggregation of the model updates of the edge devices, the edge device with the longest delay dominates the communication time of a single round. To address this issue, Ref. [52] investigated the optimal scheduling scheme for edge devices to minimize the training time of FEEL, but the communication resource optimization is not involved. In [53], the heterogeneous channel conditions of edge devices are also considered, in which the objective is to maximize the number of participating edge devices, by optimizing the scheduling scheme of edge devices and the allocation of communication resources including transmit power and bandwidth. In addition to the heterogeneity in channel conditions, the edge devices tend to be heterogeneous in computing capabilities as well. For instance, both Refs. [54, 55] comprehensively considered the heterogeneity of edge devices in terms of channel conditions and computing capabilities, and studied the optimal scheduling scheme for edge devices and the optimal communication resource allocation. The difference between [54,55] is that the considered problem in [54] only focuses on a single communication round and each communication round is treated equally. In contrast, Ref. [55] explicitly took the importance of different communication rounds into consideration and investigated the bandwidth allocation and edge device scheduling problems under long-term energy constraints. Moreover, imperfect wireless channel conditions are also investigated in [55].

On the other hand, edge devices need to utilize local computing resources for model updates. Research on computing resource management mainly focuses on two directions: optimization of the CPU/GPU frequency of edge devices [56,57] and optimization of the batch size used in model updates (e.g., stochastic gradient descent) [58, 59]. Since edge devices are usually heterogeneous in terms of computing power, the time for completing training and the energy consumed by FEEL can be largely wasteful if the computing power of different edge devices is not reasonably tuned. Both Refs. [56, 57] consider the total energy consumption of the system and the required training time at the same time in the optimization objectives to optimize the CPU frequency of different edge devices, as well as system variables such as communication resources and device scheduling. For the case where the edge devices cannot effectively adjust the computing frequency, Ref. [58] optimized the batch size of different devices to align the communication delay between different edge devices and the edge server, thus reducing the training time for FEEL. Unlike [58], which only considers the optimization within a single communication round, Ref. [59] focused on the whole FEEL training process, where the authors considered a long-time dynamic resource optimization problem, and a scheme based on Lyapunov optimization is proposed to jointly optimize the computing frequency and the batch size of each edge device in different communication rounds.

Besides the heterogeneity of communication and computation, the data heterogeneity is also typical in federated learning, as the data collected at different devices usually have different distributions depending on, e.g., the application scenarios, locations, and user behaviors. Some prior studies tended to manage the communication and computation resources in FEEL, by considering the effects of data heterogeneity. For example, in [60], the optimal client sampling strategy that tackles both system and data heterogeneity is designed to minimize the training time with convergence guarantee in a FEEL system with resource-constrained devices. Ref. [61] considered quantized FEEL with data heterogeneity, and jointly optimized the quantization level and the bandwidth allocation to minimize the training time.

3.2 Over-the-air FEEL (Air-FEEL)

Air-FEEL has emerged as a promising solution for communication-efficient edge AI [62]. As shown in Figure 6, over-the-air model/gradient aggregation is used in Air-FEEL to improve spectral efficiency of FEEL. It is shown in [63] that, compared with the conventional orthogonal multi-access, Air-FEEL can reduce the communication latency by a factor approximately equal to the number of devices without significant loss of the learning accuracy. Various research efforts have been spent on different directions in

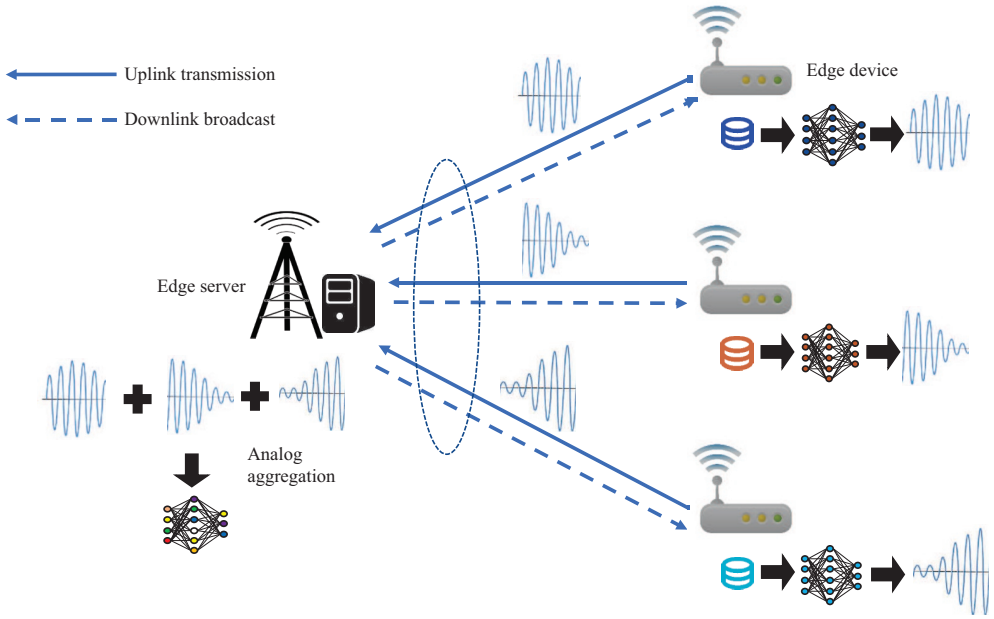

Figure 6 (Color online) Illustration of the Air-FEEL system.

Air-FEEL, such as resource management, device scheduling, and privacy preserving schemes as introduced as follows.

In Air-FEEL, edge devices can control their transmit power adaptively to reduce aggregation error for model/gradient aggregation, and thus improve learning accuracy or convergence rate. Ref. [64] developed the optimized power control to minimize the AirComp MSE, which exhibits a threshold-based structure depending on the channel conditions at different edge devices. Rather than minimizing the aggregation MSE in each round independently, Ref. [20,65] optimized power allocation across multiple global rounds to accelerate the convergence rate. Device scheduling is another efficient technique for improving Air-FEEL performance via addressing resource heterogeneity by dropping edge devices with poor communication and computation conditions. Ref. [66] presented a joint design of device scheduling and receiving beamforming in which the edge devices with weak signal strengths after receiving beamforming were dropped from the training process. Furthermore, Ref. [67] investigated device scheduling by taking into account their diverse energy constraints and computation capabilities, and an energy-aware dynamic device scheduling algorithm based on Lyapunov optimization was proposed. It is noteworthy that, in addition to the benefit of reducing multiple-access latency, Air-FEEL offers an additional advantage in improving data privacy. Although the original FEEL algorithm came with a certain level of privacy protection due to the avoidance of raw data transmission, local training data can still be inferred, to some extent, from the local model updates by modern model inversion techniques [68]. As a fix, Air-FEEL limits eavesdroppers' access to the aggregated updates, hiding each individual local update in the sea of others. A further free mask that can be used to safeguard the privacy of the data is the random disturbance that the channel noise imposes on the aggregated updates [69]. A more comprehensive overview on Air-FEEL can refer to [62].

Although AirComp is beneficial for model aggregation in Air-FEEL due to the inherent superposition property of wireless channels, Air-FEEL also suffers from the straggler issue in which the device with the weakest channel acts as a bottleneck of the model aggregation performance. To address this issue, Ref. [70] leveraged the RIS technology in Air-FEEL, and a unified communication-learning optimization problem is solved to jointly optimize device selection, over-the-air transceiver design, and RIS configuration. The aforementioned studies all focused on the centralized FEEL, in which a central edge server is required to orchestrate the training process. Ref. [71,72] considered decentralized FEEL in the scenario where an edge server is not available or reliable, in which the authors considered the precoding and decoding strategies for device-to-device communication-enabled model/gradient aggregation and proposed an AirComp-based decentralized stochastic gradient decent with gradient tracking and variance reduction algorithm to reach the consensus.

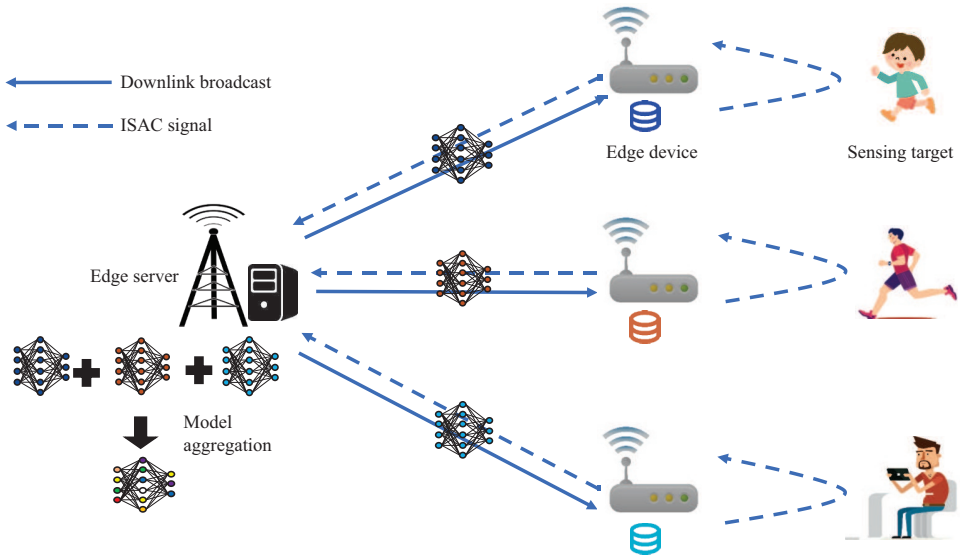

Figure 7 (Color online) Illustration of the FEEL system with ISAC.

3.3 FEEL with ISAC

FEEL also be exploited to train AI models in wireless sensing networks, and in this case, ISAC can be integrated to enhance the data acquisition and uploading processes, as shown in Figure 7. However, it is still an open problem to design edge learning systems with sensing and communication coexistence [27].

In ISAC, sensing and communication can be integrated in three different levels. At the first level, the sensing and communication signals may occupy orthogonal time-frequency resources, which do not interfere with each other functionally but compete for time-frequency resources, so the reasonable resource allocation between sensing and communication is a key issue when communication and sensing orthogonally coexist. For example, in [73], the sensing and communication work in a time-division manner, so that sensing and communication jointly compete for time resources. Based on this, a mathematical relationship between sensing time and learning performance is experimentally fitted and the trade-off between learning performance and communication rate is analyzed in [73]. Ref. [74] considered horizontal FEEL with ISAC and jointly optimized the sensing, communication, and computation resources, where sensing and communication occupy the same frequency band and different time durations, similarly in [73]. At the second integration level, sensing and communication may work on the same time-frequency resources but use separated signal waveforms, and they will interfere with each other, so how to manage interference becomes particularly important. In [75], the relationship between learning performance and sensing/communication resources is obtained by considering sensing and communication working on the same time-frequency resources and linking the learning performance to the quantity of sensed samples. Since sensing and communication may interfere with each other spatially, in [75], the beam directions of the perception and communication signals are optimized with the goal of maximizing the learning performance. In the third level, sensing and communication functions are simultaneously implemented using a shared signal waveform, which can effectively improve the system spectrum efficiency, hardware efficiency, and information processing efficiency [27, 76, 77]. Ref. [78] was the first to combine level-three ISAC with over-the-air analog aggregation technology and jointly designed beamforming for both sensing and communication signals, laying the foundation for the subsequent application of ISAC to FEEL. Unleashing the full potential of FEEL by combining with ISAC has very significant application prospects. In order to save frequency resources, the same ISAC signal is adopted for both sensing and communication in [79], and a cooperative sensing scheme based on vertical FEEL is proposed to enhance the sensing performance.

3.4 Research opportunities

Despite the research efforts discussed above for efficient FEEL, here lists some unexplored problems and challenges to motivate future studies.

• **Multi-modal data sources.** In real-world sensing systems, the edge devices involved in training may have different types of sensors, such as radio sensors or cameras, and the sensed data may have different modalities [80]. In such cases, a FEEL system with ISCC for multimodal data needs to be designed and optimized.

• **Dynamic sensing environments.** Most of the current studies consider static sensing environments, but in real scenarios the sensing environment may keep changing over time and the distribution of the sensed data samples will no longer be stationary. How to design a FEEL system with ISCC for dynamic sensing environments is also a topic worthy of in-depth study.

• **Task-oriented ISCC in FEEL.** Few current studies on FEEL have considered specific sensing processes, mostly focusing on resource optimization for communication and computation. Various types of learning tasks in FEEL often require the support of sensing, communication, and computation functions at the same time, resulting in a variety of complex relationships such as coupling, collaboration, and even competition among the three mentioned modules. Therefore, in order to improve the performance limit of FEEL, joint resource management of sensing, communication, and computation is required.

4 Edge inference

Apart from the training phase discussed above, edge inference is another important aspect for supporting the successful implementation of AI technologies at wireless edge [81]. Specifically, for edge inference, a well-trained ML model needs to be deployed at the network edge to run AI tasks in real time (such as classification, recommendation, and regression), which is beneficial to computation/storage/power-limited edge devices and delay-sensitive AI tasks [82]. Thus, edge inference has become an important technique to enable various AI applications, such as metaverse, auto-driving, and smart cities in 6G networks. For example, in auto-driving, the vehicles need to detect obstacles to avoid accidents under stringent latency constraints. To guarantee high detection accuracy, more sophisticated deep neural network (DNN) architectures are preferable for the detection, i.e., ResNet-50 [83]. However, ResNet-50 contains 50 convolutional layers, and demands nearly 100 megabytes of memory for storage. On one hand, it is non-trivial to deploy such complicated DNN on edge devices with limited computation and storage capacity. However, deploying it merely on the cloud server may induce intolerable delay and increase the risk of accidents. To deal with such a dilemma, edge inference provides a promising solution with a better trade-off among computation power, storage capacity, and communication latency.

There are three different methods to implement edge inference, namely, on-device inference [84], on-server inference [85, 86], and split inference [87–100]. For on-device inference, the computation is accomplished merely on edge devices, which is non-trivial for recent increasingly complex AI models and computation/storage/power-limited edge devices. To tackle such issue, the on-server inference is designed. However, on-server inference suffers from the communication bottleneck due to the potential high-volume data transmission over the band-limited wireless channels in the presence of uncertain channel fading, under the stringent low-latency constraint. Also, the computation resources at edge servers may still fall short when running some large-scale AI tasks. Nevertheless, potential information leakage during data uploading from the edge devices to the edge server may lead to privacy issues in edge inference. To tackle these problems, the split inference is proposed jointly considering techniques such as joint source and channel coding (JSCC) [87–92, 101], joint communication and computation resource management design [93–97], and AirComp [98, 99]. Furthermore, as a recently proposed technique, ISAC has drawn increasing research interests [27]. Intuitively, ISAC can further reduce the latency of edge inference due to the integrated data sensing and uploading processes [100]. The joint management of sensing, communication, and computation resources in this case is more difficult. In the following, we discuss the above techniques in detail.

4.1 JSCC in edge inference

Generally, when uploading features from the devices to the server to perform inference, the fluctuating wireless channels may introduce lossy transmission, which calls for proper design of JSCC for efficient feature transmission. Moreover, in some sense, JSCC can be viewed as a novel design principle incorporating both communication and computation into a joint design. With the recent development of DL, deep JSCC has been widely investigated to alleviate excessive signaling overhead as well as improve the

robustness to channel distortion. For example, for classification tasks, Refs. [87, 88] proposed a retrieval-oriented wireless image transmission framework to maximize the classification accuracy, where the JSCC framework is trained by the cross-entropy between the predictions and the ground-truth labels. Moreover, information bottleneck (IB) [101] was proposed to extract minimum features to fulfill certain tasks sufficiently. Under the guidance of IB, an image classification task was considered in [89] by designing a framework of task-oriented JSCC. By combining IB with stochastic optimization, the same task was considered in [90] to minimize energy consumption as well as service latency simultaneously. Considering edge inference with multiple edge devices, a task-oriented JSCC framework was designed in [91], where a group of edge nodes performs the classification task coordinated by an edge server. Furthermore, some initial exploration has been made in [92] to deal with the multi-modal data, where a task-oriented semantic communication scheme was proposed and the cross-entropy objective with multiuser multi-modal data fusion was considered.

4.2 Joint communication and computation resource management in edge inference

Edge inference typically involves local feature extraction and uploading to the edge server for further processing, which may encounter both communication and computation bottlenecks. Specifically, on one hand, transmitting data through wireless links naturally suffers from channel impairment, especially when the AI services have stringent low-latency requirements. On the other hand, processing data at devices and servers incurs computational delay, especially when a large number of devices require for AI services simultaneously. Furthermore, there always exists a trade-off between the inference accuracy and the computation-communication capability of the system in edge inference.

To address the above issues, joint management of communication and computation resources is considered in recent studies. For instance, in [93], the optimal control of inference accuracy and transmission cost was modeled as a Markov decision process (MDP), and their trade-off is balanced via dynamically selecting the optimal compression ratio with hard deadline requirements. By further considering the proper split of AI models in multi-user edge inference systems, Ref. [94] jointly designed the model split point selection and computational resource allocation to minimize the maximum inference latency. Also, Ref. [95] studied the trade-off between the computational cost of the on-device model and the communication overhead of uploading the extracted features to the edge server. Then they propose a three-step framework for inference, which contains model split point selection, communication-aware model compression, and task-oriented encoding mechanism for the extracted features. Due to the heavy computational burden being offloaded on the server, it is also urgent to study the computation and inference accuracy trade-off at the server side. Ref. [96] considered the early exiting technique, which allows a task exit from certain layers of a DNN without traversing the whole network. In such a way, the joint management of communication and computation resources could reduce the inference latency under various accuracy requirements. Moreover, a progressive feature transmission protocol was proposed in [97], which contains importance-aware feature selection and transmission-termination control. In such a protocol, the devices transmit the extracted features progressively according to their importance, and once the inference accuracy requirement is obtained, the transmission will stop. With such a design, the trade-off between inference accuracy and communication-computation latency can be well balanced. Moreover, RIS has recently emerged as a potential solution to provide a cost-effective way for enhancing the performance of edge inference. For example, a RIS-aided green edge inference system was considered in [86], where the set of tasks performed by each BS, uplink/downlink beamforming vectors of BSs, transmit power of edge devices, and uplink/downlink phase-shift matrices at the RIS were jointly designed to minimize the overall network power consumption.

4.3 Over-the-air edge inference

AirComp is also appealing for low-latency edge inference by seamlessly integrating communication and computation, as shown in Figure 8. The research on over-the-air edge inference is still in its early stage.

An initial study of the AirComp-based multi-device edge inference system was made in [98], where AirComp is utilized to aggregate multiple noisy feature observations of a common source to average out the feature noise for boosting the inference accuracy. The authors first characterized the influence of sensing and channel noise on inference accuracy by deriving a tractable surrogate performance metric called discriminant gain. Then the authors maximized the inference accuracy by jointly optimizing the transmit precoding and receiving beamforming. Besides the significantly enhanced spectrum efficiency, AirComp

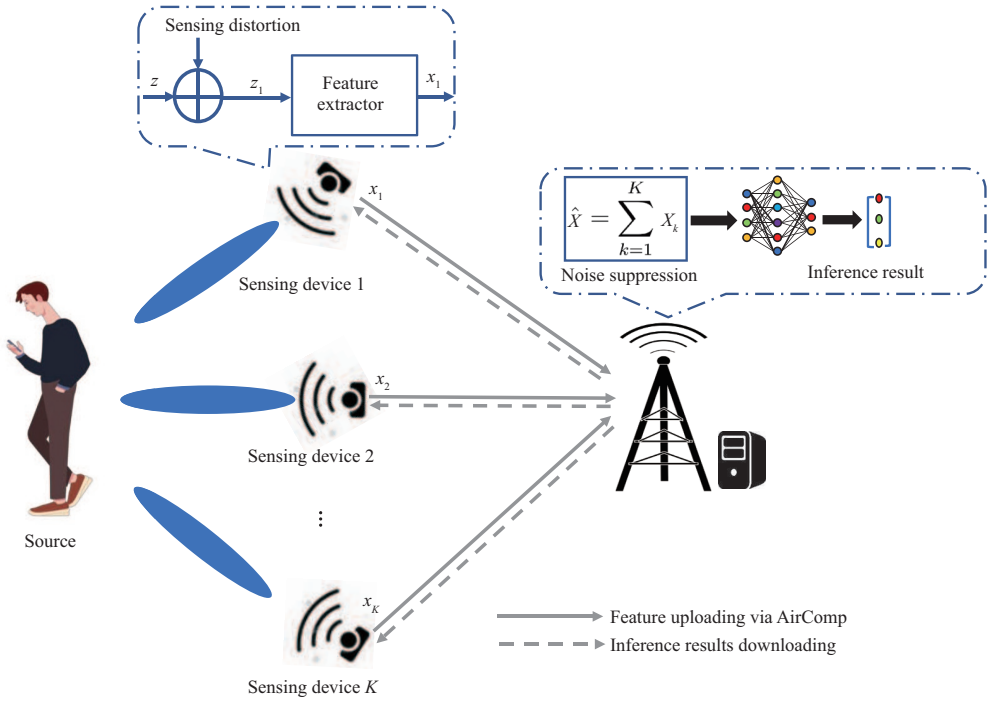

Figure 8 (Color online) Illustration of AirComp-based edge inference systems.

has additional benefit in privacy preservation. To exploit such property, Ref. [84] considered an ensemble inference framework, where each device needs to transmit its predictions to the server for further fusion to obtain final results. Apart from maximizing inference accuracy, the authors also considered maximizing the privacy of the on-device models. To this end, AirComp is exploited for privacy-enhanced outcome fusion as each individual predictive outcome is hidden in the crowd. Specifically, the authors introduced different ensemble methods, such as belief summation and majority voting, and provided privacy analysis for these AirComp-based fusion schemes. Finally, numerical results provided in [84] have shown that the proposed AirComp-based solution significantly outperforms other orthogonal transmission schemes in terms of the required communication overhead under the same target privacy guarantee. Moreover, Ref. [99] also considered the privacy issues in edge inference. Specifically, they consider the distributed inference of graph neural networks (GNNs). To deal with the possible privacy leakage problem arising from the devices exchanging information with neighbors during decentralized inference, the authors first characterized the privacy performance of the considered decentralized inference system. Then they design privacy-preserving signals and the corresponding training algorithms in combination with AirComp to further boost the privacy of the considered system.

4.4 Co-inference with ISAC

In future wireless networks, to support environment-aware intelligent applications, it is desirable to process and upload the collected data from sensing devices for inference, where sensing, communication, and computation are naturally coupled and need to be jointly designed. However, integrating ISAC with edge inference introduces several issues. First, it is non-trivial to characterize the inference performance in ISAC-enabled edge inference. Second, how to jointly design the resources for sensing, communication, and computation to proper balance the trade-off among them is challenging. To deal with the above issues, Ref. [100] studied a task-oriented ISCC-based edge inference system as shown in Figure 9, where multiple ISAC devices collect sensing data, and then upload the quantized features to the server for classification. The authors analyzed inference performance in such an ISCC-based system via deriving the tractable measure for the inference accuracy called discriminant gain, based on which, the allocation of sensing, transmit power, communication time, and quantization bits is jointly designed for the successful completion of the subsequent classification task. Finally, some interesting design insights for balancing the trade-off between sensing, communication, and computation were crystalized in [100]: the sensing power and quantization bits should be enlarged as the number of classes increases in the classification

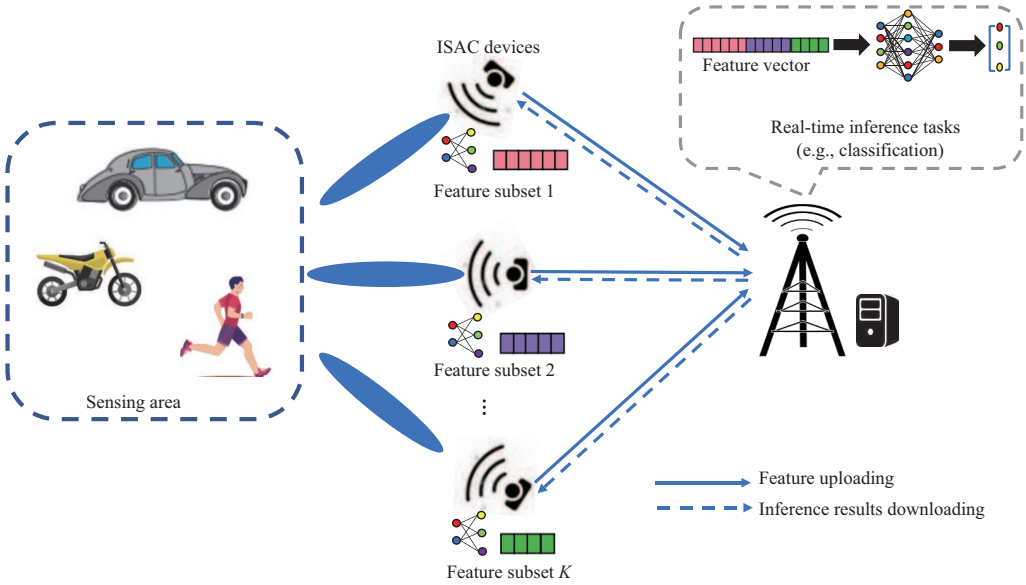

Figure 9 (Color online) Edge inference systems with multi-device sensing.

task; otherwise, more communication power should be allocated if the channel conditions of the devices are poor.

4.5 Research opportunities

Despite the research efforts discussed above for efficient edge inference, there are still many open problems and challenges unexplored yet.

- **Fundamental limits of JSCC.** Similar to Shannon's information theory, the fundamental limits of semantic-based JSCC transmission for inference need to be characterized. Moreover, it is also interesting to explore the number of optimized symbols for the successful completion of certain tasks via JSCC transmission, which can balance the delay and accuracy trade-off in JSCC-based edge inference.
- **Edge inference with multiple devices and/or multiple servers.** For edge inference systems with multiple devices and/or multiple servers, the device selection and server coordination need to be further considered. On one hand, the selection of devices for inference needs to account for the trade-off between delay and accuracy. On the other hand, a multi-server system requires flexible model deployment for large-scale AI tasks, such as heterogeneous tasks with different models requirements.
- **Task-oriented ISCC for edge inference.** In edge inference, sensing, communication, and computation may compete for resources (such as radio and hardware). How to depict the relationship between the inference performance and all the three mentioned processes is quite challenging, thus yielding the non-trivial problem of joint management of sensing, communication, and computation resources. Although Ref. [100] did an initial study in this direction, there still remain many uncharted issues warranting further investigation. Moreover, to fully exploit edge intelligence, multi-modality sensors (such as laser radar, millimeter-wave radar, and cameras) may be deployed at the wireless edge. How to process the acquired multi-modal sensing data for efficient inference is also an interesting direction to pursue.

5 Concluding remarks

In the upcoming 6G era, we will witness a paradigm shift in network functions from connecting people and things to connecting intelligence, driving the advancement of IoT in 5G to AIoT. Therefore, with the help of 6G networks, AI is expected to spread from the cloud to the network edge to provide ubiquitous AI services. However, traditional design principles of separating sensing, communication, and computation cannot meet stringent requirements for latency, reliability, and capacity.

To tackle this issue, this study presents a timely literature survey on tasked-oriented ISCC for edge intelligence. First, we introduce the motivation and basic principles of ISCC. Then, we introduce representative studies on three different scenarios, i.e., centralized edge learning, FEEL, and edge inference,

respectively, by focusing on joint communication and computing resource management, AirComp, and ISAC in each scenario. Finally, interesting research directions are proposed to motivate future work. We hope that this study can provide new insights into this interesting research topic and motivate more interdisciplinary research connecting wireless sensing, wireless communication, machine learning, and computing.

Acknowledgements The work was supported in part by National Key R&D Program of China (Grant No. 2018YFB1800800), Basic Research Project of Hetao Shenzhen-HK S&T Cooperation Zone (Grant No. HZQB-KCZYZ-2021067), National Natural Science Foundation of China (Grant Nos. U2001208, 61871137, 62001310), Science and Technology Program of Guangdong Province (Grant No. 2021A0505030002), Shenzhen Fundamental Research Program (Grant No. 20210318123512002), Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022A1515010109), and Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055).

References

1 You X H, Wang C-X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301
2 Letaief K B, Shi Y, Lu J, et al. Edge artificial intelligence for 6G: vision, enabling technologies, and applications. IEEE J Sel Areas Commun, 2022, 40: 5–36
3 Feng Z, Wei Z, Chen X, et al. Joint communication, sensing, and computation enabled 6G intelligent machine system. IEEE Network, 2021, 35: 34–42
4 Letaief K B, Chen W, Shi Y, et al. The roadmap to 6G: AI empowered wireless networks. IEEE Commun Mag, 2019, 57: 84–90
5 Shen X, Gao J, Wu W, et al. Holistic network virtualization and pervasive network intelligence for 6G. IEEE Commun Surv Tutorials, 2022, 24: 1–30
6 Ye H, Li G Y, Juang B H. Power of deep learning for channel estimation and signal detection in OFDM systems. IEEE Wireless Commun Lett, 2018, 7: 114–117
7 Cisco. Cisco Annual Internet Report (2018–2023). white-paper-c11-741490. 2020
8 Huawei Technologiy. Communications network 2030. 2022
9 Cisco. Cisco Global Cloud Index: Forecast and Methodology, 2016–2021. white-paper-c11-738085. 2018
10 Shi W, Cao J, Zhang Q, et al. Edge computing: vision and challenges. IEEE Internet Things J, 2016, 3: 637–646
11 Zhou Z, Chen X, Li E, et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE, 2019, 107: 1738–1762
12 Park J, Samarakoon S, Bennis M, et al. Wireless network intelligence at the edge. Proc IEEE, 2019, 107: 2204–2239
13 He Y, Yu G, Cai Y, et al. Integrated sensing, computation, and communication: system framework and performance optimization. 2022. ArXiv:2211.04022
14 Chen M, Liang B, Dong M. Joint offloading decision and resource allocation for multi-user multi-task mobile cloud. In: Proceedings of the 2016 IEEE International Conference on Communications (ICC), 2016. 1–6
15 Hoang D, Niyato D, Wang P. Optimal admission control policy for mobile cloud computing hotspot with cloudlet. In: Proceedings of the 2012 IEEE Wireless Communications and Networking Conference (WCNC), 2012. 3145–3149
16 Mao Y, Zhang J, Song S, et al. Power-delay tradeoff in multi-user mobile-edge computing systems. In: Proceedings of 2016 IEEE Global Communications Conference (GLOBECOM), 2016. 1–6
17 Wang F, Xu J, Wang X, et al. Joint offloading and computing optimization in wireless powered mobile-edge computing systems. IEEE Trans Wireless Commun, 2018, 17: 1784–1797
18 Cao X, Wang F, Xu J, et al. Joint computation and communication cooperation for energy-efficient mobile edge computing. IEEE Internet Things J, 2019, 6: 4188–4200
19 Zhu G, Xu J, Huang K, et al. Over-the-air computing for wireless data aggregation in massive IoT. IEEE Wireless Commun, 2021, 28: 57–65
20 Cao X, Zhu G, Xu J, et al. Optimized power control design for over-the-air federated edge learning. IEEE J Sel Areas Commun, 2022, 40: 342–358
21 Liu W, Zang X, Li Y, et al. Over-the-air computation systems: optimization, analysis and scaling laws. IEEE Trans Wireless Commun, 2020, 19: 5488–5502
22 Cao X, Zhu G, Xu J, et al. Optimized power control for over-the-air computation in fading channels. IEEE Trans Wireless Commun, 2020, 19: 7498–7513
23 Zhu G, Huang K. MIMO over-the-air computation for high-mobility multimodal sensing. IEEE Internet Things J, 2019, 6: 6089–6103
24 Fang W, Jiang Y, Shi Y, et al. Over-the-air computation via reconfigurable intelligent surface. IEEE Trans Commun, 2021, 69: 8612–8626
25 Zhang W, Xu J, Xu W, et al. Worst-case design for RIS-aided over-the-air computation with imperfect CSI. IEEE Commun Lett, 2022, 26: 2136–2140
26 Fu M, Zhou Y, Shi Y, et al. UAV aided over-the-air computation. IEEE Trans Wireless Commun, 2022, 21: 4909–4924
27 Liu F, Cui Y, Masouros C, et al. Integrated sensing and communications: toward dual-functional wireless networks for 6G and beyond. IEEE J Sel Areas Commun, 2022, 40: 1728–1767
28 Liu F, Masouros C, Petropulu A P, et al. Joint radar and communication design: applications, state-of-the-art, and the road ahead. IEEE Trans Commun, 2020, 68: 3834–3862
29 Liu F, Zhou L, Masouros C, et al. Toward dual-functional radar-communication systems: optimal waveform design. IEEE Trans Signal Process, 2018, 66: 4264–4279
30 Liu X, Huang T, Shlezinger N, et al. Joint transmit beamforming for multiuser MIMO communications and MIMO radar. IEEE Trans Signal Process, 2020, 68: 3929–3944
31 Hua H, Xu J, Han T. Optimal transmit beamforming for integrated sensing and communication. 2021. ArXiv:2104.11871
32 Liu F, Liu Y F, Li A, et al. Cramér-Rao bound optimization for joint radar-communication beamforming. IEEE Trans Signal Process, 2022, 70: 240–253

33 Lyu Z, Zhu G, Xu J. Joint maneuver and beamforming design for UAV-enabled integrated sensing and communication. IEEE Trans Wireless Commun, 2022. doi: 10.1109/TWC.2022.3211533
34 Song X, Xu J, Liu F, et al. Intelligent reflecting surface enabled sensing: Cramér-Rao bound optimization. 2022. ArXiv:2207.05611
35 Song X, Zhao D, Hua H, et al. Joint transmit and reflective beamforming for IRS-assisted integrated sensing and communication. In: Proceedings of 2022 IEEE Wireless Communications and Networking Conference (WCNC), 2022. 189–194
36 Wang X, Fei Z, Huang J, et al. Joint waveform and discrete phase shift design for RIS-assisted integrated sensing and communication system under Cramér-Rao bound constraint. IEEE Trans Veh Technol, 2022, 71: 1004–1009
37 Shi W, Xu W, You X, et al. Intelligent reflection enabling technologies for integrated and green internet-of-everything beyond 5G: communication, sensing, and security. IEEE Wireless Commun, 2022. doi: 10.1109/MWC.018.2100717
38 Li X, Liu F, Zhou Z, et al. Integrated sensing and over-the-air computation: dual-functional MIMO beamforming design. In: Proceedings of the 1st International Conference on 6G Networking (6GNet), 2022, 1–8
39 Huang Q, Chen H, Zhang Q. Joint design of sensing and communication systems for smart homes. IEEE Network, 2020, 34: 191–197
40 Liu F, Yuan W, Masouros C, et al. Radar-assisted predictive beamforming for vehicular links: communication served by sensing. IEEE Trans Wireless Commun, 2020, 19: 7704–7719
41 Xu W, Yang Z, Yang D, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing. 2022. ArXiv:2206.00422
42 Liu D, Zhu G, Zhang J, et al. Data-importance aware user scheduling for communication-efficient edge machine learning. IEEE Trans Cogn Commun Netw, 2021, 7: 265–278
43 Wang S, Wu Y C, Xia M, et al. Machine intelligence at the edge with learning centric power allocation. IEEE Trans Wireless Commun, 2020, 19: 7293–7308
44 Zhou L, Hong Y, Wang S, et al. Learning centric wireless resource allocation for edge computing: algorithm and experiment. IEEE Trans Veh Technol, 2021, 70: 1035–1040
45 Zhang H, Cisse M, Dauphin Y, et al. mixup: Beyond empirical risk minimization. 2017. ArXiv:1710.09412
46 Koda Y, Park J, Bennis M, et al. AirMixML: over-the-air data mixup for inherently privacy-preserving edge machine learning. In: Proceedings of 2021 IEEE Global Communications Conference (GLOBECOM), 2021. 1–6
47 Zhang T, Wang S, Li G, et al. Accelerating edge intelligence via integrated sensing and communication. In: Proceedings of IEEE International Conference on Communications, 2022. 1586–1592
48 Ding C, Wang J B, Zhang H, et al. Joint MIMO precoding and computation resource allocation for dual-function radar and communication systems with mobile edge computing. IEEE J Sel Areas Commun, 2022, 40: 2085–2102
49 Liang Z, Chen H, Liu Y, et al. Data sensing and offloading in edge computing networks: TDMA or NOMA? IEEE Trans Wireless Commun, 2022, 21: 4497–4508
50 Qi Y, Zhou Y, Liu Y F, et al. Traffic-aware task offloading based on convergence of communication and sensing in vehicular edge computing. IEEE Internet Things J, 2021, 8: 17762–17777
51 Roth F, Bidoul N, Rosca T, et al. Spike-based sensing and communication for highly energy-efficient sensor edge nodes. In: Proceedings of the 2nd IEEE International Symposium on Joint Communications and Sensing (JCAS), 2022. 1–6
52 Luo B, Xio W, Wang S, et al. Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In: Proceedings of IEEE Conference on Computer Communications (INFOCOM), 2022. 1–10
53 Chen H, Huang S, Zhang D, et al. Federated learning over wireless IoT networks with optimized communication and resources. IEEE Internet Things J, 2022, 9: 16592–16605
54 Chen M, Yang Z, Saad W, et al. A joint learning and communications framework for federated learning over wireless networks. IEEE Trans Wireless Commun, 2021, 20: 269–283
55 Xu J, Wang H. Client selection and bandwidth allocation in wireless federated learning networks: a long-term perspective. IEEE Trans Wireless Commun, 2021, 20: 1188–1200
56 Nguyen V D, Sharma S K, Vu T X, et al. Efficient federated learning algorithm for resource allocation in wireless IoT networks. IEEE Internet Things J, 2020, 8: 3394–3409
57 Dinh C T, Tran N H, Nguyen M N H, et al. Federated learning over wireless networks: convergence analysis and resource allocation. IEEE ACM Trans Networking, 2021, 29: 398–409
58 Ma Z, Xu Y, Xu H, et al. Adaptive batch size for federated learning in resource-constrained edge computing. IEEE Trans Mobile Comput, 2023, 22: 37–53
59 Battiloro C, Lorenzo P D, Merluzzi M, et al. Lyapunov-based optimization of edge resources for energy-efficient adaptive federated learning. IEEE Trans Green Commun Netw, 2023. doi: 10.1109/TGCN.2022.3186879
60 Luo B, Xiao W, Wang S, et al. Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In: Proceedings of IEEE Conference on Computer Communications, 2022. 1739–1748
61 Liu P, Jiang J, Zhu G, et al. Training time minimization for federated edge learning with optimized gradient quantization and bandwidth allocation. Front Inform Technol Electron Eng, 2022, 23: 1247–1263
62 Cao X, Lyu Z, Zhu G, et al. An overview on over-the-air federated edge learning. 2022. ArXiv:2208.05643
63 Zhu G, Wang Y, Huang K. Broadband analog aggregation for low-latency federated edge learning. IEEE Trans Wireless Commun, 2019, 19: 491–506
64 Zhang N, Tao M. Gradient statistics aware power control for over-the-air federated learning. IEEE Trans Wireless Commun, 2021, 20: 5115–5128
65 Cao X, Zhu G, Xu J, et al. Transmission power control for over-the-air federated averaging at network edge. IEEE J Select Areas Commun, 2022, 40: 1571–1586
66 Yang K, Jiang T, Shi Y, et al. Federated learning via over-the-air computation. IEEE Trans Wireless Commun, 2020, 19: 2022–2035
67 Sun Y, Zhou S, Niu Z, et al. Dynamic scheduling for over-the-air federated edge learning with energy constraints. IEEE J Sel Areas Commun, 2022, 40: 227–242
68 Haque A, Milstein A, Fei-Fei L. Illuminating the dark spaces of healthcare with ambient intelligence. Nature, 2020, 585: 193–202
69 Liu D, Simeone O. Privacy for free: wireless federated learning via uncoded transmission with adaptive power control. IEEE J Sel Areas Commun, 2021, 39: 170–185
70 Liu H, Yuan X, Zhang Y J A. Reconfigurable intelligent surface enabled federated learning: a unified communication-learning design approach. IEEE Trans Wireless Commun, 2021, 20: 7595–7609

71 Shi Y, Zhou Y, Shi Y. Over-the-air decentralized federated learning. In: Proceedings of 2021 IEEE International Symposium on Information Theory (ISIT), 2021. 455–460
72 Ozfatura E, Rini S, Gündüz D. Decentralized SGD with over-the-air computation. In: Proceedings of IEEE Global Communications Conference, 2020. 1–6
73 Li G, Wang S, Li J, et al. Rethinking the tradeoff in integrated sensing and communication: recognition accuracy versus communication rate. 2021. ArXiv:2107.09621
74 Liu P, Zhu G, Wang S, et al. Toward ambient intelligence: federated edge learning with task-oriented sensing, computation, and communication integration. IEEE J Sel Sig Process, 2022. doi: 10.1109/JSTSP.2022.3226836
75 Zhang T, Wang S, Li G, et al. Accelerating edge intelligence via integrated sensing and communication. 2021. ArXiv:2107.09574
76 Cui Y, Liu F, Jing X, et al. Integrating sensing and communications for ubiquitous IoT: applications, trends, and challenges. IEEE Network, 2021, 35: 158–167
77 Liu A, Huang Z, Li M, et al. A survey on fundamental limits of integrated sensing and communication. IEEE Commun Surv Tut, 2022, 24: 994–1034
78 Li X, Liu F, Zhou Z, et al. Integrated sensing, communication, and computation over-the-air: MIMO beamforming design. 2022. ArXiv:2201.12581
79 Liu P, Zhu G, Jiang W, et al. Vertical federated edge learning with distributed integrated sensing and communication. IEEE Commun Lett, 2022, 26: 2091–2095
80 Guo J, Liu Q, Chen E. A deep reinforcement learning method for multimodal data fusion in action recognition. IEEE Signal Process Lett, 2022, 29: 120–124
81 Chen M, Gündüz D, Huang K, et al. Distributed learning in wireless networks: recent progress and future challenges. IEEE J Sel Areas Commun, 2021, 39: 3579–3605
82 Xu W, Yang Z, Ng D K W, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing. 2022. ArXiv:2206.00422
83 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
84 Yilmaz S F, Hasircioglu B, Gündüz D. Over-the-air ensemble inference with model privacy. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), 2022. 1265–1270
85 Yang K, Shi Y, Yu W, et al. Energy-efficient processing and robust wireless cooperative transmission for edge inference. IEEE Internet Things J, 2020, 7: 9456–9470
86 Hua S, Zhou Y, Yang K, et al. Reconfigurable intelligent surface for green edge inference. IEEE Trans Green Commun Netw, 2021, 5: 964–979
87 Jankowski M, Gündüz D, Mikolajczyk K. Deep joint source-channel coding for wireless image retrieval. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICSAAP), 2020. 5070–5074
88 Jankowski M, Gündüz D, Mikolajczyk K. Wireless image retrieval at the edge. IEEE J Sel Areas Commun, 2020, 39: 89–100
89 Shao J, Mao Y, Zhang J. Learning task-oriented communication for edge inference: an information bottleneck approach. IEEE J Sel Areas Commun, 2021, 40: 197–211
90 Pezone F, Barbarossa S, Lorenzo P D. Goal-oriented communication for edge learning based on the information bottleneck. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICSAAP), 2020. 8832–8836
91 Shao J, Mao Y, Zhang J. Task-oriented communication for multi-device cooperative edge inference. 2021. ArXiv:2109.00172
92 Xie H, Qin Z, Li G Y. Task-oriented multi-user semantic communications for VQA. IEEE Wireless Commun Lett, 2022, 11: 553–557
93 Huang X, Zhou S. Dynamic compression ratio selection for edge inference systems with hard deadlines. IEEE Internet Things J, 2020, 7: 8800–8810
94 Tang X, Chen X, Zeng L, et al. Joint multiuser DNN partitioning and computational resource allocation for collaborative edge intelligence. IEEE Internet Things J, 2021, 8: 9511–9522
95 Shao J, Zhang J. Communication-computation trade-off in resource-constrained edge inference. IEEE Commun Mag, 2020, 58: 20–26
96 Liu Z, Lan Q, Huang K. Resource allocation for multiuser edge inference with batching and early exiting. 2020. ArXiv:2204.05223
97 Lan Q, Zeng Q, Popovski P, et al. Progressive feature transmission for split inference at the wireless edge. 2021. ArXiv:2112.07244
98 Wen D, Jiao X, Liu P, et al. Task-oriented over-the-air computation for multi-device edge AI. 2022. ArXiv:2211.01255
99 Lee M, Yu G, Dai H. Privacy-preserving decentralized inference with graph neural networks in wireless networks. 2022. ArXiv:2208.06963
100 Wen D, Liu P, Zhu G, et al. Task-oriented sensing, computation, and communication integration for multi-device edge AI. 2022. ArXiv:2207.00969
101 Tishby N, Pereira F C, Bialek W. The information bottleneck method. 2000. ArXiv:0004057