Real-Time Edge AI Model Inference: An Integrated Sensing and Communication Approach

XXX

Abstract

To enable real-time intelligent services, such as auto-driving, at the network edge, in this paper, an efficient edge *artificial-intelligent* (AI) model split inference framework is proposed. In such a system, there are one server and multiple radar sensing devices. The inference accuracy depends on the feature distortion caused by three steps: 1) each device obtains raw data via *sensing*; 2) A feature subset is generated via extraction [e.g., by *principle component analysis* (PCA)] and *quantization* on each device; 3) all subsets are *transmitted* to server, where they are cascaded to form the whole feature vector and then inputted for model inference. These three steps are highly coupled, as sensing and transmission competes for radio resources (e.g., time), and feature quantization determines the transmission load. On the other hand, classification tasks are considered in this paper. And a theoretical metric called *discriminant gain* is adopted as the theoretical measure of inference accuracy, which is defined as the distance of two classes in the Euclidean feature space under normalized covariance. Specifically, larger discriminant gain separates different classes better, and thus leads to greater inference accuracy. Thereby, to maximize the inference accuracy, new challenges, arising from the highly coupling of sensing, quantization, and communication, the device heterogeneity in terms of channel gain, quantization level, and the generated feature subsets' contribution to inference accuracy, as well as the complicated form of discriminant gain, call for task-oriented communication scheme, which should integrate the design of sensing, quantization, and communication. As a result, the discriminant gain maximization problem is non-convex. To tackle this problem, an optimal scheme of joint sensing & transmit power, time, and quantization bits allocation is proposed based on the *sum-of-ratios* method. It shows that for each radar device, more classes in the inference task requires more sensing power and quantization bits, and more transmit power should be allocated when the channel gain of the device is weak. Our theoretical analysis is verified by extensive experiments, where the AI model inference task is human motions recognition.

I. Introduction

Edge *artificial intelligence* (AI) emerges as a promising technique to support a variety of intelligent applications, such as Metaverse, auto-driving, and e-Health, at the network edge [1]–[4]. To enable these intelligence services, it's desirable to deploy well-trained machine learning models and to utilize their inference capability for making decisions. This leads to a new research paradigm called edge AI model inference, short as *edge inference* [5], [6]. Several kinds of techniques have been proposed for the efficient implementation of edge inference. The first one is called *on-device inference* (see e.g., [7]–[10]). This technique alleviates the computation load on resource limited devices via designing dedicated light models such as MobileNets, or compressing the deep models to reduce their sizes by e.g., pruning and quantization. However, as there are various kinds of AI tasks with many different models, this technique is still inefficient. To address this challenge, the technique of *on-server inference* is designed (see e.g., [11], [12]). In such a scheme, edge devices upload the data to edge server, where the model inference is performed, and then receive the feedback inference results. Although the on-server inference designs can significantly alleviate the overhead of devices, they violate the data privacy. To tackle this issue, the technique of *split inference* is proposed, which splits the AI model into two submodels (see e.g., [13]–[19]). One is deployed at the devices and generally performs feature extraction [e.g. *principle component analysis* (PCA) and convolutional layers], and the other is deployed at the edge server. As a result, both the privacy is kept by avoiding transmitting raw data and the overhead of edge devices can be reduced by offloading most computation to the edge server.

In this work, the split inference technique is adopted due to its outstanding advantages mentioned above. Different from the existing designs (see e.g., [13]–[19]), which focus on reducing the devices' computation overhead due to feature generation and overcoming the communication bottleneck caused by feature transmission, we clarify that the improvement of inference performance should be analyzed from a systematic view and call for *task oriented communication* schemes for three reasons. First, the goal of edge inference is no longer throughput maximization, but high accuracy and low latency. Besides, the inference accuracy depends on the feature vector's distortion level caused by three steps: data acquisition on devices (i.e., sensing), feature subset generation (i.e., extraction and quantization), and feature transmission to server. As the quantization (distortion) level decides the communication load, and sensing and transmission competes for radio resources (see e.g., [20]), the three factors are highly coupled. What's more important, new challenges arise from the device heterogeneity in terms of the channel gain, quantization level, and the generated feature subsets' contribution to inference accuracy. Therefore, a real-time inference task oriented communication scheme should maximize the inference accuracy by jointly designing sensing, quantization, and transmission, under a low latency requirement.

As mentioned above, sensing should be involved to enhance the inference performance, especially for real-time tasks like Metaverse and auto-driving. To this end, the technology of *integrated sensing and communication* (ISAC) is adopted in this work, which is a promising solution to enable sensing and communication shares the hardware equipment and thus reduces the cost [20]. To begin with, ISAC techniques have been widely studied in the existing literature, such as the optimal waveform design in *dual functional radar-and-communication* (DFRC) systems in [21], the power control for network ISAC systems in [22], the beamforming design for *reconfigurable intelligent surface* (RIS) assisted ISAC systems in [23], and the ISAC and computation over-the-air framework proposed in [24]. Then, based on the existing techniques, recent research manages to apply the technique of ISAC in edge AI systems. Specifically, an ISAC based centralized edge learning system is proposed in [25], which accelerates the learning process by generating and uploading as many training data as possible from the sensing devices to the edge server. Furthermore, authors in [26] propose a vertical federated learning based ISAC system for human motions recognition. However, the existing ISAC based edge AI designs focus on the training phase. There is lack of scheme for ISAC based edge inference systems. For tackling this issue, this work aims to design ISAC based solution for real-time edge inference systems.

Another main challenge of edge inference system is the lack of theoretical measure to inference accuracy. In the existing work (see [13]–[15]), specific well trained AI models are deployed with fixed inference accuracy, which limits the generalization of these designs. To address this problem, authors in [19] proposes a new metric for classification tasks, called *discriminant gain*, which measures the discernibility between two classes in the Euclidean feature space, as presented in Fig. 1. Specifically, the geometric meaning of discriminant gain between arbitrary two classes is the distance between the corresponding two classes in the feature space under normalized feature covariance. Thereby, with larger discriminant gain, the classes are can be better differentiated, which further leads to greater inference accuracy. As discriminant gain can provide a theoretical guidance for enhancing the inference accuracy, it is adopted in this work. However, challenges arising from its complicated form of covariance normalized distance, as well as the coupling of sensing, quantization, and communication, call for new task-oriented designs.

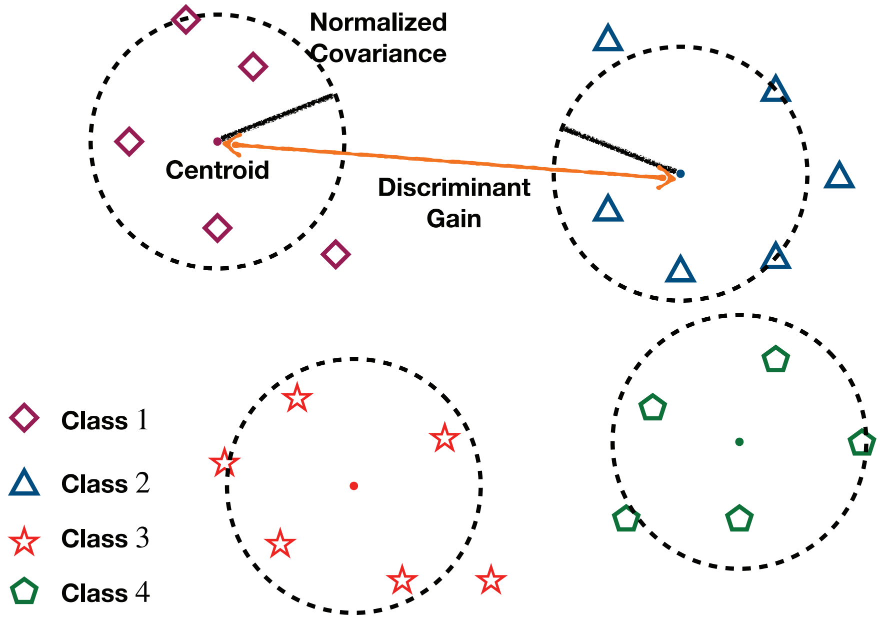


Fig. 1. Geometry of discriminant gain in the feature space.

A. Contributions

To address the challenges mentioned above, in this paper, a multi-view radar sensing based edge inference system for finishing classification tasks is considered. There are one mobile server (e.g., vehicle) and multiple radar devices equipped with DFRC systems. The mobile server requests the radar devices for helping sensing different views of the interested target (e.g., areas) for making real-time decisions (e.g., object detection in auto-driving service). The sensory data are first processed at the radar devices for feature extraction. Then, the feature elements of each device are quantized and transmitted to the mobile server, where they are further cascaded to form the the complete feature vector for AI model inference. The objective of this system is to maximize the inference accuracy measured by discriminant gain in a real-time manner, i.e., finishing the task in a given time duration. The design challenges of such a system arise from the device heterogeneity, the sum of multiple ratios form of discriminant gain, and the highly coupled design of sensing, quantization, and communication. As a result, the inference accuracy maximization problem is non-convex. To tackle this problem, an optimal solution is proposed,which jointly allocates the sensing & communication power, communication time, and quantization bits. The detailed contributions are listed below.

- **ISAC based Edge Inference Systems**: A novel multi-view radar sensing based edge inference system, integrating sensing, quantization, communication, and inference performance, is established in this paper. Specifically, the radar sensing and feature generation models are proposed and well constructed.
- **Optimal Joint Power, Time, and Quantization Bits Allocation**: To address the non-convex inference accuracy maximization problem, the method of sum-of-ratios is adopted, which iteratively solves the problem. In each iteration, a convex problem is first optimized, which minimize the weighted sum of sensing and quantization distortion under given discriminant gain, followed by the updating of the discriminant gain. Then, based on this method, an optimal joint sensing & communication power, communication time, and quantization bits allocation scheme is proposed. In the optimal solution, it is shown that the allocated sensing power and the quantization bits of each device should increase with the number of classes and the target discriminant gain. Besides, the radar device with worse channel gain should be allocated more communication power to finish the feature transmission.
- **Experimental evaluation:** Extensive simulations are conducted to evaluate the performance of our proposed optimal joint power, time, and quantization bits allocation scheme by considering multi-view human motion recognition sensing task with two inference models, i.e., *support vector machine* (SVM) and *multi-layer perception* (MLP) neural network, respectively. It is shown that maximizing the discriminant gain is equivalent to maximizing the inference accuracy for both SVM and MLP neural network. It is also shown that our proposed optimal allocation scheme achieve significantly higher inference accuracy that other two baseline schemes. The superiority of multi-view inference over single-view inference is also validated.

B. Organization

The organization of this paper is as follows. The models and metrics are introduced in Section II. Then, the inference accuracy maximization problem is formulated in Section III. The optimal joint power, time, and quantization bits allocation scheme is proposed in Section IV, followed by the experimental results in Section V. And Section VI concludes the paper.

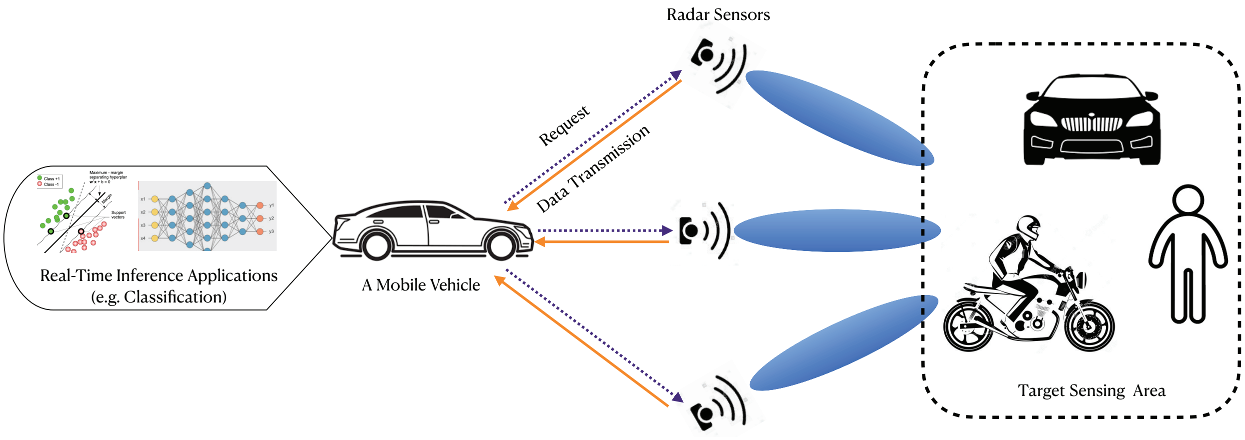


Fig. 2. Edge inference systems with multi-radar sensing.

II. Models and Metrics

In this section, the models of network, radar sensing and feature generation, quantization, and the metric of AI model inference accuracy, say discriminant gain, are introduced.

A. Network Model

The multi-view radar sensing based edge inference system is shown in Fig. 2, where there are one mobile server (e.g., vehicle) with a single antenna *access point* (AP) and K single-antenna mono-static radar devices, each of which is deployed at a fixed place (e.g., crossroads) and is equipped with a DFRC system. The mobile server needs to make a real-time decision, such as obstacle detection in the wild, via inferring a well-trained machine learning model. Its input features are collected from the radars and are the real-time sensing results. To be more specific, the server first make a request to all radars for sensing the environment. Then, the sensing result of each radar is processed and quantized locally to be a subset of features. Next, all the feature subsets are fed back to the mobile device via wireless links for finishing the reference task. Besides, the radar devices remain mute to save the energy consumption when there is no request.

The sensing and communication are performed in a time-division manner, as shown in Fig. 3. The total permitted time to finish the real-time inference task is denoted as T. For an arbitrary device, say the k-th, its sensing time is denoted as $T_{r,k}$, which is assumed to be a constant as in practical the radar's work time is a fixed value. Its communication time to transmit the features

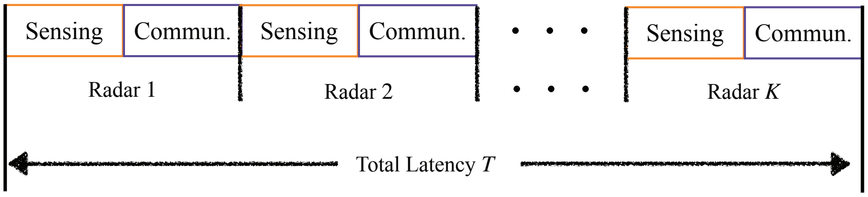


Fig. 3. Edge inference systems with multi-radar sensing.

is denoted as $T_{c,k}$. The total communication bandwidth is denoted as B. Besides, the wireless channels are assumed to be static, as the time duration T is short and smaller than the channel coherence time. The channel gain of the link between the k-th radar and the device (or AP) is denoted as $H_{c,k}$. The AP is assumed to work as a coordinator and can acquire the *global channel state information* (CSI).

B. Radar Sensing and Feature Generation Model

In this section, we first model the radar sensing channel for obtaining the sensing results. Then, the signal processing for feature generation is introduced.

1) Sensing signal: All radars transmit chirp sequences as sensing signal. Consider an arbitrary radar device, say the k-th. According to [25], its echo signal at time t can be written as

$$r_k(t) = u_k(t) + \sum_{q=1}^{Q} v_{k,q}(t) + n_r(t). \tag{1}$$

In (1), $u_k(t)$ is the desired echo signal from the line-of-sight channel and is given by

$$u_k(t) = H_{r,k} s_k(t - \tau), \tag{2}$$

where $H_{r,k}$ is the reflection coefficient including the round-trip path-loss, $s_k(t)$ is the transmitted sensing signal, and τ denotes the round-trip delay. $v_{k,q}(t)$ is the clutter signal from the q-th non-line-of-sight path and is given by

$$v_{k,q}(t) = C_{r,k,q} s_k(t - \tau_q), \tag{3}$$

where $C_{r,k,q}$ is the reflection coefficient from the q-th path; Q is the total number of non-line-of-sight reflection paths; τ_q is the signal delay of the q-th path; $n_r(t)$ is the sensing Gaussian noise. Besides, it is assumed that the values of $H_{r,k}$ and $\sum_{q=1}^{Q} C_{r,k,q}$ can be estimated before sensing.

2) Sensing signal processing: Consider an arbitrary radar device, say the k-th, the steps to process the received radar echo signals are as follows:

Signal Sampling: Assume that a sensing snapshot consisting of N_c chirps, and each chirp has a duration of T_0. For sensing snapshot m, the received signal $r_k(t)$ in (1) is sampled into a complex-valued vector $\mathbf{r}_{k,m} \in \mathbb{C}^{N_c T_0 f_s}$, where f_s is the sampling frequency. Arrange $\mathbf{r}_{k,m}$ in a two-dimensional data matrix $\mathbf{R}_{k,m} \in \mathbb{C}^{T_0 f_s \times N_c}$, in which $T_0 f_s$ is the length of the fast-time dimension, and N_c is the length of the slow-time dimension.

Data filtering: To mitigate the clutter and extract useful information, we apply a *singular value decomposition* (SVD) based linear filter to $\mathbf{R}_{k,m}$ [27]. Specifically, the data matrix after filtering is given by $\tilde{\mathbf{R}}_{k,m} = \sum_{i=r_1}^{r2} \sigma_i \mathbf{v}_i \mathbf{u}_i$, where σ_i, $\mathbf{v}_i$, and $\mathbf{u}_i$ are the i-th singular value, the i-th left-singular vector, and the i-th right-singular vector of $\mathbf{R}_{k,m}$, respectively. r_1 and r_2 are pre-determined parameters.

Feature extraction: We tend to extract features in the slow-time dimension for inference. First, we transform $\tilde{\mathbf{R}}_{k,m}$ into vector $\tilde{\mathbf{r}}_{k,m} \in \mathbb{C}^{1\times N_c}$, i.e., $\tilde{\mathbf{r}}_{k,m} = \mathbf{1}^T \tilde{\mathbf{R}}_{k,m}$. Next, the method of *principle component analysis* (PCA) is further used to extract the principle feature elements from $\tilde{\mathbf{r}}_{k,m}$, and thus make different feature elements independent[1]. The number of features remained is denoted as N_k. Since all the processing steps are linear, for an arbitrary feature element, say n_k, its magnitude, following (1), is given by

$$\bar{r}_k(n_k) = \bar{u}_k(n_k) + \sum_{q=1}^{Q} \bar{v}_{k,q}(n_k) + \bar{n}_r(n_k), \tag{4}$$

where $\bar{r}_k(n_k)$ is the desired ground-true feature, $\bar{v}_k(n_k)$ is additive information in feature element brought by the clutter signal from the q-th path, $\bar{n}_r(n_k)$ is the noise in feature element.

The magnitude of each feature element is normalized by the transmit radar sensing power, say $\sqrt{P_{r,k}}$. For an arbitrary feature element, say n_k, it is given by

$$\hat{x}(n_k) = \frac{r_k(n)}{\sqrt{P_{r,k}}} = x(n_k) + c_{r,k}(n_k) + \frac{n_r(n_k)}{\sqrt{P_{r,k}}}. \tag{5}$$

where $x(n_k) = \bar{u}_k(n_k)/\sqrt{P_{r,k}}$ is the ground-true feature and

$$c_{r,k}(n_k) = \sum_{q=1}^{Q} \frac{\bar{v}_{k,q}(n_k)}{\sqrt{P_{r,k}}}, \tag{6}$$

[1] Note that the principle eigen-space can be obtained during the model training process and is obtained from the AP.

is the normalized clutter. From (5), one can observe that the sensed feature is polluted by the clutter, say $c_{r,k}(n_k)$, and the sensing noise $n_r(n_k)$. Specifically, $c_{r,k}(n_k)$ is assumed to follow Gaussian distribution, as the number of independent reflection paths, say Q, is large. Its distribution is given as

$$c_{r,k}(n_k) \sim \mathcal{N}(0, \sigma_{c,k}^2), \tag{7}$$

where $\mathcal{N}(\cdot,\cdot)$ is the Gaussian distribution, $\sigma_{c,k}^2$ is the constant variance and can be estimated before sensing. Besides, the normalized sensing noise also has a Gaussian distribution, given by

$$n_r(n_k)/\sqrt{P_{r,k}} \sim \mathcal{N}\left(0, \sigma_r^2/P_{r,k}\right), \tag{8}$$

where σ_r^2 is the noise variance.

Then, the feature subset generated by radar device k is

$$\hat{\mathbf{x}}_k = \{\hat{x}(n_k),\ 1 \leq n_k \leq N_k\}, \tag{9}$$

where N_k is the total number of generated features. Furthermore, different feature subsets generated by different radar devices are assumed to be independent, as the radar devices are sparsely deployed and the corresponding sensing areas are non-overlapping.

C. Quantization Model

Consider an arbitrary radar device, say the k-th, whose feature subset is $\hat{\mathbf{x}}_k$. For each feature element therein, they are quantized using the same linear quantizer. Specifically, for an arbitrary feature element, say the n_k-th, according to [28], the quantized feature element is given by

$$z(n_k) = \sqrt{Q_k}\hat{x}(n_k) + d_k, \tag{10}$$

where $z(n_k)$ is the quantized feature element, $\hat{x}(n_k)$ is the radar device generated feature element defined in (5), $\sqrt{Q_k}$ is the quantization gain, d_k is the Gaussian quantization distortion, given as

$$d_k \sim \mathcal{N}(0, \delta_k^2), \tag{11}$$

and δ_k^2 is the variance. At the receiver, the quantized feature is recovered as

$$\tilde{x}(n_k) = \frac{z(n_k)}{\sqrt{Q_k}} = \hat{x}(n_k) + \frac{d_k}{\sqrt{Q_k}}, \tag{12}$$

where the notations follow that in (10). Note that in (12), higher quantization gain, say larger $\sqrt{Q_k}$, can lead to low quantization distortion in the recovered feature at the receiver.

D. Discriminant Gain

Following [19], we use the discriminant gain as the metric for measuring the classification accuracy. Consider an arbitrary feature element, say the n_k-th feature element generated by the k-th radar device, its discriminant gain is modeled as follows.

First, recall the recovered form of the n_k-th feature element at the receiver is given in (12), which, by substituting $\hat{x}(n_k)$ in (5), can be written as

$$\tilde{x}(n_k) = x(n_k) + c_{r,k}(n_k) + \frac{n_r(n_k)}{\sqrt{P_{r,k}}} + \frac{d_k}{\sqrt{Q_k}}, \tag{13}$$

where $x(n_k)$ is the ground-true feature element, $c_{r,k}(n_k)$ is the clutter defined in (7), $n_r(n_k)/\sqrt{P_{r,k}}$ is the normalized sensing noise defined in (8), d_k is the quantization distortion defined in (11), and $\sqrt{Q_k}$ is the quantization gain.

Then, the distribution of $\tilde{x}(n_k)$ is derived. According to [19], the ground-true feature element, say $x(n_k)$ has a mixed Gaussian distribution. Its probability density function is given by

$$f\left(x(n_k)\right) = \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{N}\left(\mu_{\ell,n_k}, \sigma_{n_k}^2\right), \; 1 \leq n_k \leq N_k, \; 1 \leq k \leq K. \tag{14}$$

where L is the total number of classes in the inference task, μ_{ℓ,n_k} is the centroid of the ℓ's class, and $\sigma_{n_k}^2$ is the variance. By substituting the distributions of the ground-true feature in (14), the clutter distribution in (7), the normalized sensing noise in (8), and the quantization distortion in (11), into the recovered feature element $\tilde{x}(n_k)$, its distribution can be derived as

$$f\left(\tilde{x}(n_k)\right) = \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{N}\left(\mu_{\ell,n_k}, \sigma_{n_k}^2 + \sigma_{c,k}^2 + \frac{\sigma_r^2}{P_{r,k}} + \frac{\delta_k^2}{Q_k}\right), \; 1 \leq n_k \leq N_k, \; 1 \leq k \leq K, \tag{15}$$

where the notations follow that in (7), (8), (11), and (14).

Next, according to [19], the discriminant gain of the recovered n_k-th feature element is given by

$$G(n_k) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2}{\sigma_{n_k}^2 + \sigma_{c,k}^2 + \sigma_r^2/P_{r,k} + \delta_k^2/Q_k}, \; 1 \leq n_k \leq N_k, \; 1 \leq k \leq K, \tag{16}$$

where the notations follow that in (15). Furthermore, as different feature elements are independent, the discriminant gain of the k-th feature subset is given by

$$G_k = \sum_{n_k=1}^{N_k} G(n_k), \quad 1 \leq k \leq K. \tag{17}$$

Similarly, the total discriminant of the whole feature vector, i.e., combining all feature subsets from all radar devices, is given by

$$G = \sum_{k=1}^{K} G_k = \sum_{k=1}^{K} \sum_{n_k=1}^{N_k} G(n_k), \tag{18}$$

where $G(n_k)$ is the discriminant gain defined (16).

III. Problem Formulation & Simplification

A. Problem Formulation

In the part, the problem is formulated to maximize the total discriminant gain in (18) under the constraints on latency, successful transmission, and energy. By substituting the discriminant gain of each feature element in (16) into (18), the objective can be written as

$$G = \frac{2}{L(L-1)} \sum_{k=1}^{K} \sum_{n_k=1}^{N_k} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2}{\sigma_{n_k}^2 + \sigma_{c,k}^2 + \sigma_r^2/P_{r,k} + \delta_k^2/Q_k}, \tag{19}$$

where the notations follow that in (16). In the next, the three kinds of constraints are formulated.

1) Latency Constraint: The total allocated sensing and communication time should be less than the permitted latency of the real-time inference task:

$$\text{(C1)} \quad \sum_{k=1}^{K} (T_{r,k} + T_{c,k}) \leq T, \tag{20}$$

where $T_{r,k}$ is the constant sensing time of radar device k, $T_{c,k}$ is the allocated communication time of radar device k, and T is the permitted latency to finish the task.

2) Successful Transmission Constraint: According to the information theory, the following channel capacity conditions should be satisfied for each radar device to successfully transmit the quantized feature subset to the receiver:

$$I(\tilde{\mathbf{x}}_k; \hat{\mathbf{x}}_k) \leq R_k, \ 1 \leq k \leq K, \tag{21}$$

where $\hat{\mathbf{x}}_k$ defined in (9) is the original feature subset generated by radar device k, $\tilde{\mathbf{x}}_k = \{\tilde{x}(n_k), \ 1 \leq n_k \leq N_k\}$ is the successfully recovered quantized feature subset of radar device k at the receiver, $\tilde{x}(n_k)$ is the n_k-th recovered feature element defined in (12), $I(\tilde{\mathbf{x}}_k; \hat{\mathbf{x}}_k)$ is the mutual information of the two vectors, and R_k is the channel capacity of radar device k.

In the next, the condition in (21) is simplified. First, according to the information theory, the mutual information in (21) can be derived as

$$I(\tilde{\mathbf{x}}_k; \hat{\mathbf{x}}_k) = N_k \log_2 \left(1 + \frac{Q_k}{\delta_k^2}\right), \ 1 \leq k \leq K, \tag{22}$$

where $\sqrt{Q_k}$ is the quantization gain and δ_k^2 is the variance of quantization distortion of radar device k. Besides, the channel capacity is given by

$$R_k = T_{c,k} B \log_2 \left(1 + \frac{P_{c,k} H_{c,k}}{\delta_c^2}\right), \ 1 \leq k \leq K, \tag{23}$$

where B is the system bandwidth, δ_c^2 is the channel noise power, $T_{c,k}$ is the allocated time slot, $P_{c,k}$ is the transmit power, and $H_{c,k}$ is the channel gain of radar device k. By substituting (22) and (23) into the transmission constraint in (21), it can be written as

$$\text{(C2)} \quad N_k \log_2 \left(1 + \frac{Q_k}{\delta_k^2}\right) \leq T_{c,k} B \log_2 \left(1 + \frac{P_{c,k} H_{c,k}}{\delta_c^2}\right), \ 1 \leq k \leq K, \tag{24}$$

where the notations follow that in (21), (22), (23).

3) Energy Constraint: The energy consumption of each radar device should be bounded:

$$\text{(C3)} \quad P_{r,k} T_{r,k} + P_{c,k} T_{c,k} \leq E_k, \ \ 1 \leq k \leq K, \tag{25}$$

where $P_{r,k}$ and $P_{c,k}$ are the sensing power and transmit power of radar device k, respectively, $T_{r,k}$ and $T_{c,k}$ are the sensing and communication time of radar device k, respectively, and E_k is the energy threshold of radar device k.

Under the three kinds of constraints above, the problem of maximizing discriminant gain can be formulated as

$$\text{(P1)} \quad \begin{aligned} \max_{P_{c,k}, P_{r,k}, T_{c,k}, Q_k} \quad & G = \frac{2}{L(L-1)} \sum_{k=1}^{K} \sum_{n_k=1}^{N_k} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2}{\sigma_{n_k}^2 + \sigma_{c,k}^2 + \sigma_r^2/P_{r,k} + \delta_k^2/Q_k}, \\ \text{s.t.} \quad & \sum_{k=1}^{K} (T_{r,k} + T_{c,k}) \leq T, \\ & N_k \log_2 \left(1 + \frac{Q_k}{\delta_k^2}\right) \leq T_{c,k} B \log_2 \left(1 + \frac{P_{c,k} H_{c,k}}{\delta_c^2}\right), \ 1 \leq k \leq K, \\ & P_{r,k} T_{r,k} + P_{c,k} T_{c,k} \leq E_k, \ \ 1 \leq k \leq K. \end{aligned} \tag{26}$$

Problem (P1) is a non-convex problem due to the non-convexity of the objective function and constraints (C2) and (C3). In the sequel, an equivalent simplified problem is derived.

B. Problem Simplification

To simplify Problem (P1), the following variables transformation are applied.

$$\begin{cases} S_k = \dfrac{\sigma_r^2}{P_{r,k}}, \\ D_k = \dfrac{\delta_k^2}{Q_k}, \\ E_{c,k} = P_{c,k} T_{c,k}, \end{cases} \tag{27}$$

where S_k, D_k, and $E_{c,k}$ are the normalized sensing noise power, the normalized quantization distortion power, and the communication energy consumption of radar device k, respectively. By substituting (27) into Problem (P1), it can be equivalently derived as

$$\text{(P2)} \quad \begin{aligned} \max_{E_{c,k},S_k,T_{c,k},D_k} \quad & G = \frac{2}{L(L-1)} \sum_{k=1}^{K} \sum_{n_k=1}^{N_k} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2}{\sigma_{n_k}^2 + \sigma_{c,k}^2 + S_k + D_k}, \\ \text{s.t.} \quad & \sum_{k=1}^{K} (T_{r,k} + T_{c,k}) \leq T, \\ & N_k \log_2\left(1 + \frac{1}{D_k}\right) \leq T_{c,k} B \log_2\left(1 + \frac{E_{c,k} H_{c,k}}{T_{c,k} \delta_c^2}\right), \ 1 \leq k \leq K, \\ & \frac{\sigma_r^2 T_{r,k}}{S_k} + E_{c,k} \leq E_k, \ \ 1 \leq k \leq K. \end{aligned} \tag{28}$$

In Problem (P2), all constraints are convex but the objective function still remains to be non-convex. To tackle it, a method, called sum-of-ratios, is used in the following.

IV. Optimal Joint Sensing & Transmit Power, Time, and Quantization Bits Allocation

In this section, an optimal scheme, namely joint sensing & transmit power, time, and quantization bits allocation, is proposed to solve Problem (P2). First, It is shown that Problem (P2) can be handled by an iterative algorithm, called sum-of-ratios, where in each iteration a convex sub-problem should be solved. Next, we solve the convex sub-problems of all iterations using an alternating method and obtain some insightful results. Finally, the scheme of joint sensing & transmit power, time, and quantization bits allocation is summarized.

A. The Sum-of-Ratios Method

It can be shown that that the sum-of ratios method can be applied to solve Problem (P2), as shown in the lemma below.

Lemma 1. *Problem (P2) can be optimally solved using the sum-of-ratios method.*

Proof: See Appendix A.

In the sequel, the sum-of-ratios method in [29] is used to tackle Problem (P2). First, its objective is rewritten as

$$G = \sum_{k=1}^{K} \sum_{n_k=1}^{N_k} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\mathcal{A}_{\ell,\ell',n_k}}{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)}, \tag{29}$$

where $\mathcal{A}_{\ell,\ell',n_k}$ and $\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)$ are

$$\begin{cases}\mathcal{A}_{\ell,\ell',n_k}=1,\\ \mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)=\dfrac{L(L-1)(\sigma_{n_k}^2+\sigma_{c,k}^2+S_k+D_k)}{2\left(\mu_{\ell,n_k}-\mu_{\ell',n_k}\right)^2},\end{cases}\quad \forall(\ell,\ell^{'},n_k). \tag{30}$$

Then, to solve Problem (P2), a sub-problem is created as follows.

$$\text{(P3)}\quad \begin{aligned}\max_{E_{c,k},S_k,T_{c,k},D_k}\quad & \sum_{k=1}^{K}\sum_{n_k=1}^{N_k}\sum_{\ell'=1}^{L}\sum_{\ell<\ell'} x_{\ell,\ell',n_k}\left[\mathcal{A}_{\ell,\ell',n_k}-y_{\ell,\ell',n_k}\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)\right],\\ \text{s.t.}\quad & N_k\log_2\left(1+\frac{1}{D_k}\right)\leq T_{c,k}B\log_2\left(1+\frac{E_{c,k}H_{c,k}}{T_{c,k}\delta_c^2}\right),\ 1\leq k\leq K,\\ & \sum_{k=1}^{K}(T_{c,k}+T_{r,k})\leq T,\\ & \frac{\sigma_r^2T_{r,k}}{S_k}+E_{c,k}\leq E_k,\ \ 1\leq k\leq K,\end{aligned} \tag{31}$$

where $\left\{x_{\ell,\ell',n_k}\right\}$ and $\left\{y_{\ell,\ell',n_k}\right\}$ are the introduced auxiliary variables, and $\mathcal{A}_{\ell,\ell',n_k}$ and $\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)$ are defined in (30). In Problem (P3), each term in the objective function is a scale of the sum of sensing and quantization distortion, giving its name of *weighted sum of distortion minimization problem.*

Lemma 2 (Convexity of Problem (P3))**.** *Problem (P3) is a convex problem.*

Proof: By substituting $\mathcal{A}_{\ell,\ell',n_k}$ and $\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)$ in (30), the objective function can be shown to be linear. Besides, the two constraints are shown to be convex in Appendix A. Hence, Problem (P4) is convex. This ends the proof.

Next, based on Lemma 2 and according to [29], the optimal solution of Problem (P2) can be obtained by iteratively performing the following two steps until convergence.

- *Step 1*: Optimally solving Problem (P3) with given auxiliary variables $\left\{x_{\ell,\ell',n_k}\right\}$ and $\left\{y_{\ell,\ell',n_k}\right\}$.
- *Step 2*: Updating the auxiliary variables $\left\{x_{\ell,\ell',n_k}\right\}$ and $\left\{y_{\ell,\ell',n_k}\right\}$ as

$$\begin{aligned}x_{\ell,\ell',n_k}&=\frac{1}{\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)},\ \forall(\ell,\ell^{'},n_k),\\ y_{\ell,\ell',n_k}&=\frac{\mathcal{A}_{\ell,\ell',n_k}}{\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)}=\frac{1}{\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)},\ \forall(\ell,\ell^{'},n_k),\end{aligned} \tag{32}$$

where $\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)$ is defined in (30). From the above equation, it can be observed that

$$x_{\ell,\ell',n_k}=y_{\ell,\ell',n_k}, \tag{33}$$

and they are actually the discriminant gain of feature n_k between the classes ℓ and $\ell^{'}$. In summary, the solving process iterates over addressing the weighted sum of distortion minimization problem under given discriminant gain and updating the discriminant gain.

B. An Alternating Method

In this section, an alternating algorithm is proposed to solve the convex sub-problem, say Problem (P3) with given auxiliary variables $\left\{x_{\ell,\ell',n_k}\right\}$ and $\left\{y_{\ell,\ell',n_k}\right\}$. Specifically, Problem (P3) is solved by alternately addressing the problem of time-aware joint power and quantization bits allocation and the problem of communication time allocation. In the next, the alternated two sub-problems are first introduced, followed by the summarization of the alternating algorithm.

1) Time-Aware Joint Power and Quantization Bits Allocation: In this case, the communication time, say $\{T_{c,k}\}$, is given. Hence, by substituting (33) and $\mathcal{A}_{\ell,\ell',n_k}$ in (30), Problem (P3) can be written as

$$\text{(P4)} \quad \begin{aligned} \max_{E_{c,k},S_k,D_k} \quad & \sum_{k=1}^{K}\sum_{n_k=1}^{N_k}\sum_{\ell'=1}^{L}\sum_{\ell<\ell'}\left[y_{\ell,\ell',n_k}-y_{\ell,\ell',n_k}^2\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)\right], \\ \text{s.t.} \quad & N_k\log_2\left(1+\frac{1}{D_k}\right)\leq T_{c,k}B\log_2\left(1+\frac{E_{c,k}H_{c,k}}{T_{c,k}\delta_c^2}\right), \ \ 1\leq k\leq K, \\ & \frac{\sigma_r^2T_{r,k}}{S_k}+E_{c,k}\leq E_k, \ \ 1\leq k\leq K, \end{aligned} \tag{34}$$

which is a convex problem.

KKT conditions are used to solve Problem (P4). The Lagrange function is given by

$$\begin{aligned} \mathcal{L}_{\text{P4}} = & -\sum_{k=1}^{K}\sum_{n_k=1}^{N_k}\sum_{\ell'=1}^{L}\sum_{\ell<\ell'}\left[y_{\ell,\ell',n_k}-y_{\ell,\ell',n_k}^2\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)\right], \\ & +\sum_{k=1}^{K}\alpha_k\left[N_k\log_2\left(1+\frac{1}{D_k}\right)-T_{c,k}B\log_2\left(1+\frac{E_{c,k}H_{c,k}}{T_{c,k}\delta_c^2}\right)\right], \\ & +\sum_{k=1}^{K}\beta_k\left(\frac{T_{r,k}}{S_k}+E_{c,k}-E_k\right), \end{aligned} \tag{35}$$

where $\{\alpha_k\geq 0\}$ and $\{\beta_k\geq 0\}$ are the corresponding Lagrangian multipliers, and $\mathcal{B}_{\ell,\ell',n_k}\left(S_k,D_k\right)$ is defined in (30).

The first condition can be written as

$$\frac{\partial\mathcal{L}_{\text{P4}}}{\partial S_k}=\sum_{\ell'=1}^{L}\sum_{\ell<\ell'}\left(y_{\ell,\ell',n_k}^2\times\frac{\partial\mathcal{B}_{\ell,\ell',n_k}}{\partial S_k}\right)-\frac{\beta_kT_{r,k}}{S_k^2}=0, \quad 1\leq k\leq K, \tag{36}$$

where, according to (30),

$$\frac{\partial B_{\ell,\ell',n_k}}{\partial S_k} = \frac{L(L-1)}{2\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2}. \tag{37}$$

It follows that

$$\frac{1}{S_k} = \sqrt{\sum_{\ell'=1}^{L}\sum_{\ell<\ell'} \frac{L(L-1)y^2_{\ell,\ell',n_k}}{2\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2} \times \frac{1}{\beta_k T_{r,k}}}. \tag{38}$$

Besides, by substituting S_k in (27) into (38), the following optimal sensing power allocation scheme can be obtained.

Lemma 3 (Optimal Sensing Power Allocation)**.** *The optimal allocated sensing power of radar device k should be*

$$P_{r,k} = \sigma_r^2 \times \sqrt{\sum_{\ell'=1}^{L}\sum_{\ell<\ell'} \frac{L(L-1)y^2_{\ell,\ell',n_k}}{2\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2} \times \frac{1}{\beta_k T_{r,k}}}, \quad 1 \leq k \leq K, \tag{39}$$

where $\{\beta_k\}$ are the Lagrangian multipliers.

From (39), there are several observations. For an arbitrary radar device, say the k-th, first, if the number of classes, say L, is large, or the required discriminant gains of arbitrary two classes, say $\{y_{\ell,\ell',n_k}\}$, is large, more power should be allocated for sensing. Then, if the centroid distances of arbitrary two classes, say $\{\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2\}$, are large, or the sensing distortion, say σ_r^2, is small, the required sensing power can be alleviated. Besides, long sensing time, say larger $T_{r,k}$, can also reduce the required sensing power.

The second condition is given by

$$\frac{\partial \mathcal{L}_{\text{P4}}}{\partial D_k} = \sum_{\ell'=1}^{L}\sum_{\ell<\ell'} \left(y^2_{\ell,\ell',n_k} \times \frac{\partial \mathcal{B}_{\ell,\ell',n_k}}{\partial D_k} \right) - \frac{\alpha_k N_k \ln 2}{D_k(D_k+1)} = 0, \quad 1 \leq k \leq K, \tag{40}$$

which, by substituting $\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)$ in (30), can be derived as

$$D_k = \sqrt{\frac{1}{4} + \frac{\alpha_k N_k \ln 2}{\sum\limits_{\ell'=1}^{L}\sum\limits_{\ell<\ell'} \frac{L(L-1)y^2_{\ell,\ell',n_k}}{2\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2}}} - \frac{1}{2}, \quad 1 \leq k \leq K, \tag{41}$$

where $\{\alpha_k\}$ are the Lagrangian multipliers. By substituting D_k in (27) in to (41), the following lemma can be derived.

Lemma 4 (Optimal Quantization Bits Allocation)**.** *The optimal quantization gain is*

$$Q_k = \frac{\delta_k^2}{D_k}, \quad 1 \leq k \leq K, \tag{42}$$

where δ_k^2 *is the quantization distortion and* D_k *is defined in* (41).

Several observations can be made from (42). For an arbitrary radar device, say the k-th, larger number of classes, say L, larger number of feature elements, and larger required discriminant gains, say $\{y_{\ell,\ell',n_k}\}$, call for larger quantization gain (or level). On the other hand, larger centroid distances between arbitrary two classes, say $\{(\mu_{\ell,n_k} - \mu_{\ell',n_k})^2\}$, require smaller quantization gain.

The third KKT condition can be written as

$$\frac{\partial \mathcal{L}_{\text{P4}}}{\partial E_{c,k}} = -\frac{\alpha_k B T_{c,k} H_{c,k}}{(E_{c,k} H_{c,k} + T_{c,k}\delta_c^2)\ln 2} + \beta_k = 0. \tag{43}$$

It follows that

$$E_{c,k} = \frac{\alpha_k B T_{c,k}}{\beta_k \ln 2} - \frac{T_{c,k}\delta_c^2}{H_{c,k}}. \tag{44}$$

Since $E_{c,k}$ should be non-negative, the optimal $E_{c,k}$ is given by

$$E_{c,k} = \max\left\{\frac{\alpha_k B T_{c,k}}{\beta_k \ln 2} - \frac{T_{c,k}\delta_c^2}{H_{c,k}}, \quad 0\right\}. \tag{45}$$

By substituting $E_{c,k}$ in (27) into (45), the following optimal power allocation can be derived.

Lemma 5 (Optimal Communication Power Allocation)**.** *The optimal communication power for each radar device should be*

$$P_{c,k} = \max\left\{\frac{\alpha_k B}{\beta_k \ln 2} - \frac{\delta_c^2}{H_{c,k}}, \quad 0\right\}, \quad 1 \leq k \leq K. \tag{46}$$

From (46), it can be observed the device with worse wireless link gain, say smaller $H_{c,k}$, should be allocated more communication power to finish the feature transmission.

Then, based on the results above, the primal-dual method can be used to solve Problem (P4), as summarized in Algorithm 1, where η_{α_k} and η_{β_k} are the step sizes to update the multipliers α_k and β_k, respectively.

2) Communication Time Allocation: In this case, the sensing noise $\{S_k\}$, communication energy $\{E_{c,k}\}$, and quantization distortion $\{D_k\}$ are first solved by Algorithm 1. To determine the communication time allocation, say $T_{c,k}$, a total sensing and communication time minimization problem is proposed, under the successful transmission constraint, as shown in Problem (P5).

$$\text{(P5)} \quad \begin{aligned} T^* = \min_{T_{c,k}} \quad & \sum_{k=1}^{K} (T_{c,k} + T_{r,k}) \\ \text{s.t.} \quad & N_k \log_2\left(1 + \frac{1}{D_k}\right) \leq T_{c,k} B \log_2\left(1 + \frac{E_{c,k} H_{c,k}}{T_{c,k}\delta_c^2}\right), \ 1 \leq k \leq K. \end{aligned} \tag{47}$$

Algorithm 1: Joint Power and Quantization Bits Allocation

```
Input: Channel gains $\{H_{c,k}\}$, auxiliary variables $y_{\ell,\ell',n_k}$, and the given communication latencies $\{T_{c,k}\}$.
Initialize $\{\alpha_k^{(0)}\}$, $\{\beta_k^{(0)}\}$, and $i = 0$.
Loop
    Solve $\{S_k\}$, $\{D_k\}$, and $\{E_{c,k}\}$ using (38), (41), and (45), respectively.
    Update the multipliers as
        $\begin{cases} \alpha_k^{(i+1)} = \max\left\{ \alpha_k^{(i)} + \eta_{\alpha_k} \frac{\partial \mathcal{L}_{\text{P4}}}{\partial \alpha_k}, \quad 0 \right\}, \ 1 \leq k \leq K, \\ \beta_k^{(i+1)} = \max\left\{ \beta_k^{(i)} + \eta_{\beta_k} \frac{\partial \mathcal{L}_{\text{P4}}}{\partial \beta_k}, \quad 0 \right\}, \ 1 \leq k \leq K, \end{cases}$
    $i = i + 1$.
Until Convergence
Calculate $\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)$ using (30).
Output: $\left\{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)\right\}$, $\{S_k\}$, $\{D_k\}$, and $\{E_{c,k}\}$.
```

By solving the total time T^* in Problem (P5) and comparing it with the permitted latency T, the time allocation of Problem (P3) can be solved, as described below.

- *Case of* $T^* > T$*:* In this case, the given $\{S_k\}$, $\{D_k\}$, and $\{E_{c,k}\}$ are not in the feasible region of Problem (P3). The reason is that the latency constraint therein can not be satisfied. To this end, the latency of all radar devices should be reduced to satisfy the constraint.
- *Case of* $T^* < T$*:* In this case, more time can be allocated to all radar devices to achieve discriminant gain in Problem (P3).
- *Case of* $T^* = T$*:* The current time allocation is optimal.

Based on the conclusions above, in the sequel, the solution of Problem (P5) is first derived. Then, the time update for each device is designed.

First, to solve Problem (P5), its Lagrange function can be derived as

$$\mathcal{L}_{\text{P5}} = \sum_{k=1}^{K} (T_{c,k} + T_{r,k}) + \sum_{k=1}^{K} \lambda_k \left[N_k \log_2 \left(1 + \frac{1}{D_k} \right) - T_{c,k} B \log_2 \left(1 + \frac{E_{c,k} H_{c,k}}{T_{c,k} \delta_c^2} \right) \right], \tag{48}$$

where $\{\lambda_k \geq 0\}$ are the Lagrangian multipliers. As Problem (P5) is convex, the primal-dual method can be used to obtain the optimal solution, where the optimizer are denoted as $\{T_{c,k}^*\}$.

Then, the communication time updating to fulfill the latency constraint is designed as follows.

$$T_{c,k} = T^{*}_{c,k} + \frac{\mu_k}{\sum_{k=1}^{K} \mu_k} \times (T - T^{*}), \quad 1 \leq k \leq K, \tag{49}$$

where T^{*} is the solved optimal total time, $T^{*}_{c,k}$ is the solved optimal communication time of radar device k, and μ_k is defined as

$$\mu_k = \left. \frac{\partial \mathcal{L}_{\text{P5}}}{\partial \lambda_k} \right|_{\lambda_k = \lambda_k^{*}}. \tag{50}$$

The design of communication time updating in (49) is based on two facts. One is that after the updating, the latency constraint can be satisfied and all permitted time is used to enhance the discriminant gain. The other is that the radar device with higher value of μ_k calls for more time.

Lemma 6 (Enhanced Discrimiant Gain via Additioanl Time Allocation)**.** *In the design in* (49)*, more communication time is allocated to each device. This leads to a strict increase to the value of the objective in Problem (P3), i.e., the discriminant gain can be enhanced.*

Proof: See Appendix B.

Overall, the primal dual method to solve Problem (P5) and the communication time updating are summarized in Algorithm 2, where η_{λ_k} and η_k are the step sizes, and

$$\frac{\partial \mathcal{L}_{\text{P5}}}{\partial T_{c,k}} = 1 - \lambda_k \left[B \log_2 \left(1 + \frac{E_{c,k} H_{c,k}}{T_{c,k} \delta_c^2} \right) + \frac{E_{c,k} H_{c,k}}{(E_{c,k} H_{c,k} + T_{c,k} \delta_c^2) \ln 2} \right], \; 1 \leq k \leq K. \tag{51}$$

3) Alternating Algorithm for Solving Problem (P3): By alternately using Algorithms 1 and 2, Problem (P3) can be optimally solved, as summarized in Algorithm 3. Besides, on one hand, Probelm (P3) is convex. On the other hand, the objective of Problem (P3) strictly increases after each alternating iteration, unless the convergence is achieved. Hence, according to [30], Algorithm 3 is guaranteed to have a linear convergence rate.

C. Solution to Problem (P2)

Based on the previous results, Problem (P2) can be optimally solved using the method of sum-or-ratios, together with the alternating algorithm in Algorithm 3. The detailed procedure is summarized in Algorithm 4.

Algorithm 2: Communication Time Allocation

```
Input: $\{S_k\}$, $\{E_{c,k}\}$, and $\{D_k\}$.
Initialize $\{\lambda_k^{(0)}\}$, and $i = 0$.
Loop
    Update the multipliers as
        $\lambda_k^{(i+1)} = \max\left\{\lambda_k^{(i)} + \eta_{\lambda_k}\frac{\partial \mathcal{L}_{\mathrm{P5}}}{\partial \lambda_k}, \quad 0\right\}, \ 1 \leq k \leq K.$
    Initialize $T_{c,k}^{(0)}$ and $t = 0$.
    Loop
        $T_{c,k}^{(t+1)} = \max\left\{T_{c,k}^{(t)} - \eta_k \frac{\partial \mathcal{L}_{\mathrm{P5}}}{\partial T_{c,k}^{(t)}}, \ 0\right\}.$
        $t = t + 1.$
    Until Convergence
Until Convergence
$\{T_{c,k}^{*} = T_{c,k}, \ \forall k\}$ and calculate $T^{*}$.
Update the communication time $\{T_{c,k}\}$ using (49).
Output: $\{T_{c,k}\}$.
```

Algorithm 3: Alternating Algorithm for Solving Problem (P3)

```
Input: Channel gains $\{H_{c,k}\}$ and auxiliary variables $y_{\ell,\ell',n_k}$.
Initialize communication time $\{T_{c,k}\}$.
Loop
    Solve sensing noise $\{S_k\}$, quantization distortion $\{D_k\}$, and communication energy
    $\{E_{c,k}\}$ and discriminant gains $\left\{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)\right\}$, using Algorithm 1.
    Solve communication time $\{T_{c,k}\}$ using Algorithm 2.
Until Convergence
Output: $\left\{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)\right\}$, $\{S_k\}$, $\{D_k\}$, $\{E_{c,k}\}$, and $\{T_{c,k}\}$.
```

Algorithm 4: Sum-of-Ratios Based Algorithm for Solving Problem (P2)

```
Input: Channel gains $\{H_{c,k}\}$.
Initialize auxiliary variables $\{y_{\ell,\ell',n_k}\}$.
Loop
    Solve Problem (P3) under given $\{y_{\ell,\ell',n_k}\}$, using Algorithm 3, and get $\{S_k\}$, $\{D_k\}$,
    $\{E_{c,k}\}$, $\{T_{c,k}\}$, and $\{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)\}$.
    Update the auxiliary variables as
        $y_{\ell,\ell',n_k} = \frac{1}{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)}, \forall(\ell, \ell^{'}, n_k),$
Until Convergence
Output: $\{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)\}$, $\{S_k\}$, $\{D_k\}$, $\{E_{c,k}\}$, and $\{T_{c,k}\}$.
```

TABLE I

SIMULATION PARAMETERS

Parameter	**Value**
Number of radar devices, K	3
Sensing noise variance, σ_r^2	1
Clutter variance, $\sigma_{c,k}^2$	1, 0.1, 0.5
Quantization variance, δ_k^2	1
Number of features after PCA, N_K	50
Number of classes, L	4
Permitted latency, T	1.55 s
Energy threshold, E_k	0.12 Joule
Variance of shadow fading, σ_ζ^2	8 dB
Communication channel noise power, δ_c^2	10^{-12} W
Bandwidth for communication, B	200 Hz
Bandwidth for sensing, B_s	10 MHz
Sensing carrier frequency, f_c	60 GHz
Chirp duration, T_0	$10\mu s$
Unit sensing time, $T_{r,k}$	0.5 s
Sampling rate, f_s	10 MHz

V. PERFORMANCE EVALUATION

A. Experiment Setup

1) Communication model: In this experiment, we consider a network of $K = 3$ radar devices, which are randomly located in a circular area of radius 50 meters. The distance between the circle center and the AP is 450 meters. The channel gain H_k is modeled as $H_k = |\varphi_k h_k|^2$, where φ_k and h_k are the large-scale fading propagation coefficient and small-scale fading propagation coefficient, respectively. The large-scale propagation coefficient in dB from device k to the edge server is modeled as $[\varphi_k]_{\text{dB}} = -[\text{PL}_k]_{\text{dB}} + [\zeta_k]_{\text{dB}}$, where $[\text{PL}_k]_{\text{dB}} = 128.1 + 37.6 \log_{10} \text{dist}_k$ (dist_k is the distance in kilometer) is the path loss in dB, and $[\zeta_k]_{\text{dB}}$ is the shadow fading in dB. In this simulation, $[\zeta_k]_{\text{dB}}$ is Gauss-distributed random variable with mean zero and variance σ_ζ^2. The small-scale fading is assumed to be Rayleigh fading, i.e., $h_k \sim \mathcal{CN}(0,1)$.

2) Sensing task: In our simulation, the sensing task is to identify four different human motions, i.e., *child walking*, *child pacing*, *adult walking*, and *adult pacing*. The heights of children and adults are uniformly distributed in interval $[0.9\text{m}, 1.2\text{m}]$ and $[1.6\text{m}, 1.9\text{m}]$, respectively. The speed of standing, walking, and pacing are 0 m/s, $0.5H$ m/s, and $0.25H$ m/s, respectively, where H is the height value. The heading of the moving human is set to be uniformly distributed in $[-180^\circ, 180^\circ]$.

3) Inference model: Two machine learning models, i.e., a SVM and MLP neural network, are considered for inference in the experiments, respectively. The neural network model has 2 hidden layers with 80 and 40 neurons, respectively. Both models are trained on 800 data samples without any distortion, i.e., sensing clutter, sensing noise, and quantization distortion. The inference experiments for test accuracy are implemented over 200 data samples with distortion.

Unless specifically given otherwise, other simulation parameters are stated in Table I. All experiments are implemented using Python 3.8 on a Linux server with one NVIDIA® GeForce® RTX 3090 GPU 24GB and one Intel® Xeon® Gold 5218 CPU.

B. Inference Algorithms

For comparison, we consider three allocation schemes as follows.

- *Baseline*: This scheme randomly selects a feasible point for Problem (P1).
- *Power-aware allocation*: The sensing power is first allocated as same as in the above baseline and then the other parameters are allocated by the scheme in **Algorithm 4**.
- *Optimal allocation (our proposal)*: All the parameters are allocated by the optimal scheme of joint sensing & transmit power, time, and quantization bits allocation in **Algorithm 4**.

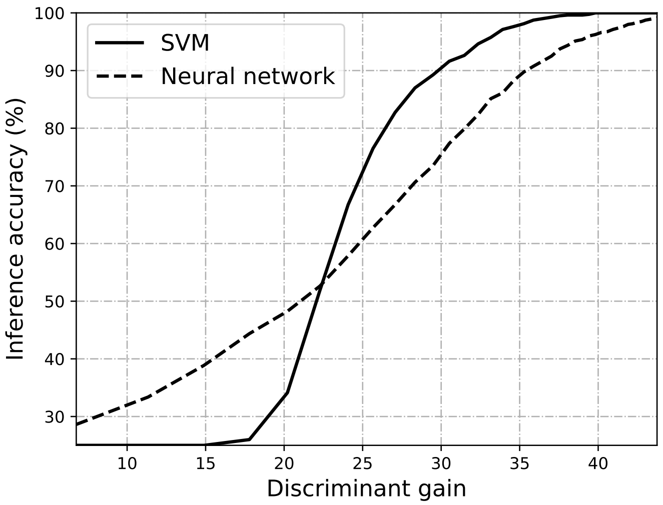


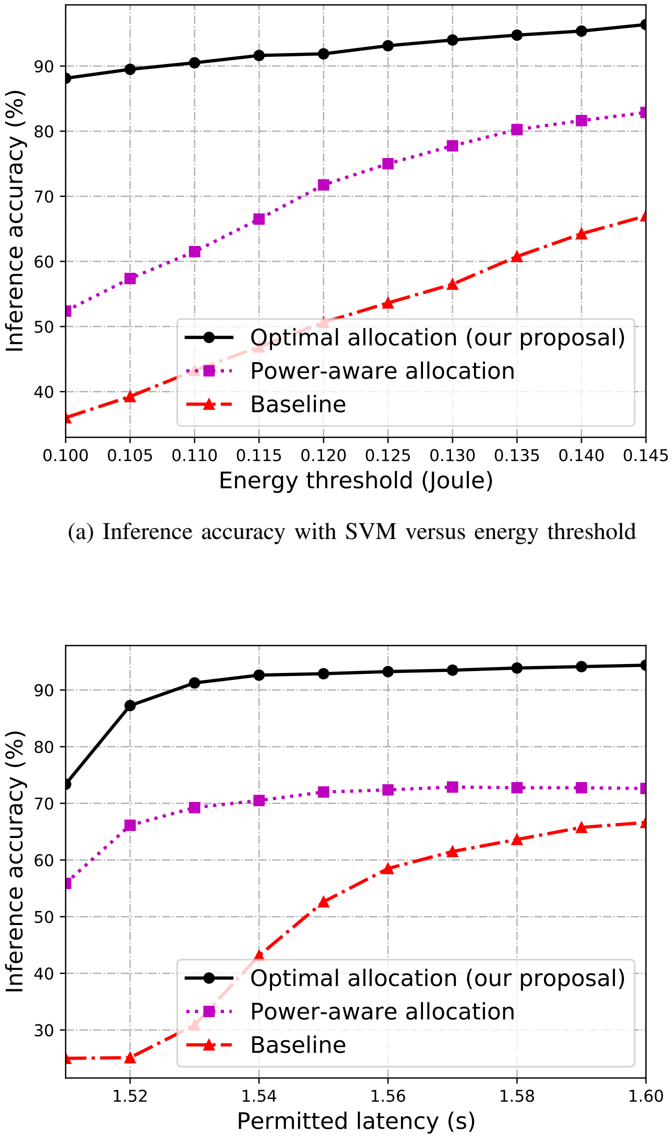
Fig. 4. Inference accuracy versus discriminant gain

C. Experimental Results

In this part, the relations between the inference accuracy and discriminant gain regarding the two models are first presented. Then, the three algorithms are compared in terms of the SVM model and the neural network, respectively. Finally, the influence of number of participated devices on the inference accuracy is shown.

1) Inference accuracy v.s. discriminant gain: The relations between the inference accuracy and discriminant gain regarding the SVM model and the MLP neural network are shown in Fig. 4. It can be observed that the inference accuracy increases as the discriminant gain grows for both models. Besides. When the discriminant gain is large, i.e., the distortion of the samples caused by sensing and quantization is small, the SVM outperforms the neural network. That's because the training of the neural network is overfitting as its model is complicated compared to its training dataset size. However, the neural network is more robust than the SVM when the discriminant gain is small, i.e., the distortion is large.

2) Inference accuracy of SVM: The inference accuracy of the SVM model is presented in Fig 5. From the figure, the performance of all schemes increases as the resources, i.e., energy threshold of each device and the permitted latency for the inference task, increase. Besides, the proposed optimal allocation scheme outperforms the other two baseline schemes. Furthermore, in the case of long permitted latency, the performance of the power-aware allocation scheme



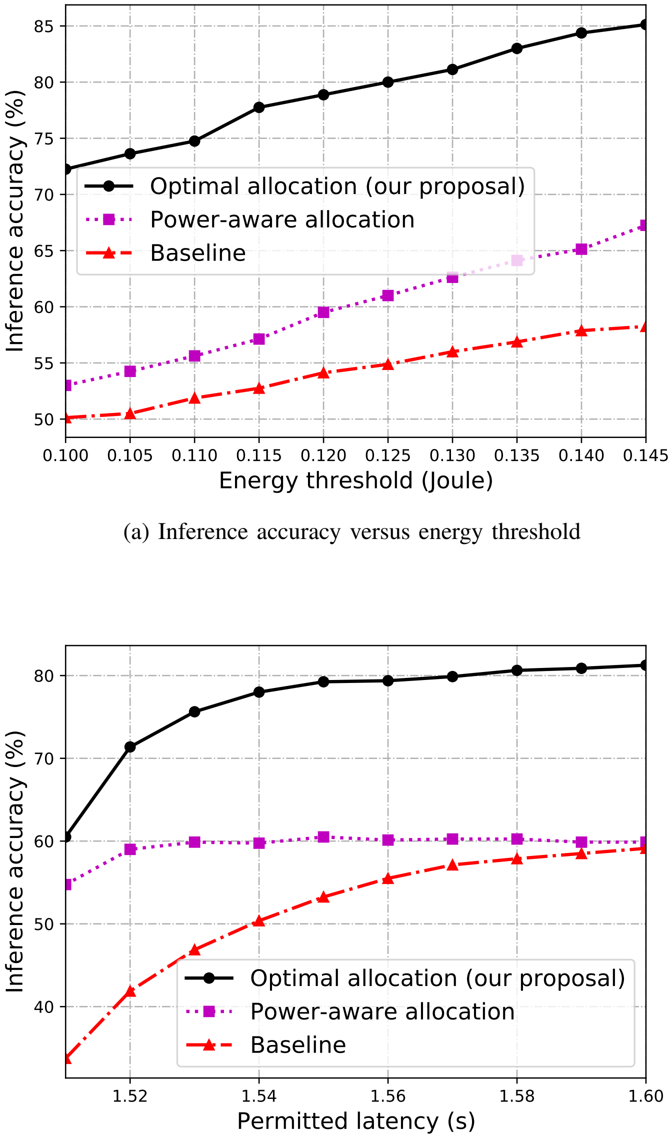
(a) Inference accuracy with SVM versus energy threshold

(b) Inference accuracy with SVM versus permitted latency

Fig. 5. Performance comparison of the SVM among different schemes

remains unchanged as the permitted latency continuously increases. The reason is that the sensing distortion is dominant in this case.

3) Inference accuracy of neural network: The inference accuracy of the MLP neural network model in terms of the energy threshold and the permitted latency is shown in Fig. 6. Again, as more resources can be allocated, the performance of all schemes increase. Besides, the proposed optimal allocation scheme achieves the best performance. And the longer permitted latency will



(a) Inference accuracy versus energy threshold

(b) Inference accuracy versus permitted latency

Fig. 6. Performance comparison of the neural network among different schemes

not lead to better performance for the power-aware allocation scheme when the latency is large, for a similar reason in the scenario of the SVM model.

4) Inference accuracy v.s. number of radar devices: In Fig. 7, the inference accuracy of both models in terms of different number of radar devices are presented. For both cases, as the number of devices increases, better inference accuracy can be achieved. The reason is that providing more features to the inference task can lead to a larger feature space, which can further

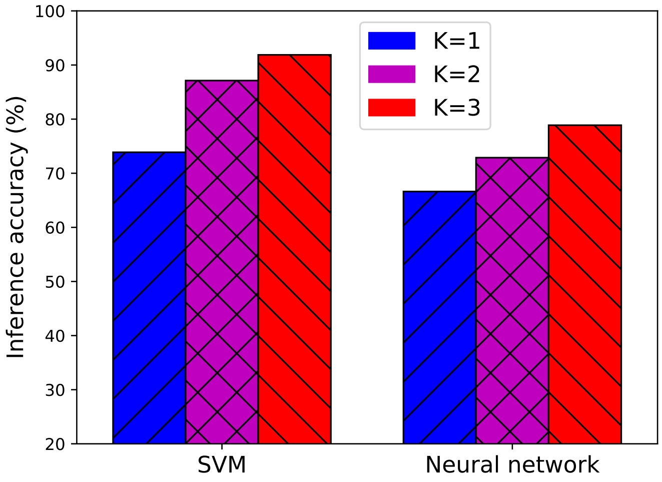


Fig. 7. Inference accuracy comparison among different models under different number of radar devices

make the distance, i.e., the discriminant gain, between arbitrary two different classes lager.

The extensive experimental results above show that the proposed optimal scheme of joint sensing & transmit power, time, and quantization bits allocation has the best performance and verify our theoretical analysis.

VI. Concluding Remarks

In this paper, an ISAC based efficient edge inference framework is built based on a multi-view DFRC system. Specifically, under the objective of inference accuracy maximization under given latency, an optimal joint sensing & transmit power, time, and quantization bits allocation scheme is proposed.

This work opens several interesting directions. One is the radar device scheduling, i.e., the feature selection, for inference accuracy maximization when the radio resources, e.g., time and frequency bands, are scarce or the permitted latency of the task is even shorter. It's desirable to design a scheme of joint device selection, and sensing and radio resource allocation. Another is to enhance the inference accuracy in the broadband systems with frequency-selective wireless channels, where the technique of *orthogonal frequency division multiple access* (OFDMA) should be applied and the binary subcarrier allocation together with sensing & transmit power allocation should be addressed.

Appendix

A. Proof of Lemma 1

The objective function of (P2) can be re-written as

$$G = \sum_{k=1}^{K} \sum_{n_k=1}^{N_k} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\mathcal{A}_{\ell,\ell',n_k}}{\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)}, \tag{52}$$

where

$$\begin{cases} \mathcal{A}_{\ell,\ell',n_k} = 1, \\ \mathcal{B}_{\ell,\ell',n_k}(S_k, D_k) = \dfrac{L(L-1)(\sigma_{n_k}^2 + \sigma_{c,k}^2 + S_k + D_k)}{2\left(\mu_{\ell,n_k} - \mu_{\ell',n_k}\right)^2}, \end{cases} \quad \forall(\ell, \ell^{'}, n_k). \tag{53}$$

In the objective, all $\{\mathcal{A}_{\ell,\ell',n_k}\}$ are constants. Besides, $\{-\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)\}$ for all $(\ell, \ell^{'}, n_k)$ are convex. Hence, Problem (P2) can be optimally solved by the sum-of-ratios method if its feasible region is convex, according to [29].

In the next, we will show that the three constraints are convex. To begin with, the first constraint in Problem (P2) is

$$\sum_{k=1}^{K} (T_{r,k} + T_{c,k}) \leq T, \tag{54}$$

which forms a linear set and hence is convex. Then, in the second constraint, say

$$N_k \log_2\left(1 + \frac{1}{D_k}\right) \leq T_{c,k} B \log_2\left(1 + \frac{E_{c,k} H_{c,k}}{T_{c,k}\delta_c^2}\right),\ 1 \leq k \leq K, \tag{55}$$

the left part is convex, as its second derivative is positive:

$$\frac{2N_k D_k + 1}{D_k^2 (D_k+1)^2 \ln 2} \geq 0,\ \forall D_k > 0. \tag{56}$$

The right part of the second constraint can be linearly transformed from $f(x, y) = x\log_2(1+y/x)$. $f(x, y)$ is concave, as the inverse of its Hessian matrix is positive, i.e.,

$$\begin{bmatrix} \dfrac{y^2}{x(x+y)^2 \ln 2}, & -\dfrac{y}{(x+y)^2 \ln 2} \\ -\dfrac{y}{(x+y)^2 \ln 2}, & \dfrac{x}{(x+y)^2 \ln 2} \end{bmatrix} > 0. \tag{57}$$

As linear transformation preserves convexity, the right part of the second constraint is a concave function. Thus, the second constraint forms a convex set. Next, the third constraint in Problem (P2), i.e.,

$$\frac{\sigma_r^2 T_{r,k}}{S_k} + E_{c,k} \leq E_k,\ \ 1 \leq k \leq K, \tag{58}$$

also forms a convex set. Hence, the feasible region of Problem (P2) is convex and the problem can be optimally solved by the sum-of-ratios method.

B. Proof of Lemma 6

In Problem (P3), the second constraint is

$$N_k \log_2\left(1+\frac{1}{D_k}\right) \leq T_{c,k} B \log_2\left(1+\frac{E_{c,k} H_{c,k}}{T_{c,k}\delta_c^2}\right),\ 1 \leq k \leq K, \tag{59}$$

whose right part is a strictly descresing function of $T_{c,k}$. That's to say, with increasing $T_{c,k}$, smaller communication energy $E_{c,k}$ can be used to satisfy this constraint for each device. Then, cosider the final constraint in Prolem (P3), given as

$$\frac{\sigma_r^2 T_{r,k}}{S_k} + E_{c,k} \leq E_k,\ \ 1 \leq k \leq K, \tag{60}$$

where smaller $E_{c,k}$ can lead to smaller sensing noise S_k. Next, according to $\mathcal{B}_{\ell,\ell',n_k}(S_k, D_k)$ defined in (30), it is a linearly increasing function of S_k. Hence, the objective function of Prolem (P3) increases, which further leads to an ehanced discriminant gain according to (29).

References

[1] 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, pp. 19–25, Jan. 2020.

[2] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 5–36, 2021.

[3] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

[4] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167–2191, 2020.

[5] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE Journal on Selected Areas in Communications*, 2021.

[6] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020.

[7] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *[Online]. Available: https://arxiv.org/abs/1510.00149*, 2015.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[9] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[10] M. Lee, G. Yu, and H. Dai, "Decentralized inference with graph neural networks in wireless communication systems," *IEEE Transactions on Mobile Computing*, 2021.

[11] K. Yang, Y. Shi, W. Yu, and Z. Ding, "Energy-efficient processing and robust wireless cooperative transmission for edge inference," *IEEE internet of things journal*, vol. 7, no. 10, pp. 9456–9470, 2020.

[12] S. Hua, Y. Zhou, K. Yang, Y. Shi, and K. Wang, "Reconfigurable intelligent surface for green edge inference," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 2, pp. 964–979, 2021.

[13] 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 *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, IEEE, 2019.

[14] X. Huang and S. Zhou, “Dynamic compression ratio selection for edge inference systems with hard deadlines,” *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8800–8810, 2020.

[15] E. Li, L. Zeng, Z. Zhou, and X. Chen, “Edge AI: On-demand accelerating deep neural network inference via edge computing,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 447–457, 2019.

[16] J. Shao and J. Zhang, “Communication-computation trade-off in resource-constrained edge inference,” *IEEE Communications Magazine*, vol. 58, no. 12, pp. 20–26, 2020.

[17] M. Jankowski, D. Gündüz, and K. Mikolajczyk, “Joint device-edge inference over wireless links with pruning,” in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, 2020.

[18] J. Shao, Y. Mao, and J. Zhang, “Learning task-oriented communication for edge inference: An information bottleneck approach,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, 2021.

[19] Q. Lan, Q. Zeng, P. Popovski, D. Gündüz, and K. Huang, “Progressive feature transmission for split inference at the wireless edge,” *[Online]. Available: https://arxiv.org/abs/2112.07244*, 2021.

[20] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, “Integrated sensing and communications: Towards dual-functional wireless networks for 6g and beyond,” *IEEE Journal on Selected Areas in Communications*, 2022.

[21] F. Liu, L. Zhou, C. Masouros, A. Li, W. Luo, and A. Petropulu, “Toward dual-functional radar-communication systems: Optimal waveform design,” *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4264–4279, 2018.

[22] Y. Huang, Y. Fang, X. Li, and J. Xu, “Coordinated power control for network integrated sensing and communication,” *[Online]. Available: https://arxiv.org/abs/2203.09032*, 2022.

[23] Y. He, Y. Cai, H. Mao, and G. Yu, “Ris-assisted communication radar coexistence: Joint beamforming design and analysis,” *arXiv preprint arXiv:2201.07399*, 2022.

[24] X. Li, F. Liu, Z. Zhou, G. Zhu, S. Wang, K. Huang, and Y. Gong, “Integrated sensing and over-the-air computation: Dual-functional mimo beamforming design,” *arXiv preprint arXiv:2201.12581*, 2022.

[25] T. Zhang, S. Wang, G. Li, F. Liu, G. Zhu, and R. Wang, “Accelerating edge intelligence via integrated sensing and communication,” *[Online]. Available: https://arxiv.org/abs/2107.09574*, 2021.

[26] P. Liu, G. Zhu, W. Jiang, W. Luo, J. Xu, and S. Cui, “Vertical federated edge learning with distributed integrated sensing and communication,” *arXiv preprint arXiv:2201.08512*, 2022.

[27] 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 *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 481–485, 2021.

[28] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, “Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks,” *IEEE Trans. Sig. Process.*, vol. 61, pp. 5646–5658, Aug. 2013.

[29] Y. Jong, “An efficient global optimization algorithm for nonlinear sum-of-ratios problem,” *Optimization Online*, pp. 1–21, 2012.

[30] N. H. Tran, W. Bao, A. Zomaya, M. N. Nguyen, and C. S. Hong, “Federated learning over wireless networks: Optimization model design and analysis,” in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 1387–1395, IEEE, 2019.