Decentralized Integration of Sensing-Communication-Computation for Multi-Task Edge AI Inference

Chenye Wang, *Student Member, IEEE,* Zeming Zhuang, *Student Member, IEEE,* Dingzhu Wen, *Member, IEEE,* Yuanming Shi, *Senior Member, IEEE,* Xin Wang, *Fellow, IEEE*

***Abstract*—Collaborative artificial intelligence (AI) inference has effectively deployed well-trained AI models at the network edge to empower immersive intelligent services such as autonomous driving and smart cities. This paper proposes an integrated sensing-computation-communication (ISCC) scheme for decentralized multi-task collaborative inference systems. The proposed scheme connects multiple devices via device-to-device (D2D) links. Each device first extracts a homogeneous feature vector from the raw sensory data obtained from the same wide view of the source target and then aggregates all local feature vectors using the over-the-air computation (AirComp) technique to complete a specific inference task. To enhance spectrum efficiency, the full-duplex communication technique is adopted, which allows all devices to transmit and receive in the same frequency band. To suppress the self-interference caused by full duplex communications and simultaneously enhance all tasks' performance, a multi-objective optimization problem is formulated, where discriminant gain is adopted as the inference performance metric. The challenges to solve this problem arise from three aspects: The impact of the self-interference (SI) channel incurred by full-duplex communication, the precoding design of each device, and the coupling among subcarrier allocation, sensing, computation, and communication processes. To tackle this problem, a *quadratic transform* and *weighted bipartite matching* based alternating maximization approach is proposed. Numerical results based on jointly completing three tasks of human motion classification, human gender recognition, and human age group classification, verify the effectiveness of the proposed method by showing that the proposed method outperforms the state-of-the-art successive convex approximation (SCA) based algorithm.**

***Index Terms*—Edge AI inference, multitask optimization, integrated sensing-communication-computation, device-to-device communication, over-the-air computation**

C. Wang and X. Wang are with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), College of Future Information Technology, Fudan University, Shanghai 200438, China (e-mail: cywang23@m.fudan.edu.cn; xwang11@fudan.edu.cn).

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). *(Corresponding author: Dingzhu Wen.)*

The work of Dingzhu Wen was supported in part by the National Natural Science Foundation of China under Grant 62401369 and in part by the Shanghai Sailing Program under Grant 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 described research work is conducted in the Core Facility Platform of Computer Science and Communication provided by ShanghaiTech University.

I. Introduction

The rapid advancement of communication and computing technologies has enabled intelligent services such as autonomous vehicles and smart factories [1]–[4]. Edge artificial intelligence (AI) inference, which runs trained models on data generated by mobile devices, supports these services by enabling low-latency decision-making [5]. Among various edge inference paradigms—on-device, on-server, and edge-device collaborative inference—the latter offers a balanced approach by deploying lightweight models on devices to extract features and offloading intensive processing to edge servers [6]. This enhances privacy and reduces communication and computation costs. A primary research focus on edge-device collaborative inference is striking a balance of the trade-off between communication and computation, such as network pruning [7], early exiting mechanisms [8], feature compression [9], progressive transmission [10], designing ultra-low-latency frameworks [11], and addressing the outage issue [12]. However, most existing methods overlook the task-oriented nature of edge inference, where the accuracy and efficiency of the inference task are the ultimate goals rather than reducing communication distortion [13]. Recent works address this by proposing multi-view pooling [14], and privacy-preserving schemes using differential privacy [15].

Nevertheless, different feature elements with the same size and distortion level may impact the inference accuracy differently [6]. The existing works [7], [8], [10], [14], [15] solely considered the data transmission stage while neglecting the impact of the data acquisition process on inference performance. To achieve this, a novel task-oriented over-the-air computation (AirComp) was proposed in [6], where each edge device obtains a noise-corrupted ground-true wide-view sensory data of the same target that is further processed by feature extraction, and the server receives an aggregated feature vector by over-the-air computation to suppress the sensing noise for inference. The authors in [16] designed task-oriented communication strategies in multi-device cooperative edge inference that leverage information bottleneck (IB) principle for task-relevant local feature extraction, incorporating distributed feature encoding, where the sensing data are observed by multiple devices from different views of the same target. Different from [6] and [16], the work [17] considered a partially observable system where the target area's local observations overlap and developed an edge-cloud coopera-

tive inference architecture that decomposes an oracle cloud inference into a group of component deep neural networks (DNNs) at the cloud and DNN-aided edge encoders. However, as highlighted in [18]–[20], the processes of sensing for data acquisition, communication for information sharing, and computation for feature extraction and decision making are intricately linked in edge AI tasks. Ignoring the design of the sensing module limits the inference performance, especially in resource-limited scenarios.

To bridge this research gap, authors in [21] first mathematically characterized the coupling mechanism of the three processes in the multi-device edge-device collaborative system and then designed a task-oriented integrated sensing-communication-computation (ISCC) scheme accordingly. This design is extended to the case of reusing one sensory data sample for completing multiple tasks [22] and the case where different devices sense the same wide view, and thus the technique of over-the-air computation (AirComp) can be adopted for communication-efficient feature aggregation [23]. In addition, authors in [24] developed an integrated sensing, communication, and computation over-the-air (ISCCO) multiple-input multiple-output (MIMO) framework and addressed joint optimization of beamformers at both the IoT devices and the server based on the semidefinite relaxation technique evaluated by the mean squared error (MSE) criterion. Other techniques include developing ISCC-based inference schemes to support mode selection among multiple inference paradigms [25], enhancing the energy efficiency on devices [26], and in unmanned aerial vehicle (UAV) networks [27], and so on.

The above ISCC designs rely on a central coordinator, which is impractical in decentralized scenarios like drone swarms or cooperative automated driving [28]. In such cases, devices need to connect and communicate via device-to-device (D2D) links (see e.g., [29], [30]) to share their features for inference tasks. Consider the scenario where each device aggregates all the local feature vectors extracted from the sensory data of each device. The basic idea of sequentially aggregating local features to all devices from others causes a high communication overhead. To mitigate this, the technique of full-duplex (FD) communication is integrated with AirComp by [31] to allow all devices to transmit and receive signals simultaneously. Taking advantage of the one-shot aggregation, the above FD AirComp technique is adopted in this work. However, the technique in [31] ignores the task-oriented property and the coupling mechanism among sensing, communication, and computation in edge inference tasks, making it difficult to achieve the high-performance requirements.

To tackle these shortages, in this work, we propose a decentralized multi-task inference framework that integrates sensing, communication, and FD AirComp. All devices extract local feature vectors from their data wirelessly sensed from the same wide view of a target and employ FD AirComp for simultaneous feature sharing and aggregation to complete inference tasks. Particularly, an orthogonal frequency division multiplexing (OFDM) based broadband channel is considered, where each subcarrier is assigned to one dimension of feature element for AirComp aggregation such that the most important feature elements are assigned to subcarriers with good channels across all devices, thereby further improving inference performance. The performance enhancement of all tasks faces two technical challenges. One is the tight coupling of sensing, communication, and on-device computation. The distortion incurred by these three processes impacts the quality of the received data of each device which determines the inference accuracy, but they compete for network resources for enhancing their respective qualities. The other arises from the competition among different tasks. Since different feature elements have varying influences on different tasks, the precoding of AirComp on each device is hard to meet the feature elements transmission requirements of all tasks with high quality. To overcome these challenges, we propose a decentralized ISCC system for multi-task collaborative inference. The key contributions are summarized as follows.

- **Novel decentralized ISCC framework for multitask collaborative inference:** In this system, each device in this decentralized ISCC system is equipped with a dual-functional-radar-communication (DFRC) system including multiple transmit and receive antennas used both for sensing and communication. All devices sense the target in the same wide view and derive noisy sensory data through transmitting a frequency modulation continuous wave (FMCW) signal. Task-specific local feature extraction is conducted at each device to facilitate the fulfillment of each task, which are then precoded before they are shared through full-duplex communication and AirComp [32] over an OFDM-based broadband channel. To enhance inference performance, each OFDM subcarrier is assigned to a unique feature element during FD AirComp. The discriminant gain metric [21] is adopted to quantify inference accuracy, theoretically characterizing the impact of each feature transmission procedure.
- **Joint Subcarrier, Sensing Power Allocation with Multicast and Receive Beamforming:** Building on the proposed framework, we formulate a joint optimization problem of subcarrier allocation, sensing power allocation, multicast beamforming, and receive beamforming under the discriminant gain metric. To address the challenges introduced by self-interference (SI) channels and the coupling between sensing, computation, and communication, we first design multicast beamforming vectors to exploit SI channels and aggregate local and received features. Then, we jointly design the receive beamformers and propose an alternating optimization approach that leverages *quadratic transform* and *weighted bipartite matching* to tackle the multiple-ratio fractional programming (FP) problem with mixed-integer non-convex constraints, yielding an efficient sub-optimal solution.
- **Performance Evaluation:** To evaluate the effectiveness of the proposed framework and algorithm, we perform extensive experiments on the University of Glasgow Radar Signature dataset [33], with different inference models, i.e., multi-layer perception (MLP) neural network and K-Nearest Neighbour (KNN) models. The experiment results validate that our proposed method outperforms the successive convex approximation (SCA) based algorithm.

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.2025.3625389

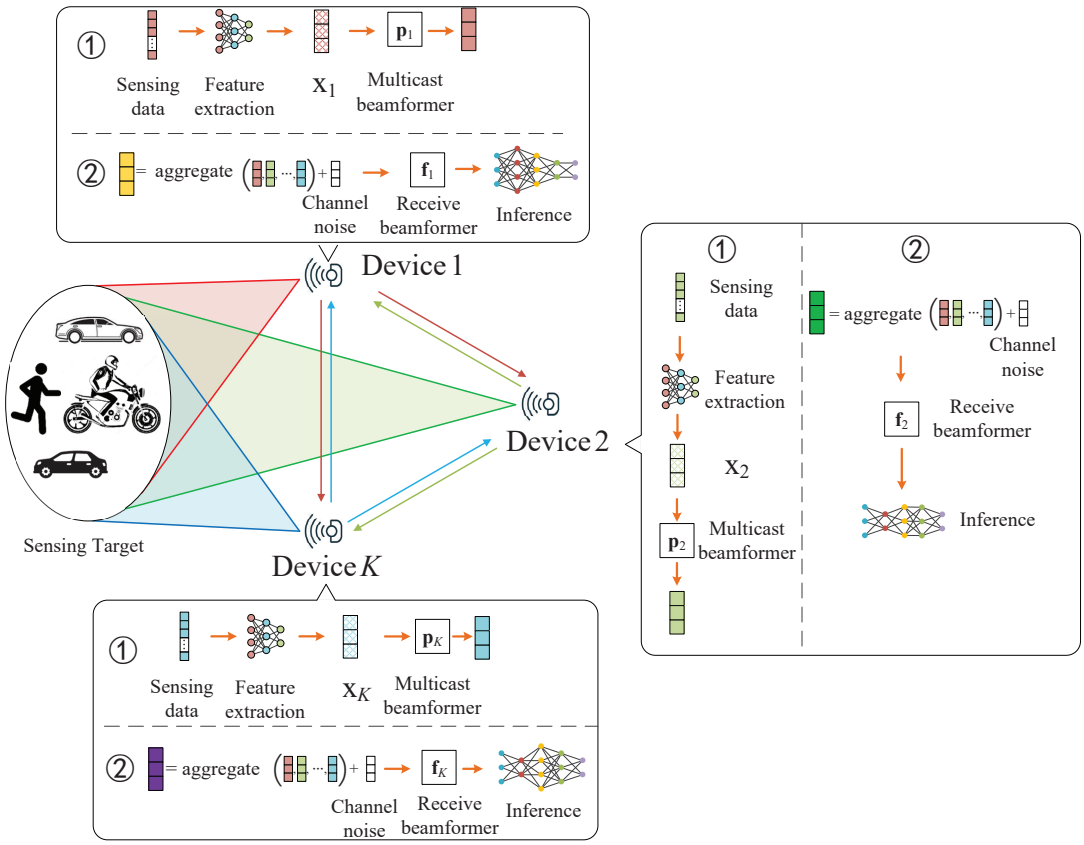


Fig. 1. The system architecture of the proposed multi-task ISCC system.

The remainder of this paper is organized as follows: Section II introduces the system model of decentralized full-duplex ISCC collaborative inference. Section III formulates the optimization problem with the objective of discriminant gain maximization. Section IV proposes a *quadratic transform* and *weighted bipartite matching* based alternating maximization method for the optimization problem. In Section V, we present numerical results to verify the effectiveness of the proposed method, followed by the conclusion in Section VI.

II. System Model

Consider a decentralized network of K ISAC devices for co-inference between devices with a central server not always available for devices, as illustrated in Fig. 1. Each device has a DFRC system containing N_t transmit antennas and N_r receive antennas. To complete the inference task, the proposed decentralized ISCC system operates through four sequential phases as shown in Fig. 2 and elaborated below.

1) Sensing: Each device transmits an FMCW signal and receives the echo signal reflected from the target.
2) Feature extraction: The raw data are processed through clutter cancellation and principal component analysis (PCA) for task-specific feature extraction, followed by the feature transmission over OFDM subcarriers.
3) Feature sharing and aggregation via FD AirComp: Utilizing full-duplex communication and the AirComp technique, every device multicasts its local features to all other devices and aggregates features from other devices simultaneously to derive a denoised global feature vector.
4) Multi-task inference: Finally, the aggregated feature vector of each device is fed into a pre-trained AI model to jointly complete the multi-task inference.

Note that the timeline diagram of different phases during latency T in Fig. 2 shows that the sensing phase and communication phase of each device k are operated in separate time frames. In particular, since FMCW is used for sensing, and there is no self-interference at the sensing phase. At the communication phase, the OFDM technique is leveraged to transmit all feature elements at the same time during feature sharing and aggregation. All M dimensions of local feature vectors are transmitted in M orthogonal subcarriers, where the m-th dimension of local feature vectors is selected and transmitted via the i-th subcarrier OFDM frequency subcarrier. Given that the duration to transmit one feature element is significantly shorter than the channel coherence time [34], channels are assumed to be static within a single time slot. All devices are assumed to have the channel state information (CSI) of links connecting to all other devices. This can be

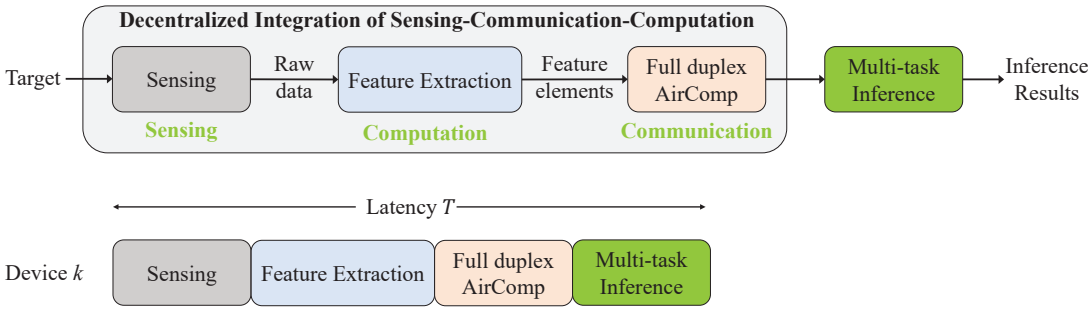


Fig. 2. Decentralized integrated sensing-computation-communication for multi-task inference framework.

achieved by a logical control node[1] collecting global information or exploiting channel reciprocity and efficient feedback [36], where the required training overhead is assumed to be negligible compared to the transmitted features. Finally, the aggregated feature vector is fed into a pre-trained AI model to perform the inference task.

A. Sensory Signal Processing and Feature Extraction

We adopt the models for sensing signal processing and feature extraction as proposed in [21]. During the radar sensing stage, each device senses the target in the same wide view by transmitting an FMCW signal $s_k(t)$ using one transmit antenna in total sensing time T_s and receives the echo signal reflected from the target using N_r receive antennas. Note that employing a single transmit antenna for radar sensing avoids the need to address mutual interference among multiple transmitters, while also reducing power consumption during the sensing phase. The received signal of ISAC device k reflected from the target is given by

$$\mathbf{r}_k(t) = \mathbf{u}_k(t) + \sum_{j=1}^{J} \mathbf{v}_{k,j}(t) + \mathbf{n}_{\mathsf{r}}(t), \tag{1}$$

where $\mathbf{u}_k(t) = \mathbf{h}_{\mathsf{s},k}(t)s_k(t-\tau)$ is the desired signal for completing the inference task with $\mathbf{h}_{s,k}(t) = [h_{s,k}^{(1)}(t), \ldots, h_{s,k}^{(N_r)}(t)]^H$ being the reflection coefficient vector of the target and τ being the round-trip delay, $\mathbf{v}_{k,j}(t) = \mathbf{c}_{\mathsf{r},k,j}(t)s_k(t-\tau_j)$ is the clutter of j-th indirect reflection path with $\mathbf{c}_{\mathsf{s},k,j}(t) = [C_{\mathsf{s},k,j}^{(1)}(t), \ldots, C_{\mathsf{s},k,j}^{(N_r)}(t)]^H$ being the round-trip coefficient vector of path j and τ_j being the delay of the j-th path, and $\mathbf{n}_{\mathsf{r}}(t)$ is the white Gaussian noise. It is assumed that $\mathbf{h}_{\mathsf{s},k}(t)$ and $\mathbf{c}_{\mathsf{s},k,j}(t)$ are estimated before the inference task. The desired signal from the echo signal reflected from the target is polluted by the additive sensing clutter incurred by higher-order (indirect) reflected paths and sensing noise. Subsequently, the following processing steps are taken at each n-th receive antenna to create a training data sample: sampling and reshaping into a matrix, SVD-based linear filter for clutter cancellation, time-frequency analysis using short-time Fourier transform (STFT), and vectorization and normalization. Following [6], [21], the PCA-based linear extractor is used to extract the local feature vector from clutter-canceled sensory data. The PCA is pre-performed at a server before the inference task using the training dataset. Then, the template of the M principal eigen-subspace is sent to all devices for extracting the local feature vectors at n-th receive antenna $\{\bar{\mathbf{r}}_k^{(n)} \in \mathbb{R}^M\}$ with M being the number of extracted feature elements. Since the clutter cancellation and feature extraction processes are linear, the m-th feature element of $\bar{\mathbf{r}}_k^{(n)}$ is given by

$$\tilde{r}_k^{(n)}(m) = \tilde{u}_k^{(n)}(m) + \sum_{j=1}^{J} \tilde{v}_{k,j}^{(n)}(m) + n_{\mathsf{r}}^{(n)}(m), m = 1, \ldots, M, \tag{2}$$

where $\tilde{u}_k^{(n)}(m)$ is the ground-truth of feature m, $\tilde{v}_{k,j}^{(n)}(m)$ is the clutter of j-th path in J paths, $n_{\mathsf{r}}^{(n)}(m)$ is the noise with Gaussian distribution $\mathcal{N}\left(0, \sigma_{\mathsf{r}}^2\right)$. Next, each feature element of device k is normalized by its sensing power $P_{\mathsf{s},k}$ and the normalized feature element m at n-th receive antenna is

$$x_k^{(n)}(m) = \frac{\tilde{r}_k^{(n)}(m)}{\sqrt{P_{\mathsf{s},k}}} = x^{(n)}(m) + c_{\mathsf{s},k}^{(n)}(m) + \frac{n_{\mathsf{r}}^{(n)}(m)}{\sqrt{P_{\mathsf{s},k}}}, \tag{3}$$

where $x^{(n)}(m) = \tilde{u}_k^{(n)}(m)/\sqrt{P_{\mathsf{s},k}}$ is the normalized ground-truth feature and $c_{\mathsf{s},k}^{(n)}(m) = \sum_{j=1}^{J}\left(\tilde{v}_{k,j}^{(n)}(m)/\sqrt{P_{\mathsf{s},k}}\right)$ denotes the normalized clutter. Since clutter is rich scattering, J is very large, and $c_{\mathsf{s},k}^{(n)}(m)$ follows zero-mean Gaussian distribution $\mathcal{N}\left(0, \sigma_{\mathsf{s},k,n}^2\right)$ according to the central limit theorem. Finally, we select the receive antenna index n^* that has minimum $\sigma_{\mathsf{s},k,n}^2$, i.e., $n^* = \arg\min \sigma_{\mathsf{s},k,n}^2$, for feature transmission

$$x_k^{(n^*)}(m) = x^{(n^*)}(m) + c_{\mathsf{s},k}^{(n^*)}(m) + \frac{n_{\mathsf{r}}^{(n^*)}(m)}{\sqrt{P_{\mathsf{s},k}}}. \tag{4}$$

For the following, we omit the superscript (n^*) for notation simplicity. Therefore,

$$x_k(m) = x(m) + c_{\mathsf{s},k}(m) + \frac{n_{\mathsf{r}}(m)}{\sqrt{P_{\mathsf{s},k}}} \tag{5}$$

is adopted for feature transmission, where $\sigma_{\mathsf{s},k} \triangleq \sigma_{\mathsf{s},k,n^*}$ is the standard deviation of $c_{\mathsf{s},k}(m)$.

Consider a classification task deployed on device k with L_k classes. Following [6], [21], the ground-truth feature vector

[1]The logical control node refers to a network device assigned to manage global information [35]. Different from a central server adopted to aggregate devices' features, the logical control node only requires little essential network information, whose signaling overhead is much smaller than feature transmission.

$\mathbf{x} = \{x(m)\}_{m=1}^{M}$ is assumed to follow a Gaussian mixture distribution. Since PCA is performed, different elements of the ground-truth feature vector are independent. Specifically, the distribution of element $x(m)$ at k-th device is given as

$$f_k(x(m)) = \frac{1}{L_k}\sum_{\ell=1}^{L_k} f_\ell(x(m)), \tag{6}$$

where $f_\ell(x(m)) = \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2\right)$ denotes the probability density function of the Gaussian component associated with the ℓ-th class, in which $\mu_{\ell,m}$ represents the centroid of class ℓ and σ_m^2 denotes the corresponding variance. These parameters are pre-estimated from the training dataset. Based on the expression in (6) and the distribution for clutter and noise, the distribution of the m-th local feature element at device k can be derived as

$$x_k(m) \sim \frac{1}{L_k}\sum_{\ell=1}^{L_k}\mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2 + \sigma_{\mathsf{s},k}^2 + \frac{\sigma_{\mathsf{r}}^2}{P_{\mathsf{s},k}}\right). \tag{7}$$

B. Broadband Decentralized AirComp

As illustrated in Fig. 1, in the decentralized co-inference system, every device needs to multicast and receive local features with all other devices to collect a denoised feature. To satisfy the demand of sharing features with every other device as described above, the technique of AirComp [32] has been used for transmitting and aggregating the feature elements over each subcarrier. Besides, conventional sequential feature aggregation across devices results in communication latency scaling linearly with the number of devices K. To this end, we leverage full-duplex communication ([31], [37]) for feature aggregation. In this setup, devices employ multicast beamforming to transmit their local features in parallel while simultaneously receiving the aggregated signals from other devices. In this way, our approach reduces the transmission latency to the order of $\mathcal{O}(K)$ compared to the conventional method. Time synchronization is assumed to be achieved through a common reference clock.

Specifically, the OFDM mechanism is leveraged where the bandwidth of the system consists of M orthogonal subcarriers aggregating M dimensions of local feature vectors. Each device is equipped with N_t transmit and N_r receive antenna arrays. However, the effects of the frequency-selective fading make different subcarriers experience different channel gains. We denote $\mathbf{H}_{j,k}^{(i)} \in \mathbb{C}^{N_r\times N_t}$ as the the channel gain on the i-th subcarrier between device j and device k. Also, we define $a_m^{(i)} \in \{0,1\}$ as the subcarrier allocation indicators where $a_m^{(i)} = 1$ represents that the i-th subcarriers is allocated for the transmission of the m-th feature dimention, otherwise $a_m^{(i)} = 0$. For each device j, the local feature element $x_j(m)$ is modulated by a multicast beamformer $\mathbf{p}_{j,m} \in \mathbb{C}^{N_t}$ before it is transmitted over the MIMO channel to all other devices. We assume all devices have the perfect CSI. The received signal vector at device k is given by

$$\begin{aligned}\mathbf{y}_k(m) = &\underbrace{\sum_{i=1}^{M} a_m^{(i)}\mathbf{H}_{k,k}^{(i)}\mathbf{p}_{k,m}x_k(m)}_{\text{Residual SI}} \\ &+ \sum_{j=1,j\neq k}^{K}\sum_{i=1}^{M} a_m^{(i)}\mathbf{H}_{j,k}^{(i)}\mathbf{p}_{j,m}x_j(m) + \mathbf{w}_k(m).\end{aligned} \tag{8}$$

In particular, $\mathbf{H}_{k,k}^{(i)}, i = 1,\ldots,M$, represents the channel gain of device k's self-interference channel. Here, $\mathbf{w}_k(m) \in \mathbb{C}^{N_r}$ is the additive white Gaussian noise which follows distribution $\mathbf{w}_k(m) \sim \mathcal{CN}(0, N_0\mathbf{I}_{N_r})$. To aggregate the features $\{x_j(m)\}_{j\neq k}$ transmitted from other devices with its local feature $x_k(m)$, the multicast beamforming is designed to exploit the self-interference channel. As stated, the channel matrix $\mathbf{H}_{j,k}^{(i)}$ remains invariant throughout the feature aggregation.

After receiving the feature, a receive beamformer $\mathbf{f}_{k,m} \in \mathbb{C}^{N_r}$ is applied to extract the feature. By taking the real part of the processed signal, the m-th aggregated feature element at device k is recovered

$$\begin{aligned}\hat{x}_k(m) &= \Re\left[\mathbf{f}_{k,m}^H\mathbf{y}_k(m)\right] \\ &= \Re\left[\sum_{j=1}^{K}\sum_{i=1}^{M}\mathbf{f}_{k,m}^H a_m^{(i)}\mathbf{H}_{j,k}^{(i)}\mathbf{p}_{j,m}x_j(m)\right] + \Re\left[\mathbf{f}_{k,m}^H\mathbf{w}_k(m)\right].\end{aligned} \tag{9}$$

The distribution of $\hat{x}_k(m)$ can also be derived as

$$f_k\left(\hat{x}_k(m)\right) = \frac{1}{L_k}\sum_{\ell=1}^{L_k} f_\ell\left(\hat{x}_k(m)\right), \tag{10}$$

where

$$f_\ell\left(\hat{x}_k(m)\right) = \mathcal{N}\left(\hat{\mu}_{\ell,k,m}, \hat{\sigma}_{k,m}^2\right) \tag{11}$$

with

$$\hat{\mu}_{\ell,k,m} = \mu_{\ell,m}\Re\left[\sum_{j=1}^{K}\sum_{i=1}^{M}\mathbf{f}_{k,m}^H a_m^{(i)}\mathbf{H}_{j,k}^{(i)}\mathbf{p}_{j,m}\right], \tag{12}$$

and

$$\begin{aligned}\hat{\sigma}_{k,m}^2 = &\sigma_m^2\left(\Re\left[\sum_{j=1}^{K}\sum_{i=1}^{M}\mathbf{f}_{k,m}^H a_m^{(i)}\mathbf{H}_{j,k}^{(i)}\mathbf{p}_{j,m}\right]\right)^2 \\ &+ \sum_{j=1}^{K}\left(\Re\left[\sum_{i=1}^{M}\mathbf{f}_{k,m}^H a_m^{(i)}\mathbf{H}_{j,k}^{(i)}\mathbf{p}_{j,m}\right]\right)^2\left(\sigma_{\mathsf{s},j}^2 + \frac{\sigma_{\mathsf{r}}^2}{P_{\mathsf{s},j}}\right) \\ &+ \frac{N_0}{2}\|\mathbf{f}_{k,m}\|^2.\end{aligned} \tag{13}$$

Note that (12) and (13) characterize the impact of FD communication and aggregation on the $\hat{x}_k(m)$. In particular, consider imperfect CSI, e.g., $\hat{\mathbf{H}}_{j,k}^{(i)} = \mathbf{H}_{j,k}^{(i)} + \Delta\mathbf{H}_{j,k}^{(i)}$, with $\hat{\mathbf{H}}_{j,k}^{(i)}$ being estimated channel and $\Delta\mathbf{H}$ being the statistic CSI error. If each element of $\hat{\mathbf{H}}_{j,k}^{(i)}$ satisfies i.i.d. $\mathcal{CN}(0, \sigma_{\Delta\mathbf{H}}^2)$ and is independent of $x_j(m)$, one can reformulate (12) and (13) for further derivation.

III. Problem Formulation

A. Discriminant Gain

In this work, we employ the summation of pair-wise discriminant gains as introduced in [23] to balance inference performance by evaluating the distribution of received features characterized in (10). For a given classification task, the pair-wise discriminant gain between any two classes (ℓ, ℓ') is quantified by the symmetric Kullback–Leibler (KL) divergence between their corresponding Gaussian distributions [10]. Specifically, for the m-th feature component received from device k, the discriminant gain between class ℓ and class ℓ' is defined as

$$\begin{aligned} G_{\ell,\ell',k,m} \triangleq & D_{KL}\left[f_\ell\left(\hat{x}_k(m)\right) \| f_{\ell'}\left(\hat{x}_k(m)\right)\right] \\ & + D_{KL}\left[f_{\ell'}\left(\hat{x}_k(m)\right) \| f_\ell\left(\hat{x}_k(m)\right)\right], \\ = & \int_{\hat{x}_k(m)} \left[f_\ell\left(\hat{x}_k(m)\right) \log\left[\frac{f_\ell\left(\hat{x}_k(m)\right)}{f_{\ell'}\left(\hat{x}_k(m)\right)}\right] \right. \\ & \left. + f_{\ell'}\left(\hat{x}_k(m)\right) \log\left[\frac{f_{\ell'}\left(\hat{x}_k(m)\right)}{f_\ell\left(\hat{x}_k(m)\right)}\right] \right] \mathrm{d}\hat{x}_k(m), \end{aligned} \tag{14}$$

where $D_{KL}\left[\mathrm{p}\|\mathrm{q}\right]$ denotes the KL divergence between probability distributions p and q. Note that the discriminant gain can be adapted to other popular tasks like regression by quantizing a regression problem into a classification problem. Specifically, instead of pinpointing the value, estimate the probability of its value belonging to a bin, i.e., classify each sample into a bin. This reformulation of regression as classification has also led to superior performance in the fields of age estimation and pose estimation [38]. Moreover, we assume that each element of the latent vector follows a Gaussian distribution, which has already been supported by [21], [22], [23]. This assumption is theoretically supported by the central limit theorem and the maximum-entropy property of the Gaussian distribution. It is also a standard modeling choice in latent-variable frameworks such as variational autoencoders and factor analysis, as it enables analytically tractable and stable inference. Although real data may deviate from Gaussianity, this approximation is widely adopted in practice and can be extended to Gaussian mixtures when greater flexibility is required.

The metric $G_{\ell,\ell',k,m}$ effectively captures how distinguishable the two classes ℓ and ℓ' are in the feature domain. A higher value of this gain suggests that the respective class distributions are more separable, which directly contributes to improved classification accuracy. Consequently, enhancing the pair-wise discriminant gain helps to better resolve the most confusable class pairs. Given that individual features $\hat{x}_k(m)$ are statistically independent, the overall discriminant gain for the feature vector $\hat{\mathbf{x}}_k = [\hat{x}_k(1), \ldots, \hat{x}_k(m), \ldots, \hat{x}_k(M)]^T$ is expressed as

$$G_k\left(\hat{\mathbf{x}}_k\right) \triangleq \sum_{m=1}^{M} \sum_{\ell'=1}^{L_k} \sum_{\ell<\ell'} G_{\ell,\ell',k,m}. \tag{15}$$

Our objective is to maximize the sum of the pair-wise discriminant gains of all devices

$$\max \quad \sum_{k=1}^{K} G_k\left(\hat{\mathbf{x}}_k\right). \tag{16}$$

In particular, the reason we use discriminant gain as the performance metric for inference tasks rather than MMSE is that the traditional MMSE criterion minimizes the overall distortion between noisy and ground-truth features but does not account for the varying significance of different features [21]. For instance, in a binary classification problem, distortions in different feature dimensions can lead to unequal impacts on classification accuracy, revealing the limitation of MMSE. To address this, the discriminant gain based on symmetric KL divergence is adopted, as it better reflects the relative importance of features by measuring the centroid distance normalized by covariance. This results in improved class separability and higher inference accuracy.

ISAC devices are typically designed for ease of deployment, which often results in limited energy and computational capabilities [21], [23]. For any given device k, the total energy consumption consists of three primary components. The first component is the sensing energy, expressed as $P_{\mathsf{s},k} T_{\mathsf{s},k}$, where $P_{\mathsf{s},k}$ is the sensing power and $T_{\mathsf{s},k}$ is the fixed sensing duration. The second component is the constant energy required for local feature extraction, denoted as $E_{\mathsf{p},k}$. The third component accounts for the transmission energy required to send the m-th feature element via AirComp, with the corresponding transmit power given by $P_{\mathsf{c},k}(m) = \mathbf{p}_{k,m}^H \mathbb{E}\left[x_k(m)\, x_k(m)^H\right] \mathbf{p}_{k,m}$.

Since the distribution of $x_k(m)$ is specified in (6), its variance is fixed and denoted by $X_k(m) = \mathbb{E}\left[x_k(m) x_k(m)^H\right]$, which is known by the devices as *a prior* information via estimation of the offline data samples.

B. Problem Formulation

The discriminant gain between class ℓ and ℓ' of the m-th received feature element on device k is formulated as

$$G_{\ell,\ell',k,m} = \frac{\left(\hat{\mu}_{\ell,k,m} - \hat{\mu}_{\ell',k,m}\right)^2}{\hat{\sigma}_{k,m}^2}, \tag{17}$$

where $\hat{\mu}_{\ell,k,m}$ and $\hat{\sigma}_{k,m}^2$ are defined in (12) and (13). Accordingly, the problem of maximizing the sum of pair-wise discriminant gains in (17) under constraints can be formulated as $\mathbf{P1}$:

$$\max_{\substack{\{P_{\mathsf{s},k}\},\{a_m^{(i)}\} \\ \{\mathbf{f}_{k,m}\},\{\mathbf{p}_{j,m}\}}} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{\ell'=1}^{L_k} \sum_{\ell<\ell'} G_{\ell,\ell',k,m} \tag{18a}$$

$$\text{s.t.} \quad P_{\mathsf{s},k} T_{\mathsf{s},k} + E_{\mathsf{p},k} + T_{\mathsf{c}} \sum_{m=1}^{M} \|\mathbf{p}_{k,m}\|^2 X_k(m) \le E_k \tag{18b}$$

$$a_m^{(i)} \in \{0,1\}, \ \forall m, i \tag{18c}$$

$$\sum_{i=1}^{M} a_m^{(i)} = 1, \ \forall m \tag{18d}$$

$$\sum_{m=1}^{M} a_m^{(i)} = 1, \ \forall i \tag{18e}$$

$$\|\mathbf{f}_{k,m}\|^2 = 1, \ \forall k, m \tag{18f}$$

where (18b) is the energy consumption constraint of device k with E_k being the energy threshold of device k and T_c is

the duration time of AirComp. The constraints (18c)∼(18e) stand for the subcarrier allocation constraints, which ensure that one subcarrier is exactly assigned to the transmission of one dimension. Also, due to the energy limitation, the receive beamforming vector $\mathbf{f}_{k,m}$ is constrained with $\|\mathbf{f}_{k,m}\|^2 = 1$ (18f) only to control the angle of arrival (AoA).

IV. Joint Subcarrier and Sensing Power Allocation with Multicast and Receive Beamforming

The objective function (18a) is to maximize the sum of several fractional functions, which is a mixed integer nonlinear sum-of-ratios problem that is difficult to solve with traditional optimization methods. In this section, we will design an effective algorithm to solve the problem formulated above. We will first transform $\mathbf{P1}$ into $\mathbf{P2}$ to simplify the formulas by leveraging subcarrier allocation constraints. Then, an alternating iterative method is developed to derive a sub-optimal solution to problem $\mathbf{P2}$.

A. Problem Transformation

It is shown in (18) that the form of the objective is highly complicated. To fully exploit the subcarrier allocation constraints, we propose the following Lemma to simplify the objective (17).

Lemma 1. *By leveraging the constraints of subcarrier allocation indicators* (18c)∼(18e)*, the equivalent discriminant gain of* (17) *can be reformulated as*

$$G_{\ell,\ell',k,m} = \frac{\left(\tilde{\mu}_{\ell,k,m} - \tilde{\mu}_{\ell',k,m}\right)^2}{\tilde{\sigma}^2_{k,m}}, \tag{19}$$

where $\left(\tilde{\mu}_{\ell,k,m} - \tilde{\mu}_{\ell',k,m}\right)^2$ *is*

$$(\mu_{\ell,m} - \mu_{\ell',m})^2 \sum_{i=1}^{M} a_m^{(i)} \cdot \left(\Re \left[\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i)} \mathbf{p}_{j,m} \right] \right)^2, \tag{20}$$

$$\begin{aligned} \tilde{\sigma}^2_{k,m} = {} & \sigma_m^2 \sum_{i=1}^{M} a_m^{(i)} \cdot \left(\Re \left[\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i)} \mathbf{p}_{j,m} \right] \right)^2 \\ & + \sum_{i=1}^{M} a_m^{(i)} \cdot \sum_{j=1}^{K} \left(\Re \left[\mathbf{f}_{k,m}^H \mathbf{H}_{j}^{(i)} \mathbf{p}_{j,m} \right] \right)^2 \left(\sigma_{\mathsf{s},j}^2 + \frac{\sigma_{\mathsf{r}}^2}{P_{\mathsf{s},j}} \right) \\ & + \frac{N_0}{2} \|\mathbf{f}_{k,m}\|^2. \end{aligned} \tag{21}$$

Proof. See in Appendix A. □

Lemma (21) decouples the subcarrier allocation indicators $a_m^{(i)}$, which transforms the $\mathbf{P1}$ into:

$$\mathbf{P2} \quad \max_{\substack{\{P_{\mathsf{s},k}\},\{a_m^{(i)}\} \\ \{\mathbf{f}_{k,m}\},\{\mathbf{p}_{j,m}\}}} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{\ell'=1}^{L_k} \sum_{\ell<\ell'} \frac{\left(\tilde{\mu}_{\ell,k,m} - \tilde{\mu}_{\ell',k,m}\right)^2}{\tilde{\sigma}^2_{k,m}} \tag{22a}$$

$$\text{s.t.} \quad \text{(18b)} \sim \text{(18f)}.$$

The transformation from $\mathbf{P1}$ to $\mathbf{P2}$ exploits the strict subcarrier allocation constraints (18c)∼(18e), which enforce a one-to-one mapping between feature elements and subcarriers. This property eliminates cross-product terms and enables the subcarrier allocation sub-problem to be reformulated as a tractable linear assignment problem. It is observed that the objective in $\mathbf{P2}$ is simplified but still appears in a sum of multiple-ratio mixed integer problem with the tightly coupled optimization variables, which is hard to deal with. We propose an alternating maximization technique to solve this problem in the following.

B. Subcarrier Allocation

For the fixed sensing power $\{P_{\mathsf{s},k}\}$, $\mathbf{f}_{k,m}$ and $\{\mathbf{p}_{j,m}\}$, we can arrive at the following problem

$$\mathbf{P3} \quad \max_{\{a_m^{(i)}\}} \quad \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{\ell'=1}^{L_k} \sum_{\ell<\ell'} \tilde{G}_{\ell,\ell',k,m} \quad \text{s.t.} \quad \text{(18c)} \sim \text{(18e)}, \tag{23}$$

where

$$\tilde{G}_{\ell,\ell',k,m} = \frac{(\mu_{\ell,m} - \mu_{\ell',m})^2 \sum_{i=1}^{M} a_m^{(i)} \tilde{d}_{k,m}^{(i)}}{\sigma_m^2 \sum_{i=1}^{M} a_m^{(i)} \tilde{d}_{k,m}^{(i)} + \sum_{i=1}^{M} a_m^{(i)} \tilde{c}_{k,m}^{(i)} + \frac{N_0}{2} \|\mathbf{f}_{k,m}\|^2} \tag{24}$$

with

$$\begin{aligned} \tilde{d}_{k,m}^{(i)} &\triangleq \left(\Re \left[\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i)} \mathbf{p}_{j,m} \right] \right)^2, \\ \tilde{c}_{k,m}^{(i)} &\triangleq \sum_{j=1}^{K} \left(\Re \left[\mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i)} \mathbf{p}_{j,m} \right] \right)^2 \left(\sigma_{\mathsf{s},j}^2 + \frac{\sigma_{\mathsf{r}}^2}{P_{\mathsf{s},j}} \right). \end{aligned} \tag{25}$$

Observing that for each pair (k,m), we can sum the ℓ, ℓ' of $\tilde{G}_{\ell,\ell',k,m}$, which yields

$$\mathbf{P4} \quad \max_{\{a_m^{(i)}\}} \quad \sum_{k=1}^{K} \sum_{m=1}^{M} \bar{G}_{k,m} \qquad s.t. \quad \text{(18c)} \sim \text{(18e)}, \tag{26}$$

where

$$\bar{G}_{k,m} = \frac{\sum_{\ell'=1}^{L_k} \sum_{\ell<\ell'} (\mu_{\ell,m} - \mu_{\ell',m})^2 \cdot \sum_{i=1}^{M} a_m^{(i)} \tilde{d}_{k,m}^{(i)}}{\sigma_m^2 \sum_{i=1}^{M} a_m^{(i)} \tilde{d}_{k,m}^{(i)} + \sum_{i=1}^{M} a_m^{(i)} \tilde{c}_{k,m}^{(i)} + \frac{N_0}{2} \|\mathbf{f}_{k,m}\|^2}. \tag{27}$$

From the above transformation, we see that both the numerator and denominator of (27) is a linear combination (plus bias) of $\{a_m^{(i)}\}$. We can rewrite (27) as

$$\bar{G}_{k,m} = \frac{\sum_{i=1}^{M} a_m^{(i)} \bar{d}_{k,m}^{(i)}}{\sum_{i=1}^{M} a_m^{(i)} \bar{c}_{k,m}^{(i)} + \frac{N_0}{2} \|\mathbf{f}_{k,m}\|^2} \tag{28}$$

where

$$\begin{aligned} \bar{d}_{k,m}^{(i)} &= \sum_{\ell'=1}^{L_k} \sum_{\ell<\ell'} (\mu_{\ell,m} - \mu_{\ell',m})^2 \tilde{d}_{k,m}^{(i)}, \\ \bar{c}_{k,m}^{(i)} &= \sigma_m^2 \tilde{d}_{k,m}^{(i)} + \tilde{c}_{k,m}^{(i)}, \end{aligned} \tag{29}$$

Note that the $\bar{G}_{k,m}$ still remains a fractional form and the objective function is still a mixed integer sum of multi ratio problem with highly coupled optimization variables. Conventional optimization methods that transform the fractional

programming into non-fractional problems like Dinkelbach or Charnes-Cooper are not suitable to solve this problem since both the denominator numerator of the objective function contain a summation over all i. To this end, we will show that the (26) can be transformed into a linear assignment problem.

Lemma 2. *By leveraging the binary assignment constraints* (18c) $\sim$ (18e)*, the objective function of* $\mathbf{P4}$ *can be transformed into a linear assignment problem:*

$$\max_{\mathbf{A}} \quad \sum_{m=1}^{M}\sum_{i=1}^{M} a_m^{(i)} w_m^{(i)} \qquad s.t. \quad (18c) \sim (18e), \tag{30}$$

where

$$w_m^{(i)} = \sum_{k=1}^{K} \frac{\bar{d}_{k,m}^{(i)}}{\bar{c}_{k,m}^{(i)} + \frac{N_0}{2}\|\mathbf{f}_{k,m}\|^2}, \tag{31}$$

and $\mathbf{A}$ *is an* $M \times M$ *binary assignment matrix containing all the entries* $a_m^{(i)}$ *that satisfies the constraints* (18c)$\sim$(18e)*.*

Proof. The proof can be found in Appendix B. □

Observing Lemma 2, we see that $\mathbf{A}$ is a *permutation* matrix since a feasible $\mathbf{A}$ has exactly one nonzero entry per row and per line and the entry is equal to 1. It is hard to work with discrete variables through a brute-force approach due to its high computational cost.

To solve this linear sum assignment problem, we leverage the idea of a graph theory model. Denote $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ as a bipartite and weighted graph where $\mathcal{U}$ and $\mathcal{V}$ are the two non-overlapping sets of M nodes such that there are no edges with both endpoints in $\mathcal{U}$ and no edges with both endpoints in $\mathcal{V}$ i.e., $|\mathcal{U}| = |\mathcal{V}| = M$ and $\mathcal{U} \cap \mathcal{V} = \varnothing$. While $\mathcal{U}$ and $\mathcal{V}$ stand for the row (element) vertex and column (subcarrier) vertex, respectively; $\mathcal{E}$ represents a set of edges connecting $\mathcal{U}$ to $\mathcal{V}$, where the cost of the $(m,i) \in \mathcal{E}$ edge is $w_m^{(i)}$. Here, a *matching* on a bipartite graph $\mathcal{G}$ is a subset of edges where no two edges are incident to the same node. The assignment problem, also known as *weighted bipartite matching problem*, is to find a *perfect matching* with the minimum total weight, i.e., find a subset of edges such that each vertex belongs to exactly one edge and the sum of the costs of corresponding edges is a minimum. Note that the maximum objective of assignment problem can be transformed by setting $\tilde{w}_m^{(i)} = \max\{w_m^{(i)}\} - w_m^{(i)}$ for $\forall m, i$ to ensure that the weights is nonnegtive, i.e.,

$$\mathbf{P5} \quad \min_{\mathbf{A}} \sum_{m=1}^{M}\sum_{i=1}^{M} a_m^{(i)} \tilde{w}_m^{(i)} \qquad s.t. \quad (18c) \sim (18e). \tag{32}$$

By introducing the dual variables $u_m, m = 1, \ldots, M$ and $v_i, i = 1, \ldots, M$, the dual problem of $\mathbf{P5}$ is

$$\max \sum_{m=1}^{M} u_m + \sum_{i=1}^{M} v_i \quad s.t.\ u_m + v_i \le \tilde{w}_m^{(i)},\ m, i = 1, \ldots, M. \tag{33}$$

On the one hand, it is easy to prove that the objective of (33) is not less than that of $\mathbf{P5}$. On the other hand, based on the complementary slackness condition, the solutions of $\mathbf{P5}$ and (33) are both optimal if and only if

$$a_m^{(i)}(\tilde{w}_m^{(i)} - u_m - v_i) = 0, \tag{34}$$

which indicates that if $u_m + v_i = \tilde{w}_m^{(i)}$ then $a_m^{(i)} = 1$, and the edge is tight. Conversely, if the current edge satisfies $u_m + v_i < \tilde{w}_m^{(i)}$, then $a_m^{(i)} = 0$. Based on (34), we can successively builds maximum (perfect) matchings on a equality graph $\mathcal{G}' = (\mathcal{U}, \mathcal{V}; \mathcal{E}')$ that satisfies $u_m + v_i = \tilde{w}_m^{(i)}, (m,i) \in \mathcal{E}'$.

To find the perfect matching on $\mathcal{G}'$, we first introduce basic graph theoretic ideas [39], [40]. An *alternating path* in a bipartite graph with respect to a matching $\mathcal{M}$ is a path whose edges are alternately in $\mathcal{M}$ and not in $\mathcal{M}$. An *augmenting path* is a simple alternating path with its initial and terminal edges being not assigned, whose unassigned edges are one more than the assigned edges. The key is searching for an augmenting path in the current partial assigned bipartite graph that only contains *rigid* edges (m,i), i.e., $u_m + v_i = \tilde{w}_m^{(i)}, (m,i) \in \mathcal{E}'$. As soon as an augmenting path $\mathcal{P}$ is found, we can increase the cardinality of the current matching $\mathcal{M}$ by one through interchanging the edges in $\mathcal{M}$ and not in $\mathcal{M}$ along $\mathcal{P}$. If there is no augmenting path with respect to $\mathcal{M}$, we find the perfect (maximal) matching on $\mathcal{G}'$. However, at the current step, if it is not possible to increase the cardinality of the current matching and the matching on M edges of $\mathcal{G}'$ is not perfect, the update to the dual variables is performed. Here, we denote $\bar{\mathcal{U}}$ as the set of vertices in $\mathcal{U}$ that were visited during the last traversal of finding the maximum matching in a equality graph, $\bar{\mathcal{V}}$ as the corresponding set of visited vertices in $\mathcal{V}$, and

$$\begin{aligned} &\Delta_i = \min_{m \in \bar{\mathcal{U}}} \{\tilde{w}_m^{(i)} - u_m - v_i\}, \quad \delta = \min_{i \notin \bar{\mathcal{V}}} \Delta_i; \\ &u_m \leftarrow u_m + \delta,\ \forall m \in \bar{\mathcal{U}}; \quad v_i \leftarrow v_i - \delta,\ \forall i \in \bar{\mathcal{V}}. \end{aligned} \tag{35}$$

Lemma 3. *This update* (35) *is designed in order to have the following effects: 1) all edges of the matching of* $\mathcal{G}'$ *will remain rigid; 2) at least one edge leaving from* $\bar{\mathcal{U}}$ *to the edges that not in* $\bar{\mathcal{V}}$ *are added to current* $\mathcal{G}'$*; 3) all edges outside the matching of* $\mathcal{G}'$ *remain rigid; 4)* $\delta > 0$*.*

Proof. See in Appendix C. □

The weighted bipartite matching approach for solving $\mathbf{P5}$ is proposed in Algorithm 1.

The computational complexity of the inner loop (lines 4 to 20) is $\mathcal{O}(M^2)$ since each iteration is performed under different cur and requires $\mathcal{O}(M)$, and computing each Δ_i requires $\mathcal{O}(M)$ time complexity. The main loop of Algorithm 1 is executed in $\mathcal{O}(M)$ times, because of selecting a different $m \notin \mathcal{T}$. Therefore, algorithm 1 has an overall $\mathcal{O}(M^3)$ complexity.

C. Alternating Maximization

For each fixed pair m, there is only one i that satisfies $a_m^{(i)} = 1$, thus we further introduced two new functions from

Algorithm 1 Weighted Bipartite Matching Approach for Subcarrier Allocation (32)

Input: Weights $\tilde{w}_m^{(i)}$
Output: Optimal subcarrier assignment $\{a_m^{(i)}\}$ obtained from $\texttt{row\_of}[i]$

Initialization: $u_m = 0$, $v_i = 0$; selected row of $i \in \mathcal{V}$ $\texttt{row\_of}[i] \leftarrow 0, \forall i$; set of matched row vertices $\mathcal{T} \leftarrow \varnothing$
while matching in $\mathcal{G}$ is not perfect **do**
  Select $m \notin \mathcal{T}$; set $\Delta_i \leftarrow \infty$, $\mathtt{p}_i \leftarrow -1$, $\bar{\mathcal{V}} \leftarrow \varnothing$, $\bar{\mathcal{U}} \leftarrow \varnothing$, $cur \leftarrow m$, $free\_col \leftarrow$ null
  while $free\_col =$ null **do**
    $\bar{\mathcal{U}} \leftarrow \bar{\mathcal{U}} \cup \{cur\}$
    for each $i \notin \bar{\mathcal{V}}$ **do**
      $r \leftarrow \tilde{w}_{cur}^{(i)} - u_{cur} - v_i$
      if $r < \Delta_i$ **then** $\Delta_i \leftarrow r$, $\mathtt{p}_i \leftarrow cur$
      if $\Delta_i = 0$ **then** $\bar{\mathcal{V}} \leftarrow \bar{\mathcal{V}} \cup \{i\}$
    end for
    if no new $i \in \bar{\mathcal{V}}$ with $\mathtt{p}_i \neq -1$ **then**
      Update u_m, v_i via (35)
      for each $i \notin \bar{\mathcal{V}}$ **do**
        $\Delta_i \leftarrow \Delta_i - \delta$
        if $\Delta_i = 0$ **then** $\bar{\mathcal{V}} \leftarrow \bar{\mathcal{V}} \cup \{i\}$
      end for
    end if
    Select $i \in \bar{\mathcal{V}}$ with $\mathtt{p}_i \neq -1$
    if $\texttt{row\_of}[i] = 0$ **then** $free\_col \leftarrow i$ **else** $cur \leftarrow \texttt{row\_of}[i]$
  end while
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{m\}$
  Reconstruct augmenting path from m to $free\_col$ and update $\texttt{row\_of}[i]$
end while

(20) and (21) as

$$\mathcal{A}^{i_m^*}_{\ell,\ell',k,m}(\mathbf{f}_{k,m}, \{\mathbf{p}_{j,m}\}_j) = (\mu_{\ell,k,m} - \mu_{\ell',k,m})^2 \left(\Re \left[\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i^*(m))} \mathbf{p}_{j,m} \right] \right)^2, \tag{36}$$

$$\mathcal{B}^{i_m^*}_{k,m}(\mathbf{f}_{k,m}, \{\mathbf{p}_{j,m}\}_j, \{P_{\mathsf{s},j}\}_j) = \sigma_m^2 \left(\Re \left[\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{((i^*(m))} \mathbf{p}_{j,m} \right] \right)^2 + \sum_{j=1}^{K} \left(\Re \left[\mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i^*(m))} \mathbf{p}_{j,m} \right] \right)^2 \left(\sigma_{\mathsf{s},j}^2 + \frac{\sigma_{\mathsf{r}}^2}{P_{\mathsf{s},j}} \right) + \frac{N_0}{2} \|\mathbf{f}_{k,m}\|^2. \tag{37}$$

where the unique index $i^*(m)$ is the one that satisfies $a_m^{(i^*(m))} = 1$. For notation simplicity, we replace $i^*(m)$ with i_m^* in the following. Substituting (36) and (37) into $\mathbf{P2}$, this problem is reformulated as follows, $\mathbf{P6}$:

$$\max_{\substack{\{\mathbf{f}_{k,m}\},\{\mathbf{p}_{j,m}\} \\ \{P_{\mathsf{s},k}\}}} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{\ell'=1}^{L_k} \sum_{\ell<\ell'} \frac{\mathcal{A}^{i_m^*}_{\ell,\ell',k,m}(\mathbf{f}_{k,m}, \{\mathbf{p}_{j,m}\})}{\mathcal{B}^{i_m^*}_{k,m}(\mathbf{f}_{k,m}, \{\mathbf{p}_{j,m}\}, \{P_{\mathsf{s},j}\})} \tag{38a}$$

$$\text{s.t.} \quad (18b), (18f)$$

Note that (18f) can be relaxed to

$$\|\mathbf{f}_{k,m}\|^2 \leq 1, \ \forall k, \ \forall m, \tag{39}$$

since the objective (38a) multiplied by a positive factor does not impact the optimization. Note that the two functions $\mathcal{A}^{i_m^*}_{\ell,\ell',k,m}(\mathbf{f}_{k,m}, \{\mathbf{p}_{j,m}\})$ and $\mathcal{B}^{i_m^*}_{k,m}(\mathbf{f}_{k,m}, \{\mathbf{p}_{j,m}\}, \{P_{\mathsf{s},j}\})$ are both convex and differentiable with respect to (w.r.t.) $\mathbf{f}_{k,m}$ and $(\{P_{\mathsf{s},j}\}, \{\mathbf{p}_{j,m}\})$, respectively. The constraint (18b) is convex and differentiable w.r.t. $(\{P_{\mathsf{s},j}\}, \{\mathbf{p}_{j,m}\})$.

However, the objective (38a) is the summation of the nonlinear multiple-ratio problem, which is hard to deal with. We propose an alternating maximization technique to solve this problem $\mathbf{P6}$ in the following.

D. Joint Sensing Power Allocation, Multicast Beamforming and Receive Beamforming Design

In this subsection, we propose a alternating method to convert (38) to two sub-problems, which alternately update the $(\{\mathbf{p}_{j,m}\}, \{P_{\mathsf{s},j}\})$ and $\{\mathbf{f}_{k,m}\}$. Specifically, the subproblems that optimizing $\{\mathbf{f}_{k,m}\}$ and $\{\mathbf{p}_{j,m}\}$ both can be sub-optimally addressed by a multiple-ratio fractional programming approach with quadratic transform.

1) Joint Sensing Power Allocation and Multicast Beamforming Design: In this case, given the subcarrier allocation indicators $\{a_m^{(i)}\}$ and receive beamforming $\mathbf{f}_{k,m}$, the *quadratic transform* technique [41, Theorem 2] is utilized to decouple the numerator and the denominator of each ratio term and convert a convex-convex multiple-ratio fractional programming problems into a sequence of convex optimization problems. The primal variables multicast beamforming $\{\mathbf{p}_{j,m}\}$ and the auxiliary variables, denoted as $\{y_{\ell,\ell',k,m}\}$, are alternately optimized till convergence, which consists of two steps.

Step 1: when $\{\mathbf{p}_{j,m}\}$ is held fixed, the optimal $y_{\ell,\ell',k,m}$ can be found in closed form as

$$y^{\star}_{\ell,\ell',k,m} = (\mathcal{B}^{i_m^*}_{k,m}(\mathbf{p}_{j,m}, P_{\mathsf{s},j}))^{-1} \varphi^{i_m^*}_{\ell,\ell',k,m}(\mathbf{p}_{j,m}) \in \mathbb{R}, \tag{40}$$

with $\varphi^{i_m^*}_{\ell,\ell',k,m}(\mathbf{p}_{j,m})\colon \mathbb{C}^{N_t} \to \mathbb{R}$ is denoted as

$$\varphi^{i_m^*}_{\ell,\ell',k,m}(\mathbf{p}_{j,m}) \triangleq (\mu_{\ell,k,m} - \mu_{\ell',k,m}) \Re \left[\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i_m^*)} \mathbf{p}_{j,m} \right]. \tag{41}$$

Step 2: when $y_{\ell,\ell',k,m}$ is fixed, solving an equivalent concave maximization sub-problem that is derived from $\mathbf{P6}$.

We demonstrate that the *quadratic transform* method can be applied to tackle multicast beamforming design as follows.

Algorithm 2 Quadratic Transform Approach for Convex-Convex FP Problem $\mathbf{P6}$

Input: Channel gain $\{\mathbf{H}_{j,k}^{(i_m^*)}\}$, Device energy $\{E_k\}$, Auxiliary variables $\{y_{\ell,\ell',k,m}\}$, $\{z_{\ell,\ell',k,m}\}$.
Output: $\{\mathbf{p}_{j,m}\}$, $\{\mathbf{f}_{k,m}\}$, $\{P_{\mathsf{s},j}\}$

Initialize auxiliary variables $\{y_{\ell,\ell',k,m}\}$ and $\{z_{\ell,\ell',k,m}\}$;
repeat
  Update $\{\mathbf{p}_{j,m}\}$ by $\mathbf{P8}$ (44) under given $\{y_{\ell,\ell',k,m}\}$;
  Update the auxiliary functions $\mathcal{A}_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m})$ and $\mathcal{B}_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m})$;
  Update the auxiliary variables as (40);
until convergence
repeat
  Update $\{\mathbf{f}_{k,m}\}$ by $\mathbf{P9}$ (45) under given $\{z_{\ell,\ell',k,m}\}$;
  Update the auxiliary functions $\mathcal{A}_{\ell,\ell',k,m}^{i_m^*}(\mathbf{f}_{k,m})$ and $\mathcal{B}_{\ell,\ell',k,m}^{i_m^*}(\mathbf{f}_{k,m})$;
  Update the auxiliary variables by (46);
until convergence
$\mathbf{f}_{k,m} \leftarrow \frac{\mathbf{f}_{k,m}}{\|\mathbf{f}_{k,m}\|}, \forall k, \forall m.$

As for multicast beamforming $\{\mathbf{p}_{j,m}\}_j$ under given $\mathbf{f}_{k,m}$ and i_m^*, the $\mathbf{P6}$ can be simplified to

$$\mathbf{P7} \quad \max_{\{\mathbf{p}_{j,m}\},\{P_{\mathsf{s},j}\}} \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\ell'=1}^{L_k}\sum_{\ell<\ell'} \frac{\mathcal{A}_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m})}{\mathcal{B}_{k,m}^{i_m^*}(\mathbf{p}_{j,m},P_{\mathsf{s},j})} \quad \text{(42a)}$$
$$\text{s.t.} \quad \text{(18b)}.$$

Lemma 4 (Quadratic Transform)**.** *The objective* (42a) *of* $\mathbf{P7}$ *can be equivalently converted into a concave maximization problem by the quadratic transform method.*

Proof. Recognize that each term in the summation of (42a) can be reformulated as

$$(\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m}))^T(\mathcal{B}_{k,m}(\mathbf{p}_{j,m},P_{\mathsf{s},j}))^{-1}\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m}) \quad (43)$$

where $\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m})$ in (41) is a linear combination of $\mathbf{p}_{j,m}$, and $\mathcal{B}_{k,m}(\mathbf{p}_{j,m},P_{\mathsf{s},j}) \in \mathbb{R}^+$ is convex with respect to $(\mathbf{p}_{j,m},P_{\mathsf{s},j})$. Replacing each term of (42a) with (43), thus $\mathbf{P7}$ satisfies the conditions of *quadratic transform* [41, Eq.(16)]. □

Based on Lemma 4, the corresponding quadratic transform for $\mathbf{P7}$ (42) is equivalent to [41, Theorem 2]:

$$\mathbf{P8} \quad \max_{\{\mathbf{p}_{j,m}\},\{P_{\mathsf{s},j}\}} \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\ell'=1}^{L_k}\sum_{\ell<\ell'}(2y_{\ell,\ell',k,m}\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m}) - y_{\ell,\ell',k,m}^2\mathcal{B}_{k,m}^{i_m^*}(\mathbf{p}_{j,m},P_{\mathsf{s},j})), \quad \text{(44a)}$$
$$\text{s.t.} \quad \text{(18b)},$$

where $\{y_{\ell,\ell',k,m}\}$ in (40) are introduced auxiliary variables. Due to the convexity of each $\mathcal{B}_{k,m}(\mathbf{p}_{j,m},P_{\mathsf{s},j})$, and the linearity of $\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{p}_{j,m})$, the quadratic transform (44a) is concave in $\{\mathbf{p}_{j,m}\}$ and $P_{\mathsf{s},j}$ for fixed $\{y_{\ell,\ell',k,m}\}$. Therefore, $\mathbf{P8}$ is concave, and the optimal $\{\mathbf{p}_{j,m}\}$ and $\{P_{\mathsf{s},j}\}$ can be efficiently obtained through numerical convex optimization.

Algorithm 3 Proposed Alternating Optimization

Input: Channel gain $\{\mathbf{H}_{j,k}^{(i_m^*)}\}$, Device energy $\{E_k\}$, Weights $\tilde{w}_m^{(i)}$.
Output: $\{\mathbf{p}_{j,m}\}$, $\{\mathbf{f}_{k,m}\}$, $\{P_{\mathsf{s},j}\}$, $\{a_m^{(i)}\}$

Initialize auxiliary variables $\{y_{\ell,\ell',k,m}\}$ and $\{z_{\ell,\ell',k,m}\}$, $\{\mathbf{p}_{j,m}\}$, $\{\mathbf{f}_{k,m}\}$, $\{P_{\mathsf{s},j}\}$;
repeat
  Update $\{a_m^{(i)}\}$ via Algorithm 1;
  Update $\{\mathbf{p}_{j,m}\}$, $\{\mathbf{f}_{k,m}\}$, $\{P_{\mathsf{s},j}\}$ via Algorithm 2;
until convergence

2) Receive Beamforming Design: Finally, similar to multicast beamforming design, with the updated sensing power $P_{\mathsf{s},j}$ and multicast beamforming vector $\mathbf{p}_{j,m}$, the receive beamforming vector $\mathbf{f}_{k,m}$ is updated with subproblem

$$\mathbf{P9} \max_{\{\mathbf{f}_{k,m}\}} \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\ell'=1}^{L_k}\sum_{\ell<\ell'}(2z_{\ell,\ell',k,m}\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{f}_{k,m}) - z_{\ell,\ell',k,m}^2\mathcal{B}_{k,m}^{i_m^*}(\mathbf{f}_{k,m})), \quad \text{(45a)}$$
$$\text{s.t.} \quad \text{(39)}.$$

with a auxiliary variable $z_{\ell,\ell',k,m}$, which is iteratively updated

$$z_{\ell,\ell',k,m}^{\star} = (\mathcal{B}_{k,m}^{i_m^*}(\mathbf{f}_{k,m}))^{-1}\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{f}_{k,m}) \in \mathbb{R}, \quad (46)$$

and $\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{f}_{k,m})\colon \mathbb{C}^{N_r} \to \mathbb{R}$ is denoted as

$$\varphi_{\ell,\ell',k,m}^{i_m^*}(\mathbf{f}_{k,m}) \triangleq (\mu_{\ell,k,m}-\mu_{\ell',k,m})\Re\left[\sum_{j=1}^{K}\mathbf{f}_{k,m}^H\mathbf{H}_{j,k}^{(i_m^*)}\mathbf{f}_{k,m}\right]. \quad (47)$$

Hence the $\mathbf{P9}$ (45) is concave maximization problem over $\{\mathbf{f}_{k,m}\}$, which can also be solved by CVXPY. We show the solution procedure in Algorithm 2.

This algorithm jointly optimizes the sensing power, multicast beamforming, and receive beamforming vectors using an iterative, alternating method. It breaks the problem into two sub-problems that are solved repeatedly until convergence. The most computationally intensive part of the Algorithm 2 is solving the convex optimization problems ($\mathbf{P8}$ and $\mathbf{P9}$), whose computational costs are on the order of $\mathcal{O}((KM(N_t+N_r))^3)$ within each iteration. This could be feasible for a small number of devices and feature dimensions. If the number of devices is too large, these devices can be assisted by a logical control node to manage the optimization problem.

3) Solution to $\mathbf{P2}$*:* Based on the alternating method described above, the detailed procedure of solving $\mathbf{P2}$ (22) contains 2 steps: for the fixed $(\{P_{\mathsf{s},k}\},\{\mathbf{p}_{j,m}\}),\{\mathbf{f}_{k,m}\}$, update $\{a_m^{(i)}\}$ via Algorithm 1; for the fixed $\{a_m^{(i)}\}$, update $(\{P_{\mathsf{s},k}\},\{\mathbf{p}_{j,m}\}),\{\mathbf{f}_{k,m}\}$ by Algorithm 2. The two steps are alternately iterated till convergence.

E. Complexity Analysis

Consider the quadratic transform method for solving joint sensing power allocation, multicast beamforming and receive beamforming design for $\mathbf{P6}$, which requires computational

complexity of $\mathcal{O}(I_0(KM(N_t+N_r))^3)$ with I_0 being the number of iterations for convergence in Algorithm 2. Provided that the complexity of Algorithm is $\mathcal{O}(M^3)$, denoting I_1 as the number of iterations in Algorithm 3, the total computational complexity of proposed alternating optimization algorithm is $\mathcal{O}(I_1(I_0(KM(N_t+N_r))^3+M^3))$.

V. Numerical Results

A. Simulation Settings

1) Network settings: In this section, we evaluate and compare the performance of our proposed decentralized ISCC method for inference tasks under different schemes. For all the following simulations, unless specified otherwise, a decentralized network consisting of three ISAC devices is considered for inference tasks, where each device is equipped with 8 transmit antennas and 8 receive antennas. The channel gains between the device j and k is modeled as $\mathbf{H}_{j,k} = |\phi_{j,k}\boldsymbol{\rho}_{j,k}|^2, \forall j \neq k$, where $[\phi_{j,k}]_{\text{dB}} = -[\mathbf{PL}_{j,k}]_{\text{dB}} + [\zeta_{j,k}]_{\text{dB}}$ represents the large scale fading propagation coefficient in dB with $[\mathbf{PL}_{j,k}]_{\text{dB}} = 128.1 + 37.6\log_{10}[d_{j,k}]_{\text{km}}$ being the path loss in dB, $d_{j,k}$ being the distance between device j and device k, which is randomly set to in the range of [d km, d+0.05 km] and by default $d = 0.4$. And $[\zeta_{j,k}]_{\text{dB}} \sim \mathcal{N}(0, \sigma_\zeta^2)$ stands for shadowing in dB. On the other hand, each element in $\boldsymbol{\rho}_{j,k}$ is assumed to satisfy a Rayleigh small-scale channel coefficient $\mathcal{CN}(0,1)$. Based on [42], a small Rician factor is used to characterize the residual self-interference channel, i.e., $\mathbf{H}_{k,k}^{(i)} \sim \mathcal{CN}(\sqrt{\frac{\sigma_{\text{SI}}^2\kappa}{1+\kappa}}\mathbf{I}_{N_r\times N_t}, \frac{\sigma_{\text{SI}}^2}{1+\kappa}\mathbf{I}_{N_r}\otimes\mathbf{I}_{N_t}), \forall k, i$ with σ_{SI}^2 used to parameterize the residual self-interference that is fixed at -60dB and κ stands for the Rician factor that is set to 3dB. The variance of both the sensing noise σ_{r}^2 and clutter signal $\sigma_{\mathsf{s},k}^2$ are set to 0.2. The channel noise variance $N_0 = 1$ and the variance of shadow fading $\sigma_\zeta^2 = 8$dB. The number of extracted feature elements $M = 4$. Both the sensing time T_s and the communication time are 1. The device energy threshold E_k is set to 25mdB, and the computation energy for each device $E_{\mathsf{p},k} = 10^{-4}$ Joule.

B. Inference tasks and models

In this experiment, the University of Glasgow Radar Signature dataset [33] is adopted to evaluate the performance of the proposed algorithm. This dataset contains the radar echo signals of different motions of 103 people in all age groups sat in nine different locations. Data corresponding to five motions are selected for the recognition task: walking, sitting down, standing up, picking up an object, and drinking water. Based on the dataset, each device is assigned a different task along with a machine learning model:

- The task of device 1 is identifying the target's motion from 5 motions with a multi-layer perceptron (MLP) neural network where the numbers of neurons in the hidden layers of the MLP are set to 80 and 40 with Adam optimizer and ReLU activation.
- The task of device 2 is to distinguish the gender of the target with a K-Nearest Neighbour (KNN) model where the value of $K = 2$.

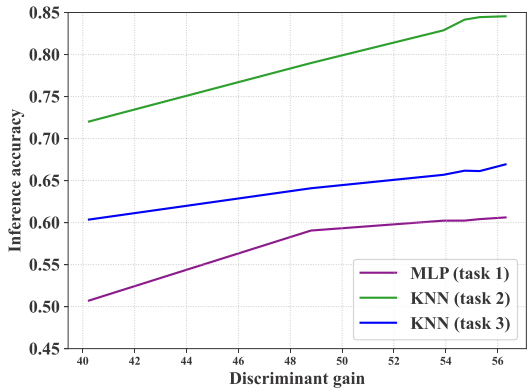


Fig. 3. Discriminant gain versus Inference accuracy.

- The task of device 3 is to separate the age group of the target (e.g.,$[0,30]$ or $[30,50]$ or $[50,]$) with a K-Nearest Neighbour model where the value of $K = 3$.

The dataset [33] comprises 1,500 samples in total, partitioned into 90% training samples and 10% test samples. In the following scenarios, the training dataset is considered the ground-truth data (free of noise) when training all ML models on a powerful server. Then the pre-trained model is transmitted to each device and used for inference therein, while the testing dataset is distorted by sensing and communication noise.

C. Inference Algorithms

To evaluate our proposed method, some benchmarking schemes are described below:

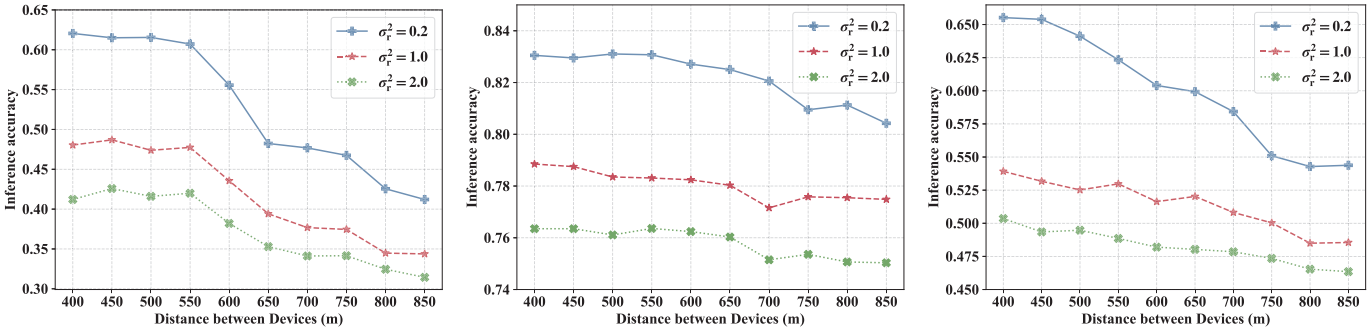
- Proposed Approach: All parameters are optimized by the proposed alternating maximization approach.
- Proposed, Subcarrier-Aware: the subcarrier allocation is randomly set, the sensing power, multicast, and receive beamforming is optimized via Algorithm 2.
- SCA-based: The successive convex approximation (SCA) based alternating algorithm proposed in [43] is used for solving $\mathbf{P1}$ (18).
- Baseline: The sensing power is allocated to a constant, the subcarriers are allocated randomly among all devices, and the multicast as well as receive beamforming vectors are set to be randomly generated during the transmission.

D. Performance Comparison

Fig. 3 illustrates the relationship between inference accuracy and discriminant gain for MLP (task 1) and KNN (tasks 2 and 3) models across the three tasks of our proposed method. The results demonstrate that accuracy consistently increases with higher discriminant gain, confirming the effectiveness of this metric for the classification task. Furthermore, it is revealed that task 2 achieves superior performance across all discriminant gain values, suggesting it is inherently simpler than tasks 1 and 3.

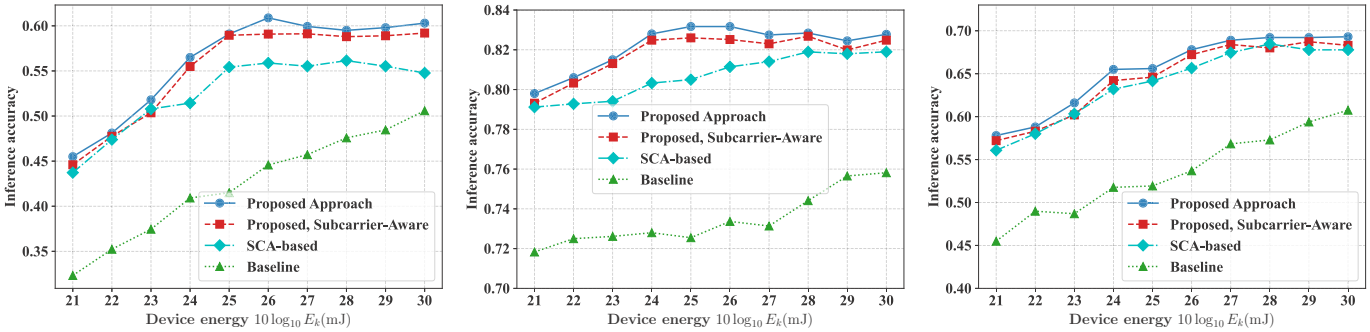
Fig. 4 demonstrates that the inference performance of all three devices generally degrades as the distance between devices d increases. It is because a larger distance between

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.2025.3625389



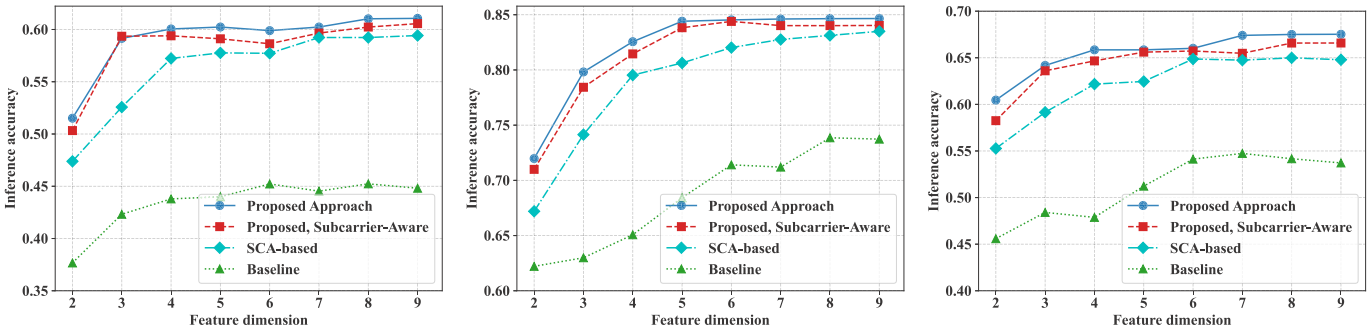
(a) Task 1. (b) Task 2. (c) Task 3.

Fig. 4. Inference accuracy of three tasks on corresponding devices versus distance between devices varying sensing noise.



(a) Task 1. (b) Task 2. (c) Task 3.

Fig. 5. Inference accuracy of three tasks on corresponding devices versus different device energy.



(a) Task 1. (b) Task 2. (c) Task 3.

Fig. 6. Inference accuracy of three tasks on corresponding devices versus different feature dimensions.

the devices leads to stronger path losses and weaker channel gains, thereby inducing larger communication distortion and worse accuracy. However, non-monotonic behavior of the proposed approach is observed at specific distance counts (e.g., $d = 0.75$km, $\sigma_r^2 = 0.2$), where the accuracy of task 2 improves while tasks 1 and 2 decline. This trade-off arises from the decentralized nature of the system, where devices independently optimize their resource allocation to balance competing multi-objectives. Additionally, we present the various sensing distortions σ_r^2's impact on inference performance. It shows that all three tasks can achieve better performance when suffering less sensing distortion.

Fig. 5 compares the performance of our proposed method with SCA-based and baseline approaches in terms of inference accuracy across three tasks under varying total device energy. The results show that a higher total energy threshold improves inference performance, as devices can allocate more sensing and communication power to mitigate clutter distortion, channel fading, and noise. Notably, our proposed method consistently achieves the highest inference accuracy across all energy levels, outperforming both the SCA-based and baseline approaches. The superiority of our method arises from the limitations of conventional SCA, which relies on iterative convex approximations of non-convex subproblems, leading to cumulative errors and suboptimal solutions. In contrast, our proposed method leverages the scalability of the quadratic

transform, ensuring more robust performance. Additionally, the baseline scheme underperforms because it does not adapt sensing power or beamforming vectors to different feature elements. The proposed subcarrier-aware approach performs worse than the proposed approach that optimizes all variables, which is due to the incomplete consideration of subcarrier allocation.

Fig. 6 compares the inference accuracy of three methods across varying feature dimensions. The results demonstrate a nonlinear relationship between feature dimensionality and inference performance. All methods exhibit improved accuracy with increasing dimensions, as additional dimensions provide richer feature representations of the target variable. Nevertheless, we observe that a feature dimension of $M = 5$ is sufficient to achieve high inference accuracy, whereas larger values of M only increase the computational and communication complexity without yielding significant accuracy gains. Notably, both the proposed method and the SCA-based approach show greater robustness compared to the baseline across all dimensional configurations. However, the proposed scheme consistently achieves superior accuracy, demonstrating its effectiveness in balancing the fundamental trade-off between the three tasks.

VI. Conclusion

This paper proposes a decentralized AirComp based ISCC system tailored for multi-task collaborative inference. Our proposed scheme facilitates simultaneous multicast and AirComp aggregation of local features of all devices through full-duplex communication, which makes the communication overhead irrelevant to the number of devices. To further enhance the inference performance, we leverage subcarrier allocation. We exploited the self-interference (SI) channel in full-duplex communication to aggregate features from one device with others, reducing the cost of computation resources. Leveraging these benefits, our proposed scheme shows a better inference performance in experiments. This new scheme paves the way for broader applications with massive ISCC devices for distributed learning.

Our work opens several research directions, including exploiting new hardware facilities like 3-D bandstop frequency selective structures [44], differential antennas [45], and time modulated antennas [46] in the ISCC design, extending the design into systems with continuous inference tasks, transforming the system into digital AirComp based framework [47].

Appendix A
Proof of Lemma 1

The subcarrier allocation constraints (18c)∼(18e) implies that for each fixed pair (k, m), only one of the $a_{k,m}^{(i)} \in \{0, 1\}$ is nonzero, i.e.,

$$a_{k,m}^{(i)} \cdot a_{k,m}^{(i')} = 0, \quad \forall i \neq i'. \tag{48}$$

Let $s_{i,k,m} \triangleq \sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i)} \mathbf{p}_{j,m} \in \mathbb{C}$, $\left(\tilde{\mu}_{\ell,k,m} - \tilde{\mu}_{\ell',k,m}\right)^2$ is

$$\begin{aligned}
&(\mu_{\ell,m} - \mu_{\ell',m})^2 \left(\Re \left[\sum_{i=1}^{M} a_{k,m}^{(i)} \cdot s_{i,k,m} \right] \right)^2 \\
=&(\mu_{\ell,m} - \mu_{\ell',m})^2 \left(\sum_{i=1}^{M} \sum_{i'=1}^{M} a_{k,m}^{(i)} a_{k,m}^{(i')} \cdot \Re\left[s_{i,k,m}\right] \Re\left[s_{i',k,m}\right] \right) \\
=&(\mu_{\ell,m} - \mu_{\ell',m})^2 \sum_{i=1}^{M} \left(a_{k,m}^{(i)} \right)^2 \cdot \left(\Re\left[s_{i,k,m}\right] \right)^2 \\
=&(\mu_{\ell,m} - \mu_{\ell',m})^2 \sum_{i=1}^{M} a_{k,m}^{(i)} \cdot \left(\Re\left[s_{i,k,m}\right] \right)^2,
\end{aligned} \tag{49}$$

where the second to last equation uses (48) and the last equation uses $\left(a_{k,m}^{(i)}\right)^2 = a_{k,m}^{(i)}$ since $a_{k,m}^{(i)} \in \{0, 1\}$. Similarly, denoting $\tilde{s}_{j,k,m}^{(i)} \triangleq \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i)} \mathbf{p}_{j,m} \in \mathbb{C}$ and $\tilde{u}_j \triangleq \sigma_{\mathsf{s},j}^2 + \frac{\sigma_{\mathsf{r}}^2}{P_{\mathsf{s},j}}$, we rewrite the second term in (13) as

$$\begin{aligned}
&\sum_{j=1}^{K} \left(\Re \left[\sum_{i=1}^{M} a_{k,m}^{(i)} \tilde{s}_{j,k,m}^{(i)} \right] \right)^2 \tilde{u}_j \\
=&\sum_{j=1}^{K} \sum_{i=1}^{M} a_{k,m}^{(i)} \left(\Re[\tilde{s}_{j,k,m}^{(i)}] \right)^2 \tilde{u}_j \\
=&\sum_{i=1}^{M} a_{k,m}^{(i)} \cdot \sum_{j=1}^{K} \left(\Re \left[\mathbf{f}_{k,m}^H \mathbf{H}_{j}^{(i)} \mathbf{p}_{j,m} \right] \right)^2 \left(\sigma_{\mathsf{s},j}^2 + \frac{\sigma_{\mathsf{r}}^2}{p_{\mathsf{s},j}} \right)
\end{aligned} \tag{50}$$

Based on (49), the first term in (13) can be rewritten as

$$\sigma_m^2 \sum_{i=1}^{M} a_{k,m}^{i} \cdot \left(\Re \left[\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}^{(i)} \mathbf{P}_{j,m} \right] \right)^2 \tag{51}$$

According to (49), (50) and (51), we can derive (20) and (21).

Appendix B
Proof of Lemma 2

The objective function of $\mathbf{P4}$ (26) is

$$\max_{\{a_m^i\}} \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{\sum_{i=1}^{M} a_m^{(i)} \bar{d}_{k,m}^{(i)}}{\sum_{i=1}^{M} a_m^{(i)} \bar{c}_{k,m}^{(i)} + \frac{N_0}{2} \|\mathbf{f}_{k,m}\|^2} \tag{52}$$

The constraint $\sum_{i=1}^{M} a_m^i = 1$, $\forall m$, implies that for any given m, a unique item, denoted as i_m^*, is assigned to it (i.e., $a_m^{(i_m^*)} = 1$). Consequently, the terms in the objective function can be simplified as follows:

$$\sum_{i=1}^{M} a_m^{(i)} \bar{d}_{k,m}^{(i)} = \bar{d}_{k,m}^{(i_m^*)}, \quad \sum_{i=1}^{M} a_m^{(i)} \bar{c}_{k,m}^{(i)} = \bar{c}_{k,m}^{(i_m^*)}. \tag{53}$$

Substituting these into the objective function (52) yields

$$\max_{\mathbf{A}} \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{\bar{d}_{k,m}^{(i_m^*)}}{\bar{c}_{k,m}^{(i_m^*)} + \frac{N_0}{2} \|\mathbf{f}_{k,m}\|^2}. \tag{54}$$

where i_m^* is the item assigned to m-th element under the assignment matrix $\mathbf{A}$. The set $\{i_1^*, \ldots, i_M^*\}$ forms a permutation of $\{1, \ldots, M\}$.

By interchanging the order of summation, we can rewrite (54) as

$$\max_{\mathbf{A}} \sum_{i=1}^{M} \left(\sum_{k=1}^{K} \frac{\bar{d}_{k,m}^{(i_m^*)}}{\bar{c}_{k,m}^{(i_m^*)} + \frac{N_0}{2}\|\mathbf{f}_{k,m}\|^2} \right). \tag{55}$$

This reformulation reveals that the problem seeks a single permutation $(i_1^*, \ldots, i_M^*)$ that maximizes the element-specific aggregate discriminant gain.

The structure of the objective function is characteristic of the standard assignment problem. The weight associated with assigning i-th subcarrier to m-th element can be defined as the sum of individual device-based ratio terms for that specific assignment:

$$w_m^{(i)} \triangleq \sum_{k=1}^{K} \frac{\bar{d}_{k,m}^{(i)}}{\bar{c}_{k,m}^{(i)} + \frac{N_0}{2}\|\mathbf{f}_{k,m}\|^2} \tag{56}$$

where the objective can also be expressed as $\sum_{m=1}^{M} w_m^{(i_m^*)}$.

Appendix C
Proof of Lemma 3

Let (m, i) be an edge in the complete bipartite graph $\mathcal{G}$. For the following cases: 1) $m \in \bar{\mathcal{U}}$, $i \in \bar{\mathcal{V}}$, the sum $u_m + v_i$ does not change, this proves that rigid edges from $\bar{\mathcal{U}}$ to $\bar{\mathcal{V}}$ remain rigid; 2) $m \in \bar{\mathcal{U}}$, $i \notin \bar{\mathcal{V}}$, the sum $u_m + v_i$ decreases by δ. By the definition of δ, the edge (m, i) which δ refers to will become rigid; 3) $m \notin \bar{\mathcal{U}}$, $i \notin \bar{\mathcal{V}}$, the sum remains unchanged; 4) $m \notin \bar{\mathcal{U}}$, $i \in \bar{\mathcal{V}}$, the resulting sum increases by δ, hence the $u_m + v_i < \tilde{w}_m^{(i)}$ is still preserved.

References

[1] 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.

[2] D. Wen, K.-J. Jeon, and K. Huang, "Federated dropout—A simple approach for enabling federated learning on resource constrained devices," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 923–927, 2022.

[3] Y. Mao, X. Yu, K. Huang, Y.-J. Angela Zhang, and J. Zhang, "Green edge AI: A contemporary survey," *Proc. IEEE*, vol. 112, no. 7, pp. 880–911, 2024.

[4] D. Gündüz, F. Chiariotti, K. Huang, A. E. Kalør, S. Kobus, and P. Popovski, "Timely and massive communication in 6G: Pragmatics, learning, and inference," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 1, pp. 27–40, 2023.

[5] M. Ji, H. Zhao, L. Jiao, S. Zhang, X. Li, Z. Qian, and B. Ye, "Edge AI inference as a service via dynamic resources from repeated auctions," *IEEE Tran. Mobile Comput.*, pp. 1–17, 2025.

[6] 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.*, vol. 23, no. 3, pp. 2039–2053, 2024.

[7] 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.

[8] 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.

[9] N. Li, A. Iosifidis, and Q. Zhang, "Attention-based feature compression for cnn inference offloading in edge computing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2023, pp. 967–972.

[10] 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.

[11] Z. Wang, A. E. Kalør, Y. Zhou, P. Popovski, and K. Huang, "Ultra-low-latency edge inference for distributed sensing," *IEEE Trans. Wireless Commun.*, early access, 2025.

[12] Z. Wang, Q. Zeng, H. Zheng, and K. Huang, "Revisiting outage for edge inference systems," 2025. [Online]. Available: https://arxiv.org/abs/2504.03686

[13] 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 Commun.*, vol. 30, no. 3, pp. 78–85, 2023.

[14] Z. Liu, Q. Lan, A. E. Kalør, P. Popovski, and K. Huang, "Over-the-air multi-view pooling for distributed sensing," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7652–7667, 2024.

[15] S. F. Yilmaz, B. Hasircioğlu, L. Qiao, and D. Gündüz, "Private collaborative edge inference via over-the-air computation," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 215–231, 2025.

[16] 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.

[17] H. Lee and S.-W. Kim, "Task-oriented edge networks: Decentralized learning over wireless fronthaul," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 15 540–15 556, 2024.

[18] D. Wen, X. Li, Y. Zhou, Y. Shi, S. Wu, and C. Jiang, "Integrated sensing-communication-computation for edge artificial intelligence," *IEEE Internet Things Mag.*, vol. 7, no. 4, pp. 14–20, 2024.

[19] D. Wen, Y. Zhou, X. Li, Y. Shi, K. Huang, and K. B. Letaief, "A survey on integrated sensing, communication, and computation," *IEEE Commun. Surv. Tut., early access*, 2024.

[20] D. Wen, S. Xie, X. Cao, Y. Cui, J. Xu, Y. Shi, and S. Cui, "Integrated sensing, communication, and computation for over-the-air federated edge learning," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2025.

[21] 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.*, vol. 23, no. 3, pp. 2486–2502, 2024.

[22] D. Wang, D. Wen, Y. He, Q. Chen, G. Zhu, and G. Yu, "Joint device scheduling and resource allocation for iscc-based multiview–multitask inference," *IEEE Internet Things J.*, vol. 11, no. 24, pp. 40 814–40 830, 2024.

[23] Z. Zhuang, D. Wen, Y. Shi, G. Zhu, S. Wu, and D. Niyato, "Integrated sensing-communication-computation for over-the-air edge AI inference," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3205–3220, 2024.

[24] X. Li, F. Liu, Z. Zhou, G. Zhu, S. Wang, K. Huang, and Y. Gong, "Integrated sensing, communication, and computation over-the-air: MIMO beamforming design," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5383–5398, 2023.

[25] S. Liu, D. Wen, D. Li, Q. Chen, G. Zhu, and Y. Shi, "Energy-efficient optimal mode selection for edge AI inference via integrated sensing-communication-computation," *IEEE Tran. Mobile Comput.*, vol. 23, no. 12, pp. 14 248–14 262, 2024.

[26] J. Yao, W. Xu, G. Zhu, K. Huang, and S. Cui, "Energy-efficient edge inference in integrated sensing, communication, and computation networks," *arXiv preprint arXiv:2503.00298*, 2025.

[27] C. Deng, X. Fang, and X. Wang, "Integrated sensing, communication, and computation with adaptive DNN splitting in multi-UAV networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 17 429–17 445, 2024.

[28] A. Sarker, C. Qiu, and H. Shen, "Connectivity maintenance for next-generation decentralized vehicle platoon networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 4, pp. 1449–1462, 2020.

[29] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, 2014.

[30] D. Wen, G. Yu, and L. Xu, "Energy-efficient mode selection and power control for device-to-device communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2016, pp. 1–7.

[31] Z. Lin, Y. Gong, and K. Huang, "Distributed over-the-air computing for fast distributed optimization: Beamforming design and convergence analysis," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 274–287, 2023.

[32] Z. Wang, Y. Zhao, Y. Zhou, Y. Shi, C. Jiang, and K. B. Letaief, "Over-the-air computation for 6G: Foundations, technologies, and applications," *IEEE Internet Things J.*, vol. 11, no. 14, pp. 24 634–24 658, 2024.

[33] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. Le Kernec, "Radar signatures of human activities," 2019.

[34] 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.

[35] H. Xu, M. Chen, Z. Meng, Y. Xu, L. Wang, and C. Qiao, "Decentralized machine learning through experience-driven method in edge networks,"

IEEE Journal on Selected Areas in Communications, vol. 40, no. 2, pp. 515–531, 2022.

[36] Z. Lin, Y. Gong, and K. Huang, "Distributed over-the-air computing for fast distributed optimization: Beamforming design and convergence analysis," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 274–287, 2023.

[37] Y. Shi, S. Xia, Y. Zhou, Y. Mao, C. Jiang, and M. Tao, "Vertical federated learning over cloud-RAN: Convergence analysis and system optimization," *IEEE Trans. Wireless Commun.*, 2023.

[38] L. Stewart, F. Bach, Q. Berthet, and J.-P. Vert, "Regression as classification: Influence of task formulation on neural network features," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, vol. 206. PMLR, 25–27 Apr 2023, pp. 11 563–11 582.

[39] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Res. Log. Quart.*, vol. 2, no. 1-2, pp. 83–97, 1955.

[40] R. Burkard, M. Dell'Amico, and S. Martello, *Assignment Problems*. Society for Industrial and Applied Mathematics, 2012.

[41] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, 2018.

[42] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4296–4307, 2012.

[43] Z. Zhuang, D. Wen, and Y. Shi, "Decentralized over-the-air computation for edge ai inference with integrated sensing and communication," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2023, pp. 4644–4649.

[44] D. Li, Y. Wu, Z. Gu, Y. Fan, X. An, H. Chen, and E. Li, "Implementation of 3-D bandstop frequency-selective structures with ultralarge angular stability utilizing narrow L-shaped strip lines," *IEEE Trans. Microw. Theory Tech.*, vol. 72, no. 4, pp. 2298–2309, 2024.

[45] Y. Zhang, "Differential antennas: Fundamentals and applications," *Electromagn. Sci.*, vol. 1, no. 1, pp. 1–17, 2023.

[46] W. Wu, Q. Chen, J.-D. Zhang, T. Huang, and D.-G. Fang, "Time modulated array antennas: A review," *Electromagn. Sci.*, vol. 2, no. 1, pp. 1–19, 2024.

[47] L. Qiao, Z. Gao, M. Boloursaz Mashhadi, and D. Gündüz, "Massive digital over-the-air computation for communication-efficient federated edge learning," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 11, pp. 3078–3094, 2024.