

UAV-Assisted Integrated Sensing, Communication, and Computation Edge Inference Network With Early-Exit Mechanism

Guofang Wu, Yejun He*, and Xiaowen Cao

State Key Laboratory of Radio Frequency Heterogeneous Integration

Sino-British Antennas and Propagation Joint Laboratory, Ministry of Science and Technology of China

Guangdong Engineering Research Center of Base Station Antennas and Propagation

Shenzhen Key Laboratory of Antennas and Propagation

College of Electronics and Information Engineering, Shenzhen University, 518000, China

E-mail: wuguofangwy@163.com, heyejun@126.com*, caoxwen@szu.edu.cn

Abstract—Integrated sensing, communication, and computation (ISCC) is a key enabling technology for 6G networks and can significantly enhance the overall performance of edge intelligent systems through the joint optimization of data acquisition, transmission, and computation processes. This paper investigates an unmanned aerial vehicle (UAV) assisted ISCC edge inference system that enables adaptive inference via an early-exit mechanism. To establish a quantitative relationship between system resource allocation and inference accuracy, discrimination gain is introduced as a performance metric, and the relationship among discrimination gain, early-exit point selection, and inference accuracy is derived based on Bayesian decision theory. On this basis, an optimization problem is formulated to minimize the total system energy consumption while maximizing the number of completed inference tasks. The problem jointly optimizes sensing power, communication power, UAV hovering location, and early-exit point selection strategy. To solve this mixed-integer nonlinear programming problem (MINLP), an efficient algorithm based on an alternating optimization framework is proposed. Simulation results demonstrate that, compared with benchmark schemes, the proposed algorithm significantly reduces system energy consumption while ensuring the task completion rate, which verifies the effectiveness of the proposed scheme.

Index Terms—Integrated sensing, communication, and computation, edge inference, early-exit, unmanned aerial vehicle, resource allocation.

I. INTRODUCTION

With the ongoing research on the sixth-generation mobile communication system (6G), integrated sensing and communication (ISAC) has become one of the key research directions. This technology enables the cooperative implementation of communication and sensing functions on the same hardware platform, thereby significantly improving the utilization efficiency of spectrum and hardware resources [1], [2]. Meanwhile, the explosive growth of Internet of Things devices has led to a large number of intelligent applications that are sensitive to computational resources. Traditional mobile terminals, constrained by limited computing capability and battery capacity, struggle to support complex artificial intelligence models [3]. Although cloud computing provides powerful centralized computing resources, the latency introduced by

long-distance data transmission often fails to satisfy real-time requirements. Mobile edge computing (MEC) deploys servers at the network edge to process computational tasks locally, which effectively reduces transmission latency and provides a feasible solution for low-latency intelligent applications [4]. ISAC and MEC are typically studied from the sensing-communication and communication-computation perspectives, respectively. At the same time, the low-altitude economy is emerging as a new growth area. Unmanned aerial vehicles (UAVs), owing to their high mobility and rapid deployment capability, can serve as aerial mobile intelligent platforms that provide flexible communication and computing services for remote areas or emergency scenarios, thereby acting as an important complement to ground-based fixed infrastructure [5].

For future 6G networks, the intelligent inference capability of terminal devices is of critical importance. To achieve high-performance inference, it is necessary to jointly optimize the entire process of sensing, communication, and computation from a system-level perspective [6]. Consequently, the integration of ISAC and MEC to enable integrated sensing, communication, and computation (ISCC) has become a promising research paradigm [7], [8]. However, conventional inference performance metrics, such as accuracy, are difficult to analytically connect with system resource allocation. To address this challenge, discrimination gain is introduced as a tractable metric to characterize inference performance by measuring class separability in the feature space. By establishing the relationship between system resources and discrimination gain, theoretical guidance for ISCC resource optimization can be obtained [9].

In addition, early-exit techniques in deep neural networks provide an effective mechanism to reduce inference latency and energy consumption by allowing adaptive termination of inference at different network depths according to task requirements [10]–[12]. This flexibility makes early-exit inference particularly suitable for resource-constrained edge environments.

Despite recent advances, most existing task-oriented ISCC designs primarily focus on data-level optimization, while the joint consideration of model structure, such as early-exit mechanisms, remains largely unexplored [9], [13]. To fill this gap, this paper proposes a UAV-assisted ISCC edge inference framework with an early-exit mechanism, which jointly optimizes system resources from both data and model structure perspectives. The main contributions of this paper are summarized as follows.

- An integrated system model covering the entire process of sensing data acquisition, wireless transmission, and edge inference is established, which incorporates sensing models, communication models, and an inference model based on the early-exit mechanism. Based on Bayesian decision theory, the relationships among discrimination gain, early-exit point selection, and inference accuracy are derived, thereby providing a theoretical foundation for system optimization.
- For multi-task scenarios, with the dual objectives of minimizing total system energy consumption and maximizing the number of completed tasks, a mixed-integer nonlinear programming problem (MINLP) is formulated by jointly optimizing sensing power, communication power, UAV hovering locations, and early-exit point selection strategies. To handle the non-convexity and complexity of the problem, an efficient algorithm based on an alternating optimization framework is proposed. By combining offline calibration and online optimization, near-optimal resource configurations are achieved. Extensive simulation results validate the superiority of the proposed scheme in terms of task completion rate, energy consumption, and latency.

The remainder of this paper is organized as follows. Section II introduces the system model and presents the derivation of the accuracy model. Section III formulates the optimization problem and presents the proposed optimization framework. Section IV provides simulation results and performance analysis. Section V concludes the paper and discusses future research directions.

II. SYSTEM MODEL

A. System Architecture

Consider an edge inference system consisting of an UAV equipped with an edge server and N edge devices, where the device set is denoted by $\mathcal{N} = 1, \dots, N$. Each edge device is capable of integrated sensing and communication. A deep neural network with multiple exit points is jointly deployed on the edge devices and the UAV. The first i_e layers of the model are always executed on the device side for unified front-end feature extraction, while the remaining layers can be executed either locally on the device or remotely on the UAV according to task requirements. When the accuracy requirement of an inference task is low, the model may exit at a shallow layer to reduce computational and energy consumption. The system framework is illustrated in Fig. 1.

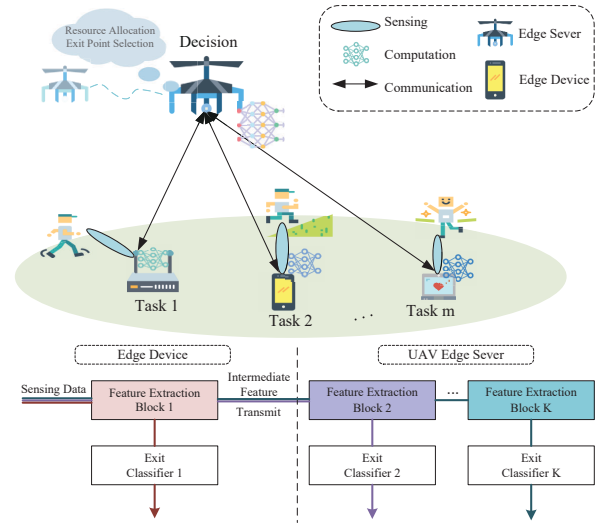


Fig. 1. System framework of the proposed UAV-assisted ISCC edge inference network.

The system operates as follows. After an edge device generates an inference task, the UAV, acting as a central controller, determines the sensing power, communication power, and inference exit point based on the task-specific latency and accuracy requirements. The edge device then performs sensing and front-end feature extraction according to the scheduling decision, and subsequently either completes the inference locally or uploads the extracted intermediate features to the UAV to finish the remaining inference process.

B. Task Model

Let the set of inference tasks be denoted by $\mathcal{M} = 1, \dots, M$. The subset of tasks generated by device $n \in \mathcal{N}$ is denoted by $\mathcal{M}_n$, where $\mathcal{M}_n \subseteq \mathcal{M}$, $\bigcup_{n=1}^N \mathcal{M}_n = \mathcal{M}$, and $\mathcal{M}_x \cap \mathcal{M}_y = \emptyset$ for $x \neq y$. Each task $m \in \mathcal{M}_n$ is associated with a latency requirement T_m^{req} and an accuracy requirement A_m^{req} .

For task m , the UAV selects an exit point $e_m \in \mathcal{K}$, where $\mathcal{K} = 1, \dots, K$ denotes the set of available exit points. When $e_m = 1$, the edge device executes all layers of the model after completing the first i_e layers. When $e_m \geq 2$, joint inference is performed, where the edge device executes the first i_e layers and transmits the extracted intermediate features to the UAV, which then executes the remaining layers up to exit point e_m .

The task completion indicator is defined as

$$\delta_m = \begin{cases} 1, & \text{if } T_m \leq T_m^{\text{req}} \text{ and } A_m \geq A_m^{\text{req}}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where T_m and A_m denote the actual latency and inference accuracy of task m , respectively.

C. Sensing Model

Each edge device transmits a frequency-modulated continuous-wave signal for target sensing. Each sensing time slot consists of D chirps, and the duration of each chirp is T_0 . Accordingly, the sensing duration of task m is given by

$T_m^s = DT_0$. The transmitted signal of device n is expressed as

$$s_n(t) = \sqrt{P_m^s} \sum_{d=0}^{D-1} \text{rect}\left(\frac{t-dT_0}{T_0}\right) \times \cos\left(2\pi f_c(t-dT_0) + \pi \frac{B_s}{T_0}(t-dT_0)^2\right), \quad (2)$$

where P_m^s denotes the sensing power, f_c is the carrier frequency, and B_s is the sensing bandwidth. The energy consumption required for the sensing process is denoted by $E_m^s = P_m^s T_m^s$. The rectangular window function is defined as

$$\text{rect}\left(\frac{t-dT_0}{T_0}\right) = \begin{cases} 1, & dT_0 < t < (d+1)T_0, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

After mixing the transmitted signal with the normalized received signal and filtering out the high-frequency components, the intermediate-frequency signal is obtained as

$$r_n(t) = e_n(t) + u_n(t) + \frac{v_{s,n}(t)}{\sqrt{P_m^s}}, \quad (4)$$

where $e_n(t)$ denotes the target signal, $u_n(t)$ represents clutter following $\mathcal{N}(0, \sigma_{n,u}^2)$, and $v_{s,n}(t)$ denotes sensing noise following $\mathcal{N}(0, \sigma_v^2)$. With a sampling rate f_s and $Q = f_s T_0$ samples per chirp, the sampled signal is transformed into a range-Doppler map $\mathbf{R}_n \in \mathbb{C}^{Q \times D}$ via fast Fourier transform along the range and Doppler dimensions.

The target is assumed to belong to one of L classes, following a Gaussian mixture distribution $f(\tilde{e}_n) = \frac{1}{L} \sum_{l=1}^L \mathcal{N}(\mu_{l,n}, \sigma_n^2)$.

After sampling the noisy received signal, the resulting observation $\tilde{r}_n$ follows

$$f(\tilde{r}_n) = \frac{1}{L} \sum_{l=1}^L \mathcal{N}\left(\mu_{l,n}, \sigma_n^2 + \sigma_{n,u}^2 + \frac{\sigma_v^2}{P_m^s}\right). \quad (5)$$

D. Computation Model

During inference, different network layers incur different computational loads. The layer type of the i -th layer is denoted by $\pi(i) \in \text{CV, MP, FC}$, corresponding to convolution, max-pooling, and fully connected layers, respectively [14]. The number of floating-point operations required by layer i is given by

$$\lambda(i) = \begin{cases} (2\gamma(i-1)\psi(i)^2 - 1)\alpha(i)\beta(i)\gamma(i), & \pi(i) = \text{CV}, \\ \alpha(i)\beta(i)\gamma(i)\psi(i)^2, & \pi(i) = \text{MP}, \\ (2n(i') - 1)n(i), & \pi(i) = \text{FC}, \end{cases} \quad (6)$$

where $\alpha(i)$, $\beta(i)$, $\gamma(i)$, and $\psi(i)$ denote the height, width, number of channels, and kernel size of the output feature map, respectively. For a fully connected layer, $n(i)$ denotes the number of neurons. The computational capacities of the edge device and the UAV are denoted by c_e and c_s , respectively.

T_m^{comp} and E_m^{comp} denote the computation time and energy consumption of task m , respectively. Given the exit point e_m , the total number of layers that need to be executed in the

inference network is denoted by I_{e_m} . When $e_m = 1$, all layers are executed on the edge device, yielding

$$T_m^{\text{comp}} = \sum_{i=1}^{I_1} \frac{\lambda(i)}{c_e}, \quad (7)$$

$$E_m^{\text{comp}} = \kappa_1 \sum_{i=1}^{I_1} \lambda(i) c_e^2. \quad (8)$$

When $e_m \geq 2$, the edge device executes the first i_e layers, and the UAV executes the remaining layers up to I_{e_m} , resulting in

$$T_m^{\text{comp}} = \sum_{i=1}^{i_e} \frac{\lambda(i)}{c_e} + \sum_{i=i_e+1}^{I_{e_m}} \frac{\lambda(i)}{c_s}, \quad (9)$$

$$E_m^{\text{comp}} = \kappa_1 \sum_{i=1}^{i_e} \lambda(i) c_e^2 + \kappa_2 \sum_{i=i_e+1}^{I_{e_m}} \lambda(i) c_s^2, \quad (10)$$

where κ_1 and κ_2 denote computation energy coefficients.

E. Communication Model

When $e_m \geq 2$, the edge device is required to transmit the intermediate feature vector $\mathbf{z}_n$, with a size of ψ_m bits, to the UAV. The location of the edge device is denoted by $\mathbf{q}_e$, and the UAV hovers at location $\mathbf{q}_u$. The Euclidean distance between the edge device and the UAV is given by $\|\mathbf{q}_u - \mathbf{q}_e\|$. Let $g_n(\mathbf{q}_u)$ denote the channel power gain between the edge device and the UAV, which is modeled as $g_n(\mathbf{q}_u) = \frac{\beta_0}{\|\mathbf{q}_u - \mathbf{q}_e\|^2}$, where β_0 represents the channel power gain at the reference distance of 1m.

The achievable communication rate between the edge device and the UAV is given by

$$R_m = B_m \log_2 \left(1 + \frac{P_m^c g_n(\mathbf{q}_u)}{\sigma_c^2} \right), \quad (11)$$

where B_m denotes the allocated bandwidth, P_m^c is the transmission power, and σ_c^2 is the noise power.

Accordingly, the transmission latency and communication energy consumption for task m are expressed as $T_m^c = \frac{\psi_m}{R_m}$, $E_m^c = P_m^c T_m^c$.

F. Classification Accuracy Analysis and Early-Exit Modeling

1) *Feature Distribution and Discrimination Gain Modeling*: The fixed front-end network consisting of the first i_e layers maps the sensed data to an intermediate feature vector denoted by $\mathbf{z}_n = \Phi(\tilde{r}_n)$. For task m , the class-conditional distribution of the extracted feature is given by

$$f_{\mathbf{z}_n} = \frac{1}{L} \sum_{l=1}^L \mathcal{N}(\mu_{l,n}, \eta^2 \Sigma_m \mathbf{I}), \quad (12)$$

where $\Sigma_m = \sigma_n^2 + \sigma_{n,u}^2 + \frac{\sigma_v^2}{P_m^s}$, and η^2 characterizes the compression effect of the front-end network on the input data distribution.

When the intermediate feature is transmitted over the wireless channel, the equivalent noise variance increases to

$$\sigma_{\text{total},m}^2 = \eta^2 \Sigma_m + \frac{\sigma_c^2}{P_m^c g_n(\mathbf{q}_u)}. \quad (13)$$

Accordingly, the minimum discrimination gain between class l and class l' for task m is defined as

$$DG_{\min,m} = \min_{l \neq l'} \frac{\|\boldsymbol{\mu}_{l,n} - \boldsymbol{\mu}_{l',n}\|^2}{\sigma_{\text{total},m}^2}. \quad (14)$$

According to Bayesian decision theory, under a two-class Gaussian model with equal prior probabilities [15], the error probability for class pair (l, l') is given by

$$P_e^{(l,l')} = Q\left(\frac{\sqrt{DG_{l,l'}}}{2}\right), \quad (15)$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt$. Since $DG_{l,l'} \geq DG_{\min,m}$ for all $l \neq l'$, we have $P_e^{(l,l')} \leq Q\left(\frac{\sqrt{DG_{\min,m}}}{2}\right)$.

For an L -class classification problem with equal prior probabilities, the overall error probability $P_e = \Pr(\hat{Y} \neq Y)$ satisfies the following inequality (by the union bound):

$$P_e \leq \frac{1}{L} \sum_{l=1, l' \neq l}^L P_e^{(l,l')} \leq (L-1)Q\left(\frac{\sqrt{DG_{\min,m}}}{2}\right). \quad (16)$$

Therefore, a lower bound on the classification accuracy is given by

$$A_m \geq \left[1 - (L-1)Q\left(\frac{\sqrt{DG_{\min,m}}}{2}\right)\right]_{(1/L)+}, \quad (17)$$

where $[\cdot]_{(1/L)+} \triangleq \max\{\cdot, 1/L\}$ prevents the bound from falling below the random-guessing probability $1/L$.

2) *Early-Exit Modeling and Accuracy Approximation*: Let ϱ_k denote the benchmark accuracy of the k -th exit point under ideal conditions. To jointly capture model capability and data quality, the inference accuracy of task m is approximated as

$$A_m \approx \frac{1}{L} + \left(\varrho_{e_m} - \frac{1}{L}\right) p_m, \quad (18)$$

where

$$p_m = \frac{\left[1 - (L-1)Q\left(\frac{\sqrt{DG_{\min,m}}}{2}\right)\right]_{(1/L)+} - \frac{1}{L}}{1 - \frac{1}{L}} \in [0, 1] \quad (19)$$

is a normalized quality factor satisfying $p_m \rightarrow 0$ when data is indiscriminable and $p_m \rightarrow 1$ under ideal conditions, ensuring $A_m \in [1/L, \varrho_{e_m}]$.

The total energy consumption of the UAV consists of the hovering energy and the computation energy incurred at the UAV side. Specifically, it is expressed as

$$E_{\text{uav}} = P_{\text{hover}} T_{\text{hover}} + \sum_{m: e_m \geq 2} E_{m,u}^{\text{comp}}, \quad (20)$$

where P_{hover} denotes the hovering power, T_{hover} is the maximum task latency within the current scheduling period, and $E_{m,u}^{\text{comp}}$ represents the computation energy consumption of task m at the UAV.

III. PROBLEM FORMULATION AND PROPOSED OPTIMIZATION FRAMEWORK

A. Problem Formulation

Given the task set $\mathcal{M}$, and subject to the latency requirement T_m^{req} and accuracy requirement A_m^{req} of each task, we jointly optimize the sensing power P_m^s , communication power P_m^c , exit point selection e_m , and UAV hovering location $\mathbf{q}_u$ to maximize the number of completed tasks while minimizing the total energy consumption. The optimization problem is formulated as

$$\begin{aligned} & \max_{\{P_m^s, P_m^c, e_m, \mathbf{q}_u\}} \sum_{m \in \mathcal{M}} (\epsilon \delta_m - \omega E_m) \\ & \text{s.t.} \quad 0 \leq P_m^s \leq P_s^{\max}, \quad \forall m \in \mathcal{M}, \\ & \quad 0 \leq P_m^c \leq P_c^{\max}, \quad \forall m \in \mathcal{M}, \\ & \quad e_m \in \{1, 2, \dots, K\}, \quad \forall m \in \mathcal{M}, \\ & \quad \mathbf{q}_u \in \mathcal{Q}, \\ & \quad E_{\text{uav}} \leq E_{\text{uav}}^{\max}, \end{aligned} \quad (21)$$

where $\epsilon > 0$ and $\omega > 0$ are weighting factors that balance the importance of task completion and energy consumption, respectively. $\mathcal{Q}$ denotes the feasible hovering region of the UAV, which is constrained by flight altitude limits and operational area requirements. $P_s^{\max}$ and $P_c^{\max}$ denote the maximum allowable sensing power and communication power, respectively, and $E_{\text{uav}}^{\max}$ represents the total energy budget of the UAV.

B. Optimization Framework

The formulated problem is a MINLP problem involving both discrete and continuous decision variables. For a given UAV hovering location $\mathbf{q}_u$ and a selected exit point $k \in \mathcal{K}$, the sensing energy consumption $E_m^s = P_m^s T_m^s$ monotonically increases with the sensing power P_m^s , while the task accuracy A_m monotonically increases with P_m^s . Therefore, within the feasible region satisfying the accuracy constraint $A_m > A_m^{\text{req}}$, the optimal sensing power that minimizes the energy consumption is the minimum power that satisfies the accuracy constraint. Consequently, the accuracy constraint is tight at the optimal point. Similarly, since the communication delay T_m^c monotonically decreases with the communication power P_m^c , the optimal communication power under the delay constraint is the minimum power that satisfies the delay budget. Hence, the delay constraint is tight at the optimal point. Based on the above monotonicity properties and the resulting tight constraints, the original problem can be efficiently solved using a hierarchical alternating optimization framework, as illustrated in Fig. 2.

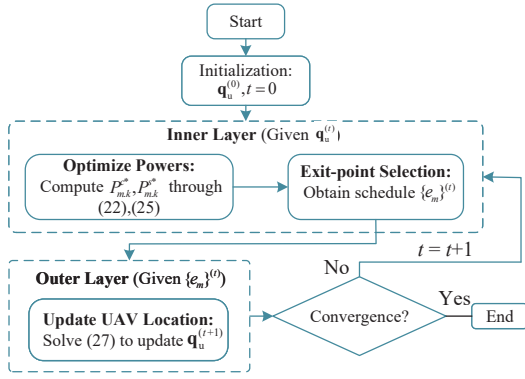


Fig. 2. Alternating optimization algorithm flowchart.

1) *Inner Layer: Task Configuration and Selection for a Given $\mathbf{q}_u$* : Given a UAV location $\mathbf{q}_u$ and the corresponding channel gain $g_n(\mathbf{q}_u)$, we compute the minimum power pair $(P_{m,k}^{s*}, P_{m,k}^{c*})$ for each task m and exit point $k \in \mathcal{K}$ satisfying latency and accuracy constraints, subsequently evaluating the associated costs and utility.

a) *Closed-form Communication Power for $k \geq 2$* : Let $T_{m,k}^{\text{fix}} \triangleq T_m^s + T_{m,k}^{\text{comp}}$ and $T_{m,k}^{\text{bud}} \triangleq T_m^{\text{req}} - T_{m,k}^{\text{fix}}$. If $T_{m,k}^{\text{bud}} \leq 0$, the exit point k is infeasible. Otherwise, the required transmission rate is given by $R_{m,k}^{\text{req}} = \frac{\psi_m}{T_{m,k}^{\text{bud}}}$. By inverting the Shannon capacity formula, the minimum communication power required to satisfy the latency constraint is obtained as

$$P_{m,k}^{c*} = \frac{\sigma_c^2}{g_n(\mathbf{q}_u)} \left(2^{\frac{R_{m,k}^{\text{req}}}{B_m}} - 1 \right), \quad k \geq 2. \quad (22)$$

If $P_{m,k}^{c*} > P_c^{\text{max}}$, the corresponding exit point is infeasible. For $k = 1$, we set $P_{m,1}^{c*} = 0$.

b) *Sensing Power Minimization*: Given k and $P_{m,k}^{c*}$, the inference accuracy is approximated by

$$A_m(k, P_m^s, P_{m,k}^{c*}, \mathbf{q}_u) = \frac{1}{L} + \left(\varrho_k - \frac{1}{L} \right) p_m, \quad (23)$$

where p_m is defined in (19) with

$$DG_{\min,m} = \min_{l \neq l'} \frac{\|\boldsymbol{\mu}_{l,n} - \boldsymbol{\mu}_{l',n}\|^2}{\eta^2 \left(\sigma_n^2 + \sigma_{n,u}^2 + \frac{\sigma_v^2}{P_m^s} + \frac{\sigma_c^2}{P_{m,k}^{c*} g_n(\mathbf{q}_u)} \right)}. \quad (24)$$

As accuracy A_m and sensing energy $E_m^s = P_m^s T_m^s$ both monotonically increase with P_m^s , the optimal sensing power lies on the constraint boundary:

$$P_{m,k}^{s*} = \min \left\{ P_m^s \in [0, P_s^{\text{max}}] : A_m(k, P_m^s, P_{m,k}^{c*}, \mathbf{q}_u) \geq A_m^{\text{req}} \right\}. \quad (25)$$

This is efficiently solved via bisection search over $[0, P_s^{\text{max}}]$. Infeasibility is declared if no solution exists.

c) *Cost and Utility Evaluation*: For each feasible exit point k , the total latency and energy consumption of task m are computed as $T_{m,k} = T_m^s + T_{m,k}^c + T_{m,k}^{\text{comp}}$, $E_{m,k} = P_{m,k}^{s*} T_m^s + P_{m,k}^{c*} T_{m,k}^c + E_{m,k}^{\text{comp}}$. The utility of task m under exit point k is then defined as $U_{m,k} = \epsilon - \omega E_{m,k}$, $\delta_{m,k} = 1$. If exit point k is infeasible, we set $U_{m,k} = -\infty$.

d) *Global Selection Under UAV Energy Constraint*: The total UAV energy comprises hovering and computation components:

$$E_{\text{uav}}(\mathbf{e}) = P_{\text{hover}} T_{\text{hover}}(\mathbf{e}) + \sum_{m: e_m \geq 2} E_{m,u}^{\text{comp}}, \quad (26)$$

where $T_{\text{hover}}(\mathbf{e}) = \max_m T_{m,e_m}$. To decouple the max operator given $\mathbf{q}_u$, we search over a hovering time threshold τ , restricting choices to $T_{m,k} \leq \tau$. Under the derived energy budget $\sum_{m: e_m \geq 2} E_{m,u}^{\text{comp}} \leq E_{\text{uav}}^{\text{max}} - P_{\text{hover}} \tau$, we solve a multiple-choice knapsack problem to maximize $\sum_m U_{m,e_m}$ using a value-density based heuristic to efficiently obtain a near-optimal solution.

2) *Outer Layer: UAV Hovering Location Update*: In iteration t , fixing $\mathbf{q}_u^{(t-1)}$, the inner layer yields the optimal schedule. To update $\mathbf{q}_u$, we minimize the smooth approximation of the hovering time:

$$\tilde{T}_{\text{hover}}(\mathbf{q}_u) = \frac{1}{\rho} \log_2 \left(\sum_{m: e_m^{(t)} \geq 2} \exp(\rho T_{m,e_m^{(t)}}(\mathbf{q}_u)) \right). \quad (27)$$

The update is performed via gradient descent with projection onto $\mathcal{Q}$, ensuring monotonic utility improvement.

IV. SIMULATION RESULTS

The simulation employs a ResNet-18 inference model with three exit points. The UAV service area is 1000 m $\times$ 1000 m, with a flight altitude of 50 m. Devices are distributed in clusters, with cluster centers located at (50, 50) and (150, 150) meters. To validate the effectiveness of the proposed joint optimization framework, the proposed method is evaluated against three baseline schemes, as described below.

- **No Early-Exit mechanism (NoEE)**: The intermediate exit points of the DNN are disabled, and all tasks are required to perform full inference at the final exit (Exit 3).
- **Fixed UAV position (NoPos)**: The UAV is fixed at the center of the service area, and only task scheduling is optimized.
- **Greedy scheduling strategy (Greedy)**: A greedy strategy is adopted to address the coupling between hovering time and task selection.

Fig. 3 illustrates the number of completed tasks versus the UAV energy budget under different strategies. It can be observed that the proposed scheme achieves the highest number of completed tasks across all energy budget levels. In contrast, the NoEE scheme performs poorly under low energy budgets, since executing the complete inference model for each task incurs high energy consumption, which significantly limits the number of tasks that can be completed when energy is constrained. As the energy budget increases, its performance gradually approaches that of the proposed scheme. The NoPos scheme suffers from suboptimal channel conditions between the UAV and edge devices due to the lack of UAV position optimization, which leads to increased transmission latency and communication energy consumption. As a result, some latency-sensitive tasks fail to meet their delay requirements.

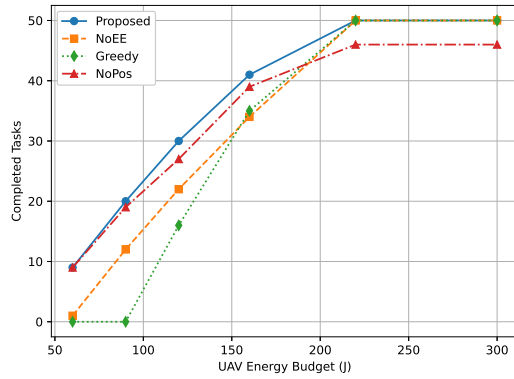


Fig. 3. Number of completed tasks versus the UAV energy budget.

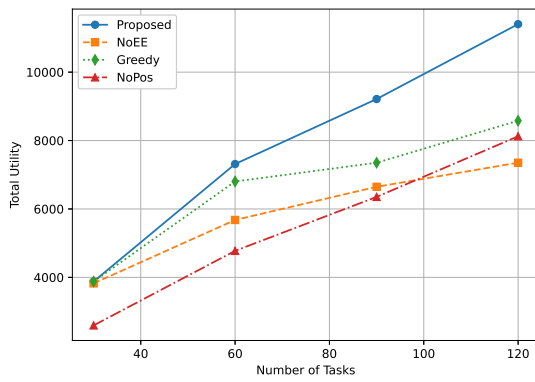


Fig. 4. Total utility under different numbers of tasks.

The Greedy scheme, which lacks a global view of resource allocation, is more likely to converge to local optima, especially under stringent energy constraints. Its performance becomes comparable to that of the proposed scheme only when the energy budget is sufficiently large.

Fig. 4 depicts the total system utility under different numbers of tasks. It is shown that the proposed scheme consistently achieves the highest system utility as the task load increases. Since the utility function jointly accounts for both task completion and energy consumption penalty, these results indicate that the proposed scheme not only completes more tasks but also provides superior energy efficiency on a per-task basis.

V. CONCLUSION

This paper establishes a UAV-assisted ISCC edge inference framework based on an early-exit mechanism. An optimization problem is formulated to minimize the total system energy consumption while maximizing the number of completed tasks through the joint optimization of device power, UAV hovering location, and exit point selection. An alternating optimization framework is developed to solve the resulting problem. Simulation results demonstrate the superiority of the proposed framework in terms of energy efficiency and task

completion performance, and show its capability to dynamically adapt system strategies according to heterogeneous task requirements. Future work will extend the proposed framework to multi-UAV collaborative scenarios.

ACKNOWLEDGMENT

This work is supported in part by National Key Research and Development Program of China under Grant 2023YFE0107900, in part by the National Natural Science Foundation of China (NSFC) under Grant U2541208 and 62501407, in part by the Key Program of Shenzhen Natural Science Foundation under Grant JCYJ20241202124219023, and in part by the Shenzhen Science and Technology Program under Grant SYSPG20241211173908022.

REFERENCES

- [1] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6g and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.
- [2] M. Lyu, H. Chen, D. Wang, G. Feng, C. Qiu, and X. Xu, "Reference signal-based waveform design for integrated sensing and communications system," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2025, pp. 6621–6626.
- [3] P. Liu, S. Tang, K. Lin, N. Ha, X. Wang, and Z. Fei, "Joint split inference design and bandwidth allocation for intelligent edge perception network," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Aug. 2025, pp. 1–5.
- [4] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [5] G. Geraci, A. Garcia-Rodriguez, M. M. Azari, A. Lozano, M. Mezzavilla, S. Chatzinotas, Y. Chen, S. Rangan, and M. D. Renzo, "What will the future of uav cellular communications be? a flight from 5g to 6g," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1304–1335, 2022.
- [6] G. Wu, Y. He, X. Cao, and C. Yuen, "Ai-enabled integrated sensing, communication, and computation survey: Techniques, status, and perspectives," *IEEE Internet Things J.*, vol. 12, no. 16, pp. 32 676–32 700, Aug. 2025.
- [7] D. Wen, Y. Zhou, X. Li, Y. Shi, K. Huang, and K. B. Letaief, "A survey on integrated sensing, communication, and computation," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 5, pp. 3058–3098, Oct. 2025.
- [8] Y. He, G. Yu, Y. Cai, and H. Luo, "Integrated sensing, computation, and communication: System framework and performance optimization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 2, pp. 1114–1128, Feb. 2024.
- [9] 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, Mar. 2024.
- [10] S. Teerapittayanon, B. McDanel, and H. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2464–2469.
- [11] Z. Liu, J. Song, C. Qiu, X. Wang, X. Chen, Q. He, and H. Sheng, "Hastening stream offloading of inference via multi-exit dnn in mobile edge computing," *IEEE Trans. Mob. Comput.*, vol. 23, no. 1, pp. 535–548, Jan. 2024.
- [12] L. Zuo, "Pre-dnnoff: On-demand dnn model offloading method for mobile edge computing," *Scalable Comput., Pract. Exp.*, vol. 25, Oct. 2024.
- [13] 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, Apr. 2024.
- [14] J. Lee, H. Lee, and W. Choi, "Wireless channel adaptive dnn split inference for resource-constrained edge devices," *IEEE Commun. Lett.*, vol. 27, no. 6, pp. 1520–1524, Jun. 2023.
- [15] Z. Wang, A. E. Kalør, Y. Zhou, P. Popovski, and K. Huang, "Ultra-low-latency edge inference for distributed sensing," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 1908–1922, 2026.