SiWiS: Fine-grained Human Detection Using Single WiFi Device

Kunzhe Song, Qijun Wang, Shichen Zhang, Huacheng Zeng
Michigan State University, USA
{songkunz,qjwang,sczhang,hzeng}@msu.edu

ABSTRACT

Sub-6GHz radio sensing offers several compelling advantages, such as resilience to poor lighting conditions, privacy preservation, and the ability to see through walls. However, in indoor environments, the sub-6GHz ISM spectrum is heavily occupied by WiFi devices, leaving little available spectrum for sensing purposes. In this paper, we introduce SiWiS, a new approach to integrate radio sensing capabilities into *individual* WiFi devices for fine-grained human activity detection. SiWiS comprises two main components: (i) a new hardware component that can be easily installed on an off-the-shelf WiFi device, and (ii) a dual-branch deep neural network (DNN) optimized for concurrent human mask segmentation and pose estimation. We have built a prototype of SiWiS and installed it on a commercial WiFi router for evaluation. Extensive experimental results demonstrate a significant performance improvement over WiFi channel state information (CSI) based sensing methods. More importantly, zero-shot experiments confirm that SiWiS can be directly transferred to *unseen* real-world environments.

CCS CONCEPTS

• **Computer systems organization** → **Sensors and actuators**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

WiFi, joint communication and sensing, OFDM waveform for sensing, human pose estimation, human mask segmentation

ACM Reference Format:
Kunzhe Song, Qijun Wang, Shichen Zhang, Huacheng Zeng. 2024. SiWiS: Fine-grained Human Detection Using Single WiFi Device. In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), November 18–22, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3636534.3690703

ACM MobiCom '24, November 18–22, 2024, Washington D.C., DC, USA

ACM ISBN 979-8-4007-0489-5/24/11
https://doi.org/10.1145/3636534.3690703

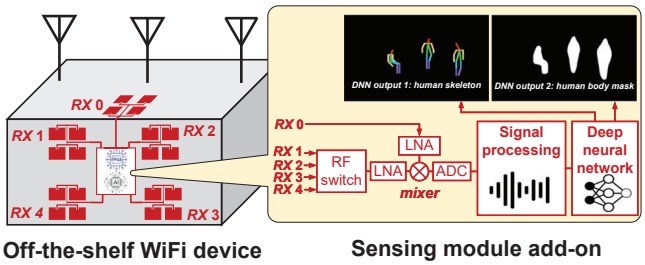


Figure 1: Schematic diagram of SiWiS. Red components are sensing module add-on. RX0 serves as Local Oscillator (LO) for the mixer, while other RX antennas are used to receive the reflective signals. No inside modifications are needed for the off-the-shelf WiFi device.

1 INTRODUCTION

Fine-grained human activity recognition is an important research area that has attracted efforts from different communities. In computer vision, many camera-based technologies [27, 34, 42, 44, 51, 55, 63, 64, 66, 76] have been developed that can detect human body keypoints with high accuracy. However, these camera-based techniques face significant challenges in large-scale deployment, including video privacy concerns and high data transmission bandwidth requirements. Furthermore, in scenarios with poor illumination, or when subjects are occluded, the performance of cameras diminishes significantly, resulting in inaccurate detections. In light of these issues, radio sensing methods [20, 25, 36, 37, 49, 50, 72–74, 77] have been regarded as a complementary approach. These methods involve collecting data through radio devices and using camera-based technologies to synchronously extract human body part annotations, which then serve as supervisory signals for the radio data. Unlike visible light, which can be obstructed by walls and objects, radio signals can penetrate these barriers and reflect off the human body. This capability allows for stable tracking of the human body in diverse scenarios, overcoming the limitations of camera-based technologies.

Among radio sensing methods, WiFi channel state information (CSI) based sensing stands out due to the prevalence and cost-effectiveness of WiFi devices. Current methods

[20, 36, 37, 46, 47, 49, 50, 70, 77] primarily utilize CSI data, obtained directly from off-the-shelf WiFi devices, to estimate the spatial coordinates of human body parts. However, CSI-based WiFi sensing has two fundamental limitations: First, CSI measurement requires at least two WiFi devices—one for transmitting and one for receiving. Due to the physical separation between the WiFi transmitter and receiver, the measured CSI inevitably suffers from carrier frequency offset (CFO), sampling time offset (STO), and carrier phase offset (CPO). While CFO and STO can be corrected, CPO cannot, which imposes a fundamental limit on the performance of CSI-based sensing methods (see §2.2). Second, the CSI measured by a WiFi device is a reflection of its surrounding environment. Due to CPO, CSI-based sensing methods are susceptible to environmental changes. This susceptibility impedes the effective transfer of trained deep learning models to new scenes (see §4.6), thereby limiting the widespread application of CSI-based sensing in real-world WiFi systems.

In this paper, we propose SiWiS, a joint hardware and software design that integrates radio sensing capabilities into ***individual*** WiFi devices for fine-grained human activity detection. SiWiS does not emit any radio signals; instead, it leverages its host WiFi device's OFDM signal for sensing. Since the transmitter and receiver are co-located on the same WiFi device, SiWiS does not suffer from CFO, STO, or CPO, thereby addressing the two fundamental limitations of CSI-based sensing methods. As shown in Fig. 1, SiWiS comprises two main components: (i) a new hardware module that can be attached to a commercial WiFi device, and (ii) a deep neural network (DNN) optimized for concurrent human mask segmentation and pose estimation. The combination of these components enables a WiFi device to detect fine-grained human activities using its OFDM signals.

For the design of SiWiS, since WiFi devices operate in time-division duplexing (TDD) mode, preventing them from receiving reflective signals while transmitting, we install an array of patch antennas on the device's surface and incorporate an RF circuit component dedicated to receiving these reflective signals. A key challenge involves the RF circuit design. The reflective signals received by SiWiS are in RF form and need to be converted to intermediate frequency (IF) for feature extraction. On one hand, traditional methods, such as using a Local Oscillator (LO) for down-conversion, face limitations due to CFO and STO between the WiFi modem and the sensing oscillator. On the other hand, WiFi modems are highly integrated with no user-accessible external interfaces, making it impossible to obtain the necessary frequency and timing clocks for the sensing circuit, even though they are physically co-located. To address this challenge, SiWiS employs a novel ***self-mixing*** architecture that involves mixing reflective WiFi signals with a "local" ambient WiFi OFDM signal. This *self-mixing* approach not only enables phase-coherent sensing but also ensures compatibility with COTS WiFi devices.

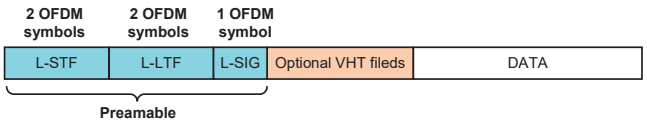


Figure 2: WiFi frame structure.

To achieve fine-grained human activity detection, SiWiS employs a dual-branch DNN designed for concurrent mask segmentation and pose estimation. It first extracts feature vectors from the input WiFi-based sensing signal using a signal encoder. Given that reflections from certain body parts may not be captured within short time intervals, potentially resulting in de-emphasized or missing key information, SiWiS incorporates a self-attention block into the signal encoder to establish connections across longer sequences of signal frames. Furthermore, to enhance the adaptability of the signal encoder to both mask segmentation and pose estimation tasks, SiWiS employs a cross-attention block to establish fine-grained spatial pixel feature connections.

We have built a prototype of SiWiS using a custom-designed PCB and optimized patch antennas, and installed it on a COTS WiFi router for experimental evaluation. Extensive experimental results demonstrate a significant improvement of SiWiS compared to CSI-based WiFi sensing methods. More importantly, zero-shot experiments confirm that SiWiS can be directly transferred to *unseen* real-world environments.

The contributions of this paper are summarized as follows:

- SiWiS represents the first approach to enable fine-grained human detection using a *single* WiFi device. This method is also applicable to other radio communication devices, such as 5G and Bluetooth, opening up new possibilities for integrated communication and sensing across a wide range of systems.
- SiWiS achieves *phase-coherent* sensing on a WiFi device by employing a *self-mixing* architecture. This feature is crucial for ensuring the system's effectiveness in unseen scenarios, significantly broadening its applicability.
- Extensive experiments validate the superior performance of SiWiS compared to CSI-based WiFi sensing methods. Additionally, zero-shot evaluation results demonstrate that SiWiS can be directly transferred to *unseen* scenarios.

2 SINGLE WIFI DEVICE FOR SENSING

2.1 Primer on WiFi CSI

WiFi (except for 802.11b) uses OFDM modulation for data transmission. In OFDM modulation, CSI refers to the complex channel coefficients across its OFDM subcarriers. Let X_k denote the original signal on OFDM subcarrier k at a WiFi

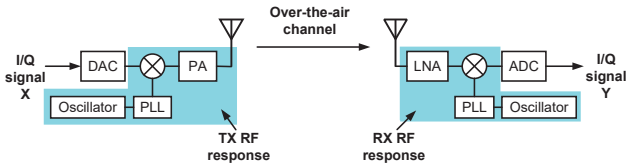


Figure 3: System model of WiFi communications.

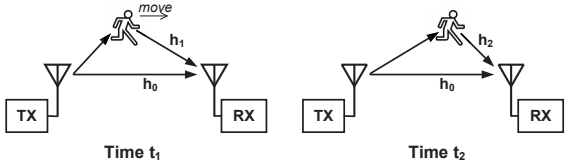


Figure 4: Relation between person movement and measured CSI.

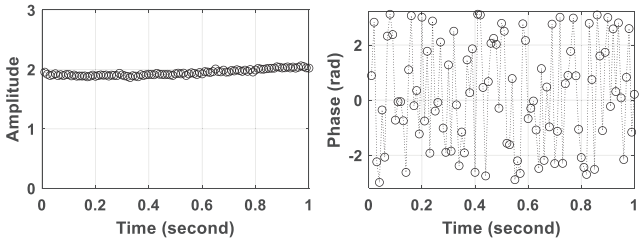


Figure 5: The amplitude (left) and phase (right) of CSI measured over 100 consecutive WiFi packets in a static scenario.

TX, and Y_k denote the received signal on OFDM subcarrier k at a WiFi RX. Then, the data transfer function can be modeled as: $Y_k = H_k X_k + W_k$, where H_k is the channel coefficient of OFDM subcarrier k between TX and RX, and W_k is the noise on OFDM subcarrier k at the RX. Denote K as the total number of valid OFDM subcarriers in WiFi. Then, $\mathbf{H} = [H_1, H_2, \ldots, H_K]$ is referred to as WiFi CSI, which plays a crucial role in the channel equalization of its data packets. For example, with channel H_k, the WiFi RX can estimate the original signal by $\hat{X}_k = \frac{Y_k}{H_k}$ if the noise is small.

To facilitate the channel estimation, WiFi signals are structured in the frame format as shown in Fig. 2. Each frame has a preamble, which is defined in the IEEE 802.11 standards and known to every WiFi device. The preamble includes legacy short training field (L-STF) and legacy long training field (L-LTF). The L-STF is used for basic synchronization and coarse frequency offset correction, and the L-LTF is used for more accurate channel estimation and fine synchronization. They were meticulously designed to maximize WiFi RX's channel estimation accuracy and packet-decoding performance.

Ideally, CSI should be determined *solely* by the over-the-air (OTA) channel between WiFi TX and WiFi RX, characterizing physical factors such as distance, reflectors, and other surrounding objects. However, due to imperfections in RF circuits, the CSI measured at a WiFi device is a reflection of three components: TX RF response, OTA channel, and RX RF response, as shown in Fig. 3. Particularly, since the RF mixers in the TX and RX are driven by different oscillators, the CSI measured at the RX inevitably suffers from three imperfections: CFO, STO, and CPO.

While CFO and STO can be estimated and corrected, CPO cannot. Fortunately, CPO does not affect the decoding of data packets because it is addressed by channel equalization. However, when CSI is used for sensing purposes, the randomness of CPO imposes a fundamental limitation on sensing performance. This is because the CSI observed by a WiFi device is not the OTA CSI. The presence of CPO prevents the estimation of the phase information of OTA CSI across a series of data packets, which would otherwise be useful for inferring object movement distances. This constitutes a key difference between CSI sensing and radar sensing. We will elaborate on this issue below.

2.2 Limitations of WiFi CSI Sensing

WiFi and CSI were originally designed for communication purposes. When repurposed for sensing, WiFi CSI encounters two fundamental limitations. First, it requires at least two WiFi devices working together to measure CSI. This complicates the system setup and limits its applicability. Second, the CSI measured by a WiFi device is not OTA CSI. Due to the existence of CPO between WiFi TX and RX, CSI-based sensing lacks *phase coherence* over time. Therefore, this approach is susceptible to environmental changes and thus not easy to be generalized to new (unseen) scenes. In what follows, we explain this limitation in detail.

Consider a simple WiFi CSI sensing case as shown in Fig. 4. Suppose that there are two OTA paths from TX to RX. One is line-of-sight (LoS) path, and the other is person-reflected path. Denote $h_0 \in \mathbb{C}$ as the LoS path channel, which remains unchanged over time. Denote $h_1 \in \mathbb{C}$ and $h_2 \in \mathbb{C}$ as the person-reflected path channels when TX sends packet 1 (at time t_1) and packet 2 (at time t_2), respectively. If there were no CFO, STO and CPO, the measured CSI from those two packets can be written as: $H_1 = h_0 + h_1$ and $H_2 = h_0 + h_2$, where H_1 and H_2 are the measured CSI at the two time moments. Then, the measured CSI change at the RX can be written as: $\Delta H = H_2 - H_1 = h_2 - h_1$. This provides two insights: (i) the measured CSI change is caused *solely* by object movements; and (ii) the measured CSI change is independent of static paths. It means that the object movement can be inferred based on the measured CSI sequence. Moreover, since the measured CSI change is not affected by the static objects in the environment, the inference can be generalized to new (unseen) scenes. This is a case of *phase-coherent* sensing.

Unfortunately, while CFO and STO can be estimated and corrected at RX, CPO can,[1] making it impossible to derive the OTA CSI based on the measured CSI. Due to the existence of CPO, the measured CSI from packets 1 and 2 should be modeled as $H_1 = (h_0 + h_1)e^{j\theta_1}$ and $H_1 = (h_0 + h_2)e^{j\theta_2}$, where θ_1 and θ_2 are unknown random phases caused by CPO. Evidently, the phase information of the measured CSI sequence is not useful at all in the inference of object movements. To illustrate this point, Fig. 5 plots the phase of our measured CSI over 100 packets (on one subcarrier) from a laptop using Intel 5300 NIC CSI tool. The 100 consecutive packets last for about one seconds. During this time period there are no object movements. We can see that the CSI phase is random across consecutive packets.

For this reason, CSI-based sensing approaches primarily rely on the CSI amplitude sequence (i.e., $|H_1|$, $|H_2|$, ...) to infer object movements. Since the measured CSI amplitude is influenced by both dynamic and static objects, these approaches have limited generalizability and transferability to unseen scenarios.

2.3 Our Design

To achieve phase-coherent sensing on a single WiFi device, SiWiS designs a new hardware component that can be easily attached on a vast majority of COTS WiFi devices such as WiFi routers, laptops, desktops, and smart TV. SiWiS employs two techniques: (i) self-mixing of WiFi OFDM signals, and (ii) patch antennas for sensing directivity. These two techniques enable individual WiFi devices to achieve Doppler-radar-like *phase-coherent* sensing capabilities.

Fig. 1 shows the hardware component of SiWiS. One dipole/patch antenna (RX0) was installed facing the WiFi communication antennas to obtain a local copy of WiFi OFDM signal, which will be used as the LO for the RF mixer. Multiple patch antennas are installed on one side of the WiFi device to receive the reflective OFDM signals from target objects. The use of patch antennas serves two purposes: (i) to reduce self-interference from WiFi TX antennas, and (ii) to maximize the strength of received signals reflected from moving objects. To reduce the hardware complexity, an RF switch is used for the sharing of a single RF chain. The received signals are first amplified using LNA (low noise amplifier) and then mixed with LO for down conversation. The output of mixers is sampled using ADC (analog-to-digital converter) for digital signal processing and learning-based inference. For this design, we have the following remarks.

[1]While the literature includes work on CPO estimation and correction, existing research primarily focuses on intra-packet CPO correction rather than inter-packet CPO correction. Therefore, current approaches are not applicable in this context.

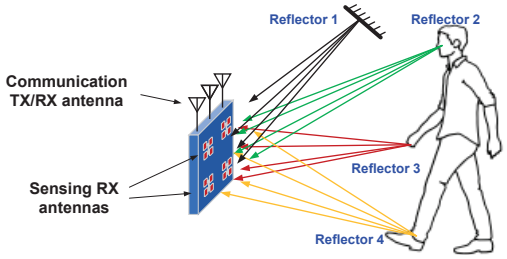


Figure 6: Reflective signals of static/moving objects.

- **Joint communication and sensing:** On one hand, the communication and sensing subsystems are physically independent. These two subsystems do not interfere with each other. On the other hand, SiWiS is a joint communication and sensing system. The sensing subsystem does not emit radio waves. Instead, it leverages ambient WiFi OFDM signals to realize its sensing functions.
- **Moving object detection:** SiWiS is fundamentally different from FMCW radar. FMCW radar can detect both static and moving objects, thanks to its FMCW waveform and its wide spectrum band (e.g., 100s MHz). However, SiWiS relies on WiFi OFDM signals for sensing. It does not have spectrum coexistence issues with WiFi communication systems. Like a Doppler radar, it is suited for detecting moving objects, not static objects.
- **Installation on COTS WiFi devices:** SiWiS can be easily installed on a COTS WiFi device such as WiFi router, smart TV, and laptop. It requires no modifications inside WiFi devices. If a WiFi device is of a small size and does not have enough surface for patch antennas, SiWiS can be placed in the close proximity of the WiFi device.

2.4 Feature Extraction for Sensing

2.4.1 Mathematical Modeling. For simplicity, we first consider one sensing antenna of SiWiS. Denote $x(t)$ as the baseband OFDM signal of a WiFi frame. Then, the WiFi RF signal can be written as $s(t) = x(t)e^{j2\pi f_c t}$, where f_c is the carrier frequency of a WiFi channel at 2.4 GHz or 5 GHz. Denote $\mathcal{R}$ and $\mathcal{M}$ as the sets of static and moving reflectors in the environment, as shown in Fig. 6. Then, the RF signal received by one sensing antenna can be written as:

$$r(t) = \sum_{i \in \mathcal{R} \cup \mathcal{M}} \alpha_i x(t - \tau_i) e^{j2\pi f_c (t - \tau_i)}, \quad (1)$$

where $\alpha_i \in \mathbb{R}^+$ and $\tau_i \in \mathbb{R}^+$ are the attenuation and delay of the reflective signal from reflector $i \in \mathcal{R} \cup \mathcal{M}$.

As shown in Fig. 1, SiWiS mixes the received RF signal (from RX1, RX2, RX3, or RX4) with the local copy from its antenna RX0. Since RX0 is physically close to WiFi's TX antennas, the LoS path is dominant compared to the non-LoS paths. Therefore, we use $s(t)$ to approximate the LO for

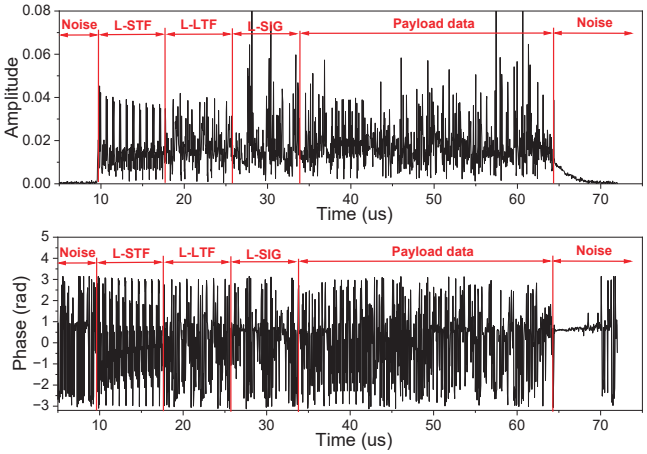


Figure 7: Illustration of the mixer's output signal $y(t)$: amplitude (up) and phase (bottom)

mixing. Then, the output of the mixer can be written as:

$$y(t) = r(t)s(t)^* = \sum_{i\in\mathcal{R}\cup\mathcal{M}} \alpha_i x(t-\tau_i)x(t)^* e^{-j2\pi f_c \tau_i}, \quad (2)$$

where $(\cdot)^*$ is complex conjugate operator.

Denote d_i as the distance between SiWiS and reflector $i \in \mathcal{R}\cup\mathcal{M}$. Then, Eqn (2) can be written as:

$$y(t) = \sum_{i\in\mathcal{R}\cup\mathcal{M}} \alpha_i x(t-\frac{2d_i}{c})x(t)^* e^{-j\frac{4\pi}{c}f_c d_i}, \quad (3)$$

where c is light speed. A sample of $y(t)$ from the RF mixer is illustrated in Fig. 7.

Suppose that the mixer's output signal is sampled with time interval Δt (50 ns for 20 MHz WiFi). Then, the digital version of $y(t)$ can be written as:

$$y(n\Delta t) = \sum_{i\in\mathcal{R}\cup\mathcal{M}} \alpha_i x(n\Delta t-\frac{2d_i}{c})x(n\Delta t)^* \, e^{-j\frac{4\pi}{c}f_c d_i}, \quad (4)$$

where n is the signal sample index in the time domain.

2.4.2 Signal Feature Analysis. Using $y(n\Delta t)$ in Eqn (4), $n = 1, 2, \ldots$, to infer the movement pattern of an object is challenging for two reasons. First, the WiFi OFDM signal $x(t)$ is time-varying, depending on its payload data. Second, the WiFi signal transmission power is not fixed and is subject to power adaptation control. These factors make it challenging to extract stable features from the mixer's output signal. To address this challenge, we take advantage of the preamble in each WiFi frame. We sum the mixer's output signal $y(n\Delta t)$ over the data samples corresponding to the L-LTF in a WiFi frame. Denote $\mathcal{B}_{\text{ltf}}$ as the set of data samples corresponding to the L-LTF in a WiFi frame. Then, we define

$$Y = \sum_{n\in\mathcal{B}_{\text{ltf}}} y(n\Delta t). \quad (5)$$

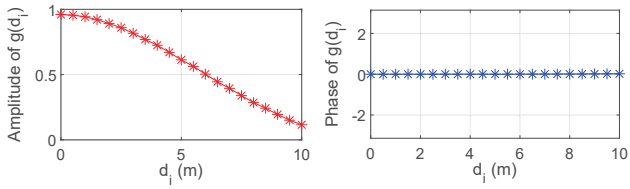


Figure 8: Amplitude (left) and phase (right) of $g(d_i)$.

Based on Eqn (4) , we have

$$Y \overset{(a)}{=} \sum_{i\in\mathcal{R}}\sum_{n\in\mathcal{B}_{\text{ltf}}} \alpha_i x(n\Delta t-\frac{2d_i}{c})x(n\Delta t)^* e^{-j\frac{4\pi}{c}f_c d_i} + \quad (6)$$

$$\sum_{i\in\mathcal{M}}\sum_{n\in\mathcal{B}_{\text{ltf}}} \alpha_i x(n\Delta t-\frac{2d_i}{c})x(n\Delta t)^* e^{-j\frac{4\pi}{c}f_c d_i} \quad (7)$$

$$\overset{(b)}{=} C_s + \sum_{i\in\mathcal{M}} \alpha_i \left(\sum_{n\in\mathcal{B}_{\text{ltf}}} x(n\Delta t-\frac{2d_i}{c})x(n\Delta t)^*\right) e^{-j\frac{4\pi}{c}f_c d_i}, \quad (8)$$

where $C_s \in \mathbb{C}$ is a constant complex number. Eqn (a) results from separating the reflections from static and mobile objects. Eqn (b) follows from the fact that the reflection from static objects remains constant after summing the L-LTF samples.

We now focus on the term in the parenthesis in Eqn (8). Define

$$g(d_i) = \sum_{n\in\mathcal{B}_{\text{ltf}}} x(n\Delta t-\frac{2d_i}{c})x(n\Delta t)^*, \quad (9)$$

where $c = 3\times10^8$ m/s, $\Delta t = 50$ ns for 20 MHz WiFi. For $n \in \mathcal{B}_{\text{ltf}}$, $x(n\Delta t)$ is the waveform of L-LTF in IEEE 802.11. Fig. 8 plots the numerical results of $g(d_i)$. Evidently, when $d_i \le 10$ m, $g(d_i)$ is a real number, i.e., $g(d_i) \in \mathbb{R}$.

Denote $[\cdot]_{\text{ac}}$ as the operation of removing the DC component of a signal vector. Then, we have

$$[Y]_{\text{ac}} = \sum_{i\in\mathcal{M}} \alpha_i \cdot g(d_i) \cdot e^{-j\frac{4\pi}{c}f_c d_i}. \quad (10)$$

Eqn (10) characterizes the relationship between SiWiS's observed signal $[Y]_{\text{ac}}$ and object distance d_i. Given that α_i, c, f_c, d_i, and $g(d_i)$ in Eqn (10) are all real numbers, we can immediately derive the following lemma.

Lemma 2.1. *If there is a single moving object of small physical size, then the phase of SiWiS' observed signal, i.e.,* $\arg\left([Y]_{\text{ac}}\right)$, *is a linear function of object distance d.*

Based on Lemma 2.1, we have the following remarks.

- **Signal phase vs. distance:** SiWiS achieves a deterministic (linear) relationship between *its observed signal phase* and *object moving distance*. This is a sharp contrast to existing CSI-based sensing approaches, where the CSI phase appears to be random as shown in Fig. 5. This feature makes it possible for SiWiS to detect sub-centimeter movements, achieving phase-coherent sensing like Doppler radars. Experimental results will be provided to validate this feature shortly.

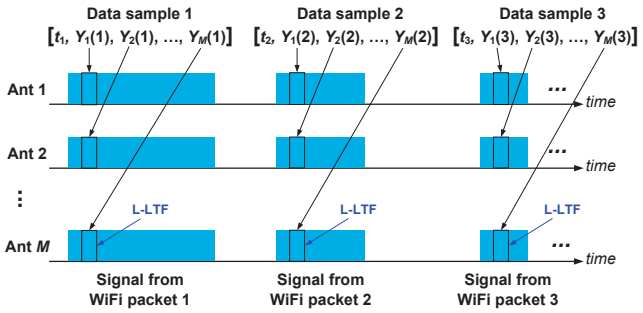


Figure 9: Illustrating the feature extraction operations.

- **WiFi's TX power adaptation:** A COTS WiFi device may use different power levels for transmitting various data packets, depending on channel condition and packet type. While this power adaptation affects the amplitude of SiWiS's observed signals, it does not impact their phase, as demonstrated by Lemma 2.1. Therefore, the phase feature is resilient to WiFi power adaptation.
- **Reflective signal separation:** Lemma 2.1 was derived under the assumption that there is a single moving object of small physical size. In practice, there may be multiple moving objects, or the object size may be large. In this case, we rely on multiple sensing antennas to differentiate the objects in the spatial domain and on the DNN to disentangle the underlying relationship between the observed signal features and the object movements.

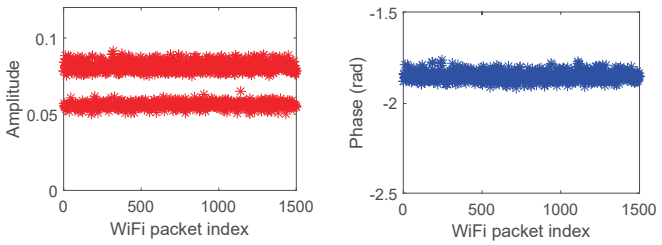
2.4.3 Feature Extraction Algorithm. Based on the above analysis, we summarize SiWiS's feature extraction operations in Fig. 9. With a bit abuse of notation, we denote $y_m(n\Delta t, k)$ as the mixer's output signal, where m is the sensing antenna index, n is the data sample index of a WiFi packet, Δt is the time sampling interval, and k is the index of detected WiFi packets. Then, we define the signal feature by letting: $Y_m(k) = \left[\sum_{n\in\mathcal{B}_{\text{ltf}}} y_m(n\Delta t, k)\right]_{\text{ac}}$. Collectively, the feature data tensor is written as:

$$\mathrm{S}_k = [t_k, Y_1(k), Y_2(k), \ldots, Y_M(k)],\ 1 \le k \le K_{\text{packet}}, \quad (11)$$

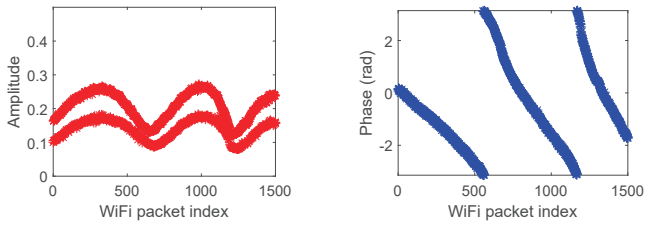
where M is the number of sensing antennas, t_k is the timestamp of WiFi packet k, and K_{packet} is the total number of detected WiFi packets. S_k, $k = 1, 2, \ldots$, is then streamed into the dual-branch DNN in §3 for human pose estimation and mask segmentation.

2.5 Feature Validation

For ease of exposition, we focus on $Y_1(k)$ in Eqn (11). We conduct experiments to validate the relationship between $Y_1(k)$ and the object movement distance in two scenarios. In the first scenario, we keep the scene static and measure $Y_1(k)$ for 1,500 WiFi packets, which last for 3 seconds. Fig. 10a presents the measured $Y_1(k)$. We can see that its amplitudes have two



(a) Observations in static scenario.



(b) Observations in moving scenario.

Figure 10: Measured amplitude (left) and phase (right) of signal $Y_1(k)$, $k = 0, 1, \ldots, 1,500$, in static and moving scenarios.

distinct levels. This was caused by WiFi's TX power adaptation. *More importantly, the phase of $Y_1(k)$ remains stable across data packets (i.e., over time). This is in a sharp contrast to the measured CSI phase shown in Fig. 5, which is random over time.*

In the second scenario, we have one person standing in front of the WiFi device and moving himself towards it. Fig. 10b plots the measured signal $Y_1(k)$ across 1,500 WiFi packets, which also last for 3 seconds. *Evidently, there is a deterministic relationship between the observed phase and the human moving distance.* The relationship is not perfectly linear. This was caused by the inconstant moving speed and the large size of human body. Over this time period, the phase change is 14.5 radians, which correspond to a distance of 11.9 cm in theory. Our manual measurement of the person's movement distance is 12.5 cm. This result roughly agrees with the theoretical value, demonstrating the usefulness of signal phase for estimating object movement distance.

Similar relationships were also observed on $Y_2(k), \ldots, Y_M(k)$. These experimental results confirm that SiWiS is a phase-coherent sensing approach. This lays a concrete foundation for the DNN design and underscores the reason why SiWiS outperforms CSI-based sensing approaches.

2.6 Interference Resilience and Removal

2.6.1 Interference from Other WiFi devices. Consider the case where multiple WiFi devices are in the same area. A question to ask is how SiWiS can determine whether its excitation OFDM signal is from its host WiFi device or from a non-host WiFi device. To answer this question, we define

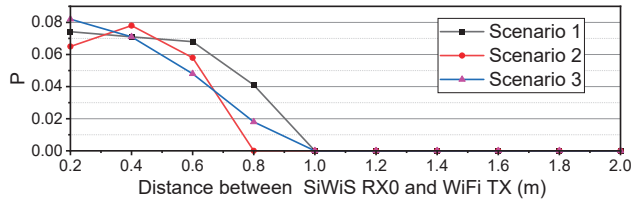


Figure 11: Measured P versus distance D.

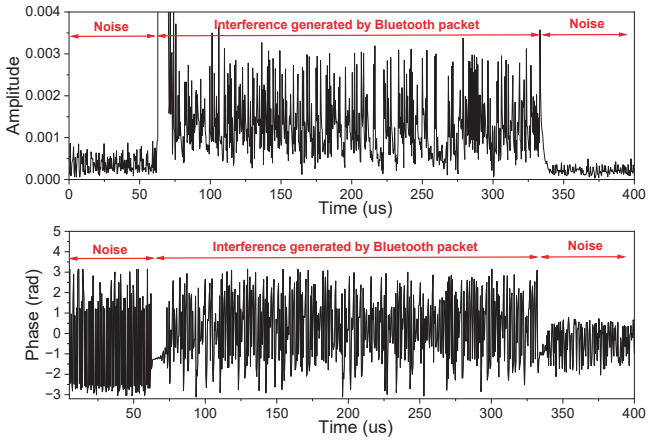


Figure 12: The mixer's output signal $y(t)$ when the excitation signal is from Bluetooth: amplitude (up) and phase (bottom)

a new metric $P = \sum_{m=1}^{M} |Y_m(k)|^2$ for SiWiS. We conducted experiments to examine the relationship between P and D, where D is the distance between SiWiS and its excitation signal source. Fig. 11 displays our measurement results across three different scenarios. It can be seen that P decreases rapidly as D increases, and $P = 0$ for $D \geq 1$ meter. This suggests that SiWiS can use the value of P to determine whether its excitation signal is from its host WiFi device. SiWiS can discard all received WiFi frames where $P \leq P_{\text{thres}}$, where P_{thres} is a threshold that can be empirically set (e.g., $P_{\text{thres}} = 0.05$ in this case). By doing this, SiWiS will effectively eliminate interference from all WiFi devices located 0.6 meters away.

2.6.2 Interference from Non-WiFi devices. If its excitation signal is not from a WiFi device, SiWiS can easily identify it based on its mixer output signal. As an example, we use Bluetooth to generate the excitation signal for SiWiS. The Bluetooth device is positioned 0.1 meters from SiWiS. Fig. 12 shows SiWiS' mixer output. Comparing to Fig. 7 (WiFi excitation), we can see that the mixer output signal is more than 20 dB weaker and lacks the L-STF and L-LTF signatures when the excitation signal originates from Bluetooth. These differences allow SiWiS to easily identify and then exclude the interference from non-WiFi devices.

3 HUMAN ACTIVITY RECOGNITION

3.1 Overview

SiWiS focuses on human activity recognition as a use case for our proposed sensing approach. Due to the complexity of RF sensing, DNNs have become mainstream for human detection [49, 50, 60, 72–74, 77]. Traditional signal-processing-based feature extraction is ineffective for complex human detection tasks because of the intricate nature of RF environments. However, DNNs can develop robust feature extractors for sensing signals through supervised learning. Furthermore, compared to "signal processing + DNN" approaches, end-to-end DNN solutions require only a few extra parameters but have the potential to enhance the performance [6]. Therefore, we employ an end-to-end DNN for this task.

SiWiS employs a cross-modal supervision approach for DNN training, transferring knowledge from vision-based human recognition models to WiFi-based models. At high level, it comprises two components: *vision processing* and *signal processing*, as shown in Fig. 13. During the training stage, we use both video frames and WiFi-based sensing signals, aligning them by their timestamps. In the inference stage, we rely solely on WiFi-based sensing signals.

3.2 Deep Neural Network Framework

3.2.1 Signal Encoder. For the input WiFi-based sensing signals, we first apply convolution layers to extract temporal features. To enhance the model's ability of capturing feature correlations over time, we employ self-attention blocks for processing longer periods of sensing signals. Additionally, to reduce computational costs, we modify the self-attention block using a bottleneck design. Preceding and following the self-attention block, we integrate two fully connected layers for dimension reduction and subsequent expansion, respectively. We then employ parameter-free identity shortcuts to connect the inputs and outputs of the module. With a reduction scale of X for the fully connected layers, the module's parameter amount is also reduced by a factor of X, significantly accelerating the model's training speed.

3.2.2 Mask Segmentation and Pose Estimation Decoder. Given that the supervisory signals for our model are heatmaps generated by the vision process network, it is necessary to upsample the sensing signal features to match the dimensions of the heatmap features. Each WiFi signal feature obtained from the signal encoder has a dimension of $C \times 1 \times 1$, where C represents the number of channels. Then, we introduce a cross-attention block designed to produce feature maps with dimensions of $C \times h \times w$. Specifically, we initialize $h \times w$ trainable pixel embeddings and employ cross-attention to compute the attention weights of each pixel embedding towards each sensing signal feature. Subsequently, we employ convolutional and upsampling layers to increase the

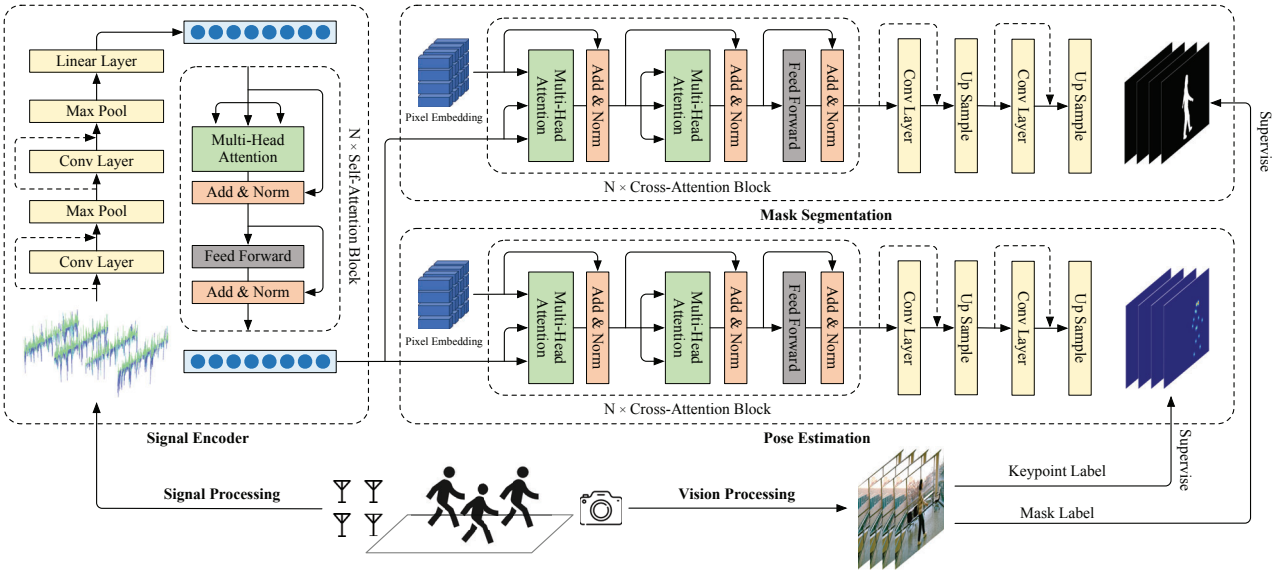


Figure 13: Deep neural network framework of our human activity recognition system. The framework consists of two components: *vision processing* and *signal processing*. In the vision process, we employ vision models to extract keypoint labels and mask labels from video frames, serving as the ground truth for supervised learning. In the signal process, we predict mask segmentation and pose estimation heatmaps from WiFi-based sensing signals.

feature map size from $h \times w$ to $h' \times w'$, aligning with the dimensions of the ground truth heatmap features. To reduce computational costs, our convolutional layer uses a bottleneck residual block from ResNet [16], which utilizes 1×1 convolutions to reduce feature channels and consequently decrease the module's parameters.

3.2.3 Loss Functions. For the task of body part mask segmentation, we initially employed the binary cross-entropy (BCE) loss [15, 30, 38]. However, we encountered limitations with BCE loss in practical applications. This was primarily due to instances where the subject was distant from the camera, resulting in the segmentation area occupying only a few pixels in the image. Consequently, this led to a class imbalance between foreground and background pixels. To address this issue, we integrated the Dice loss. The formulation of our mask segmentation loss function is expressed as follows:

$$\mathcal{L}_{mask} = \alpha_1 \sum_{i=1}^{N} \Big(y_i \log x_i + (1 - y_i) \log(1 - x_i) \Big) + \alpha_2 \left(1 - \frac{2 \sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2} \right), \tag{12}$$

where x_i represents the predicted value, and y_i is the real class label. N is the number of pixels. α_1 and α_2 denote scalar weights utilized to balance the two losses. Fig. 14 illustrates the mask segmentation outcomes w/ and w/o the integration of Dice loss. It is evident that the integration of Dice loss can significant improve the detection accuracy.

Figure 14: Dice loss improves mask segmentation. Left: RGB image; Middle: results w/o Dice loss; Right: results w/ Dice loss.

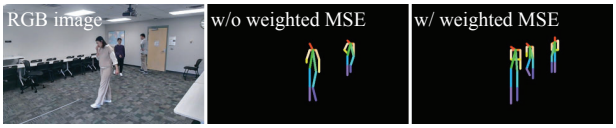


Figure 15: Weighted MSE loss improves pose estimation. Left: RGB image; Middle: results w/o weighted MSE loss; Right: results w/ weighted MSE loss.

For pose estimation, we utilize Mean Squared Error (MSE) loss [8, 11, 35, 42, 51, 55]. Since human keypoints typically occupy very few pixels in the image [50], applying MSE loss with an average regression error over all pixels may excessively emphasize background regions in the loss function. Hence, we adopt a weighted MSE loss, assigning greater importance to pixels closer to the keypoints. For multi-person pose estimation, previous studies [9, 22, 33] have shown that employing associative embedding can effectively address grouping issues with high accuracy. Following the approach outlined in [33], we use $\mathcal{L}_{group}$ for keypoint grouping. This grouping process organizes identity-free keypoints into individuals by grouping keypoints with smaller ℓ_2 distances

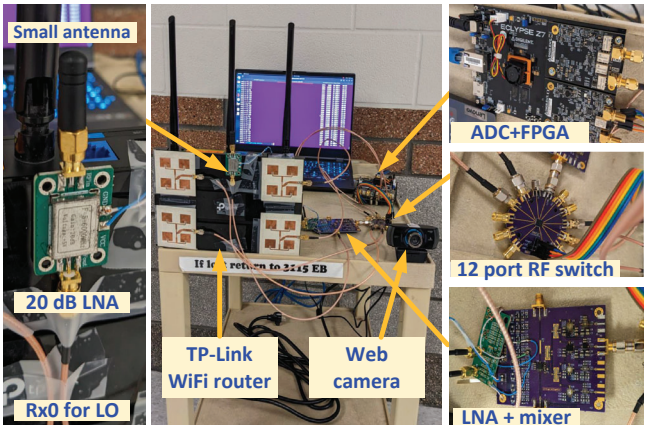


Figure 16: Prototype of SiWiS on a TP-Link WiFi router.

between their tags. The expression of our pose estimation loss function is as follows:

$$\mathcal{L}_{pose} = (y+1) \odot \| \hat{y} - y \|_2^2 + \mathcal{L}_{group}, \tag{13}$$

where $\hat{y}$ and y represent the prediction and ground truth heatmaps, respectively, and $\odot$ denotes element-wise multiplication. Fig. 15 illustrates the pose estimation outcomes w/ and w/o the integration of weighted MSE loss. It is evident that the integration of weighted MSE loss is an effective approach to improve the detection accuracy.

The overall loss function for our training process is the weighted sum of the mask segmentation loss and pose estimation loss:

$$\mathcal{L} = \mathcal{L}_{mask} + \lambda \mathcal{L}_{pose}, \tag{14}$$

where λ is a scalar weight used to balance the two losses.

4 EXPERIMENTAL EVALUATION

In this section, we conduct extensive experiments with the aim of answering the following questions.

- **Q1 (§4.4):** How does SiWiS perform when compared to the SOTA in multi-person detection?
- **Q2 (§4.5):** How does the parameters (e.g., signal duration and distance) affect the performance of SiWiS?
- **Q3 (§4.6):** How does SiWiS perform when compared to the SOTA in *unseen* scenarios?

4.1 Implementation

We have built a prototype of SiWiS on a TP-Link 802.11ac WiFi router as shown in Fig. 16 to evaluate its performance in realistic scenarios.

4.1.1 Hardware. The RF hardware comprises a small dipole antenna (RX0), four patch antennas (RX1-4), an RF switch, a custom-designed RF PCB, an ADC daughterboard, and an ECLYPSE Z7 FPGA board. The RF PCB was designed using Analog Device's HMC951A (mixer) and HMC717A (LNA).

Table 1: Statistical information of dataset. "Person #" is the number of individuals present simultaneously within a scene. "Frame #" is the number of video frames.

Person #	1	2	3	4	Total
Frame #	130,244	161,986	155,088	63,576	510,894

Figure 17: Different environments in the dataset.

It was fabricated using OSH Park substrate FR408. The four patch antennas were simulated using HFSS and fabricated on Rogers RO4350B. The FPGA controls the RF switch and the ADC daughterboard for signal sampling at 10 MSps. The FPGA sends the sampled data to a laptop via Ethernet for signal processing. The TP-link WiFi router works on channel 44. A web camera was installed on the system to capture images for DNN training.

4.1.2 Software. Our network is implemented using PyTorch. The input to the network consists of signals received from *four* antennas. Both the mask segmentation and pose estimation modules produce heatmaps with a resolution of 48×64. For $\mathcal{L}_{pose}$, the associative embedding loss is weighted by a factor of $1e-3$ relative to the MSE loss of the keypoint detection heatmaps. Training is conducted on eight V100 GPUs (32GB memory), with a learning rate of $1e-3$.

4.2 Data and Annotations

For data collection, a web camera was used to capture video frames at 10 FPS. We synchronized WiFi-based sensing signals and video frames using the timestamps recorded during signal acquisition. The number of individuals present in the video varied from 1 to 4, and no restrictions were imposed on subjects' poses or positions. Participants were free to move around and perform any actions in front of the device.

To evaluate our system, we collected data in 12 different scenes around a university campus, as shown in Fig. 17. In the experiments, our dataset was sorted according to the timestamps recorded during collection. We allocated the first 80% of the data for training and the remaining 20% for

Table 2: Evaluation Performance of Mask Segmentation and Pose Estimation. All metrics are the higher the better.

	Setting	Mask Segmentation					Pose Estimation				
		mAP↑	AP@50↑	AP@60↑	AP@70↑	AP@80↑	mAP↑	AP@50↑	AP@60↑	AP@70↑	AP@80↑
Person-in-WiFi [50]	Two 3-Ant WiFi Devices	0.3800	0.9100	0.7500	0.4000	0.0700	-	-	-	-	-
SiWiS	Single WiFi Device	0.4805	0.9452	0.8628	0.5765	0.1055	0.3469	0.8423	0.6595	0.3626	0.0816

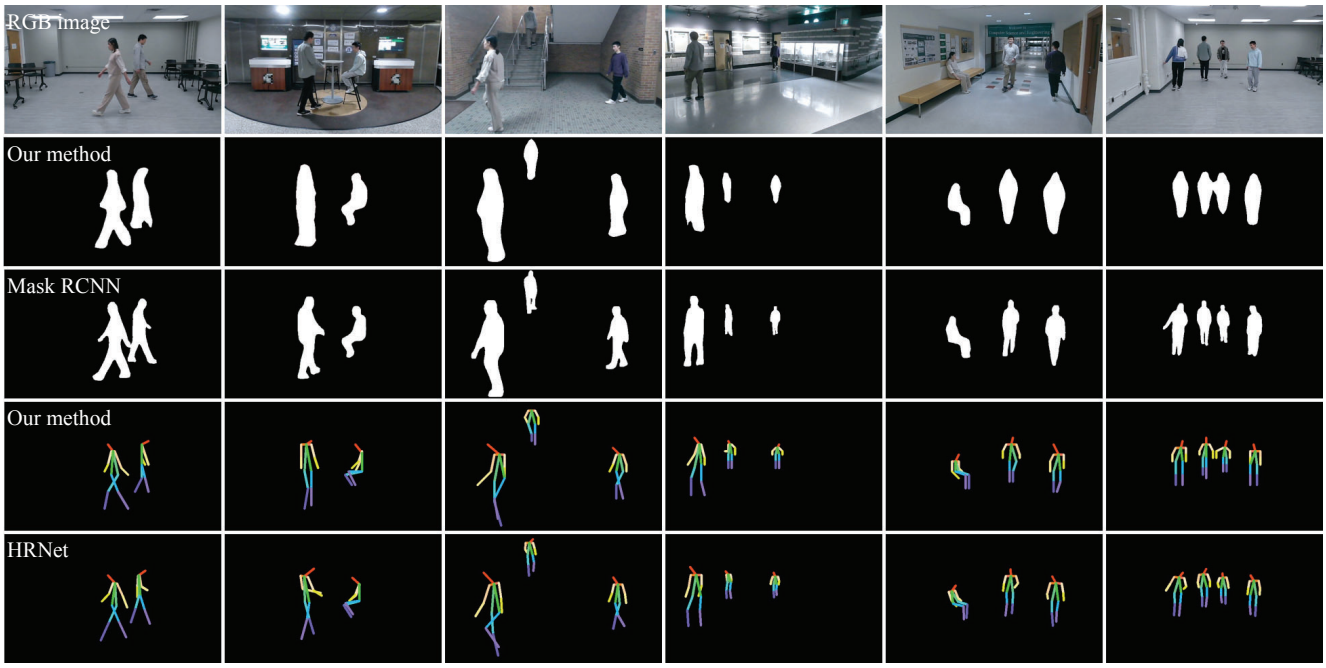


Figure 18: Mask Segmentation and Pose Estimation Results on different activities and environments.

testing. The samples in training set and test set are different in locomotion and body poses, but share the same person identities and environments. Detailed dataset statistics are provided in Table 1. The amount of training/test samples are 408,715 and 102,179, respectively.

To provide supervision for the WiFi-based sensing signals, we extracted mask segmentation and pose annotations from each video frame. We use Mask RCNN [15] to obtain body part mask segmentation and HRNet [42] to extract person keypoints. These methods are widely used as baselines in the fields of segmentation and pose estimation, respectively.

4.3 Evaluation Metrics

4.3.1 Mask Segmentation. We evaluate mask segmentation using the Intersection over Union (IoU) metric [15, 50], defined as:

$$IoU = \frac{area(S_p \cap S_{gt})}{area(S_p \cup S_{gt})} \tag{15}$$

where $S_p \cap S_{gt}$ is the intersection and $S_p \cup S_{gt}$ is the union of the predicted and ground truth masks. We calculate the average precision (AP) at a given IoU threshold a as $AP@a = Prob(IoU \geq a)$ and $mAP = 0.1 \sum_{i=0}^{9} AP@(0.5 + 0.05i)$.

4.3.2 Pose Estimation. The standard pose estimation metric is based on Object Keypoint Similarity (OKS) [2] [42, 55]:

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{16}$$

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. AP is calculated at a given OKS threshold a as $AP@a = Prob(OKS \geq a)$ and $mAP = 0.1 \sum_{i=0}^{9} AP@(0.5 + 0.05i)$.

For single-person pose estimation, we use the Percentage of Correct Keypoints (PCK) metric, following [49, 60, 77]:

$$PCK@a = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\frac{\| pd_i - gt_i \|_2^2}{\sqrt{rs^2 + lh^2}} \leq a\right), \tag{17}$$

where N is the number of joints, and $\mathbb{I}$ is an indicator function. rs and lh represents the positions of the right shoulder and left hip, respectively. pd_i and gt_i are the predicted and ground-truth coordinates. The term $\sqrt{rs^2 + lh^2}$ serves as a normalization factor based on the upper limb length, used to normalize the prediction error $\| pd_i - gt_i \|_2^2$.

[2] https://cocodataset.org/#keypoints-eval

Figure 19: Example of failure cases. SiWiS cannot effectively handle issues related to overlapping of human bodies.

4.4 Human Detection Accuracy

We compare SiWiS with Person-in-WiFi [50], the current state-of-the-art WiFi-based approach for multi-person 2D pose estimation. The experimental results are shown in Table 2. In the mask segmentation task, SiWiS outperforms Person-in-WiFi on all metrics, achieving a 10-point increase in mAP. Additionally, we also provide the evaluation results of pose estimation. In the table, the high values of AP@50 and AP@60 indicate that SiWiS can accurately detect person profiles. Notably, there was a significant performance degradation between AP@70 and AP@80, with a decrease in accuracy of 0.4710 for mask segmentation and 0.2810 for pose estimation. This issue may potentially be alleviated by adding more patch antennas to SiWiS.

Finally, we present several test results to offer a qualitative perspective on our system's effectiveness. Fig. 18 displays mask segmentation and pose estimation samples from our test dataset, comparing them with corresponding RGB images and the results obtained using Mask RCNN [15] and HRNet [42]. The results demonstrate the robust performance of SiWiS across different environments with varying activities. Additionally, we showcase some failure cases in Fig. 19, highlighting instances where errors occur in SiWiS due to overlapping of human bodies. SiWiS struggles to distinguish these cases due to the low spatial resolution caused by the limited number of antennas.

4.5 Performance Analysis

We now analyze several key factors influencing system performance and provide detailed performance of SiWiS under various conditions.

4.5.1 Signal Duration. Within a short time interval, SiWiS may not capture radio signals reflected by certain body parts, potentially leading to de-emphasized or missing key information [72]. To mitigate this, we experimented with longer sequences of signal frames. We varied the input sequence length in our experiments to evaluate SiWiS's performance. As shown in Fig. 20, the average precision is poor when using signal frames from only 0.1 seconds, and it improves as sequence length increases. The rate of improvement gradually slows down when the signal duration exceeds 1.7 seconds.

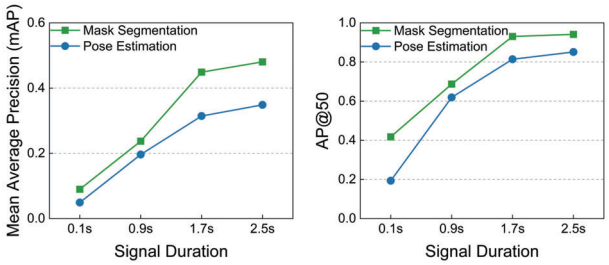


Figure 20: Impact of varying signal durations on the performance of our system.

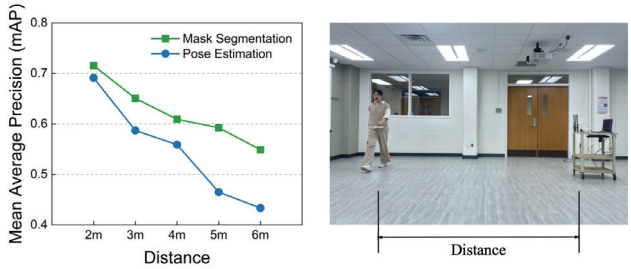


Figure 21: Impact of varying distance on the performance of our system.

4.5.2 Detection Distance. The reflective signals received by SiWiS become weaker as the object distance increases. Additionally, phenomena such as reflection and refraction can cause interference to SiWiS, and multi-path issues become more pronounced at greater distances. To assess SiWiS' performance at different distances, we conducted additional experiments. Specifically, we instructed a subject to move parallel to the device at fixed distances, collecting WiFi-based sensing signals and corresponding video frames at distances of 2m, 3m, 4m, 5m, and 6m. As shown in Fig. 21, the average precision rapidly decreases as the distance increases.

4.6 Zero-shot Performance

Different environments exhibit different radio propagation characteristics. Human pose estimation methods [37, 49, 50, 60, 77] based on WiFi CSI have faced challenges in generalizing models to new environmental settings. These methods are typically confined to fixed settings and are not directly adaptable to new (unseen or untrained) environments,

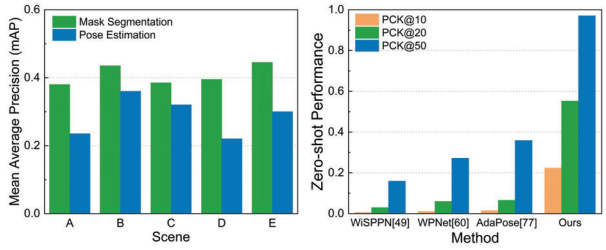


Figure 22: Left: SiWiS' zero-shot performance in five different unseen scenes. Right: Zero-shot comparison between SiWiS and other WiFi-based methods.

significantly limiting the widespread applicability of WiFi CSI-based sensing approaches. SiWiS, by achieving phase-coherent sensing through hardware innovation, fundamentally reduces the influence of surrounding environments on the WiFi signals reflected from targets. In what follows, we conducted experiments to evaluate the generalization capability of SiWiS in *unseen* scenes.

4.6.1 Dataset. To ensure a fair comparison with previous WiFi-based single-person pose estimation methods [77], we invited subjects to perform daily activities, such as raising arms, kicking legs, and squatting, within 2.5 meters of the device. Additionally, we further collected data in five different scenes, comprising three indoor and two outdoor settings. During experimentation, we alternately selected three of these scenes as the training set and the remaining two as the unseen test set. Consequently, each scene was evaluated twice as the test set.

4.6.2 Performance. We report the mean of twice evaluations as the final result for each scenario, as shown on the left side of Fig. 22. Across five different scenes, the mAP scores for mask segmentation are relatively consistent, while the scores for pose estimation vary significantly. In outdoor scenes A and D, the mAP scores for pose estimation are both around 0.2, considerably lower than those in indoor scenes.

Furthermore, we compare our work with WiFi CSI-based single-person pose estimation methods [49, 60, 77]. The experimental results are shown on the right side of Fig. 22. Evidently, SiWiS significantly surpasses them under zero-shot conditions. The results confirm the exceptional domain transfer capability of SiWiS, addressing the current challenge of WiFi CSI-based sensing methods being unable to adapt to new environments. This opens up possibilities for the widespread application of WiFi Sensing technology. Finally, we present some zero-shot results in each of the five scenes, as shown in Fig. 23.

5 LIMITATIONS AND DISCUSSIONS

In this section, we outline SiWiS' limitations and discuss possible solutions.

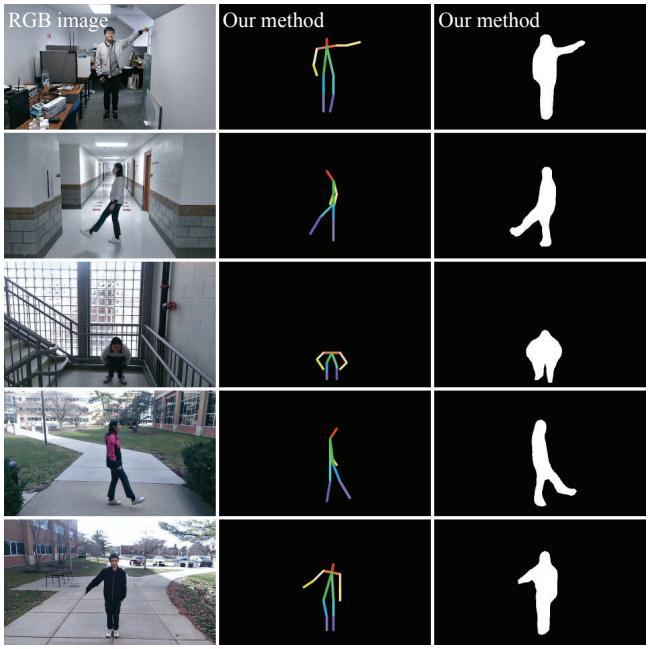


Figure 23: Zero-shot performance on different activities and environments.

Specialized Hardware: SiWiS requires the installation of a specialized RF circuit and additional antennas on a WiFi device to realize its radar-like sensing capability. This is a drawback compared to CSI-based sensing methods. To enhance the applicability of SiWiS, the specialized RF circuit could be integrated into the WiFi device's motherboard. Additionally, patch antennas can be fabricated using flexible substrate materials, which can be easily attached to the surface of WiFi devices.

Computational Complexity: SiWiS relies on a dual-branch DNN to infer human activities based on sensing signals. This learning-based approach has higher computational complexity compared to model-based approaches, which constitutes another drawback of SiWiS. To address this issue, SiWiS could outsource the DNN inference task to a cloud AI server by leveraging the Internet access of its host WiFi device. Nowadays, the Internet bandwidth of most WiFi devices is more than sufficient for this outsourcing.

Static Objects: Like a Doppler radar, SiWiS can only detect moving objects and cannot detect static objects. This is a limitation of SiWiS. However, in practice, people are rarely completely static, even when sleeping. We note that SiWiS has sub-centimeter sensitivity for detecting object movements. With such high sensitivity, SiWiS can detect individuals even if they are sitting or lying down, due to their subtle movements.

Coexistence of Multiple SiWiS: Consider the case where multiple WiFi devices are in the same area and every WiFi device is equipped with SiWiS for sensing. It is noteworthy

that all SiWiS devices will function properly in this case. This is because, as we explained in §2.6, SiWiS is active for sensing if and only if its host WiFi device is transmitting.

See Through Walls: Since WiFi signals are capable of penetrating walls, SiWiS has a great potential to see human activities through walls. The challenge lies in the video data collection for DNN training, as the camera cannot see through walls to collect supervisory data. We leave it for future work.

Antenna Number: In this work, SiWiS uses four antennas for sensing, considering the physical size of a WiFi router. It is interesting to study the impact of its antennas by examining the relationship between the number of antennas and the maximum number of recognizable individuals. This is a nontrivial problem and will be studied in our future work.

6 RELATED WORK

Computer Vision: In the field of computer vision, the task of human pose estimation [76] primarily involves estimating the coordinates of human body keypoints from images or videos. Earlier methods [34, 42, 44, 51, 64] mainly utilized CNNs for feature extraction. HRNet [42] proposed a high-resolution DNN model, which comprises parallel high-to-low resolution subnetworks. Recently, inspired by the Vision Transformers (ViT) [10], the field of human pose estimation has rapidly evolved from CNN-based networks to ViT networks [26, 27, 45, 63, 66]. HRFormer [66] builds on the foundation of HRNet, segmenting different resolution representation maps into non-overlapping small image windows. ViTPose [55] directly utilized ViT as its backbone and demonstrated that even simple network structures can achieve impressive results.

Radar Sensing: Radar sensing offer significant advantages in environmental perception and find widespread applications in daily life [54]. Compared to other wireless signals, radar has stronger anti-environmental interference and fine-grained perceptual information. Indoor positioning represents a crucial aspect of radar research, with early methods [2, 3] aiming to achieve precise indoor human body positioning using radar. Subsequently, the RF Capture system [1] was introduced, enabling the tracking of a person's 3D limb and body part positions through walls. This system laid a foundation for subsequent human body posture estimation techniques. The rapid advancement of perception technology and deep learning based on radar has facilitated the development of numerous applications, including sleep detection [68, 75], gesture recognition [21, 28, 31, 41, 43], radar imaging [1, 2, 5], physiological feature monitoring [4, 7, 29, 67], and object tracking [12, 18, 19].

WiFi Sensing: Benefiting from the advancements in deep learning, an increasing number of studies [57] have constructed various DNN-based WiFi sensing applications, such as human activity recognition [23, 32, 39, 40, 53, 56, 65, 78, 80], human identification [13, 48, 58, 59, 61, 69] and gesture recognition [14, 24, 47, 52, 62, 69–71, 79]. For instance, DeepSense [80] employs CNN modules and LSTM [17] modules to automatically identify common activities. THAT [23] studied time-over-channel features and proposed using a multi-scale convolution augmented transformer to capture range-based patterns. Widar [71] suggests the use of a body-coordinate velocity profile, which describes the power distribution over different velocities to track human motion.

Compared to the tasks mentioned above, pose estimation presents grander challenges due to the need for fine-grained prediction of multiple keypoints. Most existing WiFi Sensing methods for pose estimation utilize WiFi CSI as the input signal. WiSPPN [49] employs a setup with three antennas for both the WiFi sender and receiver, generating WiFi signals for estimating human poses. Building on this, Person-in-WiFi [50] introduces a multi-task learning approach, mapping WiFi signals to human body segmentation masks and joint coordinates. GoPose [37] used non-linearly spaced antennas on the WiFi device to expand 1D AoA estimation to 2D AoA framework. AdaPose [77] proposed the use of mapping consistency loss to tackle the challenges in cross-domain WiFi-based human pose estimation.

SiWiS advances the state-of-the-art by enabling *phase-coherent* sensing on a *single* WiFi device. It achieves a significant improvement on human detection accuracy and direct transferability in *unseen* environments.

7 CONCLUSION

In this paper, we presented SiWiS, an indoor human sensing system using a *single* WiFi device. SiWiS comprises two novel components: RF sensing hardware and dual-branch DNN model. Our hardware design enables a single WiFi device to detect the object movement by utilizing ambient WiFi OFDM signals. It achieves *phase-coherent* sensing by establishing a deterministic relation between the observed signal phase and the object movement distance. Our dual-branch DNN model was highly optimized for joint human pose estimation and mask segmentation. Extensive experimental results demonstrate a significant improvement of SiWiS compared to existing CSI-based sensing methods. More importantly, zero-shot experiments confirm that SiWiS can be effectively used in *unseen* real-world environments, addressing a fundamental challenge (transferability) in WiFi sensing.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers and our shepherd for their insightful comments. We are also grateful to Peihao Yan and Bowei Zhang for their valuable contributions to data collection for this project. This project was supported in part by NSF Grants ECCS-2225337 and CNS-2100112.

REFERENCES

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

[2] Fadel Adib, Zachary Kabelac, and Dina Katabi. 2015. {Multi-Person} Localization via {RF} Body Reflections. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*. 279–292.

[3] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 2014. 3D tracking via body radio reflections. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*. 317–329.

[4] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 837–846.

[5] Maurizio Bocca, Ossi Kaltiokallio, Neal Patwari, and Suresh Venkatasubramanian. 2013. Multiple target tracking with RF sensor networks. *IEEE Transactions on Mobile Computing* 13, 8 (2013), 1787–1800.

[6] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. 2024. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[7] Qingchao Chen, Yang Liu, Bo Tan, Karl Woodbridge, and Kevin Chetty. 2020. Respiration and activity detection based on passive radio sensing in home environments. *IEEE Access* 8 (2020), 12426–12437.

[8] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7103–7112.

[9] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5386–5395.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*. 2334–2343.

[12] Elisabetta Farella, Augusto Pieracci, Luca Benini, Laura Rocchi, and Andrea Acquaviva. 2008. Interfacing human and computer with wireless body area sensor networks: the WiMoCA solution. *Multimedia Tools and Applications* 38 (2008), 337–363.

[13] Yu Gu, Huan Yan, Mianxiong Dong, Meng Wang, Xiang Zhang, Zhi Liu, and Fuji Ren. 2021. Wione: one-shot learning for environment-robust device-free user authentication via commodity wi-fi in man–machine system. *IEEE Transactions on Computational Social Systems* 8, 3 (2021), 630–642.

[14] Yu Gu, Xiang Zhang, Yantong Wang, Meng Wang, Huan Yan, Yusheng Ji, Zhi Liu, Jianhua Li, and Mianxiong Dong. 2022. WiGRUNT: WiFi-enabled gesture recognition using dual-attention network. *IEEE Transactions on Human-Machine Systems* 52, 4 (2022), 736–746.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] Chen-Yu Hsu, Rumen Hristov, Guang-He Lee, Mingmin Zhao, and Dina Katabi. 2019. Enabling identification and behavioral sensing in homes using radio reflections. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[19] Chen-Yu Hsu, Yuchen Liu, Zachary Kabelac, Rumen Hristov, Dina Katabi, and Christine Liu. 2017. Extracting gait velocity and stride length from surrounding radio signals. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2116–2126.

[20] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using WiFi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[21] Seo Yul Kim, Hong Gul Han, Jin Woo Kim, Sanghoon Lee, and Tae Wook Kim. 2017. A hand gesture recognition sensor using reflected impulses. *IEEE Sensors Journal* 17, 10 (2017), 2975–2976.

[22] Hei Law and Jia Deng. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*. 734–750.

[23] Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. 2021. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 286–293.

[24] Chenning Li, Manni Liu, and Zhichao Cao. 2020. WiHF: Enable user identified gesture recognition with WiFi. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 586–595.

[25] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. 2019. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 872–881.

[26] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu, and Xiaolong Wang. 2021. Test-time personalization with a transformer for human pose estimation. *Advances in Neural Information Processing Systems* 34 (2021), 2583–2597.

[27] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. 2021. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*. 11313–11322.

[28] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–19.

[29] Qiwei Liu, Hanqing Guo, Junhong Xu, Honggang Wang, Aron Kageza, Saeed AlQarni, and Shaoen Wu. 2018. Non-contact non-invasive heart and respiration rates monitoring with MIMO radar sensing. In *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[31] Elishiah Miller, Zheng Li, Helena Mentis, Adrian Park, Ting Zhu, and Nilanjan Banerjee. 2020. RadSense: Enabling one hand and no hands interaction for sterile manipulation of medical images using Doppler radar. *Smart Health* 15 (2020), 100089.

[32] Parisa Fard Moshiri, Reza Shahbazian, Mohammad Nabati, and Seyed Ali Ghorashi. 2021. A CSI-based human activity recognition using deep learning. *Sensors* 21, 21 (2021), 7225.

[33] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems* 30 (2017).

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14,*

2016, Proceedings, Part VIII 14. Springer, 483–499.

[35] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. 2017. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4903–4911.

[36] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2021. Winect: 3d human pose tracking for free-form activity using commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–29.

[37] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. GoPose: 3D human pose estimation using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–25.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. 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*. Springer, 234–241.

[39] Jörg Schäfer, Baldev Raj Barrsiwal, Muyassar Kokhkharova, Hannan Adil, and Jens Liebehenschel. 2021. Human activity recognition using CSI information with nexmon. *Applied Sciences* 11, 19 (2021), 8860.

[40] Biyun Sheng, Fu Xiao, Letian Sha, and Lijuan Sun. 2020. Deep spatial–temporal model based cross-scene action recognition using commodity WiFi. *IEEE Internet of Things Journal* 7, 4 (2020), 3592–3601.

[41] Sruthy Skaria, Akram Al-Hourani, Margaret Lech, and Robin J Evans. 2019. Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks. *IEEE Sensors Journal* 19, 8 (2019), 3041–3048.

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5693–5703.

[43] Yuliang Sun, Tai Fei, Xibo Li, Alexander Warnecke, Ernst Warsitz, and Nils Pohl. 2020. Real-time radar-based gesture detection and recognition built in an edge-computing platform. *IEEE Sensors Journal* 20, 18 (2020), 10706–10716.

[44] Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1653–1660.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[46] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. 2021. Multimodal CSI-based human activity recognition using GANs. *IEEE Internet of Things Journal* 8, 24 (2021), 17345–17355.

[47] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. 2022. Airfi: empowering wifi-based passive human gesture recognition to unseen environment via domain generalization. *IEEE Transactions on Mobile Computing* (2022).

[48] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. 2022. CAUTION: A Robust WiFi-Based Human Authentication System via Few-Shot Open-Set Recognition. *IEEE Internet of Things Journal* 9, 18 (2022), 17323–17333. https://doi.org/10.1109/JIOT.2022.3156099

[49] Fei Wang, Stanislav Panev, Ziyi Dai, Jinsong Han, and Dong Huang. 2019. Can WiFi estimate person pose? *arXiv preprint arXiv:1904.00277* (2019).

[50] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained person perception using WiFi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5452–5461.

[51] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 466–481.

[52] Chunjing Xiao, Daojun Han, Yongsen Ma, and Zhiguang Qin. 2019. CsiGAN: Robust channel state information-based activity recognition with GANs. *IEEE Internet of Things Journal* 6, 6 (2019), 10191–10204.

[53] Chunjing Xiao, Yue Lei, Yongsen Ma, Fan Zhou, and Zhiguang Qin. 2020. DeepSeg: Deep-learning-based activity segmentation framework for activity recognition using WiFi. *IEEE Internet of Things Journal* 8, 7 (2020), 5669–5681.

[54] Jiaren Xiao, Bing Luo, Li Xu, Bo Li, and Zhiguo Chen. 2024. A survey on application in RF signal. *Multimedia Tools and Applications* 83, 4 (2024), 11885–11908.

[55] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems* 35 (2022), 38571–38584.

[56] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shiyang Wang, Ye Yuan, Shuochao Yao, Aidong Zhang, and Lu Su. 2020. DeepMV: Multi-view deep learning for device-free human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.

[57] Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. 2023. SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing. *Patterns* 4, 3 (2023).

[58] Jianfei Yang, Xinyan Chen, Han Zou, Dazhuo Wang, and Lihua Xie. 2022. AutoFi: Toward Automatic Wi-Fi Human Sensing via Geometric Self-Supervised Learning. *IEEE Internet of Things Journal* 10, 8 (2022), 7416–7425.

[59] Jianfei Yang, Xinyan Chen, Han Zou, Dazhuo Wang, Qianwen Xu, and Lihua Xie. 2022. EfficientFi: Toward large-scale lightweight WiFi sensing via CSI compression. *IEEE Internet of Things Journal* 9, 15 (2022), 13086–13095.

[60] Jianfei Yang, Yunjiao Zhou, He Huang, Han Zou, and Lihua Xie. 2022. MetaFi: Device-free pose estimation via commodity WiFi for metaverse avatar simulation. In *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*. IEEE, 1–6.

[61] Jianfei Yang, Han Zou, and Lihua Xie. 2022. Securesense: defending adversarial attack for secure device-free human activity recognition. *IEEE Transactions on Mobile Computing* (2022).

[62] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. 2019. Learning gestures from WiFi: A Siamese recurrent convolutional architecture. *IEEE Internet of Things Journal* 6, 6 (2019), 10763–10772.

[63] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. 2021. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11802–11812.

[64] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2017. Learning feature pyramids for human pose estimation. In *proceedings of the IEEE international conference on computer vision*. 1281–1290.

[65] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. 2017. A survey on behavior recognition using WiFi channel state information. *IEEE Communications Magazine* 55, 10 (2017), 98–104.

[66] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. 2021. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408* (2021).

[67] Shichao Yue, Hao He, Hao Wang, Hariharan Rahul, and Dina Katabi. 2018. Extracting multi-person respiration from entangled RF signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–22.

[68] Shichao Yue, Yuzhe Yang, Hao Wang, Hariharan Rahul, and Dina Katabi. 2020. BodyCompass: Monitoring sleep posture with wireless signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–25.

[69] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th annual international conference on mobile computing and networking*. 305–320.

[70] Xie Zhang, Chengpei Tang, Kang Yin, and Qingqian Ni. 2021. WiFi-based cross-domain gesture recognition via modified prototypical networks. *IEEE Internet of Things Journal* 9, 11 (2021), 8584–8596.

[71] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2021. Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8671–8688.

[72] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7356–7365.

[73] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. 2019. Through-wall human mesh recovery using radio signals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10113–10122.

[74] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 267–281.

[75] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. PMLR, 4100–4109.

[76] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *Comput. Surveys* 56, 1 (2023), 1–37.

[77] Yunjiao Zhou, Jianfei Yang, He Huang, and Lihua Xie. 2023. AdaPose: Towards Cross-Site Device-Free Human Pose Estimation with Commodity WiFi. *arXiv preprint arXiv:2309.16964* (2023).

[78] Han Zou, Jianfei Yang, Hari Prasanna Das, Huihan Liu, Yuxun Zhou, and Costas J Spanos. 2019. WiFi and vision multimodal learning for accurate and robust device-free human activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.

[79] Han Zou, Jianfei Yang, Yuxun Zhou, Lihua Xie, and Costas J Spanos. 2018. Robust WiFi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–8.

[80] Han Zou, Yuxun Zhou, Jianfei Yang, Hao Jiang, Lihua Xie, and Costas J Spanos. 2018. Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network. In *2018 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.