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

Zeming Zhuang, Dingzhu Wen and Yuanming Shi

School of Information Science and Technology, ShanghaiTech University

上海科技大学
ShanghaiTech University

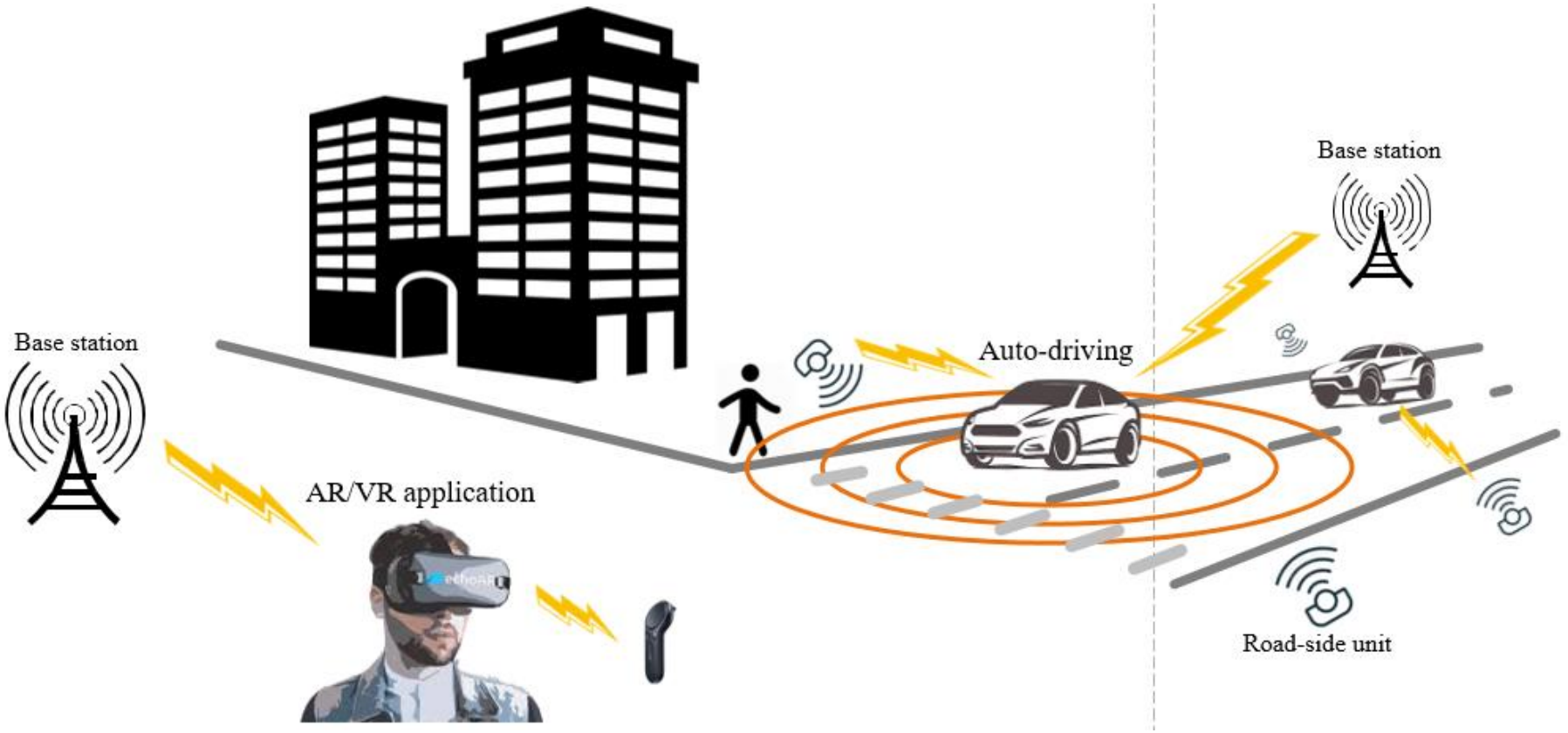
IEEE GLOBECOM®

Outline

- Background
- System Model
- Problem Formulation
- Joint Design Algorithm
- Simulation
- Conclusion

Background - Edge AI Inference

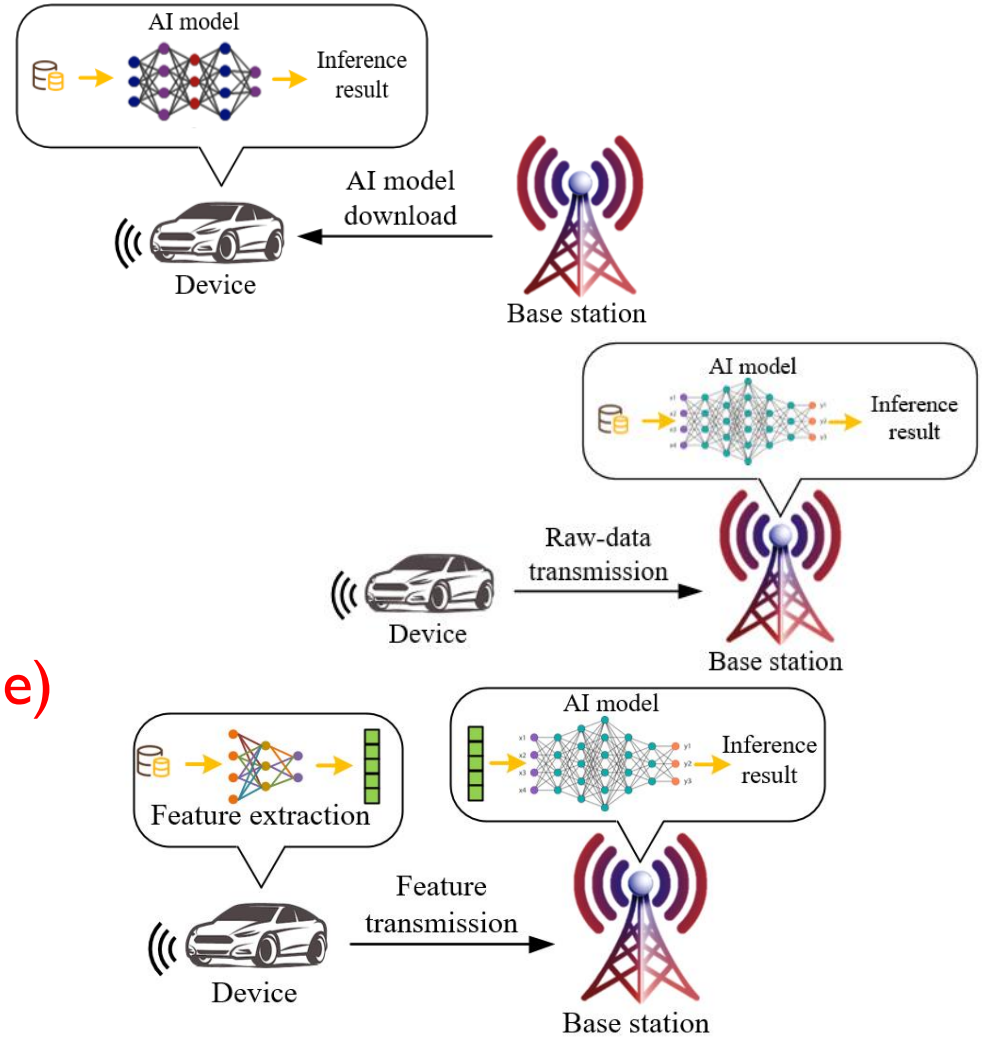
- Network Edge Intelligent Services (e.g., auto-driving and XR)
  - Fast response to the requirements
  - Intelligent decisions depends on the deployment of well-trained AI models (Edge Inference)



Background - Inference Paradigms

上海科技大学 ShanghaiTech University

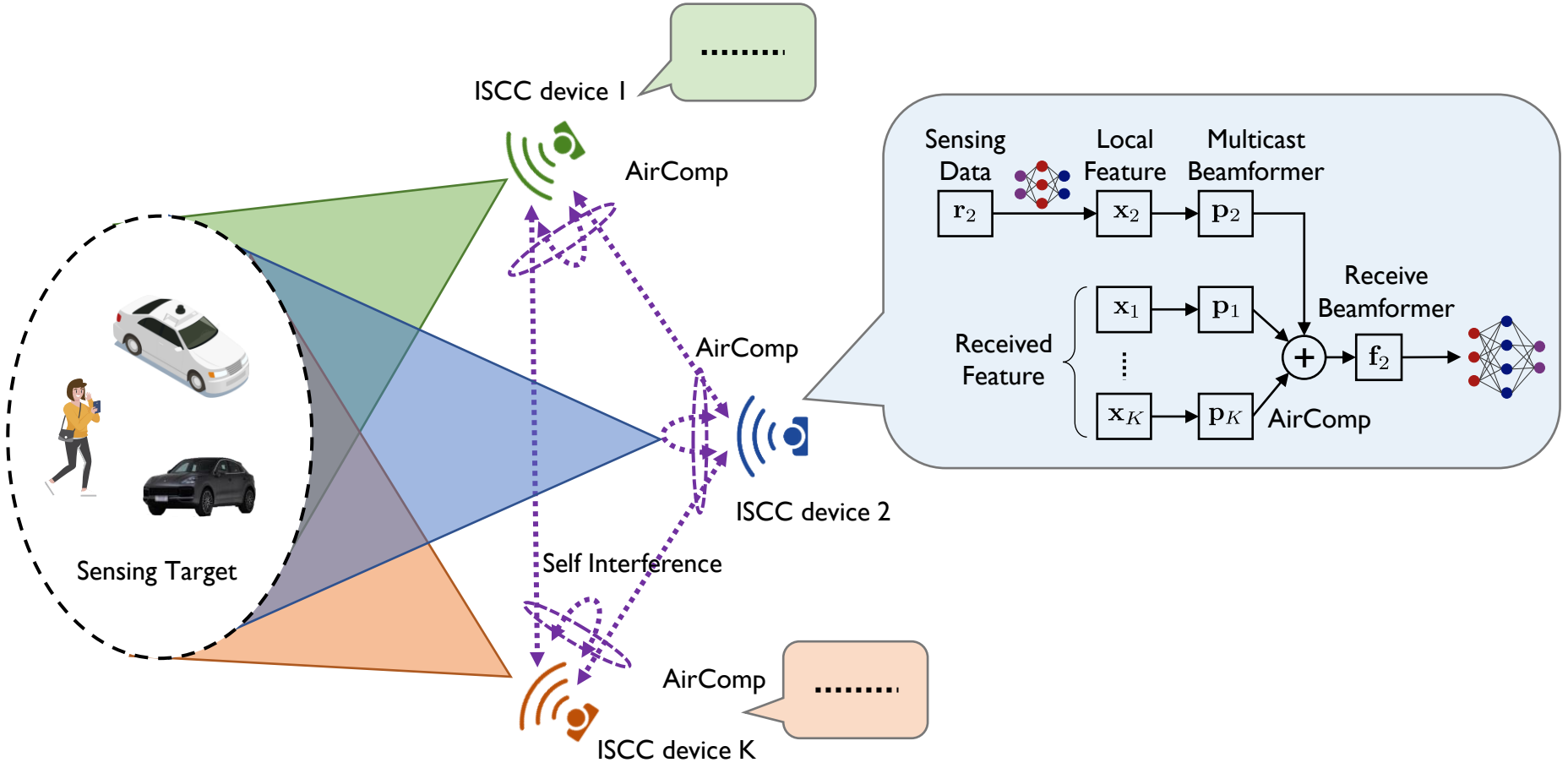
- On-device Inference
  - Execute AI model on device
  - High computation overhead
- On-server Inference
  - Communication bottleneck
  - Data privacy
- Split Inference (Device-Edge Co-inference)
  - Feature extraction on device
    - Low-dimensional feature vector
    - Preserve data privacy
  - Computation offloading



What if the server is not always available for all devices simaltaneously? Decentralized co-inference!

System Model – Network Model

- We consider a system containing K ISCC devices each equipped with a full-duplex transceiver of N_t transmit antennas and N_r receive antennas.
- Each device has its own inference task requiring features from all other devices.



Sensing Model

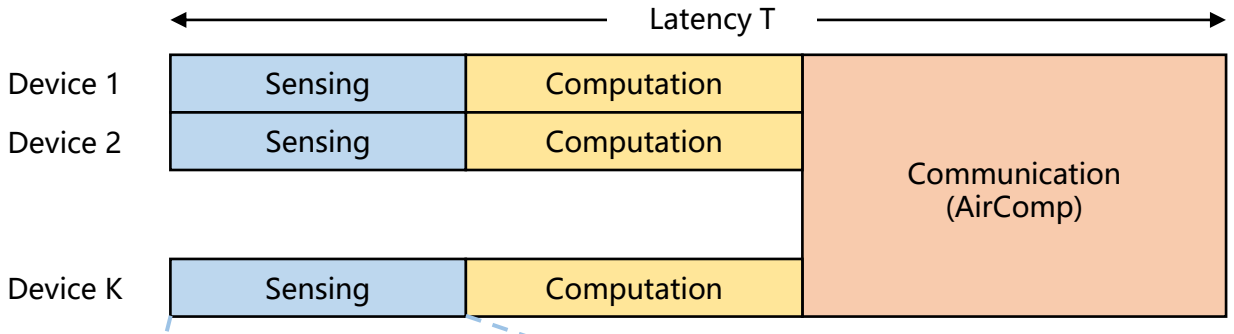
- K devices transmit orthogonal FMCW signals to the sensing target with **power $P_{s,k}$**

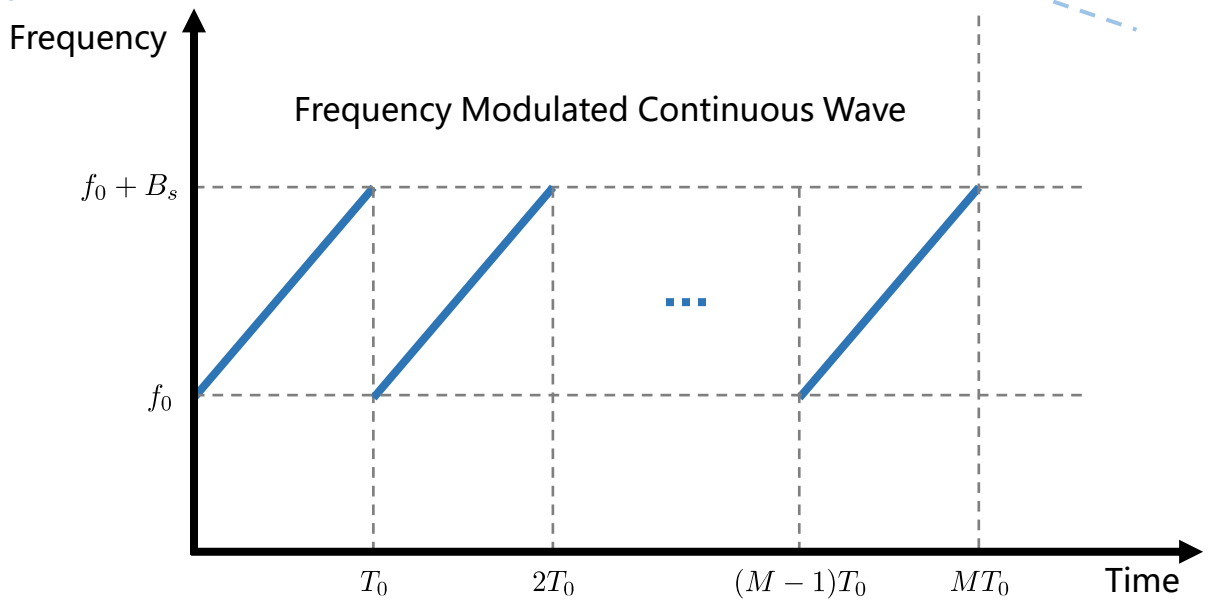
$$c_{s,k}(t) = \text{rect}\left(\frac{t}{T_0}\right) \cdot \cos\left(2\pi f_{0,k} t + 2\pi \frac{B_s}{T_0} t^2\right)$$

- Received signal at device j

$$r_k(t) = \underbrace{u_k(t)}_{\text{Direct reflect}} + \underbrace{\sum_{j=1} v_{k,j}(t)}_{\text{Clutter}} + \underbrace{n_r(t)}_{\text{Noise}}$$

- Line-of-Sight signal: $u_k(t) = H_{r,k}(t) s_k(t-\tau)$
  - $H_{r,k}(t)$ is the round-trip coefficient,
  - τ is the round-trip delay.
- Clutter of path j : $v_{k,j}(t) = C_{r,k,j}(t) s_k(t-\tau_j)$
  - Rich scattering, the overall clutter follows a zero-mean Gaussian distribution.



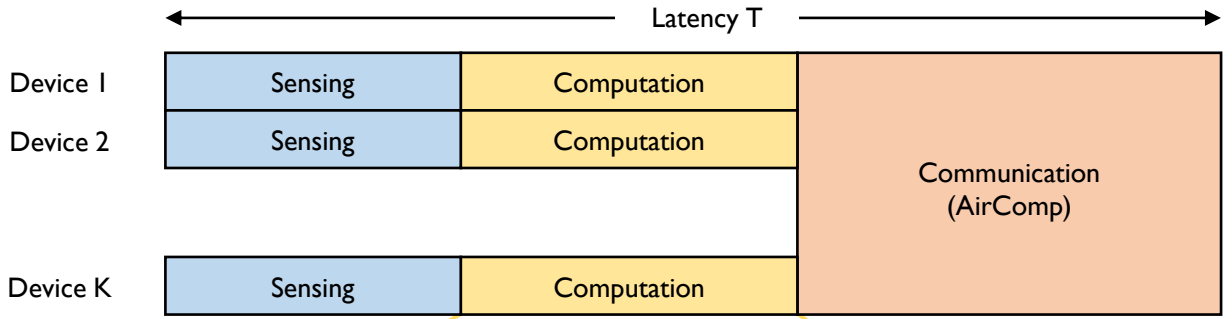


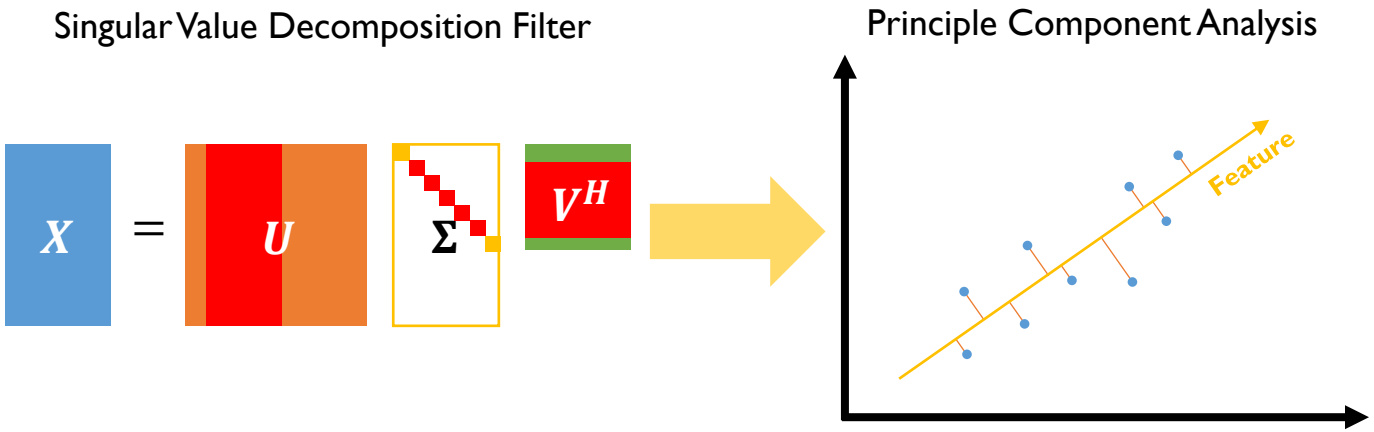
Feature Model

- Derive feature from received signal
  - Sampling the received signal
  - Data Filtering through SVD* (empirical)
  - Feature Extraction
    - Short-time Fourier transform
    - Pre-trained PCA
- Normalized wide-view feature data with noise

$$x_k(m) = \underbrace{x(m)}_{\text{Ground-truth}} + \underbrace{c_{s,k}(m) + \frac{n_r(m)}{\sqrt{P_{s,k}}}}_{\text{Clutter + Noise}}$$

- The ground-truth data is assumed to be mixed Gaussian distribution





*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 IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Sep. 2021, pp. 481–485.

Communication Model

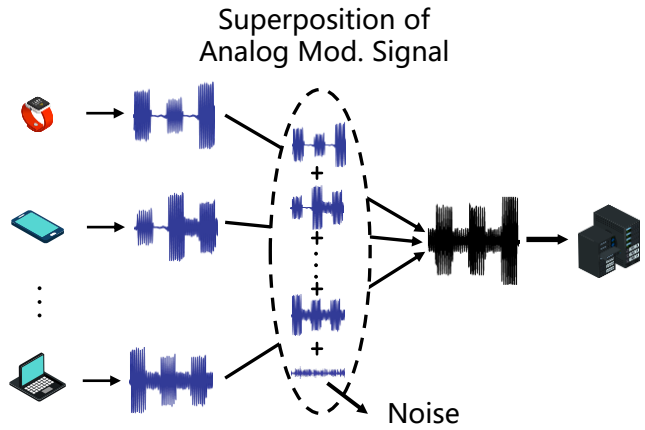
- The feature received through over-the-air computation by device $k \in \mathcal{K}$

$$\mathbf{y}_k(m) = \sum_{j \in \mathcal{K}, j \neq k} \mathbf{H}_{j,k}\mathbf{p}_{j,m}x_j(m) + \mathbf{H}_{k,k}\mathbf{p}_{k,m}x_k(m) + \mathbf{w}_k(m)$$

- $\mathbf{p}_k$ is the multicast beamformer, when $k' = k$, $\mathbf{H}_{k,k}$ is the Self-Interference channel gain
- By designing $\mathbf{p}_k$, SI channel is exploited to aggregate $\mathbf{x}_k$ with other $\mathbf{x}'_k$
- Then a receive beamformer $\mathbf{f}_{k,m}$ is applied at device k

$$\hat{x}_k(m) = \mathbf{f}_{k,m}^H \cdot \left(\sum_{j \in \mathcal{K}} \mathbf{H}_{j,k}\mathbf{p}_{j,m}x_j(m) + \mathbf{w}_k(m) \right)$$

- Then the aggregated features are used for inference task later.



Over-the-air Computation

System Model

Distribution of Extracted Feature Elements

- For the m-th element

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

Aggregated Feature Elements

$$\hat{\mu}_{\ell,k,m} = \mu_{\ell,m} \cdot \left(\sum_{j\in\mathcal{K}} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}\mathbf{p}_{j,m}\right)$$

$$\hat{\sigma}_{k,m}^2 = \sigma_{k,m}^2 \cdot \left(\sum_{j\in\mathcal{K}} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k}\mathbf{p}_{j,m}\right)^2 + \sum_{j\in\mathcal{K}} \left(\mathbf{f}_{k,m}^H \mathbf{H}_{j,k}\mathbf{p}_{j,m}\right)^2 \left(\sigma_{s,j}^2 + \frac{\sigma_r^2}{P_{s,j}}\right) + N_0 \mathbf{f}_{k,m}^H \mathbf{f}_{k,m}$$

Designing Metric

- MMSE ignores the heterogeneous distortion sensitivities of different elements
- Inference accuracy is hard to calculate instantaneously

Task-oriented Metric

- Discriminant gain is defined based on the KL divergence between two Gaussian dist. classes

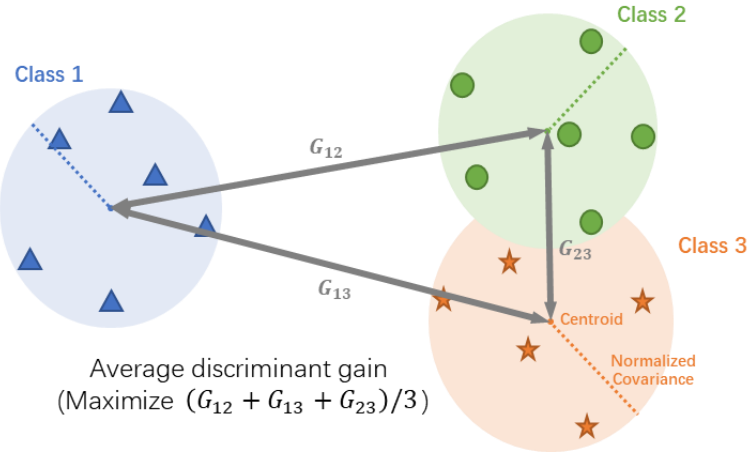
$$G_{\ell,\ell'}(\hat{x}_k(m)) = D_{KL}\left[f_\ell(\hat{x}_k(m)) \| f_{\ell'}(\hat{x}_k(m))\right] + D_{KL}\left[f_{\ell'}(\hat{x}_k(m)) \| f_\ell(\hat{x}_k(m))\right]$$

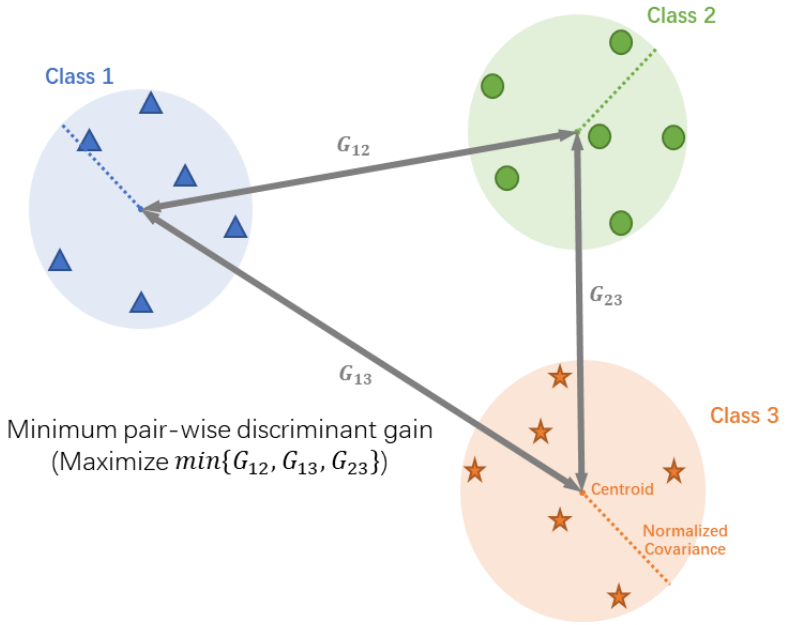
- Normally discriminant gain of every pair of classes are summed up

$$G(\mathbf{x}) = \sum_{i=1}^{L} \sum_{j<i} G_{ij}(\mathbf{x})$$

- Shortage of avg. DG: **unbalance**
  - Class 2 & 3 are too close and far away from class 1
- Solution: **Minimum Discriminant Gain**

$$G_{\min}(\hat{\mathbf{x}}_k) = \min_{1 \le \ell \ne \ell' \le L} \sum_{m=1}^{M} G_{\ell,\ell'}(\hat{x}_k(m))$$





Problem Formulation

上海科技大学 ShanghaiTech University

- Objective function: Maximize the weighted sum of minimum DG for all tasks

$$\text{maximize} \quad \sum_{k=1}^{K} \eta_k \left\{ \min_{1 \le \ell \ne \ell' \le L} \sum_{m=1}^{M} \frac{\left(\hat{\mu}_{\ell,k,m} - \hat{\mu}_{\ell',k,m}\right)^2}{\hat{\sigma}^2_{k,m}} \right\}$$

- Decision variable: Sensing power $P_{s,k}$

  Multicast beamformer $\mathbf{p}_{k,m}$

  Receive beamformer $\mathbf{f}_{k,m}$

- Long-term Power constraint: $P_{s,k}T_s + E_p + T_c \sum_{m=1}^{M} \|\mathbf{p}_{k,m}\|^2 X_k(m) \le E_k$

- Unit receive beamforming constraint: $\|\mathbf{f}_{k,m}\|^2 = 1$

Joint design of Sensing Power and Beamforming

上海科技大学 ShanghaiTech University

- Introduce α_k as a slack variable and extend the feasible region of constraints

$$\max_{\substack{\{P_{s,k}\},\{\mathbf{f}_{k,m}\},\{\mathbf{p}_{k',m}\},\\ \{\alpha_k\},\{\beta_{\ell,\ell',k,m}\}}} \quad \sum_{k=1}^{K} \eta_k \alpha_k, \tag{1}$$

$$\text{s.t.} \quad P_{s,k}T_{s,k} + E_{p,k} + T_c \sum_{m=1}^{M} \|\mathbf{p}_{k,m}\|^2 \le E_k, \quad \|\mathbf{f}_{k,m}\|^2 \le 1, \tag{2}$$

$$\alpha_k - \sum_{m=1}^{M} \beta_{\ell,\ell',k,m} \le 0, \tag{3}$$

$$\frac{(\mu_{\ell,k,m} - \mu_{\ell',k,m})^2}{\beta_{\ell,\ell',k,m}} \left(\sum_{k'=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{k'k,m} \mathbf{p}_{k',m} \right)^2 \ge Z_{\ell,\ell',k,m}\left(\{P_{s,k'}\}, \{\mathbf{p}_{k',m}\}, \mathbf{f}_{k,m}\right) \tag{4}$$

- Approximate the last constraint with Taylor expansion

$$\mathbf{P4} \quad \max_{\substack{P_{s,k},\mathbf{f}_{k,m},\mathbf{p}_{j,m},\\ \alpha_k,\beta_{\ell,\ell',k,m}}} \quad \sum_{k=1}^{K} \eta_k \alpha_k,$$

$$\text{s.t.} \quad (2),(3),$$

$$Z_{\ell,\ell',k,m}\left(\{P_{s,j}\}, \{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}\right) - Q^{[t]}_{\ell,\ell',k,m}\left(\{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}, \beta_{\ell,\ell',k,m}\right) \le 0$$

- Then alternatively
  - **P5**: Fix $\mathbf{p}_{j,m}$, $P_{s,k}$, update $\mathbf{f}_{k,m}$
  - **P6**: Fix $\mathbf{f}_{k,m}$, update $\mathbf{p}_{j,m}$, $P_{s,k}$

Joint design of Sensing Power and Beamforming

上海科技大学 ShanghaiTech University

Algorithm 1 Joint Sensing Power Allocation, Multicast and Receive Beamforming Design

Input: Channel gain $\{\mathbf{H}_{j,k}\}$, Device energy E_k

Output: $\{P_{s,k}^*\}$, $\{\mathbf{p}_{j,m}^*\}$, $\{\mathbf{f}_{k,m}^*\}$

Initialize $t=0$, $\{P_{s,k}^{[0]}\}$, $\{\mathbf{p}_{j,m}^{[0]}\}$, $\{\mathbf{f}_{k,m}^{[0]}\}$, $\alpha_k^{[0]}$, $\beta_{\ell,\ell',k,m}^{[0]}$ in feasible region of **P3**;

Initialize function $Q_{\ell,\ell',k,m}^{[0]}\left(\{\mathbf{p}_{j,m}^{[0]}\}, \mathbf{f}_{k,m}^{[0]}, \beta_{\ell,\ell',k,m}^{[0]}\right)$;

repeat

  Derive the approximated problem **P4**;

  repeat

    Update $\mathbf{f}_{k,m}$ by solving **P5**;

    Update $\mathbf{p}_{j,m}$, $P_{s,k}$ by solving **P6**;

  until convergence

  Update function $Q_{\ell,\ell',k,m}^{[t+1]}\left(\{\mathbf{p}_{j,m}^{[t+1]}\}, \mathbf{f}_{k,m}^{[t+1]}, \beta_{\ell,\ell',k,m}^{[t+1]}\right)$;

  $t \leftarrow t+1$;

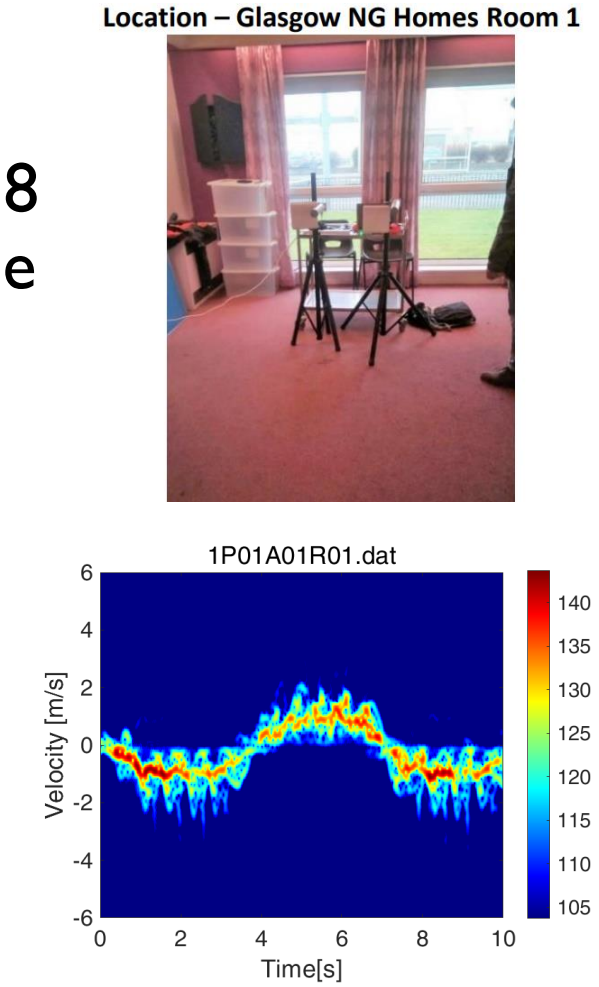
until convergence

Optimal solution $P_{s,k}^* \leftarrow P_{s,k}^{[t]}$, $c_{k,m}^* \leftarrow c_{k,m}^{[t]}$, $\mathbf{f}_m^* \leftarrow \mathbf{f}_m^{[t]}$;

Simulation

上海科技大学 ShanghaiTech University

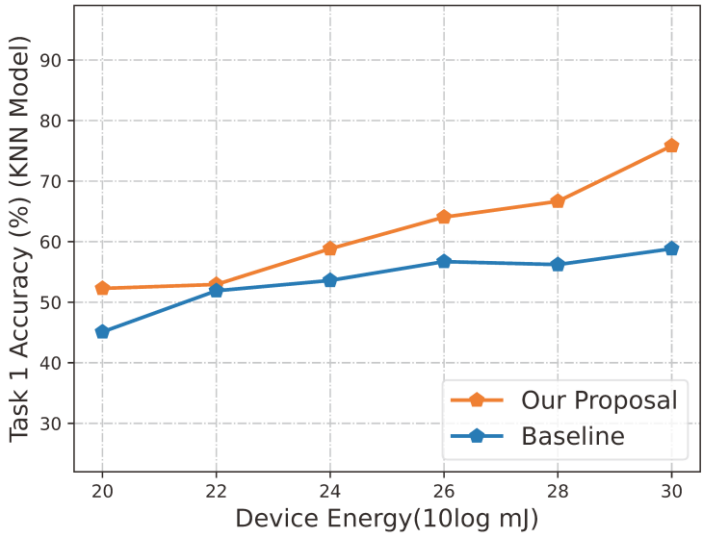
- Multi-task Radar Sensing of human activities*
- 3 edge devices each equipped with 8 receive antennas and 8 transmit antennas, retrieving wide-view sensing data of same target
- Tasks
  - Activity recognition and Person identification (Gender and Age)
  - 4 activities: sit down, stand up, pick up an object, drink water
  - 1500 samples of 54 volunteers
- Schemes
  - Joint designing ISCC scheme (proposed)
  - Fixed beamforming vectors and sensing powers

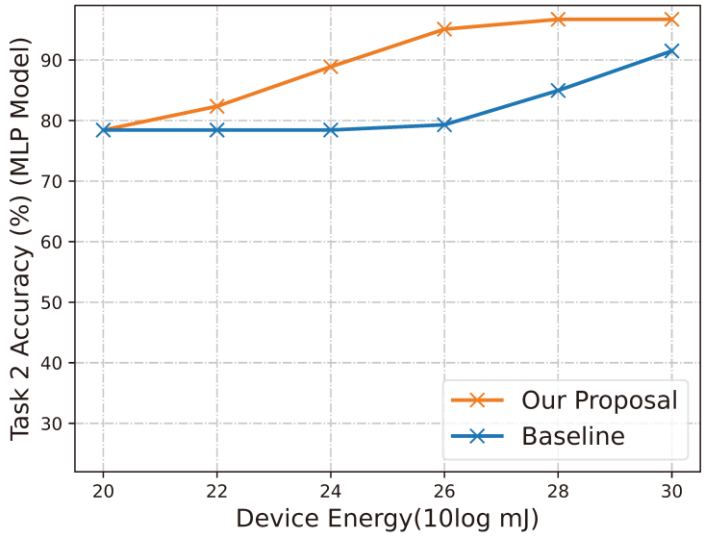


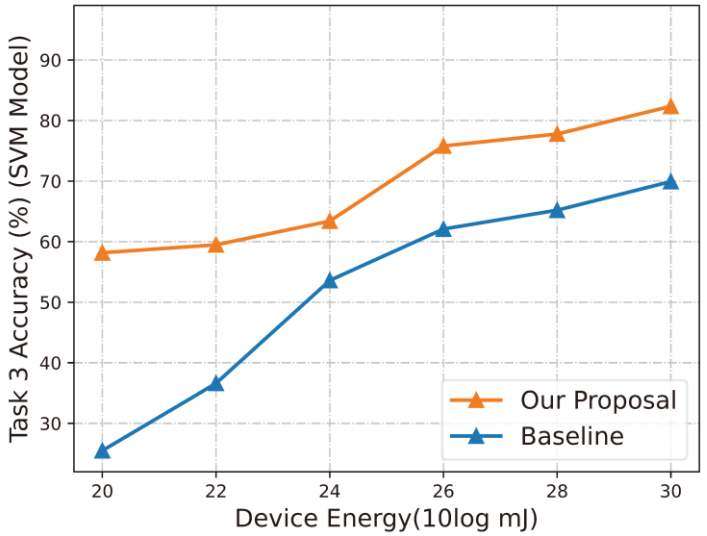
*Radar signatures of human activities, available at http://researchdata.gla.ac.uk/848/, doi: 10.5525/gla.researchdata.848

Simulation – Accuracy vs. Device Energy

- It shows that a higher upper limit of energy leads to better inference accuracy, because a higher device energy threshold indicates that the devices can allocate more power to suppress the distortion and noise.
- Moreover, our proposal outperforms the baseline since the sensing power and beamforming vectors are not adjusted to different feature elements in the baseline scheme.

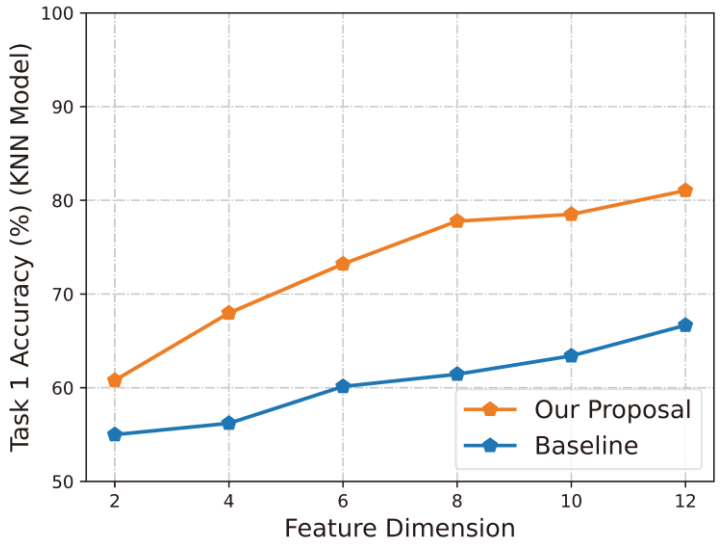


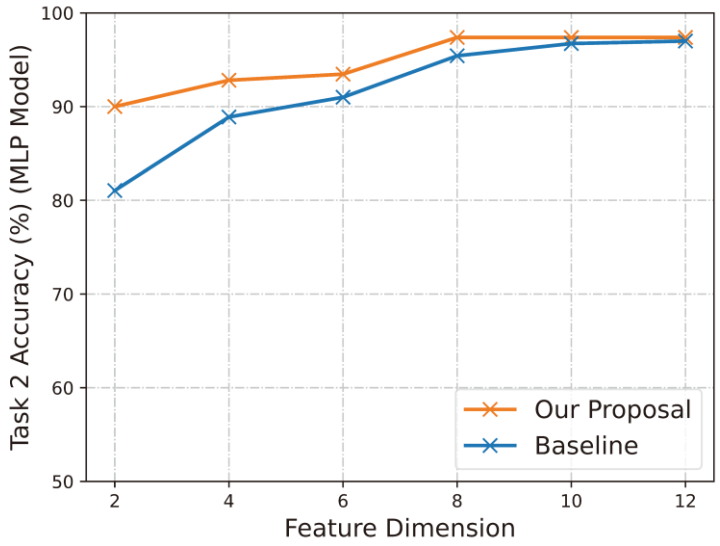


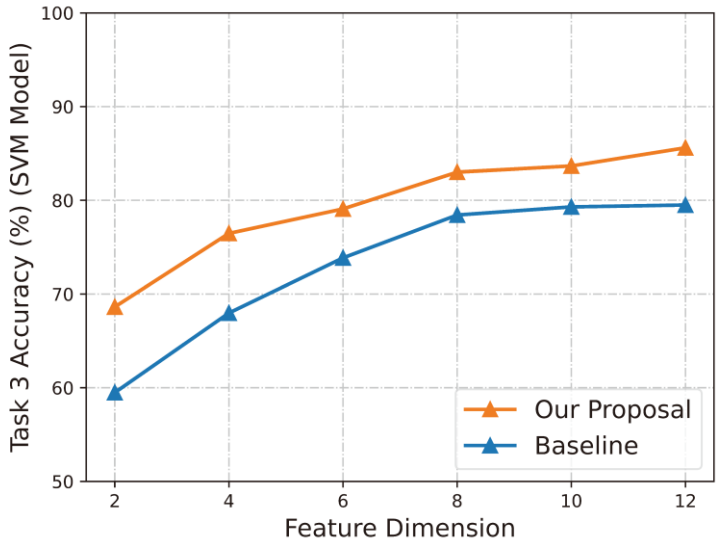


Simulation – Accuracy vs. Feature Dimension

- Accuracies of all three tasks increase as the feature dimension increases. This is because different feature dimensions are orthogonal and independent from the algorithm of PCA.
- It follows that more feature dimensions can keep more data about the sensing target, which improves the performance of inference tasks.







Conclusions

- We proposed a decentralized ISCC framework with the exploitation of the self-interference channel in full-duplex communication.
- We jointly designed the sensing power, multicast and receive beamformer under the criterion of maximizing the weighted sum of minimum discriminant gain in different tasks.
- Simulation results demonstrate that the proposed scheme can achieve higher inference accuracy than the benchmarks under various limitations of energy and feature dimensions.

Thanks