Spectrum Shortage for Radio Sensing? Leveraging Ambient 5G Signals for Human Activity Detection

Kunzhe Song, Maxime Zingraff, and Huacheng Zeng
Department of Computer Science and Engineering, Michigan State University, USA

***Abstract*—Radio sensing in the sub-10 GHz spectrum offers unique advantages over traditional vision-based systems, including the ability to see through occlusions and preserve user privacy. However, the limited availability of spectrum in this range presents significant challenges for deploying large-scale radio sensing applications. In this paper, we introduce Ambient Radio Sensing (`ARS`), a novel Integrated Sensing and Communications (ISAC) approach that addresses spectrum scarcity by repurposing over-the-air radio signals from existing wireless systems (e.g., 5G and Wi-Fi) for sensing applications, without interfering with their primary communication functions. `ARS` operates as a standalone device that passively receives communication signals, amplifies them to illuminate surrounding objects, and captures the reflected signals using a self-mixing RF architecture to extract baseband features. This hardware innovation enables robust Doppler and angular feature extraction from ambient OFDM signals. To support downstream applications, we propose a cross-modal learning framework focusing on human activity recognition, featuring a streamlined training process that leverages an off-the-shelf vision model to supervise radio model training. We have developed a prototype of `ARS` and validated its effectiveness through extensive experiments using ambient 5G signals, demonstrating accurate human skeleton estimation and body mask segmentation applications.**

***Index Terms*—5G, spectrum sharing, RF sensing, integrated sensing and communications, human activity detection**

I. Introduction

Radio sensing on sub-10 GHz spectrum bands offers several compelling advantages over camera and LiDAR sensors, particularly in challenging environments. Unlike cameras, radio waves are not affected by lighting conditions, allowing reliable operation in darkness, fog, or glare. Compared to LiDAR, radio sensing is more robust in rain, dust, and smoke, and can penetrate obstacles like walls or foliage, enabling non-line-of-sight detection. Additionally, radio sensing systems are privacy-preserving for human monitoring applications such as nursing homes, elderly care, or hospitals, as they can detect presence, movement, and vital signs without capturing identifiable visual information like faces or body details. This makes them ideal for continuous, non-intrusive monitoring while respecting individuals' privacy.

While radio sensing technologies such as FMCW radar have become very mature and cost-effective, their real-world applications remain limited. The primary obstacle lies in the scarcity of spectrum resources. The sub-10 GHz spectrum has been intensively allocated or reserved for many important applications such as TV broadcasting, cellular networks (e.g., 4G/5G), satellite communications, Wi-Fi, military and aviation radar, and public safety systems. As a result, introducing new sensing services in this spectrum range faces strict regulatory constraints and interference challenges, limiting the deployment of radio sensing technologies at scale.

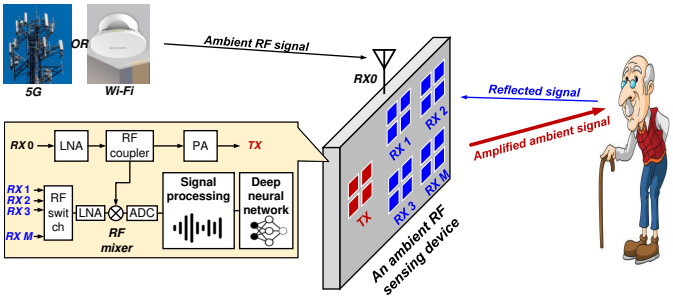

Fig. 1: Architectural diagram of `ARS`.

One approach to addressing the spectrum scarcity is to use high-frequency bands, such as the millimeter-wave (mmWave) spectrum, for sensing purposes. This approach has been widely adopted, leading to the development of a variety of mmWave FMCW radars for emerging applications such as autonomous driving, and health monitoring. The abundant bandwidth and short wavelengths of mmWave frequencies enable high-resolution sensing, making them well-suited for fine-grained environmental perception. Despite their success in various applications, mmWave radars have very limited obstacle-penetration capabilities and a short detection range.

In this paper, we propose a new approach for sub-10GHz radio sensing to address the spectrum scarcity issue. Our approach is based on the fact that radio waves from wireless communication systems such as 5G and Wi-Fi are ubiquitous, both indoors and outdoors. The key idea is to leverage the over-the-air radio waves from existing communication systems for radio sensing applications while not negatively affecting communication system performance. To validate this approach, we introduce an Ambient Radio Sensing (`ARS`) device, as shown in Fig. 1. `ARS` is a standalone radio sensing device. On the transmitter side, it receives the over-the-air radio signals from an existing communication system, such as a 5G base station, and amplifies them to illuminate surrounding objects for sensing purpose. On the receiver side, it captures the reflected signals from the environment and down-converts them to the baseband for feature extraction and learning-based object detection. Crucially, since `ARS` adopts an amplify-and-forward approach for RF signal processing in the analog domain, it does not generate interference to

existing communication systems. Instead, it potentially boosts the signal strength for the communication system.

On the hardware side, we propose a self-mixing RF architecture to generate baseband signals optimized for feature extraction. Specifically, the dipole antenna (labeled "RX0" in Fig. 1) receives over-the-air radio signals from a specific communication system and amplifies them to illuminate nearby objects using a patch antenna (labeled "TX" in Fig. 1), which is highly optimized for the specific communication spectrum band and serving as a band-pass filter to suppress the signals on unintended spectrum bands. In parallel, multiple patch antennas (labeled "RX1" to "RX M" in Fig. 1) are employed to receive the reflected signals from surrounding objects for sensing purposes. These reflected signals are first amplified and then mixed with a copy of the amplified signal from the dipole antenna, generating a baseband signal for digitization and feature extraction. This innovative self-mixing RF architecture enables **ARS** to extract coherent temporal and spatial features crucial for motion analysis.

On the algorithmic side, the OFDM waveform and limited bandwidth of existing signals present significant challenges for feature extraction. A key question is whether it is possible to extract Doppler signatures of moving objects from the baseband signal generated by our sensing device. Our answer is affirmative. Through analytical studies of 5G NR OFDM frames, we find that, through proper signal processing, the phase of the generated baseband signal exhibits an approximately linear relationship with object displacement. This deterministic relationship enables continuous estimation of Doppler signatures. Additionally, by leveraging multiple receiving antennas, **ARS** can estimate the angular direction of these signatures, providing a rich spatio-temporal representation for downstream tasks.

To demonstrate the potential of **ARS**, we implement two demanding downstream tasks: human skeleton estimation and body mask segmentation. Since radio features are inherently sparse and sensitive to environmental noise, capturing fine-grained human poses is challenging. We address this by adopting a cross-modal supervised learning approach, which distills high-fidelity knowledge from the vision domain into the radio domain. During training, an off-the-shelf vision model provides ground-truth labels from synchronized video to supervise the radio-based DNN. We have validated this framework through a functional prototype and extensive experiments, achieving state-of-the-art performance on both tasks.

The primary contributions of this work are as follows:

- It presents a novel approach to Integrated Sensing and Communications (ISAC) that addresses spectrum scarcity by repurposing ubiquitous ambient signals.
- It introduces a joint hardware-algorithmic design featuring a self-mixing RF architecture for robust feature extraction from ambient OFDM waveforms.
- It validates the practicality of ARS through extensive experiments, demonstrating superior performance in real-world human skeleton and mask estimation scenarios.

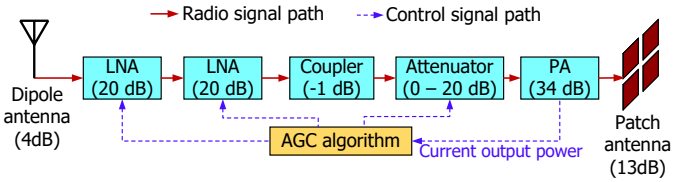

Fig. 2: Diagram of amplify-and-forward RF circuit.

II. Radar-Inspired RF Hardware

A. Hardware Design

As shown in Fig. 1, the proposed sensing device comprises two parts: (i) amplify-and-forward circuit, and (ii) self-mixing circuit.

Amplify-and-Forward Circuit: Fig. 2 shows our proposed circuit. It uses an omnidirectional dipole antenna for signal reception, multi-stage LNA/PA with automatic gain control (AGC), and a directional patch antenna for efficient signal transmission. AGC, enabled by the PA's output power pin (e.g., "DET"), prevents nonlinearity from excessive gain. Designed to forward signals at 10 dBm for sensing, the circuit operates effectively for ambient RF power from –70 to –7 dBm.

Self-Mixing Circuit: As shown in Fig. 1, M patch antennas are used to receive the signal reflected back from the nearby objects. These patch antennas go through an RF switch to share a single RF chain for circuit simplification. The reflected signal is then mixed with a copy of amplified ambient RF signal for down conversion, generating the baseband signal for feature extraction. The RF switch is controlled by an FPGA device to meet the timing requirements, generating M independent channels for object direction estimation.

The environment presents radio waves on both intended and unintended spectrum bands from diverse communication systems. Interference mitigation is challenging. We propose a joint hardware and algorithmic design for interference mitigation. On the hardware side, we optimize the patch antenna design to minimize its out-of-band gain. Patch antennas have a small bandwidth by structure. This is widely considered its drawback in many other applications, but appears to be advantageous for this application. We optimize the antenna design to align its center frequency with the spectrum communication frequency channel, minimizing its bandwidth to suppress the out-of-band interference.

B. Baseband Signal Analysis

Mathematical Modeling: Both Wi-Fi and 4/5G systems use OFDM modulation for signal transmission. Consider one OFDM symbol from the communication system. Denote $s(t)$ as the radio signal received by the dipole antenna (see Fig. 1) that corresponds to one OFDM symbol. Then, we have

$$s(t) = \sum_{k \in \mathcal{K}} X_k e^{j2\pi(f_c + k f_\Delta)t}, \tag{1}$$

where $X_k \in \mathbb{C}$ is the QAM symbol, k is OFDM subcarrier index and $\mathcal{K}$ is the set of valid subcarriers, f_c is the carrier/center frequency of the radio signal, f_Δ is the subcarrier spacing.

The received signal $s(t)$ from the dipole antenna is amplified and forwarded to the "TX" antenna for transmission. The transmitted signal can be written as $\beta s(t)$, where $\beta \in \mathbb{C}$ represents the amplification and phase shift introduced by the amplify-and-forward circuit.

The objects in the proximity are illuminated by two signal sources: **ARS**'s TX antenna and the ambient RF signal (from the original communication system). Since the amplified signal is much stronger than the ambient signal itself, we ignore the ambient RF signal for illumination. Therefore, by denoting $r(t)$ as the received signal from the "RX" antenna, we have

$$r(t) = \beta \sum_{l \in \mathcal{L}} \alpha_l \cdot s(t - \tau_l), \tag{2}$$

where $\mathcal{L}$ is the set of multi-path reflectors, α_l and τ_l are the attenuation coefficient and time delay of path l.

The received signal is mixed with the amplified ambient RF signal, generating baseband signal for feature extraction. The baseband signal, i.e., the output of the RF mixer, can be written as:

$$y(t) = r(t) \cdot s(t)^* = |\beta|^2 \sum_{l \in \mathcal{L}} \alpha_l \cdot s(t - \tau_l) \cdot s(t)^*, \tag{3}$$

where $(\cdot)^*$ denotes conjugate operator.

Plugging Eqn (1) into the above equation, we can obtain $y(t)$ as follows:

$$y(t) = |\beta|^2 \sum_{l \in \mathcal{L}} \alpha_l \cdot \underbrace{\left[\sum_{k \in \mathcal{K}} |X_k|^2 e^{-j2\pi(f_c + k f_\Delta)\tau_l} \right]}_{\text{Part-A}} + |\beta|^2 \sum_{l \in \mathcal{L}} \alpha_l \underbrace{\left[\sum_{k \in \mathcal{K}} X_k e^{-j2\pi(f_c + k f_\Delta)\tau_l} \right]}_{\text{Part-B}} \underbrace{\left[\sum_{k' \in \mathcal{K}}^{k \neq k'} X_{k'}^* e^{j2\pi(k - k')f_\Delta t} \right]^*}_{\text{Part-C}}. \tag{4}$$

Denote $\mathcal{F}(\cdot)$ as the frequency of a signal. Then, we have $\mathcal{F}(\text{Part-A}) = 0$, $\mathcal{F}(\text{Part-B}) = 0$, and $\mathcal{F}(\text{Part-C}) \geq f_\Delta$.

We apply a low-pass filter to $y(t)$ and denote the output signal as $z(t)$, which can be written as:

$$z(t) = \lfloor y(t) \rfloor_{f_\Delta/2} = |\beta|^2 \sum_{l \in \mathcal{L}} \alpha_l \cdot \left[\sum_{k \in \mathcal{K}} |X_k|^2 e^{-j2\pi(f_c + k f_\Delta)\tau_l} \right], \tag{5}$$

where $\lfloor \cdot \rfloor_B$ denotes a low pass filter with cutoff frequency B.

Denote $z(t) = \sum_{l \in \mathcal{L}} z_l(t)$, where $z_l(t)$ is contributed by the l-th object. Then, we have

$$z_l(t) = |\beta|^2 \alpha_l \cdot \sum_{k \in \mathcal{K}} |X_k|^2 e^{-j2\pi(f_c + k f_\Delta)\tau_l}. \tag{6}$$

Denote d_l as the distance between **ARS** and the l-th object. We have $\tau_l = d_l / c$, where c is the light speed. Consequently, we have

$$z_l(t) = |\beta|^2 \alpha_l \cdot \sum_{k \in \mathcal{K}} |X_k|^2 e^{-j2\pi(f_c + k f_\Delta)\frac{d_l}{c}}. \tag{7}$$

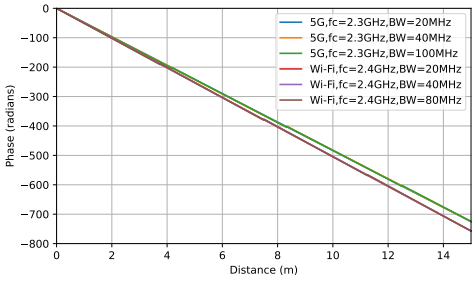

Fig. 3: Relationship between an object-moving distance and the signal phase rotation.

Eqn (7) characterizes the relationship between the measured baseband signal and the object distance. Particularly, we are interested in the relationship between $\angle z_l$ and the object distance d_l. In communication systems, OFDM subcarrier spacing is much smaller than the carrier frequency, i.e., $f_\Delta \ll f_c$. Therefore, we approximate Eqn (7) by:

$$z_l(t) \approx \left[|\beta|^2 \alpha_l \sum_{k \in \mathcal{K}} |X_k|^2 \right] e^{-j2\pi f_c \frac{d_l}{c}}. \tag{8}$$

Eqn (8) indicates the linear relationship between d_l (object distance) and $\angle z_l(t)$. Based on Eqn (8), the instantaneous velocity of the l-th object, denoted as $v_l(t)$, can be estimated by:

$$v_l(t) = \frac{unwrap(\angle z_l(t)) - unwrap(\angle z_l(t - T_\Delta))}{T_\Delta}, \tag{9}$$

where $unwrap(\cdot)$ is a function used to fix phase discontinuities in angle data. T_Δ is a small time step.

Eqn (9) shows the linear relationship between an object's moving distance and the phase change, and thus lays the foundation for the design of our Doppler signature extraction.

Numerical Results: The derivation from Eqn (7) to Eqn (8) is based on the assumption that f_Δ is negligible compared to f_c. In 5G and Wi-Fi communication systems, despite the fact that $f_\Delta \ll f_c$, f_Δ is not negligible. In Wi-Fi systems, $f_\Delta = 312.5$ kHz while $f_c = 2.4$ GHz or $f_c = 5$ GHz. In 4G/5G systems, $f_\Delta = 15$ kHz or $f_\Delta = 30$ kHz while its f_c is typically larger than 1.8 GHz. We thus conduct numerical studies to evaluate the relation between $\angle z_l(t)$ and d_l in Eqn (8). Fig. 3 presents our numerical results. It confirms the approximately linear relation between the angle of $z_l(t)$ and object moving distance d_l.

III. Data Processing

The above analysis reveals the Doppler signatures of moving objects in the observed signal. However, extracting meaningful features for human activity detection presents several challenges that were not accounted for in the prior analysis: (i) the signal received by **ARS**'s RX0 may undergo multipath propagation, (ii) the ambient 5G signal is inherently noisy, (iii) the use of adaptive modulation and coding schemes (MCS) in 5G introduces variability, and (iv) interference from adjacent systems such as Wi-Fi can distort the signal. In this section,

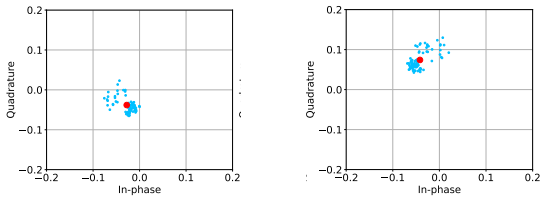

Fig. 4: Signal constellation before (blue dot scatters) and after (a single red point) sanitization.

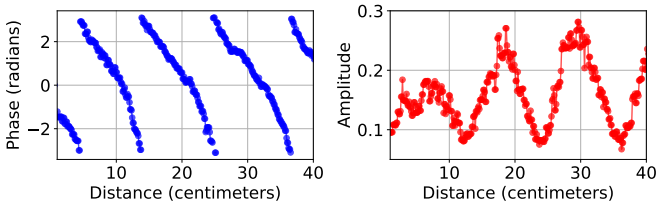

Fig. 5: Observed signal phase and amplitude when a person moves toward the detector at a roughly constant speed.

we first introduce a heuristic algorithm for signal sanitization to mitigate these challenges, followed by a differential beamforming scheme for generating heatmap images.

A. Signal Sanitization

To ensure reliable human activity detection from ambient 5G signals, we design a multi-stage signal processing pipeline that enhances signal quality by addressing noise, bias, and outliers. Our approach is based on the fact that the signal sampling rate is 2 MSPS, which is much higher than what is required for human activity detection. This high sampling rate introduces redundancy that can be exploited for effective signal sanitization.

Noise Suppression and Bias Correction. The blue dot scatters in Fig. 4 shows the constellations of the raw baseband signal samples from the RF hardware (i.e., the output of self-mixer). Depending upon the ambient 5G signal strength and other factors, the raw signal samples may appear to widespread or well-concentratrated. To mitigate these issues, we first apply a Butterworth low-pass filter, which attenuates high-frequency noise while preserving relevant signal features. Next, we correct sample bias by identifying stationary signal components clustered near a common offset. We use the k-means clustering algorithm to group constellation points into three clusters and identify the one closest to the origin. This cluster is presumed to represent static signal components. All constellation points are then translated to re-center this cluster at the origin, thereby normalizing the spatial bias.

Outlier Detection and Removal. Despite initial noise reduction, residual outliers may still present, primarily due to interference from radio systems on adjacent spectrum bands. These outliers can distort downstream learning models and obscure motion-related patterns. We implement a two-step strategy for outlier removal: (i) all points within a defined radius around the origin are discarded, and (ii) we employ the Local Outlier Factor (LOF) algorithm to eliminate spatially sparse and anomalous points. LOF compares the local density of a point to its neighbors using k-nearest neighbor metrics, and flags points with significantly lower density as outliers. This technique effectively preserves regions of meaningful, homogeneous signal activity while suppressing noise-induced anomalies.

Sample Projection. After denoising and outlier removal, the data often still contains some scattered points. For efficient heatmap generation and model input simplification, we reduce the constellation diagram to a single representative projection. We apply k-means clustering to all remaining points and retain only the centroid of the dominant cluster. This yields a compact, informative representation of the signal dynamics associated with each time frame. The single red point in Fig. 4 represents the output of the sanitization.

Experimental Validation. To validate the relationship between movement distance and phase rotation, we conducted experiments using our testbed (see Fig. 9 in §V-A). We measured the signal amplitude and phase as a person moved toward the antennas at a constant speed. The carrier frequency of the 5G signal was 2.35 GHz. Fig. 5 shows the measurement results. Notably, the phase exhibits a clear linear relationship with distance, confirming coherent phase detection and the presence of Doppler features. As expected, the signal amplitude also change with the person's movement.

B. Differential Beamforming

Here, beamforming refers to the use of a two-dimensional antenna array to estimate the direction of incoming signals, namely, the azimuth angle θ and elevation angle ϕ as illustrated in Fig. 6. This technique, along with its variants, is widely used in radar imaging to infer the spatial structure of a scene. In our work, we focus on detecting motion rather than reconstructing static scenes, and therefore refer to our approach as differential beamforming.

Consider the antenna array as shown in Fig. 6. One antenna is for transmitting and the other antennas for receiving. Denote the coordinate of the TX antenna as the origin, i.e., $\vec{p}_{TX} = (0,0,0)$. Denote $\mathcal{R}$ as the set of RX antenna array, and $\vec{p}_i$ as the coordinate of RX antenna $i \in \mathcal{R}$.

Consider a small object or a small part of a big object characterized by azimuth angle θ and elevation angle ϕ as shown in Fig. 6. To estimate the signal strength from this direction, we define a unit direction vector by letting:

$$\vec{u}(\theta,\phi) = \big(\cos(\theta)\cos(\phi),\quad \cos(\theta)\sin(\phi),\quad \sin(\phi)\big)\,. \tag{10}$$

Based on this unit vector, the expected phase offset of the signal received by RX antenna $i \in \mathcal{R}$ can be written as:

$$w_i(\theta,\phi) = \exp\left(-j\frac{2\pi}{\lambda}\langle\vec{u}(\theta,\phi),\ \vec{p}_i\rangle\right), \tag{11}$$

where λ is the wavelength of 5G/Wi-Fi carrier frequency, and $\langle\cdot,\cdot\rangle$ is the inner product.

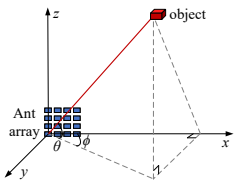

Fig. 6: Illustration of spatial beamforming.

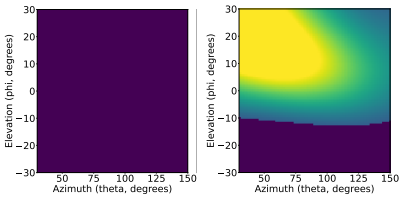

Fig. 7: Heatmap from object-static (left) and object-moving (right) cases.

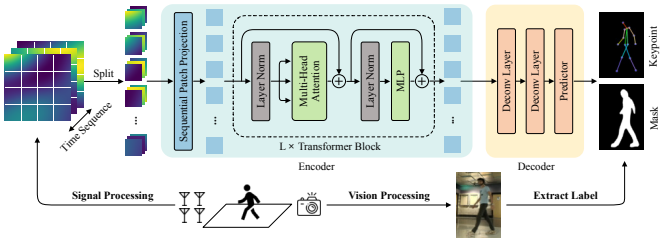

Fig. 8: Deep neural network framework of **ARS**.

Recall that $z_i(t)$ represents the output baseband signal from Rx antenna $i \in \mathcal{R}$ at time t. Then, the differential beamforming operation can be written as:

$$I(\theta, \phi, t) = \sum_{i \in \mathcal{R}} w_i(\theta, \phi) \cdot [z_i(t) - z_i(t - T_\Delta)]^*, \quad (12)$$

where T_Δ is a fixed time gap, which is empirically chosen. We note that, by taking the difference of $z(t)$ over a time period T_Δ, we suppress static components and concentrate on moving objects such as human, thereby enhancing motion detection sensitivity.

Referring to Fig. 6, we use the amplitude of $I(\theta, \phi, t)$ to estimate the strength of incoming signal from the direction (θ, ϕ) at time t. By querying $I(\theta, \phi, t)$ with all possible θ and ϕ values, we obtain the heatmap image of the moving objects at time t, i.e.,

$$\mathbf{X}(t) = \Big[\, |I(\theta, \phi, t)| : \forall\theta, \forall\phi \Big], \quad (13)$$

where $\mathbf{X}$ is the heatmap of moving object at time t.

Fig. 7 presents two examples of heatmaps generated using Eqn (13). The data was collected using a radio setup consisting of one TX antenna and eight RX antennas (see Fig. 9 in §V-A). The system operates on an ambient 5G signal with a carrier frequency of 2.35 GHz and a bandwidth of 40 MHz. Fig. 7(a) shows the heatmap generated in a static environment with no movement, while Fig. 7(b) shows the heatmap when a person is moving on the right side of the sensing area. As shown, the heatmap is able to capture movement at a coarse level. In the next section, we introduce a DNN model to refine this coarse representation for more accurate activity detection.

IV. Human Activity Recognition

In this section, we evaluate the capabilities of **ARS** through two challenging downstream tasks: human skeleton estimation and body mask segmentation. These tasks serve as a benchmark to demonstrate the potential of leveraging ambient 5G signals for fine-grained human activity detection.

A. Overview

Radio signal features are inherently sparse, noisy, and sensitive to environmental dynamics, which presents significant challenges for accurately perceiving fine-grained human activity. To address this, we employ a powerful deep neural network (DNN) capable of learning complex representations from radio features. Our approach centers on a cross-modal supervised learning framework that distills high-fidelity knowledge from the vision domain into the radio domain. During the training phase, we equip the **ARS** system with a co-located video camera to provide precise ground-truth labels, such as human skeletons and body masks, which supervise the learning of the radio-based model.

The proposed architectural framework, illustrated in Fig. 8, consists of two primary pipelines: signal processing and vision processing. The signal processing pipeline features a Split module, an Encoder, and a Decoder; it processes a sequence of radio heatmap frames (as defined in Eqn (13)) to output the estimated human skeletal structure and body mask. Simultaneously, the vision pipeline leverages off-the-shelf computer vision algorithms to extract ground-truth annotations from synchronized video frames. Crucially, once the model is trained, the video camera is discarded. In the inference stage, only ambient radio heatmap features are required to perform accurate and privacy-preserving human activity estimation.

B. Structure Framework

In practice, RF signal reflections from specific human body parts may not be captured by the receivers consistently, resulting in intermittent information loss. To mitigate the impact of such temporal voids, we exploit the inherent temporal correlation within sequences of radio frames. Given the proven effectiveness of self-attention in capturing long-range dependencies, we leverage it to jointly model the spatio-temporal relationships within the radio frame sequence. Specifically, we propose a sequential patch projection method to structure the input data for this joint analysis. Given a sequence of RF frames $\mathbf{X} \in \mathbb{R}^{F \times H \times W}$, where (H, W) denotes the spatial dimension of each radio heatmap specified in Eqn (13) and F is the number of frames, we first partition the data into $N = HW/P^2$ non-overlapping spatial patches:

$$\text{Split}(\mathbf{X}) = (\mathbf{X}_P^1, \mathbf{X}_P^2, \ldots, \mathbf{X}_P^N) : \mathbb{R}^{F \times H \times W} \mapsto \mathbb{R}^{F \cdot P^2}, \quad (14)$$

where (P, P) is the patch size, and N is the total number of patches. The data is then reshaped into a new sequence $\mathbf{X}_P$, a step that ensures each element encapsulates the complete temporal trajectory for a single spatial location.

$$\mathbf{X}_P = [\mathbf{X}_P^1, \mathbf{X}_P^2, \cdots, \mathbf{X}_P^N] : \mathbb{R}^{N \times (F \cdot P^2)}. \quad (15)$$

These spatio-temporal patches are subsequently flattened and linearly projected into C-dimensional embeddings via a trainable layer $f_\theta(\cdot)$, yielding the initial representation $\mathbf{Z}_0$:

$$\mathbf{Z}_0 = f_\theta(\mathbf{X}_P) : \mathbb{R}^{N \times (F \cdot P^2)} \mapsto \mathbb{R}^{N \times C}. \quad (16)$$

This sequence is then processed by a series of L transformer blocks, each comprising a multi-head self-attention (MHSA) layer and a multi-layer perception (MLP). To ensure stable training, layer normalization is applied before each layer, followed by residual connections. The computation for the l-th transformer block is as follows:

$$\mathbf{Z}'_l = \mathbf{Z}_{l-1} + \text{MHSA}(\text{LN}(\mathbf{Z}_{l-1})), \quad l = 1, \ldots, L. \tag{17}$$

$$\mathbf{Z}_l = \mathbf{Z}'_l + \text{MLP}(\text{LN}(\mathbf{Z}'_l)), \quad l = 1, \ldots, L. \tag{18}$$

where $\mathbf{Z}_l$ represents the output of the l-th transformer block. We denote the output of the final transformer block as $\mathbf{Z}_{\text{out}} \in \mathbb{R}^{N \times C}$, serves as the representation for downstream tasks.

To perform dense prediction from these extracted features, we design a universal decoder architecture comprising consecutive deconvolution layers. We first unflatten the output embedding sequence $\mathbf{Z}_{\text{out}}$ back into a feature map $\mathbf{F}_0 \in \mathbb{R}^{C \times \frac{H}{P} \times \frac{W}{P}}$ to restore the original 2D spatial arrangement. This feature map is then progressively upsampled; each deconvolution layer doubles the spatial resolution while halving the number of channels. After i such layers, the resulting feature map $\mathbf{F}_i$ reaches dimensions of:

$$\mathbf{F}_i \in \mathbb{R}^{\frac{C}{2^i} \times \frac{2^i H}{P} \times \frac{2^i W}{P}}, \quad \text{for } i \in \{1, 2\}. \tag{19}$$

Finally, a 1×1 convolution layer projects the map to C_{out} channels, where the output depth is tailored to the requirements of the specific task.

C. Loss Function

1) Mask Segmentation: We employ a composite loss $\mathcal{L}_{\text{mask}}$ that combines Binary Cross-Entropy (BCE) with Dice loss. While BCE handles pixel-wise classification [1]–[3], the Dice component is incorporated to mitigate the severe class imbalance between the sparse human masks and the background [4]. The balance between these terms is controlled by hyperparameters α_1 and α_2, with ϵ ensuring numerical stability:

$$\begin{aligned}\mathcal{L}_{\text{mask}} = &- \alpha_1 \sum_{i=1}^{N} \Big(y_i \log(x_i) + (1 - y_i) \log(1 - x_i) \Big) \\ &+ \alpha_2 \left(1 - \frac{2 \sum_{i=1}^{N} x_i y_i + \epsilon}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2 + \epsilon} \right),\end{aligned} \tag{20}$$

where x_i is the predicted probability of the i-th pixel belonging to the mask, and y_i is the corresponding ground truth label. N represents the total number of pixels.

2) Keypoint Estimation: For keypoint estimation, we utilize a two-level weighted Mean Squared Error (MSE) loss $\mathcal{L}_{\text{keypoint}}$:

$$\mathcal{L}_{\text{keypoint}} = \frac{1}{K} \sum_{k=1}^{K} w_k \cdot \text{mean} \left((\mathbf{Y}_k + 1) \odot (\hat{\mathbf{Y}}_k - \mathbf{Y}_k)^2 \right), \tag{21}$$

where K is the total number of keypoints, $\hat{\mathbf{Y}}_k$ and $\mathbf{Y}_k$ are the predicted and ground truth heatmaps for the k-th keypoint, respectively. Unlike standard MSE, this approach prioritizes pixels in the immediate vicinity of joints and assigns joint-specific weights w_k to emphasize critical skeletal structures.

To organize identity-free keypoints into individual instances, we adopt an associative embedding strategy [5]–[7] governed by the grouping loss $\mathcal{L}_{\text{group}}$:

$$\begin{aligned}\mathcal{L}_{\text{group}} = &\lambda_1 \frac{1}{|P|} \sum_{p \in P} \frac{1}{|K_p|} \sum_{k \in K_p} (e_{p,k} - \bar{e}_p)^2 \\ &+ \lambda_2 \frac{1}{|P|(|P| - 1)} \sum_{p_i \in P} \sum_{p_j \in P, i \neq j} \max(0, \delta - |\bar{e}_{p_i} - \bar{e}_{p_j}|),\end{aligned} \tag{22}$$

where P is the set of persons, and K_p is the set of visible keypoints for person $p \in P$. This loss enforces intra-instance cohesion by clustering joint tags around their mean $\bar{e}_p$, while maintaining inter-instance separation through a margin hyperparameter δ. The objectives are balanced via λ_1 and λ_2.

V. Experimental Evaluation

A. Implementation

1) Hardware: We have developed a prototype of **ARS**, as illustrated in Fig. 9, to evaluate its performance in realistic scenarios. The hardware consists of an RF circuit, a dipole antenna, a patch antenna array, and an FPGA board. As shown in Fig. 1, the RF board comprises two main components: an amplify-and-forward circuit and a self-mixing circuit. The amplify-and-forward section is implemented using the MMZ25332 amplifier, which provides a 30 dB power gain over the 1.8–2.7 GHz frequency range. A Mini-Circuits D17W+ signal coupler is used to derive the LO signal for the self-mixing circuit, which integrates a Skyworks SKY13418 for eight-channel RF switching, a Qorvo QPL9096 low-noise amplifier (LNA), and an Analog Devices LT5575 quadrature mixer. The printed circuit board (PCB) was fabricated on an OSH Park FR408 substrate. To digitize the baseband signals, we use the ECLYPSE Z7 FPGA board, which features two ADC channels operating at a sampling rate of 2 MSPS. The patch antennas were simulated using HFSS and fabricated on Rogers RO4350B substrate.

2) Ambient 5G Signals: To avoid potential interference with commercial 5G services provided by AT&T, Verizon, and T-Mobile, we set up a private 5G base station using commercial O-RAN equipment, including a Benetel O-RU, srsRAN O-DU and O-CU, and Open5GS core. Smartphones are connected to this 5G base station operating on spectrum band n40 (TDD, 2300–2400 MHz). This spectrum band was used under our FCC Experimental License with Call Sign #WA3XEP. **ARS** leverages the radio signals from this private 5G base station for sensing applications.

3) Software: We implement our DNN framework in PyTorch, taking 21 RF frames of size 100×100 as input. These are divided into 10×10 patches, linearly embedded into a 256-dimensional space, and processed by a 2-layer Transformer encoder. The decoder output channels are set to 1 for binary segmentation and 26 for keypoint estimation. Training uses the Adam optimizer ($5\text{e}{-4}$ LR) on four NVIDIA A6000 GPUs.

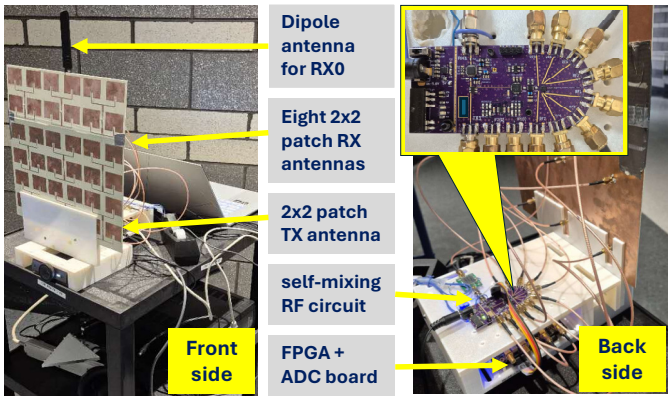

Fig. 9: Our implementation of **ARS**.

B. Data and Annotations

Our data collection testbed consists of four RF sensing device and a co-located web camera for capturing ground truth. We collected a comprehensive dataset across 8 diverse indoor and outdoor scenes on a university campus. During each collection session, the camera recorded time-synchronized video at 10 FPS. To ensure a rich variety of data, participants were instructed to move and perform arbitrary actions freely within the sensing area, imposing no restrictions on their poses or positions. We partitioned our dataset chronologically, allocating the first 80% of the collected data for the training set and the remaining 20% for the test set. This resulted in 206,734 training samples and 51,677 test samples, respectively.

To generate supervision signals for our cross-modal training, we automatically extracted annotations from the synchronized video frames. Specifically, we employed Mask R-CNN [1] to obtain binary body segmentation masks and HRNet [8] to extract 2D human keypoints.

C. Evaluation Metrics

1) Mask Segmentation: We evaluate mask segmentation performance using the standard Intersection over Union (IoU) metric [1], [9]. IoU quantifies the overlap between the set of predicted pixels S_p and the set of ground truth pixels S_{gt}. We calculate the average precision (AP) at a given threshold α as $\text{AP@}\alpha = \text{Prob}(\text{IoU} \geq \alpha)$ and $\text{AP} = 0.1 \sum_{i=0}^{9} \text{AP@}(0.5 + 0.05i)$.

2) Keypoint Estimation: We follow the COCO benchmark protocol and report Average Precision (AP) and Average Recall (AR) based on Object Keypoint Similarity (OKS) [8], [11]:

$$\text{OKS} = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{23}$$

where d_i is the Euclidean distance between detected keypoints and ground truth, v_i is the visibility flag, s is the object scale, and k_i is a per-keypoint constant that controls falloff. The main AP metric is computed by averaging AP scores over 10 OKS thresholds: $\text{AP} = 0.1 \sum_{i=0}^{9} \text{AP@}(0.5 + 0.05i)$, and the main AR metric is calculated based on averaging AR scores at the same OKS thresholds: $\text{AR} = 0.1 \sum_{i=0}^{9} \text{AR@}(0.5 + 0.05i)$.

To further dissect localization accuracy at the joint level, we employ Percentage of Correct Keypoints (PCK) metric [12]:

$$\text{PCK}_k@\alpha = \frac{\sum_{i=1}^{N} \mathbb{I}\left(\|\mathbf{p}_{ik} - \mathbf{g}_{ik}\|_2 \leq \alpha \cdot L_i\right) \cdot v_{ik}}{\sum_{i=1}^{N} v_{ik}} \times 100\%, \tag{24}$$

where N is the total number of human instances, and v_{ik} is a visibility flag. $\mathbf{p}_{ik}$ and $\mathbf{g}_{ik}$ are the predicted and ground truth coordinates of keypoint k for instance i. The scale factor $L_i = \sqrt{h(B_{gt_i})^2 + w(B_{gt_i})^2}$ is defined as the diagonal length of the ground truth bounding box B_{gt_i} of instance i.

D. Evaluation Performance

We evaluate the performance of **ARS** by comparing it with two state-of-the-art baselines: Person-in-WiFi [9] and SiWiS [10]. As detailed in Table I, **ARS** demonstrates superior performance across both mask segmentation and keypoint estimation tasks.

A key observation from our results is that **ARS** significantly widens its performance gap over the baselines as the evaluation criteria become more stringent. While all methods perform competitively at lower precision thresholds, **ARS** maintains high fidelity even under demanding requirements. For instance, in mask segmentation, **ARS** achieves a nearly 150% improvement in AP@.80 compared to SiWiS. This superior high-precision performance validates the effectiveness of our self-mixing architecture in capturing fine-grained spatial features that are typically lost in traditional CSI-based sensing. For the keypoint estimation task, **ARS** consistently outperforms the baselines in both Average Precision (AP) and Average Recall (AR). The substantial improvement in these primary metrics indicates that our cross-modal learning framework effectively distills complex body structures from noisy ambient signals, enabling precise localization of individual joints rather than mere presence detection.

To further dissect the system's capabilities, we analyze the Percentage of Correct Keypoints (PCK) across different body parts (Table II). Our findings reveal a clear performance hierarchy where torso keypoints, such as shoulders and hips, consistently achieve higher localization accuracy compared to limb extremities like wrists and ankles. This disparity is attributed to two key physical factors: primarily, the larger surface area of the torso provides stronger and more stable RF reflections ; furthermore, core body parts exhibit greater kinematic stability than the rapid and erratic motions typical of hands and feet, which allows our Transformer-based encoder to track them more reliably.

Fig. 10 visualizes representative results across various indoor scenes. Even in complex environments with multipath interference, the skeletons and masks generated by **ARS** remain highly congruent with the vision-based ground truths. These qualitative samples confirm that **ARS** can serve as a robust, privacy-preserving alternative to camera sensors for fine-grained activity monitoring.

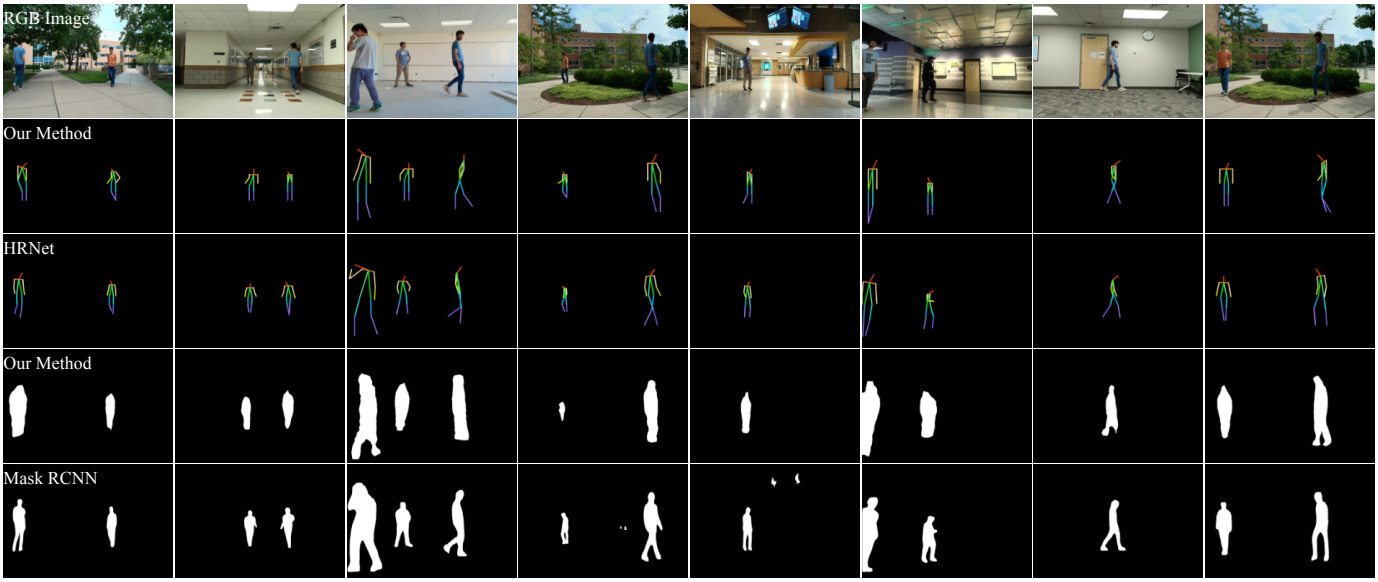

Fig. 10: Qualitative results for keypoint estimation and mask segmentation.

TABLE I: Quantitative comparison of **ARS** with SOTA methods.

	Mask Segmentation					Keypoint Estimation					
	AP$^{\uparrow}$	AP@.50$^{\uparrow}$	AP@.60$^{\uparrow}$	AP@.70$^{\uparrow}$	AP@.80$^{\uparrow}$	AP$^{\uparrow}$	AP@.50$^{\uparrow}$	AP@.60$^{\uparrow}$	AP@.70$^{\uparrow}$	AP@.80$^{\uparrow}$	AR$^{\uparrow}$
Person-in-WiFi [9]	0.3800	0.9100	0.7500	0.4000	0.0700	-	-	-	-	-	-
SiWiS [10]	0.4805	0.9452	0.8628	0.5765	0.1055	0.3469	0.8423	0.6595	0.3626	0.0816	0.4559
ARS	0.5097	0.9326	0.8311	0.6614	0.2624	0.3997	0.8995	0.7534	0.4462	0.1140	0.5071

TABLE II: Percentage of Correct Keypoints (PCK) for different keypoints. The diagonal length of the ground truth bounding box is used as the normalization factor for PCK.

Metric	Nos	Sho	Elb	Wri	Hip	Kne	Ank
PCK@.01$^{\uparrow}$	10.46	12.41	11.67	9.75	12.30	11.13	9.04
PCK@.02$^{\uparrow}$	34.95	40.33	39.27	32.92	40.17	37.20	31.96
PCK@.03$^{\uparrow}$	58.76	66.12	64.27	56.03	66.64	61.67	52.71
PCK@.04$^{\uparrow}$	76.23	82.90	80.68	73.09	83.50	78.94	68.71
PCK@.05$^{\uparrow}$	86.62	91.87	89.38	83.67	92.09	88.65	79.89

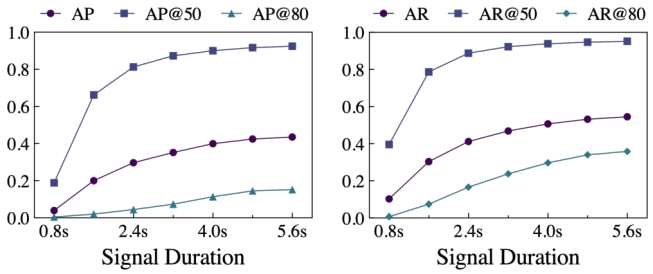

Fig. 11: Impact of input signal duration on keypoint estimation performance. System accuracy increases sharply with longer signal durations before plateauing around 4.0 seconds.

E. Performance Analysis

We now analyze the key factors influencing the performance of **ARS**, focusing on how signal duration and detection distance affect sensing accuracy.

Impact of Signal Duration. The duration of the input signal sequence is a fundamental factor in capturing the spatio-temporal dependencies of human motion. Our evaluation, conducted across durations ranging from 0.8s to 5.6s, reveals that very brief intervals are insufficient for the model to reconstruct complete poses, primarily because short windows may not capture reflections from all body parts during a kinematic cycle [13]. As illustrated in Fig. 11, system accuracy increases sharply as the duration extends toward 4.0s, after which the performance gains begin to saturate. This plateau suggests that a 4.0-second window provides a sufficient temporal receptive field for the Transformer-based encoder to effectively characterize typical human activity patterns. Consequently, we adopt 4.0s as the standard configuration to maintain a balance between sensing fidelity and computational latency.

Impact of Detection Distance. Detection distance presents a significant challenge for RF-based sensing due to the dual physical constraints of signal attenuation and complex multipath propagation. As the subject moves from 2m to 6m away from the **ARS** device, we observe a downward trend in both keypoint estimation and mask segmentation performance (Fig. 12). This degradation is directly linked to the reduction in effective Signal-to-Noise Ratio (SNR) at longer ranges. A more granular analysis of individual joints further reveals a differential impact of distance: limb extremities exhibit a more

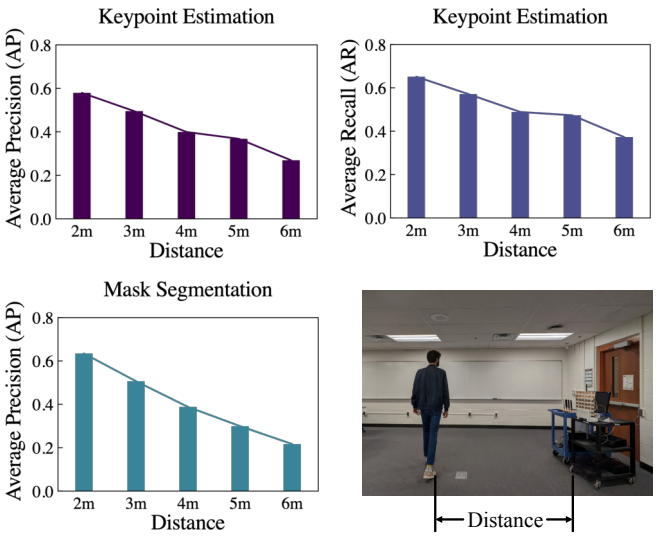

Fig. 12: Impact of detection distance on system performance. The bottom-right image depicts the experimental scenario.

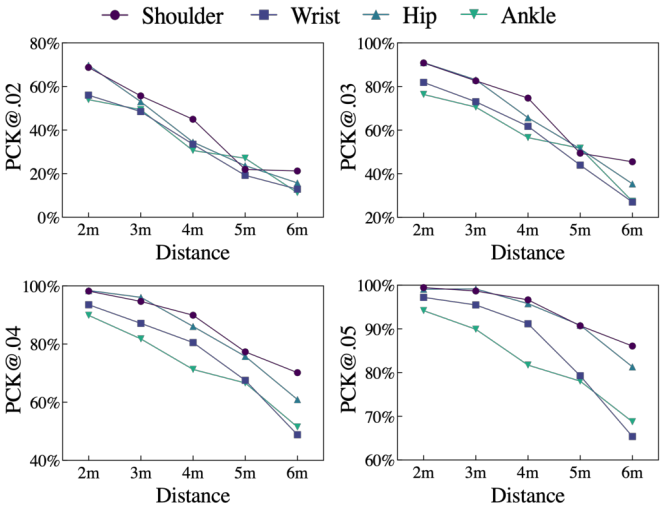

Fig. 13: Differential impact of distance on keypoint localization accuracy. Extremities show a more pronounced drop in PCK score compared to the more stable torso keypoints.

pronounced drop in localization accuracy compared to the torso. Specifically, while the larger and more stable shoulder and hip keypoints remain relatively robust, the accuracy for wrists and ankles diminishes sharply as distance increases. This confirms that the weaker reflections and more subtle movements of smaller body parts are the first to be compromised as the signal environment becomes more challenging.

VI. Related Work

ARS is an ISAC technique designed to enable ubiquitous sensing operations. It is related to the below areas.

ISAC for 5G and Beyond: ISAC is a key enabling technology for 5G and beyond [14]. A variety of techniques have been explored to realize ISAC in future cellular networks, including joint waveform design [15], [16], dual-function resource optimization [17]–[20], localization and tracking applications [21], and efforts toward standardization [22]. **ARS** is a new approach to ISAC and fundamentally differs from existing efforts.

Radar Sensing: Radar sensing offers significant advantages in environmental perception and has found widespread applications on mmWave bands [23]. The rapid advancement of deep learning has enabled a wide range of radar applications, including sleep monitoring [24], gesture recognition [25], radar imaging [26], and physiological signal monitoring [27]. However, radar sensing typically requires the use of dedicated wideband spectrum. The scarcity of available spectrum below 10 GHz limits the scalability and widespread deployment of radar applications in this frequency range. **ARS** addresses the spectrum scarcity issue.

Radio Sensing in WiFi: Channel State Information (CSI) in WiFi has been extensively studied for human activity detection. With the rapid advancement of deep learning, a growing number of studies [28] have developed various DNN-based WiFi CSI sensing applications, such as human activity recognition [29], gesture recognition [30], and human pose estimation [31]. However, WiFi CSI-based sensing faces fundamental limitations in practical deployments, including the lack of reliable temporal features due to frequency misalignment between transmitting and receiving devices, and the need for multiple coordinated devices. **ARS** is designed to complement and overcome these limitations in WiFi CSI-based sensing.

Human Activity Detection. There exists a large body of research on human activity detection, employing various approaches such as radio-based sensing [9], [13], [31]–[33] and camera-based computer vision [8], [11], [34]–[36]. In particular, the computer vision community has made rapid progress in human activity recognition by developing increasingly sophisticated methods, ranging from convolutional neural networks (CNNs) to vision transformers (ViTs), to achieve unprecedented levels of accuracy. By leveraging these advances in computer vision, **ARS** can continue to improve its detection performance.

VII. Conclusion

In this paper, we presented **ARS**, a novel sensing approach that enables standalone radio devices to leverage over-the-air RF signals from existing communication systems. Through joint hardware–algorithm co-design, **ARS** uses a self-mixing RF architecture to extract temporal and spatial features of human motion. A cross-modal learning framework further enables vision-supervised training of a radio-based DNN for activity recognition. Prototype experiments demonstrate its effectiveness in human skeleton estimation and body mask segmentation. This work introduces a scalable, spectrum-exempt sensing paradigm for future radio systems.

Acknowledgment

The authors sincerely thank the anonymous reviewers for their constructive comments. This work was supported in part by NSF Grant ECCS-2434001.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, Ieee, 2016.

[5] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 734–750, 2018.

[6] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *Advances in neural information processing systems*, vol. 30, 2017.

[7] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5386–5395, 2020.

[8] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

[9] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-wifi: Fine-grained person perception using wifi," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5452–5461, 2019.

[10] K. Song, Q. Wang, S. Zhang, and H. Zeng, "Siwis: Fine-grained human detection using single wifi device," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '24, (New York, NY, USA), p. 1439–1454, Association for Computing Machinery, 2024.

[11] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38571–38584, 2022.

[12] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[13] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7356–7365, 2018.

[14] 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 journal on selected areas in communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[15] W. Zhou, R. Zhang, G. Chen, and W. Wu, "Integrated sensing and communication waveform design: A survey," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1930–1949, 2022.

[16] Z. Xiao and Y. Zeng, "Waveform design and performance analysis for full-duplex integrated sensing and communication," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1823–1837, 2022.

[17] J. Du, Y. Tang, X. Wei, J. Xiong, J. Zhu, H. Yin, C. Zhang, and H. Chen, "An overview of resource allocation in integrated sensing and communication," in *2023 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pp. 1–6, IEEE, 2023.

[18] L. Zhao, D. Wu, L. Zhou, and Y. Qian, "Radio resource allocation for integrated sensing, communication, and computation networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 10, pp. 8675–8687, 2022.

[19] Z. He, W. Xu, H. Shen, D. W. K. Ng, Y. C. Eldar, and X. You, "Full-duplex communication for isac: Joint beamforming and power optimization," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 9, pp. 2920–2936, 2023.

[20] F. Dong, F. Liu, Y. Cui, W. Wang, K. Han, and Z. Wang, "Sensing as a service in 6g perceptive networks: A unified framework for isac resource allocation," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3522–3536, 2022.

[21] Z. Zhang, H. Ren, C. Pan, S. Hong, D. Wang, J. Wang, and X. You, "Target localization in cooperative isac systems: A scheme based on 5g nr ofdm signals," *IEEE Transactions on Communications*, 2024.

[22] X. Luo, Q. Lin, R. Zhang, H.-H. Chen, X. Wang, and M. Huang, "Isac–a survey on its layered architecture, technologies, standardizations, prototypes and testbeds," *IEEE Communications Surveys & Tutorials*, 2025.

[23] J. Xiao, B. Luo, L. Xu, B. Li, and Z. Chen, "A survey on application in rf signal," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 11885–11908, 2024.

[24] S. Yue, Y. Yang, H. Wang, H. Rahul, and D. Katabi, "Bodycompass: Monitoring sleep posture with wireless signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–25, 2020.

[25] S. Y. Kim, H. G. Han, J. W. Kim, S. Lee, and T. W. Kim, "A hand gesture recognition sensor using reflected impulses," *IEEE Sensors Journal*, vol. 17, no. 10, pp. 2975–2976, 2017.

[26] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the human figure through a wall," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–13, 2015.

[27] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 837–846, 2015.

[28] J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, "Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing," *Patterns*, vol. 4, no. 3, 2023.

[29] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, 2017.

[30] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Learning gestures from wifi: A siamese recurrent convolutional architecture," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10763–10772, 2019.

[31] Y. Zhou, J. Yang, H. Huang, and L. Xie, "Adapose: Towards cross-site device-free human pose estimation with commodity wifi," *arXiv preprint arXiv:2309.16964*, 2023.

[32] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pp. 267–281, 2018.

[33] M. Zhao, Y. Liu, A. Raghu, T. Li, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human mesh recovery using radio signals," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10113–10122, 2019.

[34] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.

[35] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *proceedings of the IEEE international conference on computer vision*, pp. 1281–1290, 2017.

[36] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," *arXiv preprint arXiv:2110.09408*, 2021.