Integrated Sensing-Communication-Computation for Over-the-Air Edge AI Inference

Zeming Zhuang, *Student Member, IEEE,* Dingzhu Wen, *Member, IEEE,* Yuanming Shi, *Senior Member, IEEE,* Guangxu Zhu, *Member, IEEE,* Sheng Wu, *Member, IEEE,* and Dusit Niyato, *Fellow, IEEE*

***Abstract*—Edge-device co-inference refers to deploying well-trained artificial intelligent (AI) models at the network edge under the cooperation of devices and edge servers for providing ambient intelligent services. For enhancing the utilization of limited network resources in edge-device co-inference tasks from a systematic view, we propose a task-oriented scheme of integrated sensing, computation and communication (ISCC) in this work. In this system, all devices sense a target from the same wide view to obtain homogeneous noise-corrupted sensory data, from which the local feature vectors are extracted. All local feature vectors are aggregated at the server using over-the-air computation (AirComp) in a broadband channel with the orthogonal-frequency-division-multiplexing technique for suppressing the sensing and channel noise. The aggregated denoised global feature vector is further input to a server-side AI model for completing the downstream inference task. A novel task-oriented design criterion, called maximum minimum pair-wise discriminant gain, is adopted for classification tasks. It extends the distance of the closest class pair in the feature space, leading to a balanced and enhanced inference accuracy. Under this criterion, a problem of joint sensing power assignment, transmit precoding and receive beamforming is formulated. The challenge lies in three aspects: the coupling between sensing and AirComp, the joint optimization of all feature dimensions' AirComp aggregation over a broadband channel, and the complicated form of the maximum minimum pair-wise discriminant gain. To solve this problem, a task-oriented ISCC scheme with AirComp is proposed. Experiments based on a human motion recognition task are conducted to verify the advantages of the proposed scheme over the existing scheme and a baseline.**

The work of Dingzhu Wen was supported by Shanghai Sailing Program under Grants No. 23YF1427400. The work of Yuanming Shi was supported in part by the Natural Science Foundation of Shanghai under Grant No. 21ZR1442700, the National Nature Science Foundation of China under Grant 62271318, and the Shanghai Rising-Star Program under Grant No. 22QA1406100. The work of Guangxu Zhu was supported by National Natural Science Foundation of China under Grant 62001310 and Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010109, the Internal Project Fund from Shenzhen Research Institute of Big Data under Grants J00120230001. The work of Sheng Wu was supported in part by the National Natural Science Foundation of China under Grant 62022019; and in part by the Open Foundation of State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, under Grant SKLNST-2021-1-17. The work of Dusit Niyato was supported by the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research & Development Programme, DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019), and Singapore MOE Tier 1 (RG87/22). *(Corresponding authors: D. Wen and Y. Shi)*

Z. Zhuang, D. Wen, and Y. Shi are with Network Intelligence Center, School of Information Science and Technology, ShanghaiTech University, Shanghai, China (e-mail: {zhuangzm, wendzh, shiym}@shanghaitech.edu.cn).

G. Zhu is with Shenzhen Research Institute of Big Data, Shenzhen, China (e-mail: gxzhu@sribd.cn).

S. Wu is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: thuraya@bupt.edu.cn).

D. Niyato is with School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: dniyato@ntu.edu.sg).

I. Introduction

The next generation of wireless technology (6G) will go far beyond just communication services to push forward an era of true Intelligence of Everything (IoE) for providing immersive intelligent services like auto-driving, Metaverse, smart city, etc. [1]–[6]. However, the realization of these services highly depends on utilizing the inference capability of well-trained AI models at the network edge for intelligent decision making. This gives rise to a new research topic called edge AI inference, or edge inference [7]–[10].

The implementation of edge inference includes three paradigms, i.e., on-device inference, on-server inference and edge-device co-inference. In on-device inference, well-trained AI models are downloaded by edge devices for executing inference tasks, leading to heavy computation overhead (see, [11]–[13]). To alleviate the computation bottleneck at devices, the on-server inference uploads the raw data samples from devices to an edge server, where large-scale AI models are deployed for inference (see, [14]–[16]). This, however, violates the data privacy of edge devices. To further address the privacy issue, the edge-device co-inference emerges as a promising solution (see, [17]–[20]). It divides an AI model into two parts. The front-end part has a smaller size and is deployed at devices for feature extraction. The computation-intensive back-end part is deployed at the server, which leverages the received local feature vectors to complete the remaining inference task. As a result, computation is offloaded to the edge server and the avoidance of raw data transmission keeps devices' data privacy. Hence, the edge-device co-inference paradigm is adopted in this work.

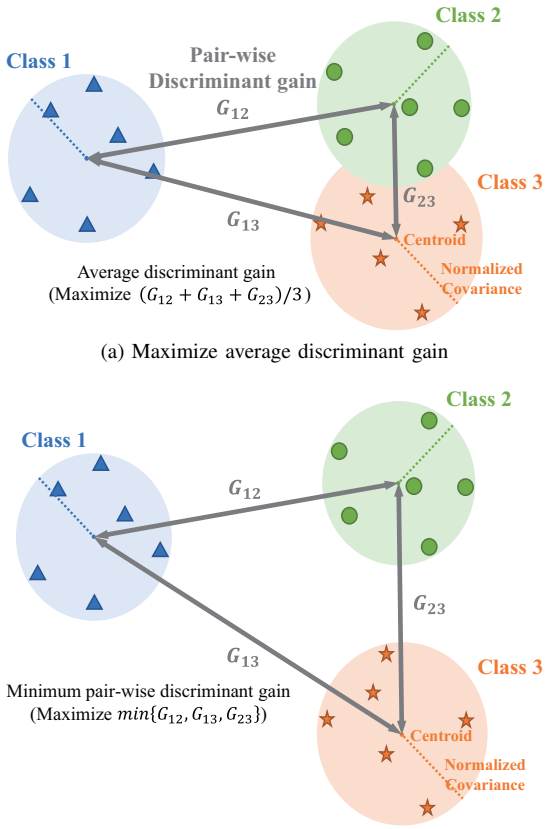
Recently, the edge-device co-inference has experienced a rapid advancement. The first research focus is to balance the trade-off between communication and computation. In [17], [21], the neural network was pruned at training phase to avoid the huge communication overhead caused by in-layer data amplification phenomenon. A suitable split layer selection method was developed in [22] together with the scheme for encoding/decoding the intermediate feature vector by an automated machine learning (AutoML) framework. Besides,

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3306465

methods of setting early exiting points in neural networks were proposed in [7], [23], [24] to balance the communication and computation overhead under a given empirical inference accuracy threshold. The authors in [19] further combined the methods of early exiting, model partitioning and data quantization to improve the inference performance. A joint source and channel coding (JSCC) approach was developed in [25] to map feature vectors into channel symbols. Nevertheless, as stated by [26]–[29], edge inference features a task-oriented property where the effectiveness and efficiency of the inference task execution are of crucial significance. As a result, the conventional design criteria including communication capacity or signal-to-noise ratio (SNR) of received signals work no longer well, as they cannot differentiate the feature elements with the same size and distortion level but different contributions on inference accuracy [28]. To address this issue, this work proposes to directly use the inference accuracy as the design criterion.

One main challenge of designing task-oriented schemes is that the instantaneous inference accuracy is unknown and has no mathematical model. To address this issue, the authors in [30] proposed an approximate but tractable metric, called discriminant gain. By considering classification tasks and based on the assumption that the feature vector follows a Gaussian mixture distribution with each Gaussian component corresponding to one class, a pair-wise discriminant gain for two arbitrary classes (called a class pair) is defined as the symmetric Kullback-Leibler (KL) divergence of their distributions. With a larger pair-wise discriminant gain, the two classes can be easily differentiated in the feature space, leading to an enhanced achievable inference accuracy. Existing works (see, [28], [30], [31]) use the average of all pair-wise discriminant gains as the design objective. This, however, causes an unbalanced inference accuracy of different classes and degrades the overall inference performance. As shown in Fig. 1(a), under this design goal, one particular class (i.e., Class 1) may be far separated from all other classes (i.e., Classes 2 and 3), which could be very close to each other in the feature space. To address this issue, in this work, we target maximizing the minimum pair-wise discriminant gain, which guarantees the closest class pair can be well separated in the feature space, as shown in Fig. 1(b).

On the other hand, although the previous works can enhance the inference performance, they optimize the edge-device co-inference systems from a partial view (i.e., the perspectives of communication or computation or both), which ignores the influence of the data acquisition process on inference performance and focuses on task offloading, model partitioning or data compressing (see, [19], [30], [32]). Also, many existing works on multi-device ISAC framework have been proposed and developed [33] (e.g., UAV deployment [34], data redundancy exploitation and sensing-communication switching [35]). However, they cannot achieve the full potential for enhancing the inference performance. As stated in [31], the fulfillment of an edge-device co-inference task requires the cooperation of sensing for data acquisition, computation for feature extraction and communication for feature transmission, at edge devices. The inference accuracy depends on the



(a) Maximize average discriminant gain

(b) Maximize minimum pair-wise discriminant gain

Fig. 1. Average discriminant gain maximization v.s. minimum pair-wise discriminant gain maximization.

feature distortion level caused during the data acquisition, computation and communication three processes. Besides, they compete for network resources including time and energy for suppressing their own distortion. Hence, edge-device co-inference calls for integrated sensing, communication and computation (ISCC) schemes [31]. To this end, a task-oriented scheme was proposed in [31] for maximizing the inference accuracy. However, the aforementioned work investigates the scenario of narrow-view sensing, which refers to that all devices perceive disjoint small ranges of a source target to obtain high-quality low-dimensional sensory data. There is a lack of ISCC schemes for handling the scenario of wide-view sensing, where each device perceives the same wide range of a source target and acquires noise-corrupted high-dimensional sensory data. To fill this gap, we propose a task-oriented scheme that integrates sensing and over-the-air computation (AirComp) for wide-view sensing based edge-device co-inference systems.

In this paper, a multi-device based ISCC system is considered to support edge-device co-inference tasks in many application scenarios such as ensuring security and reducing energy consumption in smart home (see, [36]), autonomous driving (see, [37]) and traffic monitoring in Vehicle-to-Everything (V2X) (see, [38]). Each device is equipped with a single antenna and a dual-functional-radar-communication (DFRC) transceiver used both for sensing and communication. First, all devices transmit a frequency modulation continuous wave

(FMCW) signal in an orthogonal frequency band to sense the same wide view of the source target for obtaining homogeneous sensory raw data. Then, a singular value decomposition (SVD) based linear filter is adopted for clutter cancellation and a principal component analysis (PCA) based extractor is exploited for extracting a low-dimensional local feature vector at each device. For further suppressing the sensing noise power and enhancing the communication efficiency, all local feature vectors are aggregated at the edge server via the technique of AirComp. Specifically, AirComp allows all devices simultaneously to transmit the same dimension of all local feature vectors over the same frequency band, leading to a significant enhancement of communication efficiency (see, [39]–[42]). By leveraging the waveform superposition property, a weighted sum of all local feature elements is directly calculated instead of decoding the value of each one individually. This work jointly considers the aggregation of all elements over an orthogonal frequency division multiplexing (OFDM) based broadband channel. Based on the novel design criterion called maximum minimum pair-wise discriminant gain, we propose the joint sensing power assignment, transmit precoding and receive beamforming problem. The challenges to solving this problem arise from three aspects: the coupling between sensing and AirComp, the joint optimization of all feature elements and the complicated form of the maximum minimum pair-wise discriminant gain. To address this problem, we propose the task-oriented ISCC scheme with AirComp. The detailed contributions of this work are summarized as follows.

- **Novel Design Metric of Maximum Minimum Pair-Wise Discriminant Gain**: To overcome the limitation of unbalanced and low inference accuracy resulting from the existing metric of average pair-wise discriminant gain (see, [28], [30], [31]), we adopt a novel design criterion called maximum minimum pair-wise discriminant gain in this work. It maximizes the discriminant gain between the closest class pair. Consequently, the least distinguishable class pair can be well separated in the feature space. This leads to a balanced and enhanced achievable inference accuracy.
- **AirComp based ISCC Framework for Edge-Device Co-Inference**: An AirComp based ISCC framework is established to complete edge-device co-inference tasks. The modules of sensing (including sensing waveform design and SVD based clutter cancellation), on-device computation (i.e., PCA based feature extraction) and AirComp (local feature vectors aggregation) are efficiently constructed. Particularly, an OFDM based broadband channel is used for the aggregation of all local feature vectors. Over an arbitrary frequency subcarrier, the same dimension of all local feature vectors is aggregated. The aggregation of different dimensions is over different subcarriers. The influences of each module on the design metric, i.e., minimum pair-wise discriminant gain, are mathematically characterized in closed-form expressions.
- **Task-Oriented ISCC Scheme with AirComp**: Under the criterion of maximum minimum pair-wise discriminant gain, we formulate the problem of joint sensing power assignment, transmit precoding and receive beamforming. We then propose the task-oriented ISCC scheme to address this problem, which first conducts variables transformation to derive an equivalent problem with a difference-of-convex (d.c.) form and then solves the d.c. problem based on the typical method of successive convex approximation (SCA) [43]. Compared with the existing AirComp based scheme in [28], where the optimization of different feature elements is separately designed and the sensing stage is not considered, the sensing, on-device computation and AirComp of all feature elements are jointly optimized in our proposed scheme. This provides two extra degrees of freedom to enhance the inference performance. On one hand, the system is optimized from a systematic view that coordinates the design of sensing, computation and communication by fully considering their coupling mechanism and competence in inference tasks. On the other hand, the joint design of all feature dimensions allows adaptive resource allocation among different feature dimensions, i.e., more resources can be assigned to the more important feature dimensions of the inference task.
- **Performance Evaluation**: Extensive experiments are performed to evaluate our proposed framework and algorithm based on the wireless sensing simulator proposed in [44]. A wide-view human motion recognition task is considered with two inference models: a multi-layer perception (MLP) neural network and a support vector machine (SVM) model. To begin with, the inference accuracy is shown to be monotonically increasing with the maximum minimum pair-wise discriminant gain, which verifies the efficiency of the adopted design criterion. Then, the proposed scheme is shown to outperform the state-of-the-art scheme and a baseline scheme.

The rest of this paper is organized as follows. Section II describes the proposed task-oriented AirComp based ISCC framework. Section III introduces the novel designed metric, presents and simplifies the minimax problem under long-term energy constraint. Section IV reformulates the problem with variable transformation and proposes an SCA based algorithm. The simulation results of the proposed scheme are shown in Section V by comparing with other baseline schemes. Finally, we conclude this paper in Section VI.

II. System Model and Problem Formulation

A. Network Model

Consider a single network to support edge-device co-inference tasks, as shown in Fig. 2. There is one edge server equipped with an N_r-antenna access point (AP) and K edge devices, each of which is equipped with a dual-functional-radar-communication (DFRC) system. Many types of radar are used for sensing in different scenarios including pulsed radar, continuous-wave radar, OFDM radar, OTFS radar, FMCW radar, etc [45]. Pulsed radar and continuous-wave radar are low-efficiency due to the avoidance of self-interference. The OFDM radar and OTFS radar suffer from co-channel interference from the communication systems [45], [46]. In the

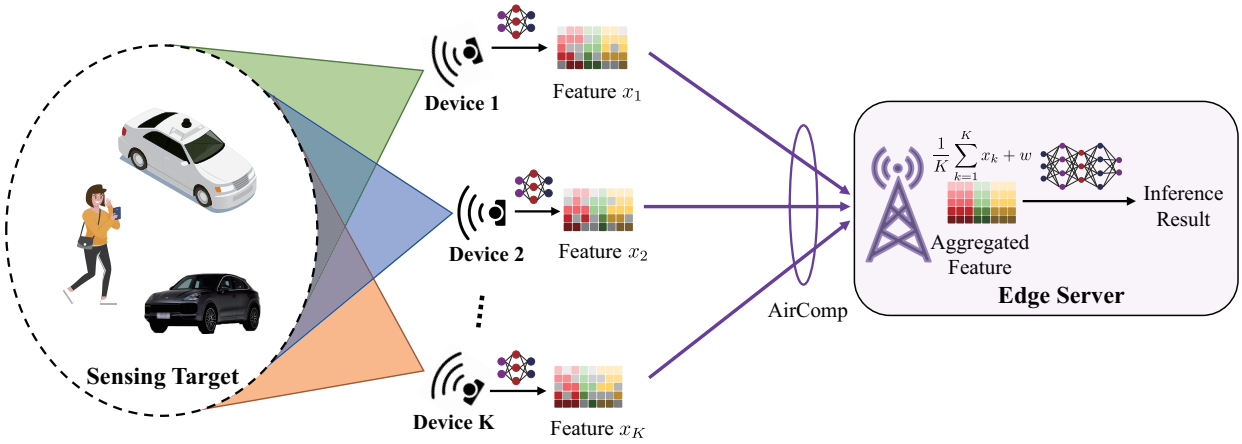


Fig. 2. The system architecture of proposed ISCC framework.

FMCW radar adopted in this paper, a dedicated frequency band is utilized for sensing and the frequency of the sensing signal is modulated as a linear function of time. As a result, there is no co-channel interference and self-interference [45], [46]. The workflow to complete an edge inference task is shown in Fig. 3. All devices perceive the same wide view of a source target and obtains homogeneous sensory data, from which the local feature vectors are extracted. The dimension of each local feature vector is denoted as M. The sensing frequency bands of different devices are orthogonal. Then, all local feature vectors are aggregated to derive a denoised global feature vector at the edge server using the technique of AirComp. Finally, the global feature vector is input into a server-side AI model to complete the whole inference task.

The sensing, computation and AirComp processes operate sequentially at all devices, as shown in Fig. 3. Particularly, to aggregate all feature elements using AirComp, OFDM is leveraged. M frequency subcarriers are used to aggregate all the M dimensions of the local feature vectors. Over each subcarrier, an element of the same feature dimension is transmitted by all devices and is aggregated at the edge server to get a global denoised one. As the time length of transmitting one feature element is much shorter than the channel coherence-time duration [47], static channels are assumed during one time slot. The edge server serves as a central coordinator and has the ability to acquire the channel state information (CSI) of all involved links.

B. Sensing Signal Processing and Feature Extraction

We adopt the models of sensing signal processing and feature extraction proposed in [31]. As shown in Fig. 3, during the radar sensing stage, each device transmits the FMCW signal of N up-ramp chirps for sensing. Each chirp has a time duration of $T_0 = T_s/N$ with T_s being the total sensing time. For device k, the sensing signal of one chirp is formulated as

$$c_{s,k}(t) = \text{rect}\left(\frac{t}{T_0}\right)\cdot\cos\left(2\pi f_{k,0}t + 2\pi\frac{B_s}{T_0}t^2\right),\ 1 \le k \le K, \tag{1}$$

where $\text{rect}(\cdot)$ is the rectangular pulse function with amplitude 1 and pulse length 1 centered at $t = 0$, $f_{k,0}$ is the starting frequency of sensing signal, B_s is the bandwidth of the sensing signal. It follows that the signal of the whole sensing duration is

$$s_k(t) = \sum_{n=0}^{N-1} c_{s,k}(t - nT_0), \tag{2}$$

$$\begin{aligned} s_k(t) = &\sum_{n=0}^{N-1} \text{rect}\left(\frac{t - nT_0}{T_0}\right) \\ &\times \cos\left(2\pi f_0(t - nT_0) + 2\pi\frac{B_s}{T_0}(t - nT_0)^2\right). \end{aligned} \tag{3}$$

Then the reflected signals from the direct and indirect paths are received by each device. The desirable echo signal is the one directly reflected from the target, given by

$$u_k(t) = H_{s,k}(t)s_k(t - \tau),\ 1 \le k \le K, \tag{4}$$

where $H_{s,k}(t)$ is the reflection matrix of the target including the round-trip path-loss, τ is the round-trip delay. The echo signal indirectly reflected through the j-th indirect path is

$$v_{k,j}(t) = C_{r,k,j}(t)s_k(t - \tau_j),\ 1 \le k \le K, \tag{5}$$

where $C_{s,k,j}(t)$ is the round-trip coefficient of path j, τ_j is the delay of the j-th path. Note that $H_{s,k}(t)$ and $\{C_{s,k,j}(t)\}$ can be pre-estimated by each device and fed back to the edge server before the inference task. Thereby, the received signal of ISAC device k is given by

$$r_k(t) = u_k(t) + \sum_{j=1}^{J} v_{k,j}(t) + n_r(t),\ 1 \le k \le K, \tag{6}$$

where $u_k(t)$ is the desired signal for completing the inference task, $\sum_{j=1}^{J} v_{k,j}(t)$ is the clutter of J indirect reflection paths and $n_r(t)$ is the white Gaussian noise. In (6), the useful signal $u_k(t)$ is polluted by the additive sensing clutter and noise. In the sequel, the clutter cancellation procedure is introduced.

1) Clutter cancellation: First, the received signal of device k is sampled at a frequency of f_s into a complex feature vector $\mathbf{r}_k \in \mathbb{C}^{NT_0f_s}$. The data sample vector $\mathbf{r}_k$ contains both the ranging and velocity information of the target. Thus, for deriving the information of sensing target, $\mathbf{r}_k$ is transformed into a complex matrix $\mathbf{R}_k \in \mathbb{C}^{T_0f_s \times N}$, the column dimension

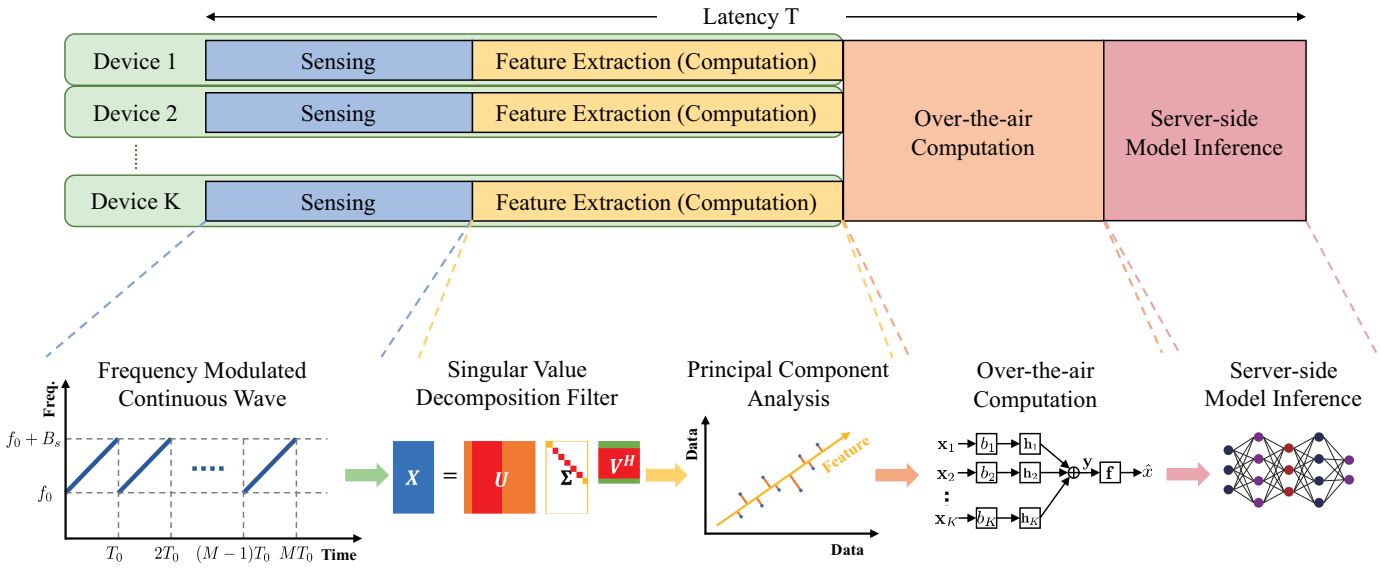


Fig. 3. The workflow for completing an inference task.

of which is usually used for ranging and the row dimension contains the feature in the Doppler spectrum shift. Each column of $\mathbf{R}_k$ represents the data samples in one chirp containing the distance information of the target and each row of $\mathbf{R}_k$ reflects the motion of the target among different chirps, where the velocity of the target can be extracted from the Doppler shift. Then, the SVD based linear filter proposed in [48] is utilized for clutter cancellation. To be specific, the SVD of $\mathbf{R}_k$ is

$$\mathbf{R}_k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H = \sum_{i=1}^{I} \mathbf{u}_i \sigma_i \mathbf{v}_i^H, \ 1 \le k \le K, \tag{7}$$

where $I = \min\{T_0 f_s, N\}$, $\mathbf{u}_i$, σ_i and $\mathbf{v}_i$ are the i-th left singular vector, singular value and right singular vector of $\mathbf{R}_k$, respectively, $\mathbf{V}^H$ is the conjugate transpose of $\mathbf{V}$. Clutter cancellation is performed by deleting the principal and least dimensions of $\mathbf{R}_k$. As a result, the data matrix after filtering is

$$\tilde{\mathbf{R}}_k = \sum_{i=r_1}^{r_2} \mathbf{u}_i \sigma_i \mathbf{v}_i^H, \ 1 \le k \le K, \tag{8}$$

where $1 \le r_1$ and $r_2 \le I$ are empirical parameters with respect to different kinds of radar sensors. Since only the information in row dimension, i.e., the Doppler spectrum shift, is needed for the inference task, $\tilde{\mathbf{R}}_k$ is compressed into a vector $\bar{\mathbf{r}}_k \in \mathbb{C}^N$. Its i-th element is given by

$$\bar{r}_k^i = \sum_{j=1}^{T_0 f_s} \tilde{R}_k^{j,i}, \ 1 \le k \le K, \tag{9}$$

where $\tilde{R}_k^{j,i}$ is the (j,i)-th element of matrix $\tilde{\mathbf{R}}_k$. Then the real part and the imaginary part of $\bar{r}_k$ is cascaded into a real vector $\tilde{\mathbf{r}}_k \in \mathbb{R}^{2N}$

$$\tilde{\mathbf{r}}_k = [\mathfrak{Re}(\bar{\mathbf{r}}_k), \mathfrak{Im}(\bar{\mathbf{r}}_k)] \tag{10}$$

2) Feature extraction: Following [28], [30], [31], the PCA based linear extractor is used to extract the local feature vector from clutter-cancelled sensory data $\tilde{\mathbf{r}}_k \in \mathbb{R}^{2N}$. The PCA is performed at the edge server before the inference task using the training dataset. Then, the template of the M principal eigen-subspace is broadcast to all devices for extracting the local feature vectors $\{\tilde{\mathbf{r}}_k \in \mathbb{R}^M\}$ with M being the number of extracted feature elements. Since the clutter cancellation and feature extraction processes are linear and based on (6), the m-th feature element of $\tilde{\mathbf{r}}_k$ is given by

$$\tilde{r}_k(m) = \tilde{u}_k(m) + \sum_{j=1}^{J} \tilde{v}_{k,j}(m) + n_r(m), \tag{11}$$

where $\tilde{u}_k(m)$ is the ground-truth of feature m, $\tilde{v}_{k,j}(m)$ is the clutter from path j, $n_r(m)$ is the noise in Gaussian distribution, defined by

$$n_r(m) \sim \mathcal{N}\left(0, \sigma_r^2\right), \ 1 \le m \le M. \tag{12}$$

Next, each feature element of device k is normalized by its sensing power $P_{s,k}$ and the normalized feature element m is given by

$$x_k(m) = \frac{\tilde{r}_k(m)}{\sqrt{P_{s,k}}} = x(m) + \tilde{c}_{s,k}(m) + \frac{n_r(m)}{\sqrt{P_{s,k}}}, \tag{13}$$

where $x(m) = \tilde{u}_k(m)/\sqrt{P_{s,k}}$ is the normalized ground-truth feature and

$$\tilde{c}_{s,k}(m) = \sum_{j=1}^{J} \frac{\tilde{v}_{k,j}(m)}{\sqrt{P_{s,k}}}, \ 1 \le m \le M, \ 1 \le k \le K, \tag{14}$$

is the normalized clutter. Since clutter is rich scattering and its number of paths J is very large, these individual clutter elements are assumed to be independent and identically distributed with finite variance. Thus $\tilde{c}_{s,k}(m)$ follows a Gaussian distribution according to the Central Limit Theorem (CLT), given by

$$\tilde{c}_{s,k}(m) \sim \mathcal{N}\left(\mu_{s,k}, \sigma_{s,k}^2\right), \ 1 \le m \le M, \ 1 \le k \le K, \tag{15}$$

where $\mu_{s,k}$ is the mean of clutter and can be pre-estimated and $\sigma_{s,k}^2$ is the clutter variance. Then the pre-estimated mean of $\tilde{c}_{s,k}(m)$ is eliminated to derive a zero-mean residual clutter element $c_{s,k}(m) = \tilde{c}_{s,k}(m) - \mu_{s,k}$. The CLT states that the sum or mean of a large number of independent and identically distributed random variables will approximate a Gaussian distribution, regardless of the shape of the original distribution, as long as the original variables have finite variance. Thereby, the local feature vector of device k can be written as

$$\mathbf{x}_k = \mathbf{x} + \mathbf{c}_{s,k} + \frac{\mathbf{n}_r}{\sqrt{P_{s,k}}},\ 1 \le k \le K, \tag{16}$$

where $\mathbf{x} = \{x(m)\}_{m=1}^M$, $\mathbf{c}_{s,k}(m) = \{c_{s,k}\}_{m=1}^M$ and $\mathbf{n}_r = \{n_r(m)\}_{m=1}^M$.

C. Feature Distribution

Consider a classification task with L classes. Following [28], [30], [31], the ground-truth feature vector $\mathbf{x}$ is assumed to follow a Gaussian mixture distribution. Since PCA is performed, different elements of ground-truth feature vector are independent. Consider an arbitrary element $x(m)$, its distribution is given as

$$f(x(m)) = \frac{1}{L}\sum_{\ell=1}^{L} f_\ell(x(m)),\ 1 \le m \le M, \tag{17}$$

where $f_\ell(x(m)) = \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2\right)$ is the probability density function of the Gaussian component corresponding to the ℓ-th class, $\mu_{\ell,m}$ is the centroid of class ℓ and σ_m^2 is the variance. These parameters are pre-estimated using the training dataset. Based on (17) and the clutter distribution in (15) and the noise distribution in (12), the distribution of the local feature element $x_k(m)$ can be derived as in the following lemma.

Lemma 1. *The distribution of local feature elements $x_k(m)$ can be derived as*

$$x_k(m) \sim \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2 + \sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right),\ 1 \le k \le K. \tag{18}$$

Proof. See Appendix A. □

D. Broadband Over-the-air Computation

In the edge-device co-inference system shown in Fig. 2. The edge server needs to aggregate all local feature vectors to obtain a global denoised one. If the conventional orthogonal multiple access technique such as TDMA is used, the consumed resource blocks linearly increase with the number of devices, leading to heavy communication overhead. To address this communication bottleneck, the technique of AirComp (see [39]–[42]) is adopted for the feature vector aggregation. As shown in Fig. 4, over the same subcarrier, it allows all devices simultaneously transmit the same feature dimension. At the server, the waveform superposition property is leveraged to directly derive a weighted sum of the elements from all devices. As a result, the communication overhead remains unchanged as the number of devices varies, leading to a significant enhancement of communication efficiency.

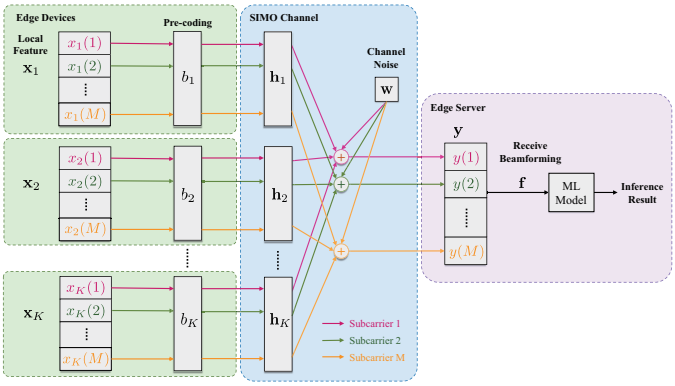


Fig. 4. The signal diagram of over-the-air computation with OFDM.

Specifically, consider an arbitrary subcarrier to aggregate an arbitrary feature dimension m. At each device, the local feature element $x_k(m)$ is first pre-coded with $b_{k,m}$ and then transmitted over the single-input-multiple-output (SIMO) channel, the aggregated received signal at the server is given by

$$\mathbf{y}(m) = \sum_{k=1}^{K} \mathbf{h}_{k,m} b_{k,m} x_k(m) + \mathbf{w}(m), \tag{19}$$

where $\mathbf{h}_{k,m} \in \mathbb{C}^{N_r}$ is the channel gain of device k, $b_{k,m}$ is the pre-coding complex scalar of $x_k(m)$, $\mathbf{w}(m)$ is the additive white Gaussian noise following the distribution of $\mathcal{N}(\mathbf{0}, N_0\mathbf{I})$ and N_0 is the channel noise variance, $\mathbf{I} \in \mathbb{R}^{N_r \times N_r}$ is the identity matrix. As mentioned, the channel vector $\mathbf{h}_{k,m}$ remains static for aggregating all feature elements. After receiving the signal, a receive beamforming vector $\mathbf{f}_m \in \mathbb{C}^{N_r}$ is added by the edge server to extract the feature vector

$$\hat{x}(m) = \mathbf{f}_m^H \mathbf{y}(m) = \mathbf{f}_m^H \sum_{k=1}^{K} \mathbf{h}_{k,m} b_{k,m} x_k(m) + \mathbf{f}_m^H \mathbf{w}(m). \tag{20}$$

For similar reasons as (18), the distribution of $\hat{x}(m)$ can be further derived as

$$f\left(\hat{x}(m)\right) = \frac{1}{L}\sum_{\ell=1}^{L} f_\ell(\hat{x}(m)),\ 1 \le m \le M, \tag{21}$$

where

$$f_\ell(\hat{x}(m)) = \mathbf{f}_m^H \sum_{k=1}^{K} \mathbf{h}_{k,m} b_{k,m} f_\ell(x_k(m)) + \mathbf{f}_m^H f(\mathbf{w}(m)), \tag{22}$$

and $f_\ell(x_k(m)) = \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2 + \sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right)$ and $f(\mathbf{w}(m)) = \mathcal{N}(\mathbf{0}, N_0\mathbf{I})$ are the distributions of the ℓ-th component of local feature in device k and the Gaussian white noise in wireless channel.

Then, all dimensions of the local feature vectors are aggregated in a similar way over M subcarriers, as shown in Fig. 4. Thereby, the overall received feature vector is $\hat{\mathbf{x}} = [\hat{x}(1), ..., \hat{x}(m), ..., \hat{x}(M)]^T$. Since PCA is performed at each device, different elements of each local feature vector are independent. As a result, the distributions of different elements in the received feature vector $\hat{\mathbf{x}}$ are independent, since each feature element $\hat{x}(m)$ only depends on the corresponding local

feature elements $\{x_k(m)\}$ and the white Gaussian channel noise according to (20).

III. Problem Formulation and Simplification

In this section, a novel design criterion called minimum pair-wise discriminant gain is adopted, based on which, the problem is formulated.

A. Minimum Pair-Wise Discriminant Gain

As mentioned, the design criterion adopted in this work is maximum inference accuracy instead of the conventional minimum mean square error (MMSE), as the latter cannot distinguish the importance levels of different elements to the inference task [28]. However, the instantaneous inference accuracy is unknown and does not have a mathematical model at the design stage. To this end, an approximate but tractable metric called discriminant gain is adopted as an alternative. Based on the received feature distribution in (21), a pair-wise discriminant gain of an arbitrary class pair $(\ell, \ell^{'})$ is defined as the symmetric KL divergence of their corresponding Gaussian components [30], [49]. Specifically, considering the m-th feature element, its pair-wise discriminant gain in terms of the class pair $(\ell, \ell^{'})$ is given by

$$\begin{aligned} G_{\ell,\ell'}(\hat{x}(m)) \triangleq & D_{KL}\left[f_\ell\left(\hat{x}(m)\right) \| f_{\ell'}\left(\hat{x}(m)\right)\right] \\ & + D_{KL}\left[f_{\ell'}\left(\hat{x}(m)\right) \| f_\ell\left(\hat{x}(m)\right)\right], \\ = & \int_{\hat{x}(m)} \left[f_\ell\left(\hat{x}(m)\right) \log\left[\frac{f_\ell\left(\hat{x}(m)\right)}{f_{\ell'}\left(\hat{x}(m)\right)}\right] \right. \\ & \left. + f_{\ell'}\left(\hat{x}(m)\right) \log\left[\frac{f_{\ell'}\left(\hat{x}(m)\right)}{f_\ell\left(\hat{x}(m)\right)}\right] \right] \mathrm{d}\hat{x}(m), \end{aligned} \tag{23}$$

where $D_{KL}\left[\mathrm{p}\|\mathrm{q}\right]$ represents the KL divergence between distributions p and q. As mentioned, different feature elements in the received feature vector $\hat{\mathbf{x}}$ are independent. It follows that the pair-wise discriminant gain of $\hat{\mathbf{x}}$ is derived as

$$\begin{aligned} G_{\ell,\ell'}(\hat{\mathbf{x}}) &= D_{KL}\left[f_\ell(\hat{\mathbf{x}}) \| f_{\ell'}(\hat{\mathbf{x}})\right] + D_{KL}\left[f_{\ell'}(\hat{\mathbf{x}}) \| f_\ell(\hat{\mathbf{x}})\right] \\ &= \sum_{m=1}^{M} G_{\ell,\ell'}\left(\hat{x}(m)\right), \ \forall(\ell, \ell^{'}). \end{aligned} \tag{24}$$

With a larger pair-wise discriminant gain, the corresponding pair of classes are better separated in the feature space, thus resulting in an improved achievable inference accuracy.

In existing literatures [28], [30], [31], maximizing the average of all pair-wise discriminant gains as defined in (25) is used as the design criterion, i.e.,

$$G\left(\hat{\mathbf{x}}\right) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} G_{\ell,\ell'}(\hat{\mathbf{x}}). \tag{25}$$

However, under this design goal, the values of one or several pair-wise discriminant gains can be dominant, while other pair-wise discriminant gains are very small. That says, only a subset of class pairs is well separated but the others cannot be differentiated [see Fig. 1(a) for example]. This leads to an unbalanced and low inference accuracy. To overcome this limitation, this work proposes to maximize the minimum pair-wise discriminant gain of all pairs, defined as

$$\begin{aligned} G_{\min}\left(\hat{\mathbf{x}}\right) &= \min_{1\le\ell\neq\ell'\le L} G_{\ell,\ell'}(\hat{\mathbf{x}}) \\ &= \min_{1\le\ell\neq\ell'\le L} \sum_{m=1}^{M} G_{\ell,\ell'}\left(\hat{x}(m)\right), \ \forall(\ell, \ell^{'}). \end{aligned} \tag{26}$$

By maximizing the minimum pair-wise discriminant gain in (26), the closest class pair in the feature space can be well separated, leading to a balanced and enhanced inference accuracy.

B. Problem Formulation

The maximization of the minimum pair-wise discriminant gain defined in (26) is constrained by the energy threshold of each device. Consider an arbitrary device k, its sensing energy consumption is $P_{s,k}T_{s,k}$ with $P_{s,k}$ being the sensing power and $T_{s,k}$ being the fixed sensing time. Its energy consumption for on-device feature extraction is denoted as $E_{p,k}$, which is a constant. For AirComp, the power of device k to transmit the m-th feature element is

$$P_{c,k}(m) = b_{k,m}\mathbb{E}\left[x_k(m)x_k(m)^H\right] b_{k,m}^H, \ \forall(m,k). \tag{27}$$

In (27), since the distribution of $x_k(m)$ is known [Please refer to (17)], its variance is determined and is denoted as $X_k(m) = \mathbb{E}\left[x_k(m)x_k(m)^H\right]$. It follows that the energy consumption of the whole AirComp process is

$$E_{c,k} = T_c P_{c,k}(m) = T_c \sum_{m=1}^{M} b_{k,m} b_{k,m}^H X_k(m), \ 1 \le k \le K, \tag{28}$$

where T_c is the AirComp transmission time for each element. Therefore, the energy consumption constraint of device k can be derived as

$$P_{s,k}T_{s,k} + E_{p,k} + T_c \sum_{m=1}^{M} b_{k,m} b_{k,m}^H X_k(m) \le E_k, \ 1 \le k \le K, \tag{29}$$

where E_k is the energy threshold of device k.

Accordingly, the problem of maximizing the minimum pair-wise discriminant gain under the energy consumption constraint can be formulated as

$$\begin{aligned} \mathbf{P1}: \max_{\substack{\{P_{s,k}\},\\ \{b_{k,m}\},\{\mathbf{f}_m\}}} & \min_{1\le\ell\neq\ell'\le L} \sum_{m=1}^{M} G_{\ell,\ell'}\left(\hat{x}(m)\right), \\ \text{s.t.} \quad & P_{s,k}T_{s,k} + E_{p,k} + T_c \sum_{m=1}^{M} b_{k,m} b_{k,m}^H X_k(m) \\ & \le E_k, \ 1 \le k \le K. \end{aligned} \tag{30}$$

C. Problem Simplification

Since the distributions of the received elements $\{\hat{x}(m)\}$ in (21) are complex, the minimum pair-wise discriminant gain defined based on these distributions, i.e., the objective of $\mathbf{P1}$ is a complicated non-convex function. Besides, the energy constraint in $\mathbf{P1}$ is also non-convex. To address this complicated

non-convex problem, a conventional approach (see, [28], [40], [41]) is applied to simplify it by pre-determining the precoders as

$$\mathbf{f}_m^H \mathbf{h}_{k,m} b_{k,m} = c_{k,m},\ 1 \le m \le M,\ 1 \le k \le K, \tag{31}$$

where $c_{k,m} \in \mathbb{R}^+$ represents the received signal power of element m from device k. Accordingly, the precoder $b_{k,m}$ can be written in a function of $c_{k,m}$ by multiplying $(\mathbf{f}_m^H \mathbf{h}_{k,m})^H$ on both sides of equation (31):

$$(\mathbf{f}_m^H \mathbf{h}_{k,m})^H \mathbf{f}_m^H \mathbf{h}_{k,m} b_{k,m} = \mathbf{h}_{k,m}^H \mathbf{f}_m c_{k,m},\ \forall (m,k). \tag{32}$$

Then, $b_{k,m}$ is derived as

$$b_{k,m} = \frac{c_{k,m} \mathbf{h}_{k,m}^H \mathbf{f}_m}{\mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m}} = \frac{c_{k,m}}{\mathbf{f}_m^H \mathbf{h}_{k,m}},\ \forall (m,k). \tag{33}$$

By substituting $b_{k,m}$ in (33) into the received feature element in (20), we have

$$\hat{x}(m) = \sum_{k=1}^{K} c_{k,m} x_k(m) + \mathbf{f}_m^H \mathbf{w}(m),\ \forall (m,k), \tag{34}$$

which, by substituting the local feature elements $\{x_k(m)\}$ in (13), is further derived as

$$\begin{aligned} \hat{x}(m) = &\left(\sum_{k=1}^{K} c_{k,m} \right) x(m) \\ &+ \sum_{k=1}^{K} c_{k,m} \left(c_{s,k}(m) + \frac{n_r(m)}{\sqrt{P_{s,k}}} \right) + \mathbf{f}_m^H \mathbf{w}(m). \end{aligned} \tag{35}$$

It follows that the distribution of $\hat{x}(m)$ can be derived as

$$\begin{aligned} f\left(\hat{x}(m)\right) &= \frac{1}{L} \sum_{\ell=1}^{L} f_\ell(\hat{x}(m)) \\ &= \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{N}\left(\hat{\mu}_{\ell,m}, \hat{\sigma}_m^2\right),\ 1 \le m \le M, \end{aligned} \tag{36}$$

Since the transformations in (33) are all linear and $x_\ell(m) \sim \mathcal{N}(\mu_{\ell,m}, \sigma_m^2)$, $c_{s,k}(m) \sim \mathcal{N}(0, \sigma_{s,k}^2)$ and $n_r(m) \sim \mathcal{N}(0, \sigma_r^2)$ are following independent Gaussian distributions, the distribution of $\hat{x}(m)$ can be derived in a closed form. The mean of the ℓ-th class component is given as follows:

$$\hat{\mu}_{\ell,m} = \left(\sum_{k=1}^{K} c_{k,m} \right) \mu_{\ell,m}, \tag{37}$$

and the variance of the ℓ-th class component is given as follows:

$$\begin{aligned} \hat{\sigma}_m^2 = &\left(\sum_{k=1}^{K} c_{k,m} \right)^2 \sigma_m^2 \\ &+ \sum_{k=1}^{K} c_{k,m}^2 \left(\sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}} \right) + N_0 \mathbf{f}_m^H \mathbf{f}_m. \end{aligned} \tag{38}$$

As a result, the pair-wise discriminant gain $G_{\ell,\ell'}\left(\hat{x}(m)\right)$ can be derived as

$$G_{\ell,\ell'}\left(\hat{x}(m)\right) = \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2} = \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2 \left(\sum_{k=1}^{K} c_{k,m} \right)^2}{\sigma_m^2 \left(\sum_{k=1}^{K} c_{k,m} \right)^2 + \sum_{k=1}^{K} c_{k,m}^2 \left(\sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}} \right) + N_0 \mathbf{f}_m^H \mathbf{f}_m}. \tag{39}$$

Besides, by substituting the precoders in (33) into the energy constraint in $\mathbf{P1}$, it can be re-formulated as

$$P_{s,k} T_{s,k} + E_{p,k} + T_c \sum_{m=1}^{M} \frac{c_{k,m}^2 X_k(m)}{\mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m}} \le E_k,\ 1 \le k \le K. \tag{40}$$

In summary, with the precoders defined in (33), $\mathbf{P1}$ is simplified as

$$\mathbf{P2} \quad \begin{aligned} \max_{\substack{\{P_{s,k}\},\\ \{c_{k,m}\},\{\mathbf{f}_m\}}} \quad & \min_{1 \le \ell \ne \ell' \le L} \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2}, \\ \text{s.t.} \quad & P_{s,k} T_{s,k} + E_{p,k} + T_c \\ & \times \sum_{m=1}^{M} \frac{c_{k,m}^2 X_k(m)}{\mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m}} \le E_k,\ 1 \le k \le K. \end{aligned} \tag{41}$$

IV. Joint Sensing Power Assignment, Transmit Precoding and Receive Beamforming

Although $\mathbf{P2}$ has a simplified form, it is still difficult to solve due to the minimax form and the complicated non-convex fractional functions in both objective and constraints. To address this problem, in the sequel, variables transformation is conducted to decouple the minimax objective function and to derive an equivalent problem with the d.c. form, based on which, the typical method of SCA is utilized to obtain a sub-optimal solution.

A. Variables Transformation

To begin with, the following variable is defined to decouple the minimax objective function:

$$\alpha = \min_{1 \le \ell \ne \ell' \le L} \sum_{m=1}^{M} G_{\ell,\ell'}\left(\hat{x}(m)\right). \tag{42}$$

It follows that all pair-wise discriminant gains should be no less than α:

$$\sum_{m=1}^{M} \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2} \ge \alpha,\ 1 \le \ell \ne \ell' \le L. \tag{43}$$

Accordingly, $\mathbf{P2}$ is equivalent to the problem that maximizes α under the constraints of the original energy consumption and pair-wise discriminant gains in (43), i.e.,

$$\mathbf{P3}\quad \begin{aligned} \max_{\substack{\{P_{s,k}\},\alpha,\\ \{c_{k,m}\},\{\mathbf{f}_m\}}} \quad & \alpha, \\ \text{s.t.} \quad & P_{s,k}T_{s,k} + E_{p,k} + T_c \\ & \times \sum_{m=1}^{M} \frac{c_{k,m}^2 X_k(m)}{\mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m}} \le E_k,\ 1 \le k \le K, \\ & \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2} \ge \alpha, \qquad 1 \le \ell \ne \ell' \le L. \end{aligned}$$

Then, to further address the non-convex ratios in the energy consumption constraint (the first constraint), the following variables are introduced:

$$u_{k,m} = \frac{c_{k,m}^2}{\mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m}} \ge 0,\ 1 \le m \le M. \tag{44}$$

By substituting (44), the energy constraint in $\mathbf{P3}$ for each device k is equivalently decomposed into the following two constraints:

$$P_{s,k}T_{s,k} + E_{p,k} + T_c \sum_{m=1}^{M} u_{k,m} X_k(m) \le E_k,\ 1 \le k \le K, \tag{45}$$

and

$$c_{k,m}^2 = \mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m} u_{k,m},\ 1 \le m \le M. \tag{46}$$

Next, we extend the feasible region of the equality constraint (46) as in (47) while keeping the same optimal solution to $\mathbf{P3}$, as shown in Lemma 2.

$$c_{k,m}^2 \le \mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m} u_{k,m},\ 1 \le m \le M. \tag{47}$$

Lemma 2. *A new problem $\mathbf{P3}'$ which extends the feasible region of* (46) *to* (47) *and keeps the same objective function, the constraint in* (45) *and the pair-wise discriminant constraint (the second constraint in $\mathbf{P3}$), reaches the same optimum as $\mathbf{P3}$.*

Proof. See Appendix B. □

To further address the non-convex pair-wise discriminant gain constraint (the second constraint) in $\mathbf{P3}$, a set of variables $\{v_{\ell,\ell',m}\}$ are introduced as follows:

$$\begin{aligned} &\frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{v_{\ell,\ell',m}} \left(\sum_{k=1}^{K} c_{k,m}\right)^2 \\ &= \sigma_m^2 \left(\sum_{k=1}^{K} c_{k,m}\right)^2 + \sum_{k=1}^{K} c_{k,m}^2 \left(\sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right) + N_0 \mathbf{f}_m^H \mathbf{f}_m. \end{aligned} \tag{48}$$

It follows that the pair-wise discriminant gain constraint (the second constraint) in $\mathbf{P3}$ can be equivalently decomposed as

$$\sum_{m=1}^{M} v_{\ell,\ell',m} \ge \alpha. \tag{49}$$

For similar reasons to (46) and Lemma 2, the feasible region of the constraint in (48) can be extended as that in (50) without changing the optimal solution of $\mathbf{P3}$.

$$\begin{aligned} &\frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{v_{\ell,\ell',m}} \left(\sum_{k=1}^{K} c_{k,m}\right)^2 \\ &\ge \sigma_m^2 \left(\sum_{k=1}^{K} c_{k,m}\right)^2 + \sum_{k=1}^{K} c_{k,m}^2 \left(\sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right) + N_0 \mathbf{f}_m^H \mathbf{f}_m. \end{aligned} \tag{50}$$

In summary, $\mathbf{P3}$ can be equivalently derived as the following form:

$$\mathbf{P4}\quad \begin{aligned} \max_{\substack{\{P_{s,k}\},\{c_{k,m}\},\\ \{\mathbf{f}_m\},\{u_{k,m}\},\\ \{v_{\ell,\ell',m}\},\alpha}} \quad & \alpha, \\ \text{s.t.} \quad & P_{s,k}T_{s,k} + E_{p,k} + T_c \sum_{m=1}^{M} u_{k,m} X_k(m) \\ & \qquad - E_k \le 0,\ 1 \le k \le K, \\ & \alpha - \sum_{m=1}^{M} v_{\ell,\ell',m} \le 0, \qquad 1 \le \ell \ne \ell' \le L, \\ & \frac{c_{k,m}^2}{u_{k,m}} - R_{k,m}\left(\mathbf{f}_m\right) \le 0,\ \forall (k,m), \\ & Z_m\left(\{P_{s,k}\},\{c_{k,m}\},\mathbf{f}_m\right) \\ & \qquad - Q_{\ell,\ell',m}\left(\{c_{k,m}\},v_{\ell,\ell',m}\right) \le 0, \end{aligned}$$

where

$$\begin{cases} R_{k,m}\left(\mathbf{f}_m\right) = \mathbf{h}_{k,m}^H \mathbf{f}_m \mathbf{f}_m^H \mathbf{h}_{k,m}, \\ Z_m\left(\{P_{s,k}\},\{c_{k,m}\},\mathbf{f}_m\right) = \sigma_m^2 \left(\sum_{k=1}^{K} c_{k,m}\right)^2 \\ \qquad\qquad + \sum_{k=1}^{K} c_{k,m}^2 \left(\sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right) + N_0 \mathbf{f}_m^H \mathbf{f}_m, \\ Q_{\ell,\ell',m}\left(\{c_{k,m}\},v_{\ell,\ell',m}\right) = \dfrac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{v_{\ell,\ell',m}} \left(\sum_{k=1}^{K} c_{k,m}\right)^2. \end{cases}$$

Although $\mathbf{P4}$ is still non-convex, it is in the d.c. form, as shown in Lemma 3.

Lemma 3. *Problem $\mathbf{P4}$ is the d.c. problem.*

Proof. See Appendix C. □

B. SCA based Algorithm

To solve $\mathbf{P4}$, the method of SCA is adopted, which iterates between the following two steps until convergence to obtain a suboptimal solution, where all Karush-Kuhn-Tucker (KKT) conditions of $\mathbf{P4}$ are satisfied.

- *Convex approximation*: Based on a reference point, a convex approximation of $\mathbf{P4}$ is derived using Taylor expansion. The feasible region of the approximated problem is a subset of that of $\mathbf{P4}$. This guarantees that its solution is feasible for $\mathbf{P4}$.

- *Reference point update*: The approximated problem is optimally solved and the solution is used as the new reference point for the next iteration.

In the sequel, the detailed procedures to solve $\mathbf{P4}$ are presented.

1) Convex approximation: We first randomly initialize the optimization variables and set the counter $t = 0$. Then, for an arbitrary iteration, i.e., $t > 0$, the convex approximation of $\mathbf{P4}$ is described as follows.

According to Lemma 3, $R_{k,m}(\mathbf{f}_m)$ and $Q_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})$ are both differentiable convex functions. Therefore, they are no less than their first-order Taylor expansions with the reference point being the optimal solution in the $(t-1)$-th iteration, i.e.,

$$R_{k,m}(\mathbf{f}_m) \geq \hat{R}^{[t]}_{k,m}(\mathbf{f}_m), \tag{51}$$

$$Q_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m}) \geq \hat{Q}^{[t]}_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m}), \tag{52}$$

where $\hat{R}^{[t]}_{k,m}(\mathbf{f}_m)$ and $\hat{Q}^{[t]}_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})$ are the first-order Taylor expansions at $\mathbf{f}^{[t]}_m$ and $\left(\{c^{[t]}_{k,m}\}, v^{[t]}_{\ell,\ell',m}\right)$ respectively. They are given by

$$\hat{R}^{[t]}_{k,m}(\mathbf{f}_m) = R_{k,m}(\mathbf{f}^{[t]}_m) + \left(\mathbf{f}_m - \mathbf{f}^{[t]}_m\right)^H \mathbf{A}^{[t]}_{k,m}, \tag{53}$$

$$\begin{aligned}\hat{Q}^{[t]}_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m}) &= Q_{\ell,\ell',m}\left(\left\{c^{[t]}_{k,m}\right\}, v^{[t]}_{\ell,\ell',m}\right) \\ &+ B^{[t]}_{\ell,\ell',m}\left(v_{\ell,\ell',m} - v^{[t]}_{\ell,\ell',m}\right) + \sum_{k=1}^{K} C^{[t]}_{k,m}\left(c_{k,m} - c^{[t]}_{k,m}\right),\end{aligned} \tag{54}$$

where

$$\begin{cases} \mathbf{A}^{[t]}_{k,m} = \left.\dfrac{\partial R}{\partial \mathbf{f}_m}\right|_{\mathbf{f}_m = \mathbf{f}^{[t]}_m} = 2\mathbf{h}_{k,m}\mathbf{h}^H_{k,m}\mathbf{f}^{[t]}_m, \\ B^{[t]}_{\ell,\ell',m} = \left.\dfrac{\partial Q}{\partial v_{\ell,\ell',m}}\right|_{v_{\ell,\ell',m} = v^{[t]}_{\ell,\ell',m}} \\ \qquad\quad = -\left(\dfrac{(\mu_{\ell,m} - \mu_{\ell',m})\sum_{k=1}^{K} c^{[t]}_{k,m}}{v^{[t]}_{\ell,\ell',m}}\right)^2, \\ C^{[t]}_{k,m} = \left.\dfrac{\partial Q}{\partial c_{k,m}}\right|_{c_{k,m} = c^{[t]}_{k,m}} \\ \qquad\quad = \dfrac{2\left(\sum_{k=1}^{K} c^{[t]}_{k,m}\right)(\mu_{\ell,m} - \mu_{\ell',m})^2}{v^{[t]}_{\ell,\ell',m}}. \end{cases} \tag{55}$$

By replacing $R_{k,m}(\mathbf{f}_m)$ and $Q_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})$ with $\hat{R}^{[t]}_{k,m}(\mathbf{f}_m)$ and $\hat{Q}^{[t]}_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})$ respectively, an approximated convex problem of $\mathbf{P4}$ can be derived as

$$\mathbf{P5}: \begin{aligned} \max_{\substack{\{P_{s,k}\},\{c_{k,m}\},\\ \{\mathbf{f}_m\},\{u_{k,m}\},\\ \{v_{\ell,\ell',m}\},\alpha}} \quad & \alpha, \\ \text{s.t.} \quad & P_{s,k}T_{s,k} + E_{p,k} + T_c\sum_{m=1}^{M} u_{k,m}X_k(m) \\ & \qquad - E_k \leq 0,\ 1 \leq k \leq K, \\ & \alpha - \sum_{m=1}^{M} v_{\ell,\ell',m} \leq 0, \qquad 1 \leq \ell \neq \ell' \leq L, \\ & \frac{c^2_{k,m}}{u_{k,m}} - \hat{R}^{[t]}_{k,m}(\mathbf{f}_m) \leq 0,\ \forall (k,m), \\ & Z_m(\{P_{s,k}\}, \{c_{k,m}\}, \mathbf{f}_m) \\ & \qquad - \hat{Q}^{[t]}_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m}) \leq 0, \end{aligned}$$

where $Z_m(\{P_{s,k}\}, \{c_{k,m}\}, \mathbf{f}_m)$, $\hat{R}^{[t]}_{k,m}(\mathbf{f}_m)$ and $\hat{Q}^{[t]}_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})$ are the same as those defined in $\mathbf{P4}$.

2) Solution to $\mathbf{P5}$*:* The primal-dual method is used to optimally solve $\mathbf{P5}$. First, the Lagrangian function of $\mathbf{P5}$ is given by

$$\begin{aligned} \mathcal{L}_{\mathbf{P5}} = & -\alpha \\ & + \sum_{k=1}^{K}\beta_k\left(P_{s,k}T_{s,k} + E_{p,k} + T_c\sum_{m=1}^{M} u_{k,m}X_k(m) - E_k\right) \\ & + \sum_{\ell'=1}^{L}\sum_{\ell\neq\ell'}\gamma_{\ell,\ell'}\left(\alpha - \sum_{m=1}^{M} v_{\ell,\ell',m}\right) \\ & + \sum_{k=1}^{K}\sum_{m=1}^{M}\theta_{k,m}\left[\frac{c^2_{k,m}}{u_{k,m}} - \hat{R}^{[t]}_{k,m}(\mathbf{f}_m)\right] \\ & + \sum_{m=1}^{M}\sum_{\ell'=1}^{L}\sum_{\ell\neq\ell'}\lambda_{\ell,\ell',m}\Big[Z_m(\{P_{s,k}\}, \{c_{k,m}\}, \mathbf{f}_m) \\ & \qquad - \hat{Q}^{[t]}_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})\Big], \end{aligned} \tag{56}$$

where $\beta_k, \gamma_{\ell,\ell'}, \theta_{k,m}$ and $\lambda_{\ell,\ell',m}$ are all positive Lagrange multipliers. Then, some useful KKT conditions are given by

$$\frac{\partial \mathcal{L}_{\mathbf{P5}}}{\partial P_{s,k}} = \beta_k T_{s,k} - \frac{\sigma^2_r c^2_{k,m}}{P^2_{s,k}} = 0, \tag{57}$$

$$\frac{\partial \mathcal{L}_{\mathbf{P5}}}{\partial \mathbf{f}_m} = 2N_0\mathbf{f}_m - \sum_{k=1}^{K} 2\theta_{k,m}\mathbf{h}_{k,m}\mathbf{h}^H_{k,m}\mathbf{f}^{[t]}_m = 0, \tag{58}$$

$$\begin{aligned} \frac{\partial \mathcal{L}_{\mathbf{P5}}}{\partial c_{k,m}} = & \frac{2\theta_{k,m}}{u_{k,m}} + \sum_{\ell'=1}^{L}\sum_{\ell\neq\ell'}\lambda_{\ell,\ell',m}\Bigg(\frac{2\sigma^2_r c_{k,m}}{P_{s,k}} + 2c_{k,m}\sigma^2_{s,k} \\ & + 2\sigma^2_m\sum_{k=1}^{K} c_{k,m} - \frac{2\left(\sum_{k=1}^{K} c^{[t]}_{k,m}\right)(\mu_{\ell,m} - \mu_{\ell',m})^2}{v^{[t]}_{\ell,\ell',m}}\Bigg), \end{aligned} \tag{59}$$

Algorithm 1 Primal-dual method for solving $\mathbf{P5}$ in SCA iteration t

Input: Channel gain $\{\mathbf{h}_{k,m}\}$, feature elements' class centroid $\{\mu_{\ell,m}\}$ and variance $\{\sigma_m^2\}$, communication latency $\{T_{s,k}\}$ and other given parameters derived from iteration $t-1$

Output: $\{P_{s,k}^{[t+1]}, c_{k,m}^{[t+1]}, \mathbf{f}_m^{[t+1]}, \alpha^{[t+1]}, u_{k,m}^{[t+1]}, v_{\ell,\ell',m}^{[t+1]}\}$

1: Initialize $\{\beta_k\}$, $\{\gamma_{\ell,\ell'}\}$, $\{\theta_{k,m}\}$, $\{\lambda_{\ell,\ell',m}\}$, the step size $\{\eta_{\beta_k}^{(0)}\}$, $\{\eta_{\gamma_{\ell,\ell'}}^{(0)}\}$, $\{\eta_{\theta_{k,m}}^{(0)}\}$, $\{\eta_{\lambda_{\ell,\ell',m}}^{(0)}\}$ and $i \leftarrow 0$;
2: **while** not convergence **do**
3: Derive $\{P_{s,k}^{[t+1]}\}$, $\{\mathbf{f}_m^{[t+1]}\}$ and $\{c_{k,m}^{[t+1]}\}$ using (60),(61) and (62), respectively;
4: Update the multipliers as

$$\begin{cases} \beta_k^{(i+1)} = \max\left\{\beta_k^{(i)} + \eta_{\beta_k}\dfrac{\partial \mathcal{L}_{\mathbf{P5}}}{\partial \beta_k}, 0\right\}, \\ \gamma_{\ell,\ell'}^{(i+1)} = \max\left\{\gamma_{\ell,\ell'}^{(i)} + \eta_{\gamma_{\ell,\ell'}}\dfrac{\partial \mathcal{L}_{\mathbf{P5}}}{\partial \gamma_{\ell,\ell'}}, 0\right\}, \\ \theta_{k,m}^{(i+1)} = \max\left\{\theta_{k,m}^{(i)} + \eta_{\theta_{k,m}}\dfrac{\partial \mathcal{L}_{\mathbf{P5}}}{\partial \theta_{k,m}}, 0\right\}, \\ \lambda_{\ell,\ell',m}^{(i+1)} = \max\left\{\lambda_{\ell,\ell',m}^{(i)} + \eta_{\lambda_{\ell,\ell',m}}\dfrac{\partial \mathcal{L}_{\mathbf{P5}}}{\partial \lambda_{\ell,\ell',m}}, 0\right\}; \end{cases}$$

5: $i \leftarrow i+1$;
6: **end while**

which can be respectively derived as below to reach the optimal value of $P_{s,k}^{[t+1]}$, $\mathbf{f}_m^{[t+1]}$ and $c_{k,m}^{[t+1]}$

$$P_{s,k}^{[t+1]} = \frac{\sigma_r c_{k,m}}{\sqrt{\beta_k T_{s,k}}}, \tag{60}$$

$$\mathbf{f}_m^{[t+1]} = \frac{1}{2N_0}\sum_{k=1}^{K} 2\theta_{k,m}\mathbf{h}_{k,m}\mathbf{h}_{k,m}^H\mathbf{f}_m^{[t]}, \tag{61}$$

$$c_{k,m}^{[t+1]} = \left[\sum_{\ell'=1}^{L}\sum_{\ell\neq\ell'}\lambda_{\ell,\ell',m}\left(\frac{\sum_{k=1}^{K} c_{k,m}^{[t]}(\mu_{\ell,m}-\mu_{\ell',m})^2}{v_{\ell,\ell',m}^{[t]}} - \sigma_m^2\sum_{k'\neq k} c_{k',m}\right) - \frac{2\theta_{k,m}}{u_{k,m}}\right] \Big/ \left[\left(\frac{\sigma_r^2}{P_{s,k}} + \sigma_{s,k}^2 + \sigma_m^2\right)\left(\sum_{\ell'=1}^{L}\sum_{\ell\neq\ell'}\lambda_{\ell,\ell',m}\right)\right]. \tag{62}$$

Based on the results above, the multipliers $\beta_k, \gamma_{\ell,\ell'}, \theta_{k,m}$ and $\lambda_{\ell,\ell',m}$ can be updated with their stepsizes $\eta_{\beta_k}, \eta_{\gamma_{\ell,\ell'}}, \eta_{\theta_{k,m}}$ and $\eta_{\lambda_{\ell,\ell',m}}$ to solve the problem in the next round, respectively. The primal-dual method is presented in Algorithm 1. Compared directly adopting the typical algorithms in existing toolbox like CVX, Algorithm 1 enjoys the benefits of using the closed-form solutions in (60), (61) and (62). Therefore, the computational complexity of Algorithm 1 is reduced to $\mathcal{O}(I(11N_rKM + N_rL^2M + \frac{1}{2}L^2K^2M))$ with the assumption that Algorithm 1 converges after I loops of computing.

As a result, the optimal solution of $\mathbf{P5}$ can be obtained and is denoted as $P_{s,k}^{[t+1]}$, $c_{k,m}^{[t+1]}$, $\mathbf{f}_m^{[t+1]}$, $\alpha^{[t+1]}$, $u_{k,m}^{[t+1]}$, $v_{\ell,\ell',m}^{[t+1]}$, which are used as the reference points for the $(t+1)$-th iteration.

Algorithm 2 Joint Sensing Power, Transmit Precoder and Receive Beamformer Design

Input: Channel gain $\{\mathbf{h}_{k,m}\}$, device energy $\{E_k\}$, sensing time $\{T_{s,k}\}$, communication time T_c, computation energy $\{E_{p,k}\}$

Output: $\{P_{s,k}^*\}$, $\{c_{k,m}^*\}$, $\{\mathbf{f}_m^*\}$ and α^*

1: Initialize $t \leftarrow 0$, $\{P_{s,k}^{[0]}\}$, $\{c_{k,m}^{[0]}\}$, $\{\mathbf{f}_m^{[0]}\}$ in feasible region of $\mathbf{P4}$;
2: Calculate the initial value of $\{v_{\ell,\ell',m}^{[0]}\}$;
3: Initialize the auxiliary function $\hat{R}_{k,m}^{[0]}(\mathbf{f}_m)$ and $\hat{Q}_{\ell,\ell',m}^{[0]}(\{c_{k,m}\}, v_{\ell,\ell',m})$;
4: **while** not convergence **do**
5: Derive $\mathbf{P5}$ by relaxing $\mathbf{P4}$ with $\hat{R}_{k,m}^{[t]}(\mathbf{f}_m)$ and $\hat{Q}_{\ell,\ell',m}^{[t]}(\{c_{k,m}\}, v_{\ell,\ell',m})$;
6: Solve $\mathbf{P5}$ with Algorithm 1 to get optimum $\{P_{s,k}^{[t+1]}, c_{k,m}^{[t+1]}, \mathbf{f}_m^{[t+1]}, \alpha^{[t+1]}, u_{k,m}^{[t+1]}, v_{\ell,\ell',m}^{[t+1]}\}$;
7: Calculate the new auxiliary function $\hat{R}_{k,m}^{[t+1]}(\mathbf{f}_m)$ and $\hat{Q}_{\ell,\ell',m}^{[t+1]}(\{c_{k,m}\}, v_{\ell,\ell',m})$;
8: $t \leftarrow t+1$;
9: **end while**
10: The optimal solution $P_{s,k}^* \leftarrow P_{s,k}^{[t]}$, $c_{k,m}^* \leftarrow c_{k,m}^{[t]}$, $\mathbf{f}_m^* \leftarrow \mathbf{f}_m^{[t]}$ and $\alpha^* \leftarrow \alpha^{[t]}$;

3) Solution to $\mathbf{P3}$*:* Based on the solution to $\mathbf{P5}$ and the SCA method described before, the solution procedure to $\mathbf{P3}$ is summarized in Algorithm 2.

V. Simulation Results

A. Simulation Setup

1) Network settings: A single-cell network is used to complete edge-device co-inference tasks. There is one edge server equipped with an 8-antenna AP located at the center and K single-antenna devices randomly located in a ring with radius in the range of $[R,\ R+0.05]$ kilometers. By default, R is set as 0.45 and K is set as 3 unless specified otherwise. The channel gain of the link between the edge server and device k is modeled as $\mathbf{h}_k = |\varphi_k\boldsymbol{\rho}_k|^2$. $[\varphi_k]_{\mathrm{dB}} = -[\mathbf{PL}_k]_{\mathrm{dB}} + [\zeta_k]_{\mathrm{dB}}$ is the large-scale fading channel coefficient, where $[\mathbf{PL}_k]_{\mathrm{dB}} = 128.1 + 37.6\log_{10} d_k$ is the path loss in dB, d_k is the distance between device k and the edge server, $[\zeta_k]_{\mathrm{dB}}$ is the shadowing in dB which follows the Gaussian distribution of $\mathcal{N}(0, \sigma_\zeta^2)$. On the other hand, $\boldsymbol{\rho}_k \sim \mathcal{CN}(0, \mathbf{I})$ stands for the small-scale fading channel coefficient, where Rayleigh small-scale fading is considered in the simulation. The variances of sensing noise σ_r^2 and clutter signal $\sigma_{s,k}^2$ are both set to 0.2. The channel noise variance N_0 is set to 1 and the variance of shadow fading $\sigma_\zeta^2 = 8$ dB.

2) Inference tasks: A human motion recognition task is selected to evaluate the performance of the proposed algorithm. The aim of this task is to distinguish 4 human motions, i.e., adult pacing, adult walking, child pacing and child walking, where the heights of adults are uniformly randomized between $[1.6\text{m}, 1.9\text{m}]$ and the heights of children follow the uniform distribution in $[0.9\text{m}, 1.2\text{m}]$. The facing directions of adults

and children are considered to be uniformly distributed in the range $[-180°, 180°]$ and the speed of moving is divided into three classes, with 0 m/s, $0.25H$ m/s and $0.5H$ m/s representing standing, pacing and walking where H is the height of each individual. The sensing time T_s and communication time T_c of devices are set to 1 second and the computation energy $E_{p,k}$ is set to 0.1 Joule. The dataset of radar sensing signals used for training and testing is generated by the wireless sensing simulator proposed in [44].

3) Inference Models: To identify the motion from local features, two machine learning models are adopted: a support vector machine (SVM) model and a multi-layer perceptron (MLP) neural network. In this experiment, the MLP network is trained with Adam optimizer [50], with the numbers of neurons in the hidden layers of MLP set to 80 and 40. The dataset generated by the simulator proposed in [44] is separated into a training dataset containing 6400 samples and a test dataset which contains 1600 samples. The training dataset is considered as the ground-truth data (free of noise) to train both of the two ML models. The testing dataset is distorted by clutter and noise through the sensing process and communication process determined by the three schemes mentioned below.

4) Inference algorithms: To verify the priority of the proposed scheme, three algorithms are compared in the experiments, as listed below.

- *Our proposal*: All parameters are allocated by the proposed scheme in Algorithm 2.
- *Existing AirComp scheme*: The sensing power is allocated randomly and other parameters are allocated following the AirComp scheme in [28].
- *Baseline*: The sensing power is allocated randomly, the receive beamforming is set to a constant of all elements' transmission and a maximum steering power is allocated under the energy constraint (40).

All experiments are implemented using Python 3.8.5 on a Windows 10 server with one NVIDIA® GeForce® GTX 1070 GPU 8GB and one Intel® Core™ i7-8700 CPU.

B. Performance Comparison

In this part, the relations between the inference accuracy and the minimum pair-wise discriminant gain are firstly presented. Then, the impact of the cell radius on inference accuracy is analyzed. Finally, the three algorithms are compared in terms of the SVM model and the MLP model with different numbers of devices and different device energy thresholds, respectively.

1) Relation between inference accuracy and minimum pair-wise discriminant gain: The relations between the inference accuracy and the minimum pair-wise discriminant gain for both two machine learning models are illustrated in Fig. 5. It shows that the inference accuracy grows from 25% to 95% as the minimum pair-wise discriminant gain increases for both AI models. Also, the SVM model reaches a higher inference accuracy than the MLP network, particularly, the accuracy of the SVM model gets nearly 40% when the minimum pair-wise discriminant gain is 10 while the accuracy of the MLP model is still 25%.

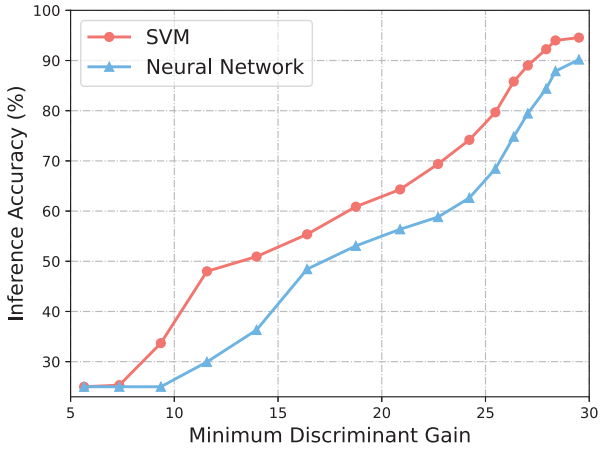


Fig. 5. Inference accuracy versus minimum pair-wise discriminant gain on different models

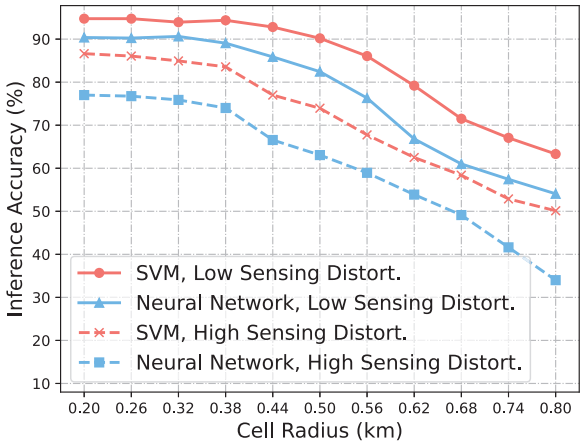
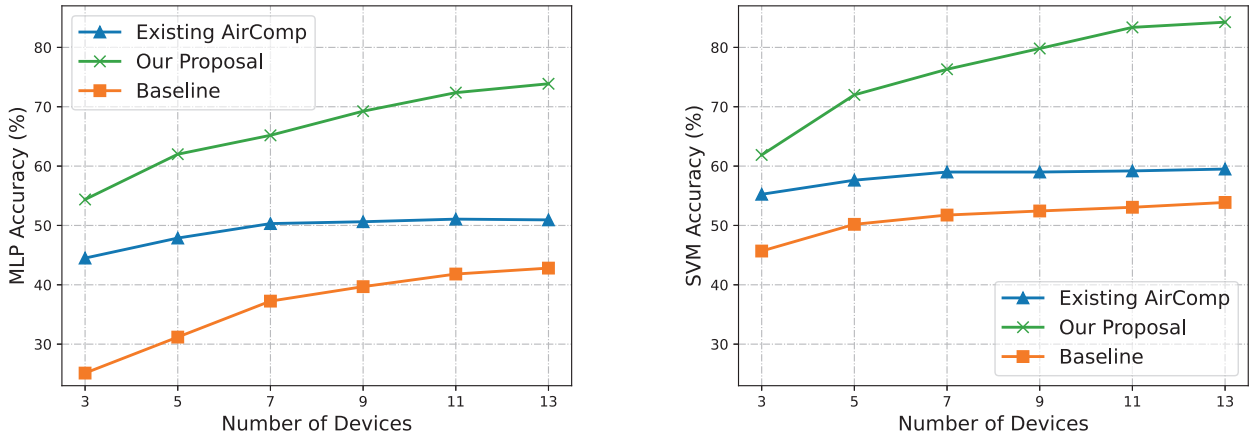


Fig. 6. Inference accuracy versus cell radius on different models

2) Relation between inference accuracy and cell radius: Fig. 6 presents the change of inference accuracy under different cell radiuses. It shows that the inference accuracies of both machine learning models decrease when the cell radius R increases from $200m$ to $800m$. That's because the distances between the devices and the edge server turn to be larger with a larger R, leading to stronger path losses and weaker channel gains. This causes a larger communication distortion level and reduces the inference accuracy. Besides, Fig. 6 also illustrates the effect of sensing distortion on inference accuracy. Both two machine learning models perform better in the case of low sensing distortion ($\sigma_r^2 = 0.2$) than in the case of high sensing distortion ($\sigma_r^2 = 1.4$).

3) Inference accuracy v.s. number of devices: The inference accuracies of the three schemes versus different number of devices are presented in Fig. 7. The performance of all three schemes increases as the number of devices increases for both machine learning models. It is because using more devices and aggregating their local features can reduce both the sensing



(a) Inference accuracy of MLP versus the number of devices

(b) Inference accuracy of SVM versus the number of devices

Fig. 7. Inference accuracy of different algorithms with different numbers of devices on MLP and SVM models

distortion and communication noise. Besides, the proposed ISCC scheme outperforms the existing AirComp scheme proposed in [28]. The reasons are three folds. First, the proposed scheme adopts a more reasonable metric, say the minimum pair-wise discriminant gain, instead of the average pair-wise discriminant gain used in the existing AirComp scheme, leading to a balanced and enhanced achievable inference accuracy. Besides, the sensing stage of the inference task, which is separately designed in the existing AirComp scheme, is jointly designed in this work. Furthermore, rather than separately optimizing the aggregation of the feature elements in the existing scheme, they are jointly optimized, which allows more resources being assigned to the important elements. In addition, the inference accuracies of all scheme gradually saturate, since involving more devices has little contribution on suppressing the sensing and channel noise when the number of devices is large. The inference accuracy of the existing AirComp scheme saturates first because it's achievable inference accuracy is lower than that of the proposed scheme but it can well suppress the channel noise with a small number of devices.

4) Inference accuracy v.s. device energy: Fig. 8 shows the impact of the device total energy on the accuracies of inference task in three schemes. It is shown that as a higher device energy is permitted, all of the three schemes have better inference accuracy since a higher device energy threshold means the devices can set larger sensing power and communication power to suppress the corresponding noise. In addition, the proposed scheme has a better performance than other two schemes for similar reasons as mentioned before.

VI. Conclusion

In this paper, an AirComp based ISCC scheme was proposed for edge-device co-inference tasks. Compared to existing schemes, the proposed scheme enjoyed advantages from three aspects. To begin with, a novel design criterion, called maximum minimum pair-wise discriminant gain, was adopted, which enlarged the distance of the closest pair in the feature space, resulting in a balanced and enhanced achievable inference accuracy. Besides, the sensing, computation and communication processes were jointly investigated from a systematic view, allowing more flexible resource coordination and sharing among the three modules. Moreover, the aggregation of all feature elements was jointly designed, enabling adaptive resource allocation among different feature elements. Benefiting from the above three advantages, the proposed scheme enjoyed a more reasonable design goal and better resource utilization, thus leading to better inference performance compared to existing schemes as verified by the experiments.

This work opens several interesting directions for task-oriented ISCC scheme designs. One is to enhance the inference accuracy over time-variant channels or device scheduling under limited communication resources. Another is to design the scheme with some devices only acquiring part of the sensory view.

Appendix

A. Proof of Lemma 1

As mentioned in (17), the ground-true feature element can be written as the average of L independent Gaussian random variables

$$x(m) = \frac{1}{L}\sum_{\ell=1}^{L} x_\ell(m), \tag{63}$$

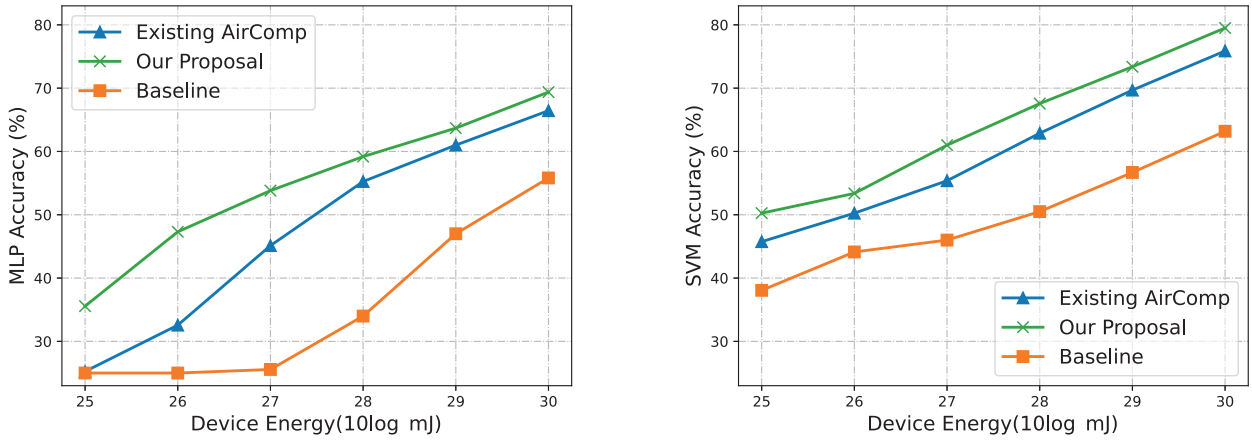
where

$$x_\ell(m) \sim \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2\right). \tag{64}$$

Then by substituting it into (13), the local feature element can be rewritten as

$$\begin{aligned} x_k(m) &= \frac{1}{L}\sum_{\ell=1}^{L} x_\ell(m) + c_{s,k}(m) + \frac{n_r(m)}{\sqrt{P_{s,k}}} \\ &= \frac{1}{L}\sum_{\ell=1}^{L} x_{\ell,k}(m), \end{aligned} \tag{65}$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3306465



(a) Inference accuracy of MLP versus device total energy

(b) Inference accuracy of SVM versus device total energy

Fig. 8. Inference accuracy of different algorithms with different device total energy on MLP and SVM models

where $x_{\ell,k}(m) = x_\ell(m) + c_{s,k}(m) + n_r(m)/\sqrt{P_{s,k}}$. Thus, according to (12), (15) and (64), we can obtain the distribution of $x_{\ell,k}(m)$

$$x_{\ell,k}(m) \sim \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2 + \sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right), \ \forall(k,\ell). \quad (66)$$

Finally, the distribution of local feature element m of device k is given by

$$x_k(m) \sim \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2 + \sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right), \ 1 \le k \le K. \quad (67)$$

B. Proof of Lemma 2

Denote the optimal solution of the new problem $\mathbf{P3}'$ as $\left\{\{P_{s,k}^*\}, \{c_{k,m}^*\}, \{\mathbf{f}_m^*\}, \alpha^*\right\}$. Assume $\exists m' \in [1, M]$ so that $\mathbf{f}_{m'}^*$ satisfy the following strict inequality:

$$c_{k,m'}^{*\,2} < \mathbf{h}_{k,m'}^{*\,H}\mathbf{f}_{m'}^*\mathbf{f}_{m'}^{*\,H}\mathbf{h}_{k,m'}^* u_{k,m'}. \quad (68)$$

Then, based on the continuity of quadratic function on the right-hand side of (68) and for a fixed $c_{k,m'}^*$, there always exists a number $\eta > 0$ such that

$$\mathbf{f}_{m'-}^* = (1-\eta)\mathbf{f}_{m'}^* \prec \mathbf{f}_{m'}^*, \quad (69)$$

which leads to

$$\begin{aligned} c_{k,m'}^{*\,2} &< \mathbf{h}_{k,m'}^{*\,H}\mathbf{f}_{m'-}^*\mathbf{f}_{m'-}^{*\,H}\mathbf{h}_{k,m'}^* u_{k,m'} \\ &< \mathbf{h}_{k,m'}^{*\,H}\mathbf{f}_{m'}^*\mathbf{f}_{m'}^{*\,H}\mathbf{h}_{k,m'}^* u_{k,m'}, \end{aligned} \quad (70)$$

where $\mathbf{x} \prec \mathbf{y}$ represents $\mathbf{x}$ is element-wise less than $\mathbf{y}$. By substituting $\mathbf{f}_{m'-}^*$ for the pair-wise discriminant gain constraint, the value of α can be increased to derive a better optimal value of $\mathbf{P3}$, which means that $\mathbf{f}_{m'-}^*$ is the optimal solution instead of $\mathbf{f}_{m'}^*$. However, this is a contradiction of the fact that $\mathbf{f}_m^*$ is the optimal solution of $\mathbf{P3}'$. Thus, the problem extended the constraint (46) achieves the same optimal solution as $\mathbf{P3}$.

C. Proof of Lemma 3

It is quite apparent that the objective function, the first and second constraints of $\mathbf{P4}$ are all affine functions. Additionally, $R_{k,m}(\mathbf{f}_m)$ is quadratic, which are convex and differentiable. Thus, we only need to prove that $c_{k,m}^2/u_{k,m}$, $Z_m(\{P_{s,k}\}, \{c_{k,m}\}, \mathbf{f}_m)$ and $Q_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})$ are convex and differentiable.

Denote $f(x,y) = x^2/y$ with a positive y, we can derive the Hessian matrix

$$\mathbf{H}_f = \begin{bmatrix} \frac{2}{y} & -\frac{2x}{y^2} \\ -\frac{2x}{y^2} & \frac{2x^2}{y^3} \end{bmatrix}, \quad (71)$$

where the eigenvalues are

$$\lambda_1 = 0, \quad \lambda_2 = \frac{2(x^2+y^2)}{y^3}.$$

Since $y > 0$, both eigenvalues of $\mathbf{H}_f$ are non-negative, which indicates the Hessian matrix is positive semidefinite and thus $f(x,y)$ is convex.

By taking $x = c_{k,m}$ and $y = u_{k,m}$, it can be proved that $c_{k,m}^2/u_{k,m}$ is convex. Function $Z_m(\{P_{s,k}\}, \{c_{k,m}\}, \mathbf{f}_m)$ is composed of three parts, the first part of which is the sum of $f(x,y)$ with $x = c_{k,m}$ and $y = P_{s,k}$ and the latter two parts are both quadratic. It follows that $Z_m(\{P_{s,k}\}, \{c_{k,m}\}, \mathbf{f}_m)$ is convex and differentiable since linear transformation does not violate the convexity. Similar to $Z_m(\{P_{s,k}\}, \{c_{k,m}\}, \mathbf{f}_m)$, function $Q_{\ell,\ell',m}(\{c_{k,m}\}, v_{\ell,\ell',m})$ can also be transformed from $f(x,y)$, which proves the convexity and differentiability. Thus, the third and fourth constraints are in the form of difference of convex functions and $\mathbf{P4}$ is a d.c. problem.

REFERENCES

[1] 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, 2019.

[2] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, 2020.

[3] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, "6G internet of things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, 2022.

[4] 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, 2020.

[5] M. Xu, H. Du, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, A. Jamalipour, D. I. Kim, V. Leung *et al.*, "Unleashing the power of edge-cloud generative ai in mobile networks: A survey of AIGC services," *arXiv preprint arXiv:2303.16129*, 2023.

[6] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Letaief, "Federated learning via intelligent reflecting surface," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 808–822, 2022.

[7] Z. Liu, Q. Lan, and K. Huang, "Resource allocation for multiuser edge inference with batching and early exiting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1186–1200, 2023.

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

[9] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, 2022.

[10] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6G: Vision, principles, and technologies," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 78–85, 2023.

[11] M. Lee, G. Yu, and H. Dai, "Decentralized inference with graph neural networks in wireless communication systems," *IEEE Trans. Mobile Comput.*, vol. 22, no. 5, pp. 2582–2598, 2023.

[12] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[13] B. Lu, J. Yang, J. Xu, and S. Ren, "Improving QoE of deep neural network inference on edge devices: A bandit approach," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21 409–21 420, 2022.

[14] 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, 2020.

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

[16] X. Yang, S. Hua, Y. Shi, H. Wang, J. Zhang, and K. B. Letaief, "Sparse optimization for green edge AI inference," *J. Commun. Inf. Netw.*, vol. 5, no. 1, pp. 1–15, 2020.

[17] W. Shi, Y. Hou, S. Zhou, Z. Niu, Y. Zhang, and L. Geng, "Improving device-edge cooperative inference of deep learning via 2-step pruning," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2019, pp. 1–6.

[18] 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)*, 2021, pp. 8488–8492.

[19] T. Niu, Y. Teng, Z. Han, and P. Zou, "An adaptive device-edge co-inference framework based on soft actor-critic," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 2571–2576.

[20] J. Yan, S. Bi, and Y.-J. A. Zhang, "Optimal model placement and online model splitting for device-edge co-inference," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8354–8367, 2022.

[21] S. H. Shabbeer Basha, S. N. Gowda, and J. Dakala, "A simple hybrid filter pruning for efficient edge inference," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 3398–3402.

[22] X. Zhang, J. Shao, Y. Mao, and J. Zhang, "Communication-computation efficient device-edge co-inference via AutoML," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2021, pp. 01–06.

[23] 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, 2020.

[24] Y. Wang, J. Shen, T.-K. Hu, P. Xu, T. Nguyen, R. Baraniuk, Z. Wang, and Y. Lin, "Dual dynamic inference: Enabling more efficient, adaptive, and controllable deep inference," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 4, pp. 623–633, 2020.

[25] 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, 2021.

[26] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, 2022.

[27] ——, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 73–87, 2023.

[28] D. Wen, X. Jiao, P. Liu, G. Zhu, Y. Shi, and K. Huang, "Task-oriented over-the-air computation for multi-device edge AI," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[29] G. Zhu, Z. Lyu, X. Jiao, P. Liu, M. Chen, J. Xu, S. Cui, and P. Zhang, "Pushing AI to wireless network edge: an overview on integrated sensing, communication, and computation towards 6G," *Sci. China Inf. Sci.*, vol. 66, no. 3, p. 130301, 2023.

[30] Q. Lan, Q. Zeng, P. Popovski, D. Gündüz, and K. Huang, "Progressive feature transmission for split classification at the wireless edge," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3837–3852, 2023.

[31] D. Wen, P. Liu, G. Zhu, Y. Shi, J. Xu, Y. C. Eldar, and S. Cui, "Task-oriented sensing, computation, and communication integration for multi-device edge AI," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[32] Y. Hu, Z. Li, Y. Chen, Y. Cheng, Z. Cao, and J. Liu, "Content-aware adaptive device-cloud collaborative inference for object detection," *IEEE Internet Things J.*, pp. 1–1, 2023.

[33] D. Wen, X. Li, Y. Zhou, Y. Shi, S. Wu, and C. Jiang, "Integrated sensing-communication-computation for edge artificial intelligence," *arXiv preprint arXiv:2306.01162*, 2023.

[34] Y. Tang, G. Zhu, W. Xu, M. H. Cheung, T.-M. Lok, and S. Cui, "Integrated sensing, computation, and communication for UAV-assisted federated edge learning," *arXiv preprint arXiv:2306.02990*, 2023.

[35] G. Li, S. Wang, K. Ye, M. Wen, D. W. K. Ng, and M. Di Renzo, "Multi-point integrated sensing and communication: Fusion model and functionality selection," *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2660–2664, 2022.

[36] H. Yar, A. S. Imran, Z. A. Khan, M. Sajjad, and Z. Kastrati, "Towards smart home automation using IoT-enabled edge-computing paradigm," *Sensors*, vol. 21, no. 14, 2021.

[37] X. Cheng, D. Duan, S. Gao, and L. Yang, "Integrated sensing and communications (ISAC) for vehicular communication networks (VCN)," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 23 441–23 451, 2022.

[38] Z. Du, F. Liu, Y. Li, W. Yuan, Y. Cui, Z. Zhang, C. Masouros, and B. Ai, "Towards ISAC-empowered vehicular networks: Framework, advances, and opportunities," *arXiv preprint arXiv:2305.00681*, 2023.

[39] Z. Wang, Y. Zhao, Y. Zhou, Y. Shi, C. Jiang, and K. B. Letaief, "Over-the-air computation: Foundations, technologies, and applications," *arXiv preprint arXiv:2210.10524*, 2022.

[40] 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, 2019.

[41] D. Wen, G. Zhu, and K. Huang, "Reduced-dimension design of MIMO over-the-air computing for data aggregation in clustered IoT networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5255–5268, 2019.

[42] 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, 2021.

[43] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Ph.D. dissertation, University of Minnesota, 2014.

[44] G. Li, S. Wang, J. Li, R. Wang, X. Peng, and T. X. Han, "Wireless sensing with deep spectrogram network and primitive based autoregressive hybrid channel model," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2021, pp. 481–485.

[45] J. A. Zhang, M. L. Rahman, K. Wu, X. Huang, Y. J. Guo, S. Chen, and J. Yuan, "Enabling joint communication and radar sensing in mobile networks—a survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 306–345, 2022.

[46] G. K. Carvajal, M. F. Keskin, C. Aydogdu, O. Eriksson, H. Herbertsson, H. Hellsten, E. Nilsson, M. Rydström, K. Vänas, and H. Wymeersch, "Comparison of automotive FMCW and OFDM radar under interference," in *Proc. IEEE Radar Conf. (RadarConf)*, 2020, pp. 1–6.

[47] 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, 2020.

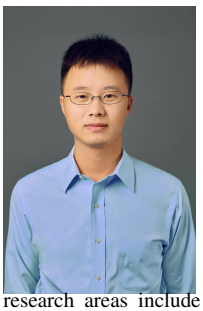
[48] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[49] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

Zeming Zhuang (Student Member, IEEE) received the B.S. degree in electrical information engineering from ShanghaiTech University, Shanghai, China, in 2021. He is currently pursuing the master's degree with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include wireless communication, edge artificial intelligent inference and integrated sensing-communication-computation.

Dingzhu Wen (Member, IEEE) is currently an assistant professor of the School of Information Science and Technology at ShanghaiTech University. He received the Bachelor degree and the Master degree from the Department (School) of Information Science and Electronic Engineering of Zhejiang University in 2014 and 2017, respectively, and received the Ph. D. degree from the Department of Electrical and Electronic Engineering of The University of Hong Kong in 2021. His research interests include federated edge learning, edge artificial intelligent inference, integrated sensing-communication-computation, over-the-air computation, in-band full-duplex communications, and device-to-device communications. He served as the session chairs of APEMC 2022 and IEEE ICC 2023, the TPC members of IEEE ICC 2023, IEEE GlobeCom 2022, IEEE VTC-Fall 2020, and IEEE WCSP 2018, and the TPC co-chair of IEEE PIMRC 2023 workshop on "Edge Learning for 5G Mobile Networks and Beyond" and the co-chair of IEEE VTC 2023-Fall workshop on "Task-Oriented Communications and Networking for 6G". He was elected as the exemplary reviewer of IEEE Transactions on Communications in 2023.



Yuanming Shi (S'13-M'15-SM'20) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011. He received 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 in ShanghaiTech University, where he is currently a tenured Associate Professor. He visited University of California, Berkeley, CA, USA, from October 2016 to February 2017. His research areas include federated learning, edge AI, and satellite networks. He was a recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2016, the Young Author Best Paper Award by the IEEE Signal Processing Society in 2016, the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2021, and the Chinese Institute of Electronics First Prize in Natural Science in 2022. He is also an editor of IEEE Transactions on Wireless Communications, IEEE Journal on Selected Areas in Communications, and Journal of Communications and Information Networks.

Guangxu Zhu (Member, IEEE) received the B.Eng. and M.Eng. degrees from Zhejiang University, and the Ph.D. degree from The University of Hong Kong in 2019. He is now a research scientist with the Shenzhen Research Institute of Big Data. His research interests include edge intelligence, distributed machine learning, and integrated sensing and communications. He is a recipient of the 2022 "AI 2000 Most Influential Scholar Award Honorable Mention", the UCOM 2023 Young Scientist Award, the Hong Kong Postgraduate Fellowship (HKPF), the Best Paper Award from WCSP 2013, and the First Prize of National "Bloom Cup" 5G Industrial Competition in 2022. He served as a track/symposium/workshop co-chair of many IEEE conferences including IEEE PIMRC 2021, WCSP 2023, IEEE Globecom 2023, and ICASSP 2024.

Sheng Wu (Member, IEEE) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, in 2014. He was a Post-Doctoral Researcher with the Tsinghua Space Center, Tsinghua University. He is currently a Professor with the Beijing University of Posts and Telecommunications. He has published more than 80 journals and conference papers, such as IEEE Journal on Selected Areas in Communications, IEEE Transactions on Wireless Communications, and IEEE Transactions on Communications. He also holds more than 30 granted patents. His research interests include iterative detection and decoding, channel estimation, massive MIMO, and satellite communications. He has received the First Prize from the Science and Technology Award of Chinese Institute of Electronics in 2017, the Silver Medal from the 46th Geneva International Exhibition of Inventions in 2018, the Second Prize from the National Technological Invention Award of China in 2018, and the Natural Science Foundation of China Excellent Young Scientists Fund Award in 2020.

Dusit Niyato (M'09-SM'15-F'17) is a professor in the School of Computer Science and Engineering, at Nanyang Technological University, Singapore. He received B.Eng. from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand in 1999 and Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada in 2008. His research interests are in the areas of sustainability, edge intelligence, decentralized machine learning, and incentive mechanism design.